Mobile Phone Data for Development

Analysis of mobile phone datasets for the development of Ivory Coast

May 1-3, 2013

Selected contributions to the D4D challenge sponsored by Orange

www.netmob.org

// EDITORS

Vincent BLONDEL
Nicolas de CORDES
Adeline DECUYPER
Pierre DEVILLE
Jacques RAGUENEZ
Zbigniew SMOREDA

▶ D4D Scientific Committee

Professor Vincent BLONDEL (Chairman)
University of Louvain (UCL), Louvain-La-Neuve, Belgium

Professor Francis AKINDES
Université de Bouaké, Bouaké, Ivory Coast

Mr William HOFFMAN
Head of Telecom industry, World Economic Forum, New York, USA

Mrs Marie-Noëlle JÉGO-LAVEISSIÈRE
Head of Orange Labs, Paris, France

Mr Robert KIRKPATRICK
Head of Global Pulse, United Nations, New York, USA

Mr Chris LOCKE
Managing director GSMA Development Fund, GSMA, London, UK

Professor Alex (Sandy) PENTLAND
Medialab, MIT, Cambridge, USA

SOCIAL AND ECONOMIC DEVELOPMENT

1. **Mobile Communications Reveal the Regional Economy in Côte d'Ivoire**
   H. MAO, X. SHUAI, Y.-Y. AHN, J. BOLLEN

2. **Ubiquitous Sensing for Mapping Poverty in Developing Countries**
   C. SMITH, A. MASHADI, L. CAPRA

3. **Analyzing Social Divisions Using Cell Phone Data**
   O. BUCICOVSCHI, R. W. DOUGLASS, D. A. MEYER, M. RAM, D. RIDEOUT, D. SONG

4. **Development, Information and Social Connectivity in Cote d'Ivoire**
   C. ANDRIS, L. M. A. BETTENCOURT

5. **Can Fires, Night Lights, and Mobile Phones reveal behavioral fingerprints useful for Development ?**
   D. PASTOR-ESCUREDO, T. SAVY, M. A. LUENGO-OROZ

6. **Exploration and Analysis of Massive Mobile Phone Data: A Layered Visual Analytics approach**
   S. van den ELZEN, J. BLAAS, D. HOLTEN, J.-K. BUENEN, J. J. VAN WIJK, R. SPOUSTA, A. MIAO, S. SALA, S. CHAN

7. **Exploring the Multilevel Community Structure in the D4D Dataset**
   X. LIU, T. MURATA, K. WAKITA

8. **Social Capital for Economic Development : Application of Time Series Cluster Analysis on Personal Network Structures**
   B. LIM, D. DORAN, V. MENDIRATTA, M. RODRIGUEZ, D. KLABJAN

9. **Estimating Human Dynamics in Cote d'Ivoire Through D4D Call Detail Records**
   K. WAKITA, R. KAWASAKI

10. **« Calling Abidjan » - Improving Population Estimations with Mobile Communication Data**
    H. STERLY, B. HENNIG, K. DONGO

11. **Understanding ethnical interactions on Ivory Coast**
    A. J. MORALES, W. CREIXELL, J. BORONDO, J.C. LOSADA, R.M. BENITO

12. **Impacts of External Shocks in Commodity-Dependent Low-Income Countries : Insights from mobile phone call detail records from Cote D'Ivoire**
    A. FAJEBE, P. BRECKE

13. **Regional patterns of socio-economic activity in Côte d'Ivoire**
    M. DUSI, M. AHMED, R. CAPORICCI, N. CHEESEMAN

14. **Regional Development – Capturing a nation's sporting interest through call detail analysis**
    D. MC GOWAN, N. HURLEY

15. **Rapid Assessment of Population Movements in Crises : The Potential and Limitations of Using Nighttime Satellite Imagery and Mobile Phone Data**
    N. BHARTI, X. LU, L. BENGTSSON, E. WETTER, A. TATEM

16. **Analysing and mapping population movements from anonymous cellphone activity data**
    H. GLASS, I. KIRKPATRICK, A. SCHIFF

17. **Spotted : Connecting People, Locations and Real-World Events in a Cellular Network**
    R. TRESTIAN, F. ZAMAN, G.-M. MUNTEAN

18. **Towards an early warning system : the effect of weather on mobile phone usage. A case study in Abidjan**
    J. PEDRO CRAVEIRO, F. M. V. RAMOS, E. KANJO, N. EL MAWASS

19. **Does conflict affect human mobility and cellphone usage ? Evidence from Côte d'Ivoire**
    S. LINARDI, S. KALYANARAMAN, D. BERGER

DATA MINING

20. Symbolic clustering of users and antennae
    M. CERINSEK, J. BODLAJ, V. BATAGELJ
21. Discovering common structures in mobile call data : An efficient way to clustering ego graphs
    S. AGHA MUHAMMAD, K. VAN LAERHOVEN
22. Data Analysis and Mining of Mobile Phone Dataset
    C. BADENES OLMEDO, S. MUÑOZ HERNANDEZ
23. First steps for a Synthetic Population of Ivory Coast
    A. APOLLONI, A. CAMACHO, K. EAMES, J. W. EDMUNDS, S. FUNK
24. Place Identification and Prediction in the D4D Data Set using Machine Learning
    N. GHOURCHIAN, D. PRECUP
25. Patterns of Cell Towers in Mobile Cellular Network
    J. XIONG, G. RANJAN, L. CHEN, Z.-L. ZHANG
26. Mobile Phone Data Analysis of Cote d'Ivorie
    D. GÜNDOGDU
27. Properties of Dynamic Networks
    R. ANAND, C. K. REDDY
28. Constrained link prediction on the D4D dataset
    B. ZONG, P. BOGDANOV, A. K. SINGH
29. Interactive Visualization of Cellphone Network Data Using D3 : The Case of Ivory Coast
    M. RODRIGUEZ, V. MENDIRATTA, B. LIM, D. DORAN, D. KLABJAN
30. NVizABLE : A Web-Based Network Visualization Interface
    J. SMITH, J. STEVENS, M. Y. IDRIS
31. Mobile Data Delivery through Opportunistic Communications among Cellular Users :
    A Case Study for the D4D Challenge
    Y. ZHU, C. ZHANG, Y. WANG
32. EEMC : An Energy-Efficient Mobile Crowdsensing Mechanism by Reusing Call/SMS Connections
    H. XIONG, L. WANG, D. ZHANG

MOBILITY/TRANSPORT

33. The Differing Tribal and Infrastructural Influences on Mobility in Developing and
    Industrialized Regions
    A. AMINI, K. KUNG, C. KANG, S. SOBOLEVSKY, C. RATTI
34. Approaching the Limits of Predictability in Human Mobility: A Study of 500,000 Mobile Phone
    Users in Cote d'Ivoire after the 2011 Civil War
    X. LU, E. WETTER, N. BHARTI, A. TATEM, L. BENGTSSON
35. Identification and Characterization of Human Behavior Patterns from Mobile Phone Data
    P. PARASKEVOPOULOS, T.-C. DINH, Z. DASHDORJ, T. PALPANAS, L. SERAFINI
36. Egocentric and population-density patterns of cellphone communication in Ivory Coast
    P. SCHMITT, M. VIGIL, M. ZHELEVA, E. M. BELDING
37. Multi-perspective analysis of D4D fine resolution data
    G. ANDRIENKO, N. ANDRIENKO, G. FUCHS
38. AllAboard : a system for exploring urban mobility and optimizing public transport using
    cellphone data
    M. BERLINGERIO, F. CALABRESE, G. DI LORENZO, R. NAIR, F. PINELLI, M. L. SBODIO

39. Mobility Modeling for Transport Efficiency – Analysis of Travel Characteristics Based on Mobile Phone Data
V. ANGELAKIS, D. GUNDLEGÅRD, B. RAJNA, C. RYDERGREN, K. VROTSOU, R. CARLSSON, J. FORGEAT, T. H. HU, E. L. LIU, S. MORITZ, S. ZHAO, Y. ZHENG

40. MP4-A Project : Mobility Planning For Africa
M. NANNI, R. TRASARTI, B. FURLETTI, L. GABRIELLI, P. VAN DER MEDE, J. DE BRUIJN, E. DE ROMPH, G. BRUIL

41. Crowdsourcing Physical Package Delivery Using the Existing Routine Mobility of a Local Population
J. MCINERNEY, A. ROGERS, N. R. JENNINGS

42. Towards a recommender system for bush taxis
S. GAMBS, M.-O. KILLIJIAN, M. NÚÑEZ DEL PRADO CORTEZ, M. TRAORÉ

43. Real-time streaming mobility analytics
A. GARZÓ, I. PETRÁS, C. ISTVÁN SIDLÓ, A. A. BENCZÚR

44. Visualization of traffic
J. BODLAJ, M. CERINSEK, V. BATAGELJ

45. Daily Commuting in Ivory Coast : Development Opportunities
M. MAMEI, L. FERRARI

46. Building a minimal traffic model from mobile phone data
M. ZILSKE, K. NAGEL

47. A Tale of Peoples'Movement Patterns in Developing Countries
K. YADAV, A. KUMAR, V. NAIK, A. SINGH

48. Commuting Dynamics 4 Change
R. MAESTRE, R. LARIO, M. MUÑOZ, R. ABAD, J. GONZALEZ, A. MARTÍN, E. PEREZ, JL. FDEZ-PACHECO

49. The geography and carbon footprint of mobile phone use in Cote d'Ivoire
V. SALNIKOV, D. SCHIEN, H. YOUN, R. LAMBIOTTE, M. T. GASTNER

50. Analysis of New Strategies for Resources Allocation and Infrastructure Development in Côte d'Ivoire by Mapping Telecommunication Densities
Y. HUI, M. LIU, P. HUI

51. Social, Disconnected or In between : Mobile Data reveals urban mood
E. KANJO, N. EL MAWASS, J. PEDRO CRAVEIRO, F. M. V. RAMOS

52. Studying Intercity Travels and Traffic Using Cellular Network Data
W. WU, E. YEOW CHEU, Y. FENG, D. NGAN LE, G. ENG YAP, X. LI

53. Human Mobility Flows in the City of Abidjan
D. NABOULSI, M. FIORE, R. STANICA

54. Revealing the pulse of human dynamics in a country from mobile phone data
S. SCEPANOVIC, P. BEN HUI, A. YLA-JAASKI

55. Profiling workers' activity-travel behavior based on mobile phone data
F. LIU, D. JANSSENS, G. WETS, M. COOLS

56. Extracting Large Scale Social Relational Dynamics from Mobile Communications Data
J. HUCK, P. COULTON, D. WHYATT

57. Mobility and communication patterns in Ivory Coast
M. MITROVIC, V. PALCHYKOV, H.-H. JO, J. SARAMÄKI

58. Detecting Mobility Patterns in Mobile Phone Data from the Ivory Coast
M. F. DIXON, S. P. AIELLO, F. FAPOHUNDA, W. GOLDSTEIN

59. Predicting Human Mobility Patterns in Cities
X-Y. YAN, C. ZHAO, W. WANG

60. Combining call records and road data for strategic disaster response planning
Z. HUANG, U. KUMAR

HEALTH/EPIDEMICS

61. Mitigating Epidemics through Mobile Micro-measures
M. KAFSI, E. KAZEMI, L. MAYSTRE, L. YARTSEVA, M. GROSSGLAUSER, P. THIRAN

62. Exploiting Cellular Data for Disease Containment and Information Campaigns Strategies in Country-Wide Epidemics
A. LIMA, M. DE DOMENICO, V. PEJOVIC, M. MUSOLESI

63. Linking the Human Mobility and Connectivity Patterns with Spatial HIV distribution
K. GAVRIC, S. BRDAR, D. CULIBRK, V. CRNOJEVIC

64. Using Mobile Phone Data to Supercharge Epidemic Models of Cholera Transmission in Africa : A Case Study of Côte d'Ivoir
A. S. AZMAN, E. A. URQUHART, B. ZAITCHIK, J. LESSLER

65. Information Dissemination using Human Mobility in Realistic Environment- (E-Inspire)
R. AGARWAL, V. GAUTHIER, M. BECKER

66. Design and implémentation of a tool for the Correlation between the rate of prévalence of a pathology and the flow of communication between diverse localities
T. DJOTIO NDIÉ, Z. NGANMENI, S. J. NOUHO NOUTAT

67. Human mobility and communication patterns in Côte d'Ivoire : A network perspective for malaria control
E. A. ENNS, J. H. AMUASI

68. Exploring Community Structure to Understand Disease Spread and Control Using Mobile Call Detail Records
M. SARAVANAN, P. KARTHIKEYAN, A. AARTHI, M. KIRUTHIKA, S. SUGANYA

69. Large-scale Measurements of Network Topology and Disease Spread : A Pilot Evaluation Using Mobile Phone Data in Côte d'Ivoire
R. CHUNARA, E. O. NSOESIE

70. Are gravity models appropriate for estimating the spatial spread of malaria ?
A. WESOLOWSKI, C. O. BUCKEE

71. On Models Characterizing Cellular Social Networks
D. DEKA, S. VISHWANATH

72. Neighborhood structures in socio-demographic and HIV infection conditions Indication to the potential of mHealth for tackling HIV/AIDS in Ivory Coast
A. ARAI, T. HORANONT, A. WITAYANGKURN, R. SHIBASAKI

73. Disease Outbreak Detection by Mobile Network Monitoring : a case study with the D4D datasets
N. BALDO, P. CLOSAS

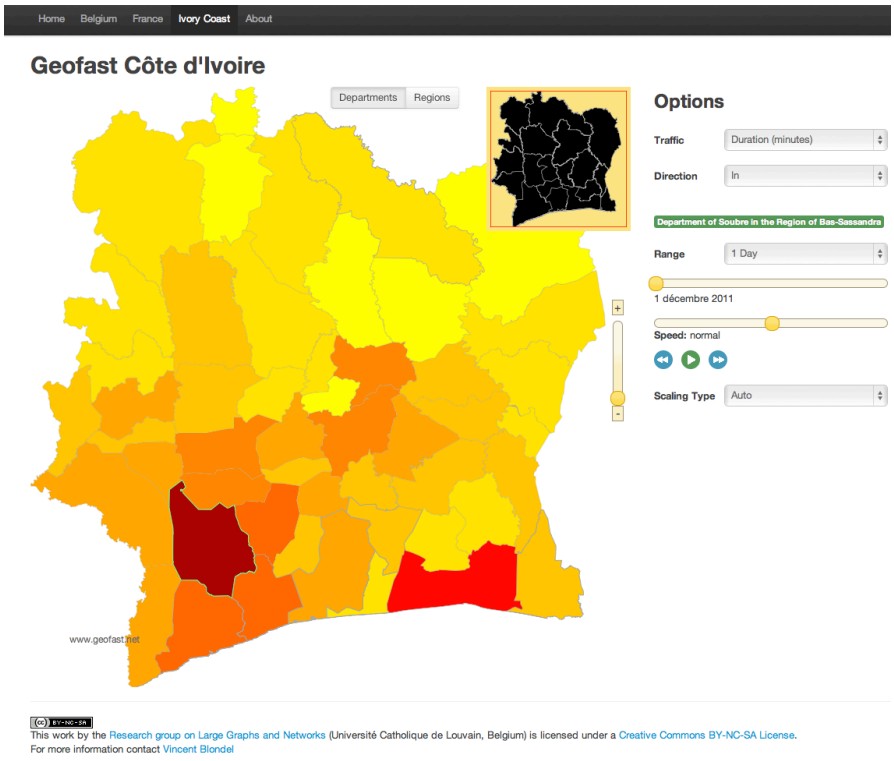
74. Applying Mobile Datasets in Computational Public Health Research
J. P. LEIDIG, Y. KUTSUMI, K. A. O'HEARN, C. M. SAUER, J. SCRIPPS, G. WOLFFE

DATA FOR DEVELOPMENT: THE D4D CHALLENGE ON MOBILE PHONE DATA

VINCENT D. BLONDEL*, MARKUS ESCH*, CONNIE CHAN*, FABRICE CLEROT†, PIERRE DEVILLE*, ETIENNE HUENS*, FRÉDÉRIC MORLOT†, ZBIGNIEW SMOREDA†, AND CEZARY ZIEMLICKI†

Abstract. The Orange "Data for Development" (D4D) challenge is an open data challenge on anonymous call patterns of Orange's mobile phone users in Ivory Coast. The goal of the challenge is to help address society development questions in novel ways by contributing to the socio-economic development and well-being of the Ivory Coast population. Participants to the challenge are given access to four mobile phone datasets and the purpose of this paper is to describe the four datasets. The website `http://www.d4d.orange.com` contains more information about the participation rules.

The datasets are based on anonymized Call Detail Records (CDR) of phone calls and SMS exchanges between five million of Orange's customers in Ivory Coast between December 1, 2011 and April 28, 2012. The datasets are: (a) antenna-to-antenna traffic on an hourly basis, (b) individual trajectories for 50,000 customers for two week time windows with antenna location information, (3) individual trajectories for 500,000 customers over the entire observation period with sub-prefecture location information, and (4) a sample of communication graphs for 5,000 customers.



The geofast web interface `www.geofast.net` for the visualisation of mobile phone communications (countries available: France, Belgium, Ivory Coast).

*University of Louvain, B-1348 Louvain-la-Neuve, Belgium. `vincent.blondel@uclouvain.be`
†Orange Labs, France

1. Introduction. The availability of detailed mobility traces and mobile phone communication data for large populations has already had a significant impact on research in behavioral science. Some researchers consider such datasets as an opportunity to refine the analysis of human behavior [5], while others question the usefulness of such datasets to draw conclusions on collective human behavior [1, 2, 5].

Digital traces left by mobile phone users often reveal sensitive private individual information. It is therefore natural to limit access to such data. Limited access to data of scientific interest is however a potential source of a "new digital divide" in the scientific community, as described in [2]. In order to improve the availability of large mobile phone datasets and to foster research in this area, the Orange Group decided to provide anonymized datasets from Ivory Coast for the purpose of scientific research. With around five million customers, Orange has a significant market share in Ivory Coast, whose total population is estimated to be 20 million individuals. In addition to the scientific benefit, the project intends to foster development in Ivory Coast by establishing new collaborations with African scientists and by providing behavioral data that has not yet been collected by the national statistics agency [3].

2. Other datasets on Ivory Coast. Researchers participating in the D4D challenge are encouraged to combine the D4D mobile phone datasets with other datasets and source of information. These sources include the following.

African Development Bank Group. The African Development Bank (AfDB) Groups mission is to help reduce poverty, improve living conditions for Africans and mobilize resources for the continents economic and social development. With this objective in mind, the institution aims at assisting African countries in their efforts to achieve sustainable economic development and social progress.
`http://www.afdb.org/en/`

African Economic Outlook. Economic, social and political developments of African countries, with the expertise of the African Development Bank, the OECD Development Centre, the United Nations Economic Commission for Africa, the United Nations Development Programme and a network of African think tanks and research centres.
http://www.africaneconomicoutlook.org/en/

Africa and Middle East Telecom News. Africa and Middle East Telecom-Week' tracks the fixed, broadband and mobile phone markets in Africa and Middle East.
`http://www.africantelecomsnews.com/`

Africa Renewal on Line. The Africa Renewal magazine is produced by the United Nations organism and provides up-to-date information and analysis of the major economic and development challenges facing Africa today. It works with the media in Africa and beyond to promote the work of the United Nations, Africa and the international community to bring peace and development to Africa.
`http://www.un.org/french/ecosocdev/geninfo/afrec/vol25n`

Africa Research Program. Data set consists of an aggregate of a number of the most commonly used publicly available variables used in the study of African political economy.
`http://africa.gov.harvard.edu/`

African Union. Pan African Organization.
http://au.int/en/resources/documents

Africover. Geographic data produced by the the Africover Project and the participating countries. The information available in the national Multipurpose Africover Databases on Environmental Resources (MADE) is composed by a main geographic information layer (i.e. land cover) and several additional layers that vary for each country (e.g. roads, rivers and water bodies, etc.); the available data produced by Africover is listed for each country in the Africover Data table (here only full resolution data sets and public domain spatially aggregated data sets are listed. Thematic aggregations are available starting from the metadata of these data sets).
http://www.africover.org/

Afristat. Observatoire Economique et Statistique dAfrique Subsaharienne
http://www.afristat.org/publication/acces-direct-aux-donnees

Banque Centrale des Etats de l'Afrique de l'Ouest. Financial and economic data.
http://edenpub.bceao.int/

Center For International Development. This page is a depository for data developed through research at the Center for International Development at Harvard University (CID). Often the data are associated with a research paper and, thus, the paper is also available for downloading.
http://www.cid.harvard.edu/ciddata/ciddata.html

CIA - The World Factbook. The World Factbook provides information on the history, people, government, economy, geography, communications, transportation, military, and transnational issues for 267 world entities.
https://www.cia.gov/library/publications/the-world-factbook/

Factset: A compilation of various international economic data sets.
http://www.factset.com

Famine Early Warning Systems Network. The Famine Early Warning Systems Network (FEWS NET) is a US AID-funded activity that collaborates with international, regional and national partners to provide timely and rigorous early warning and vulnerability information on emerging and evolving food security issues.
http://www.fews.net/Pages/

Food and Agriculture Organization of the United Nations. FAO's mandate is to raise levels of nutrition, improve agricultural productivity, better the lives of rural populations and contribute to the growth of the world economy.
http://www.fao.org/corp/statistics/en/

Global Distribution of Poverty. A website with a collection of subnational, spatially explicit, poverty data sets. This page is maintained by The Poverty Mapping Project at CIESIN (The Center for International Earth Science Information Network) at the Earth Institute at Columbia University.
http://sedac.ciesin.columbia.edu/povmap/

International Census. Global population trends, links to historical population estimates, population clocks, and estimates of population, births, and deaths occurring each year, day, hour, or second.
http://www.census.gov/ipc/www/idb/

Investir en zone France. Economic data about african french-speaking countries (glossary, economic indicators, maps, etc.).
http://www.izf.net/bdd-entreprise/

ITU - Telecommunication Development Sector.
http://www.itu.int/net/ITU-D/

Measure DHS. Information about population, health and nutrition programs.
http://www.measuredhs.com/

Measuring the Information Society. ICT Indicators for Development. ICT measurement is a tool for policymakers, to assess the status of ICT in developing countries and craft policies to maximize the benefits of ICT for those countries.
http://new.unctad.org/

Princeton Data and Statistical Services: Data on Africa. A compilation of datasets on Africa.
http://dss1.princeton.edu/cgi-bin/dataresources/newdataresources.cgi?term=14

Research ICT Africa Network. The Research ICT Africa Network conducts research on ICT policy and regulation that facilitates evidence-based and informed policy making for improved access, use and application of ICT for social development and economic growth.
http://www.researchictafrica.net/home.php

The International Aid Transparency Initiative. The International Aid Transparency Initiative aims to make information about aid spending easier to access, use and understand.
http://www.aidtransparency.net/

United Nations Data. The United Nations Statistics Division (UNSD) launched a new internet based data service for the global user community. It brings UN statistical databases within easy reach of users through a single entry point.
http://data.un.org/

United Nations Economic Commission for Africa. ECA's mandate is to promote the economic and social development of its member States, foster intra-regional integration, and promote international cooperation for Africa's development.
http://www.uneca.org/

US Census International Bureau Programs. The U.S. Census Bureau conducts demographic, economic, and geographic studies of countries around the world.
http://www.census.gov/population/international/

World Bank Data. The World Bank provides free and open access to a comprehensive set of data about development in countries around the globe.
`http://data.worldbank.org/`

World Trade Organization. Interactive access to the most up-to-date WTO trade statistics.
`http://www.wto.org/`

Mobile and Development Intelligence GSMA. MDI is an Open Data portal for the developing world mobile industry. A challenge facing mobile industry stakeholders in the developing world is the lack of publicly available data and analysis to support their business decision making and to clarify the socio-economic impact of mobile. MDI will fill this information gap and will aggregate and host data from multiple sources such as the World Bank, UN, member operators and from vendors and development organisations.
`http://mobiledevelopmentintelligence.com/`

Centre sur les politiques internationales des TIC pour les pays de l'Afrique de l'Ouest. CIPACO has been initiated by Panos Institute West Africa (PIWA - a regional West African NGO), in order to strengthen the capacity of African stakeholders for an effective participation in ICT decision-making processes.
`http://www.cipaco.org/index.php`

Institut National de la Statistique - République de Côte d'Ivoire. General information about the country data.
`http://www.ins.ci/`

3. Data Preprocessing. The data was collected for 150 days, from December 1, 2011 until April 28, 2012. The original set of Call Detail Records (CDRs) contains 2.5 billion calls and SMS exchanges between around five million users. CDRs have the following standard format: `timestamp, caller_id, callee_id, call_duration, antenna_code`. The customer identifiers were anonymized by Orange Ivory Coast and all subsequent data processing was completed by Orange Labs in Paris.

In order to have a homogeneous data sample, customers that subscribed or resigned from Orange during the observation period have been removed. Additionally, incoming and outgoing calls have been paired in order to eliminate double counts (i.e. an incoming call for an individual is an outgoing call for the correspondent).

The provided datasets contain the geographical positions of cell phone antennas. Orange considers the exact antenna location as sensitive information and therefore the locations have been slightly blurred so as to protect Orange's commercial interests.

For technical reasons, the antenna identifiers are not always available. Instead of removing the corresponding communications, the code -1 was given to antenna with missing identifier. This happens for a significant number of calls (about one in four).

The datasets covers a total of 3600 hours. Due to technical reasons data is sometimes missing in the datasets; missing data covers a total period of about 100 hours.

4. Published Datasets. All datasets are available in Tabulation Separated Values (TSV) plain text format.

4.1. Antenna-to-antenna (SET1). For this dataset, the number of calls as well as the duration of calls between any pair of antennas have been aggregated hour by hour. Calls spanning multiple time slots are considered to be in the time slot they started in. Antennas are uniquely identified by an antenna id and a geographic location. This data is available for the entire observation period. Communication between Orange customers and customers of other providers have been removed.

The antenna-to-antenna traffic data is provided in the files *SET1TSV_0.TSV* to *SET1TSV_9.TSV*. The 10 files each correspond to 14 days. Each line in a TSV file provides the number of calls as well as the total duration of calls between a pair of antennas for a given hour.

The DDL code for this data is:

```
CREATE TABLE H_A_FLOWS (
date_hour TIMESTAMP,
originating_ant INTEGER,
terminating_ant INTEGER,
nb_voice_calls INTEGER,
duration_voice_calls INTEGER
);
```

Example of data:

```
2012-04-28 23:00:00 1236 786 2 96
2012-04-28 23:00:00 1236 804 1 539
2012-04-28 23:00:00 1236 867 3 1778
2012-04-28 23:00:00 1236 939 1 1
2012-04-28 23:00:00 1236 1020 6 108
2012-04-28 23:00:00 1236 1065 1 1047
2012-04-28 23:00:00 1236 1191 1 67
2012-04-28 23:00:00 1236 1236 18 2212
2012-04-28 23:00:00 1237 323 1 636
2012-04-28 23:00:00 1237 710 1 252
```

This first dataset can be visualized with Geofast `www.geofast.net`. Geofast is a web-based tool for the interactive exploration of mobile phone data. The data is aggregated on different administrative levels and users are able to select administrative regions and visualize the amount of communication traffic on selected days.

4.2. Individual Trajectories: High Spatial Resolution Data (SET2). Individual movement trajectories can be approximated from the geographic location of the cell phone antennas during calls. Limited knowledge of an individual's trajectory is often sufficient for identification and the individual can then be traced during the entire observation period. Two obvious solutions to reduce the possibility of identification are to reduce the spatial resolution or to publish trajectories only for limited periods of time. Since long term observation data as well as trajectories with a high spatial resolution have interesting scientific applications, two different datasets are published in order to balance privacy protection and scientific interest.

The first dataset contains high resolution trajectories of $50,000$ randomly sampled individuals over two-week periods. The second dataset contains the trajectories of $50,000$ randomly sampled individuals for the entire observation period but with reduced spatial resolution. We describe the first dataset in this section and the second

dataset in the next section.

The original data has been split into consecutive two-week periods. In each time period, 50,000 of the customers are randomly selected and are assigned anonymized identifiers. To protect privacy new random identifiers are chosen in every time period. Time stamps are rounded to the minute.

This dataset is in the archive **SET2** and contains the files *POS_SAMPLE_0.TSV* to *POS_SAMPLE_9.TSV*.

The DDL code for the data is:

```
CREATE TABLE POS_SAMPLE_0(
user_id INTEGER,
connection_datetime TIMESTAMP,
antenna_id INTEGER
);
```

Example of data in *POS_SAMPLE_0.TSV*:

```
437690 2011-12-10 10:51:00 980
316462 2011-12-10 16:12:00 607
277814 2011-12-10 20:48:00 560
419518 2011-12-10 10:05:00 -1
18945 2011-12-10 11:32:00 401
283750 2011-12-10 10:16:00 10
11813 2011-12-10 10:08:00 970
92418 2011-12-10 21:08:00 -1
287887 2011-12-10 09:48:00 583
```

The coordinates of the antenna positions are given in the files *ANT_POS.TSV*. The DDL code for the data is:

```
   CREATE TABLE ANT_POS(
antenna_id INTEGER,
longitude FLOAT,
latitude FLOAT,
);
```

Example of data in *ANT_POS.TSV*:

```
1 -4.143452 5.342044
2 -3.913602 5.341612
3 -3.967045 5.263331
4 -4.070007 5.451365
5 -3.496235 6.729410
6 -3.485944 6.729422
7 -3.981175 5.273144
8 -3.911705 5.858010
9 -4.014445 5.421120
```

4.3. Individual Trajectories: Long Term Data (SET3). In this dataset, the trajectories of 500,000 randomly selected individuals is provided for the entire observation period but with reduced spatial resolution. The spatial resolution is

reduced by publishing the sub-prefectures of the antennas rather than the antennas' identifiers. The published dataset also contains the geographic center of the sub-prefectures. The 255 sub-prefectures of Ivory Coast along with Orange's cell phone towers are shown in Figure 4.1.

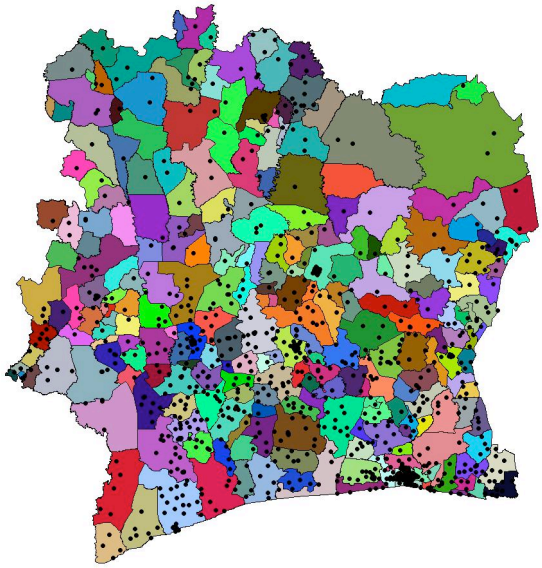


Fig. 4.1: Orange's cell phone towers in Ivory Coast and sub-prefectures administrative regions. Note that some sub-prefectures do not have cell phone towers.

This dataset is in the archive **SET3** and contains the files from *SUBPREF_POS_SAMPLE_A.TSV* to *SUBPREF_POS_SAMPLE_J.TSV*, and the file *SUBPREF_POS_LONLAT.TSV*.

The DDL code for *SUBPREF_POS_LONLAT.TSV* is:

```
   CREATE TABLE SUBPREF_POS_LONLAT(
subpref_id INTEGER,
longitude FLOAT,
latitude FLOAT,
);
```

Example of data in *SUBPREF_POS_LONLAT.TSV*:

```
1 -3.260397 6.906417
2 -3.632290 6.907771
3 -3.397551 6.426104
4 -3.662953 6.660800
5 -3.440788 6.937723
6 -3.291995 6.328551
7 -3.366372 7.182663
8 -3.498494 7.166416
9 -3.149608 7.015214
```

The DDL code for *SUBPREF_POS_SAMPLE.TSV* is:

```
CREATE TABLE SUBPREF_POS_SAMPLE(
user_id INTEGER,
connection_datetime TIMESTAMP,
subpref_id INTEGER
);
```

Example of data in *SUBPREF_POS_SAMPLE_A.TSV*:

```
134931 2011-12-02 10:50:00 60
89571 2011-12-02 10:49:00 39
457232 2011-12-02 16:05:00 60
155864 2011-12-02 09:26:00 60
280671 2011-12-02 13:24:00 -1
13689 2011-12-02 20:34:00 97
171642 2011-12-02 22:36:00 60
247694 2011-12-02 15:11:00 60
376500 2011-12-02 09:49:00 58
294553 2011-12-02 20:45:00 185
```

4.4. Communication Subgraphs (SET4). Our aim with this dataset is to allow the analysis of communication graphs. The dataset contains the communication subgraphs for 5,000 randomly selected individuals (egos). For these individuals, communications within their second order neighborhood have been divided into periods of two weeks spanning the entire observation period. For constructing an ego-centered graph, one consider first and second order neighbors of the ego and communications between all individuals (we do however not include communications between second order neighbors). The anonymized identifiers assigned to the individuals are identical for all time slots but are unique for each subgraph. That is, a customer who is part of the communication graph of two different customers has a different identifier in the two graphs (see Figure 4.2). We therefore have a total of 5,000 connected graphs in every time period. The egos have been given identifiers between 1 and 10,000 and neighbor labelling starts from 20,000.

Phone calls that follow a public phone usage pattern have been excluded from the randomly selected individuals. In Ivory Coast, it is common for some mobile phone owners to provide their phone to people on the street for a fee. This usage is characterized by a large number of outgoing calls but little mobility. We have removed from our selection of egos the customers identified as public phone providers.

The files are in the archive **SET4**. The communication subgraph data is published in the files *GRAPHS_0.TSV* to *GRAPHS_9.TSV*. Each of the files contains the aggregated communication graphs within the second order neighborhood of the randomly selected individuals, divided into two-week periods, starting on December 5, 2011. For every pair of individuals we indicate if there has been a communication between the two, we do not provided the number of communications, total communication time or the direction of the communication.

The DDL code for those data is:

```
CREATE TABLE GRAPHS_0(
source_user_id INTEGER,
```

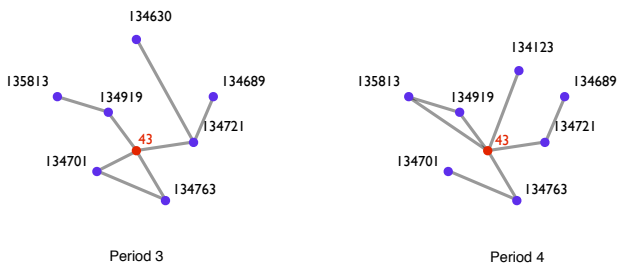


Fig. 4.2: Ego-centered graphs. Identifiers remain unchanged during successive periods and individual appearing in two different ego-centered graphs are given different identifiers.

```
destination_user_id INTEGER,
);
```

Example of data in *GRAPHS_0.TSV*:

```
1052  20002
20002 20022
20018 20019
1052  20019
20019 20030
20019 20031
20129 20119
20132 20119
20134 20119
20102 20135
```

REFERENCES

[1] D. Berry. The computational turn: Thinking about the digital humanities. Culture Machine, 12: 1-22, 2011.
[2] d. boyd and K. Crawford. Six provocations for big data. Computer, 123(1):1-17, 2011.
[3] N. Eagle. Engineering a common good: Fair use of aggregated, anonymized behavioral data. IEEE Engaging Data, Boston, MA 2009.
[4] B. Latour. Tardes idea of quantification. In: M. Candea (ed.) The social after Gabriel Tarde: Debates and assessments. London - New York: Routledge 2010.
[5] D. Lazer, A. Pentland, L. Adamic, S. Aral, A.L. Barabasi, D. Brewer, N. Christakis et al. Computational social science. Science, 323(February):721-723, 2009.

Mobile Communications Reveal the Regional Economy in Côte d'Ivoire

Huina Mao[1,2], Xin Shuai[1,2], Yong-Yeol Ahn[1], Johan Bollen[1],
1 School of Informatics and Computing, Indiana University, Bloomington, Indiana, United States
2 authors made equal contributions.
E-mail: {huinmao,xshuai,yyahn,jbollen}@indiana.edu

Abstract

We investigate the relation between mobile phone usage and regional economic development in Côte d'Ivoire. Most previous studies on mobile phone communication focused on developed countries. Their results may be difficult to extend to developing countries. Here we study mobile communication in a developing nation, namely Côte d'Ivoire. Our main motivations are, first, that the effects of information technology as a tool in reducing poverty are not well understood, and, second, that developing countries often do not have the infrastructure to collect and maintain detailed socio-economic data. To tackle these issues we examine the relationships between mobile communication patterns and economic development. We first define several indicators of mobile phone usage and analyze their correlations with available indicators of economic status. We then examine communication patterns between rich and poor areas. We find that mobile communication data provides an accurate and detailed picture of economic development in the country. Rich areas communicate to a great degree with each other and split into distinct smaller communities, whereas poor areas do not exhibit internal structure so much but tend to merge into one large community. Mobile communication data may provide a cost-effective way to analyze the current status of economic development within a nation, thereby allowing governments and aid organizations to respond swiftly and effectively to changing conditions om the ground.

1 Introduction

Economic inequality is a global problem. As part of their efforts to alleviate poverty and support socio-economic development many governments have focused on extending the adoption of telecommunication technologys. Telecommunication technology, in particular mobile telephony and land lines, are expected to provide access to crucial information, for example with regards to job opportunity, that the poor people can leverage to facilitate their social and economic development.

By the end of 2011, worldwide penetration of mobile devices exceeded 87%, equivalent to 6 billion mobile subscriptions. Mobile networks are generating data of unprecedented scale, which not only benefits mobile service providers, but can support advances in the behavioral and social sciences, and economics. Existing literature in sociology and economics suggests that social structure affects economic development. For instance, UK call data revealed that the diversity of social communication network of a region is strongly correlated with its income [1], a result that hints at the benefit of weak ties [2] and diversity [3,4]. Recent finding also suggests that as the size of city grows, communication within the city grows in a super-linear fashion [5,6].

Mobile communication data has enabled a plethora of intriguing studies like the ones referenced above. Unfortunately, most of this work is focused on developed countries, possibly due to the lack of relevant mobile and socio-economic data for developing nations. It is thus unclear whether existing findings in this domain can be generalized to developing nations, defined by the World Bank as those with "low income" and "lower-middle income". This is particularly unfortunate since previous work suggests that wider mobile phone penetration can boost the GDP of developing countries [7,8]. In addition, the economic impact of mobile technology is more pronounced in developing countries compared to developed countries[1], because they have seen faster increases of mobile telephony adoption rates in recent years.

[1] `http://www.ictregulationtoolkit.org/en/Document.3532.pdf`

To investigate the impact of mobile phone technology on developing countries, we study large-scale mobile communication data from Côte d'Ivoire, whose income level is at the lower-middle level in the world[2]. Half of the country's population is poor. There is a strong divide between rich coastal regions (south) and poor ones (north). Its economic capital, Abidjan, further distinguishes itself from other areas. Its annual per capita income is about three times higher than the north [9]. Inside Abidjan, there is another clear divide between rich and poor communities, ranging from the wealthiest neighborhood in Cocody to the slums of Adjamé.

Detailed socio-economic indicators for Côte d'Ivoire are scarce. The best economic data we could find is at the level of 10 development poles[3]. However, in Côte d'Ivoire, the mobile phone penetration was around 79%. The Internet penetration rate is however rather low, i.e. 6% in 2010. We are interested in whether mobile communication data can fill the gaps in Côte d'Ivoire economic data, enabling researchers and governments to study the status of the Côte d'Ivoire economy with existing mobile infrastructure. The main contributions and findings of our work include:

- We develop a variety of indicators from mobile data, and correlate these with social-economic statistics. We find that *CallRank* (i.e. PageRank) of mobile communication networks can help us to identify economic centers at both the national and city scale. In addition, the degree to which an area initiating mobile communications with other regions is highly correlated to an area's local annual income and poverty rate.

- We examine the scaling properties of mobile data indicators with respect to the population. We find that mobile communication activity scales super-linearly with population, indicating a "digital divide" effect in which mobile phone usage is overly concentrated in rich areas and populous areas.

- We apply the Louvain method to identify communities in a map of Côte d'Ivoire based on mobile activity, and then compare this map to administrative boundaries. We find that rich areas split into distinct small communities, while poor areas do not show such structure within but rather tend to merge into one large community.

- We investigate the "rich-club" effect in the communication graph. We find that rich areas are much more likely to communicate with other rich areas than poor areas. A "rich-club" is observed in the South and South West areas of Côte d'Ivoire, and it is isolated from other poor areas like the North and West areas.

2 Background Information about Côte d'Ivoire

2.1 Administration and Economic Development

Côte d'Ivoire, one of the sub-Saharan countries in Africa, is located in Western Africa. Its administrative structure consists of 19 regions and 81 departments, which are subdivided into 255 sub-prefectures [4]. Yamoussoukro is the official political capital, while Abidjan is the economic capital. Table 1 lists the top 10 biggest cities in Côte d'Ivoire, ranked by population. In Figure 1, these ten cities are marked in red in the map by their corresponding rank IDs. The total population of the nation is about 20 million individuals in 2011 according to World Bank. One fourth of the population lives in urban cities, namely Abidjan, Abobo, and Bouaké.

Côte d'Ivoire heavily depends on agriculture, and is the world's largest producer and exporter of cocoa. By 2012, its GDP per capita ranked 199^{th} in the world and it can thus be considered a relatively

[2] `http://data.worldbank.org/country/cote-divoire`

[3] `http://www.imf.org/external/pubs/ft/scr/2009/cr09156.pdf`

[4] According to 1998 Census, Côte d'Ivoire has 19 divisions at the region level, 50 at the department level, and 185 at the sub-prefecture level. The map with these divisions is used in our paper.

Table 1: Top 10 Major Cities in Côte d'Ivoire

Rank	City Name	Region	Population
1	Abidjan	Lagunes	3,677,115
2	Abobo	Lagunes	900,000
3	Bouaké	Vallée du Bandama	567,481
4	Daloa	Haut-Sassandra	215,652
5	San-Pédro	Bas-Sassandra	196,751
6	Yamoussoukro	Lacs	194,530
7	Korhogo	Savanes	167,359
8	Man	Dix-Huit Montagnes	139,341
9	Divo	Sud-Bandama	127,867
10	Gagnoa	Fromager	123,184

Table 2: Regions and Economic Indicators of 10 Development Poles in Côte d'Ivoire

Development Pole	Capital	Regions	Poverty Rate	Annual Income⋆
City of Abidjan	Abidjan	Abidjan	21.0	561,575
Center-North	Bouaké	Vallée du Bandama	57.0	281,660
Center-West	Daloa	Haut-Sassandra,Marahoué,Fromager	62.9	243,236
North-East	Bondoukou	Zanzan	54.7	301,966
North	Korhogo	Savanes	77.3	191,540
West	Man	Dix-Huit Montagnes,Moyen-Cavally	63.2	256,319
South	Abidjan	Sud-Bandama,Lagunes,Agnéby,Sud-Comoé	44.6	334,147
South-West	San Pédro	Bas-Sassandra	45.5	348,257
Center	Yamoussoukro	N'zi-Comoé,Lacs	56	287,080
Center-East	Abengourou	Moyen-Comoé	53.7	289,126
North-West	Odienné	Bafing,Denguélé,Worodougou	57.9	284,393

⋆Annual Income is in Central African CFA frac (CFAF).

poor country[5]. Also, the economic development is highly uneven across the country. The whole country is divided into 10 development poles based on its regional economic level [9] . Figure 1 shows the annual income distribution, where darker color indicates worse poverty. Table 2 lists the development poles and their corresponding economic indicators inculding poverty rate and annual average per capita income. From Figure 1, we observe a clear divide between the north and the south, where Northern regions are poorer and Southern regions are richer. Specifically, the richest areas include the South-West and South, whose annual average per capita income is over CFAF 334,000, while the poorest areas contains Centre-North, the West, the North-West, the Centre-West and the North, whose annual average per capita income is between CFAF 191,540 and CFAF 284,393. The rest of areas at the middle level are the Centre-East, the North-East, and the Centre (see details from [9]).

Over 3.6 million people live in the economic capital, Abidjan. The annual average per capita income of Abidjan is the highest in the country, i.e. about three times that of the North which is the poorest. Abidjan accounts for about 50% of the Côte d'Ivoire GDP. Given the important role of Abidjan, we also drill down into the city and analyze data for its communities ("communes"). The District of Abidjan is inside the Lagunes region, which consists of Abidjan-Ville and three external sub-prefectures, including Anyama, Bingerville, and Songon. The ten communities inside Abidjan-Ville are Abobo, Adjamé, Attécoubé, Cocody, Koumassi, Marcory, Plateau, Port-Bouët, Treichville, and Yopougon. The map structure of Abidjan is shown in Figure 2.

[5] https://www.cia.gov/library/publications/the-world-factbook/geos/iv.html

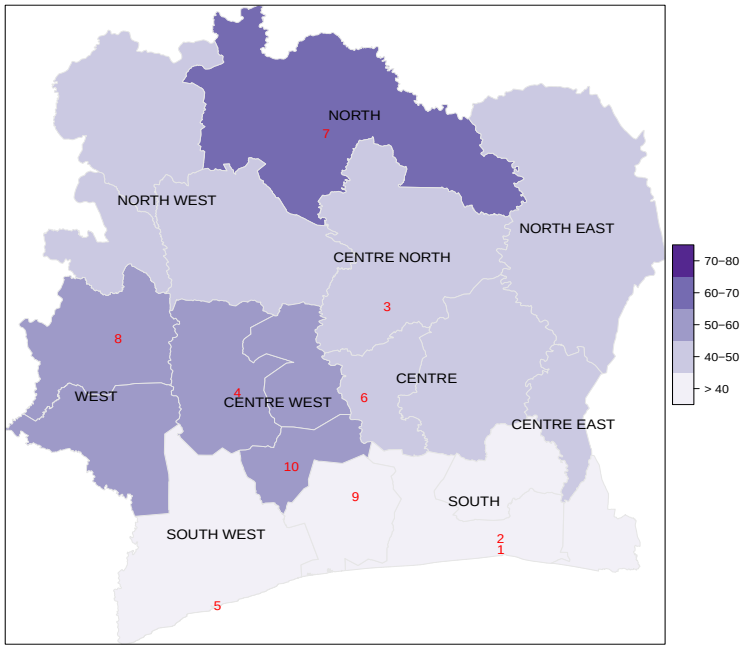


Figure 1: Ten Development Poles in Côte d'Ivoire.
1: Abidjan; 2: Abobo; 3:Bouaké; 4:Daloa; 5:San-Pédro; 6: Yamoussoukro; 7: Korhogo; 8: Man; 9: Divo; 10: Gagnoa

Although Abidjan is the richest area in Côte d'Ivoire, it is characterized by the income inequalities. As noted by a 1997 report [10]: "37% of households in Abidjan earned 80% of total income, and 40% of households earned under 7%. 14% live in slums and 58% live in courtyard dwellings." Cocody is the wealthiest commune in Abidjan. Plateau is the business district and central government area, and most of its residents are whites. Many slums are distributed in Adjamé, and Marcory and Treichville are also poor areas. Yopougon and Abobo are the largest and second largest communes. Yopougon is most populous.

2.2 Mobile Communication In Côte d'Ivoire

Côte d'Ivoire has millions of mobile subscriptions. The national mobile penetration rate is about 78%. which provide us a large amount of data for our socio-economic study on Côte d'Ivoire with respect to its telecommunication infrastructure.

Our mobile phone data is recorded for Orange consumers and provided through the Data for Development("D4D") Challenge. Orange has about five million customers in Côte d'Ivoire which is about one

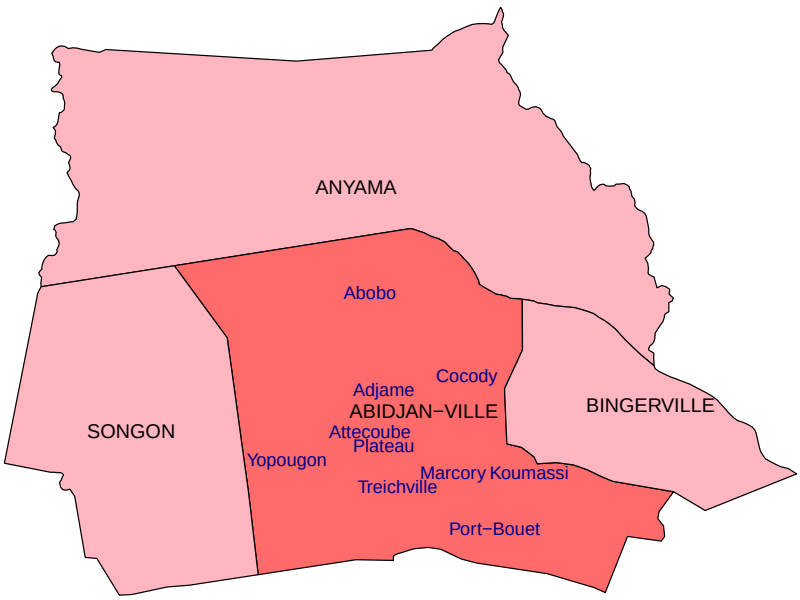


Figure 2: Capital City, Abidjan.

third of the mobile market. It is a significant sample of all Côte d'Ivoire mobile users that we assume is not a strongly biased, lacking data to the contrary.

The data includes about 2.5 billion calls and SMS exchanges. The sample period is from December 1, 2011 and April 28, 2012. There are four different data sets released by the challenge, including (1) hourly antenna-to-antenna traffic, (2) individual trajectories for 50,000 customers for two week time window with the resolution of individual antenna, (3) individual trajectories for 50,000 customers over the entire observation period with the resolution of sub-prefecture, and (4) a sample of communication graphs for 5,000 customers. In our project, we mainly use the first and second data sets as well as the geographical information (latitudes and longitudes) of the antennae and sub-prefectures.

Côte d'Ivoire has 1,231 unique cellular telephone towers, which are distributed across its 50 departments (see Figure 3). The Pearson's correlation between the department's number of towers and its population is significantly positive ($\gamma = 0.87, p \ll 0.01$), but is not necessarily highly correlated with the geological area ($\gamma = 0.31, p = 0.03$). In addition, we observe that the tower density is biased by the regional economy, i.e. there are many more towers in the southern part (rich area and cities) than the northern areas. Specifically, the South and South West has 695 towers (i.e. 56% of all the Orange antennae), while the North only has 46 antennae. The Bafing region in the North West (a poor area) has the least number of towers, i.e. 9. Likewise, among 396 towers in the capital city, Abidjan, Cocody, the wealthiest region, has the largest number of towers (96) and Adjamé, a poor commune, 20 towers. Among these three sub-prefectures outside Abidjan-Ville, there are only 9 towers, even though their areas

are large.

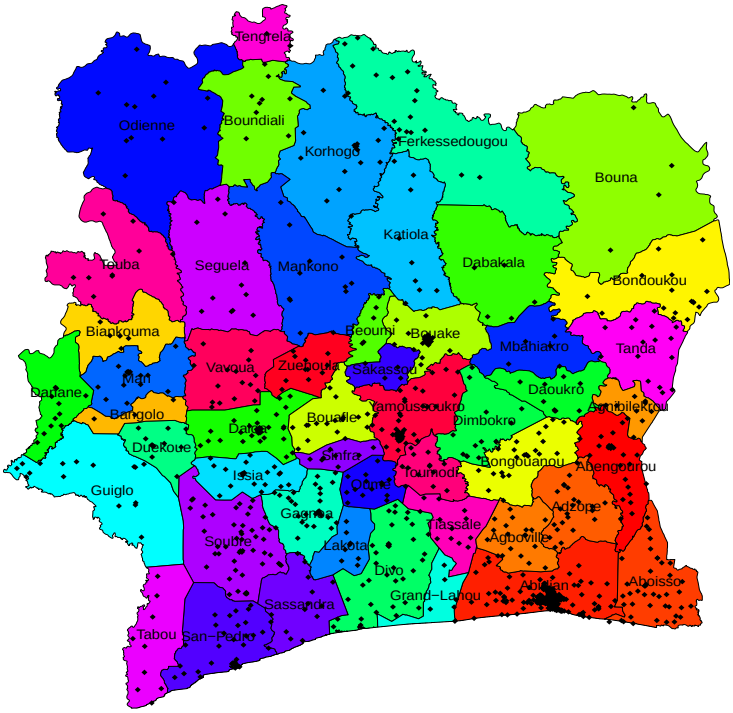


Figure 3: The Distribution of the Orange Cell Phone Towers over 50 Departments in Côte d'Ivoire.

It is not surprising to find an unequal distribution of mobile phone antenna towers among the rich and poor areas. This often observed (and lamented) stark difference between the access to Information and Communication Technology between different socio-economic groups is often referred to as the "digital divide". A digital divide reflects underlying economic development levels, which may further exacerbate economic equality and poverty. One of the important goals of supporting the adoption of new communication technology, such as mobile phones, is to mitigate the digital divide. It is therefore important to understand the relations between mobile phone usage and economic development, and the communication patterns among different socio-economic communities, especially with regards to wealth and poverty levels. The rest of the paper focuses on these issues.

2.3 Network Construction and Terminology

All of our analysis start from constructing a network that represents our mobile data. Assume that $A = [a_1, a_2, ..., a_n]$ denotes all antenna, $R = [r_1, r_2, ..., r_m]$ denotes all regions, $D = [d_1, d_2, ..., d_l]$ denotes all departments, and $P = [p_1, p_2, ..., p_k]$ denotes all development poles in Côte d'Ivoire. As we mentioned

before, $n = 1,231$, $m = 19$, $l = 50$ and $k = 10$. We construct two types of networks, i.e. calling record and trajectory record, on four different scales, i.e. antenna-level, department-level, region-level, and pole-level.

Calling Record Network: In Dataset 1, the number of calls as well as the duration of calls between any pair of antennae in 5 months period were aggregated hour by hour. We construct three types of networks according to different weighting schema at the antenna level. $G_a^n = (A, N_a)$ is a directed weighted network where $N_a = \{n(a_i, a_j)\}$ denotes the total number of calls between a_i and a_j. $G_a^d = (A, D_a)$ is also a directed weighted network where $D_a = \{d(a_i, a_j)\}$ denotes the total duration of calls between a_i and a_j. We further map antennae into different levels of administrative areas based on their geo-location and aggregate the mobile communication flow (in terms of numbers or duration of calls) between pairs of administrative areas. For instance, $G_d^n = (D, N_d)$ is a directed weighted network of department-to-department communication, where $N_d = \{n(d_i, d_j) = \sum_{(a_u,a_v):a_u \in d_i \wedge a_v \in d_j} n(a_u, a_v)\}$ denotes the total number of calls between department d_i and d_j, aggregated from their antenna records. Here the mobile communication within the same department is ignored. Similarly, we generate G_r^n, G_p^n, G_d^d, G_r^d and G_p^d, respectively.

Trajectory Record Network: In Dataset 2, individual movement trajectories are approximated by the geographic location of the cell phone antennae during calls. We define a movement trajectory of user u_i given a period of time as a sequence of antennae its mobile phone connected to $S(u_i) = [a_{s_1}, a_{s_2}, ..., a_{s_n}]$. Then for each pair $(a_{s_i}, a_{s_{i+1}})$ we build an edge from a_{s_i} to $a_{s_{i+1}}$ if $a_{s_i} \neq a_{s_{i+1}}$ and use $w(a_{s_i}, a_{s_{i+1}})$ to count the frequency of such movements in $S(u_i)$. Finally, we obtain a weighted directed trajectory record network $G_a^t = (A, T_a)$, where $T_a = \sum_{u \in U} w(a_{s_i}, a_{s_{i+1}})$ denotes the collective trajectories of all users $U = [u_1, u_2, ..., u_z]$. Especially, $z = 50,000 * 2 * 5 = 500,000$ (there may be overlapping users, but it does not affect our collective trajectory), since 50,000 users are randomly sampled for each bi-week period during the whole five months period. Similar to the calling record network, we can aggregate antennae into different level of administrative areas and construct G_d^t, G_r^t and G_p^t respectively.

In summary, we construct twelve networks, i.e calling record networks weighted by number of calls, calling record networks weighted by duration of calls, and trajectory record networks weighted by number of adjacent transitions, based on antenna-level, department level, region level and pole level. In addition, we introduce a special *Abidjan Mobile Network* G_c^* which aggregates the flow of antenna-to-antenna at the commune level in the capital city, Abidjan and is used for capital case study. All these networks provide rich information about mobile communication patterns in Côte d'Ivoire and will be selectively used in the following analysis.

3 Mobile Communication and Economy

In this section, we first visualize the flow on G_a^n, G_a^d and G_a^t; then develop and calculate several network indicators from G_p^n, G_p^d and G_p^t, and correlate them with real social-economic indicators. Finally, we examine the scaling properties of phone calls from G_d^n, G_d^d and G_d^t with respect to the population size.

3.1 Visualizing flow of calls

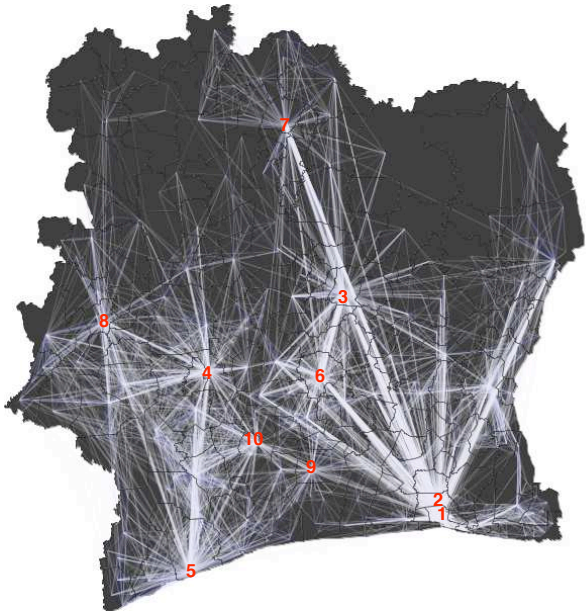
In the visualization, the brightness of the edges indicates the weight defined by the number or duration of phone calls in G_a^n, G_a^d, as well as the flux of people between connected nodes in G_a^t. To improve visualization clarity, we remove low-weighted edges.

First, we normalize the weight of each edge by:

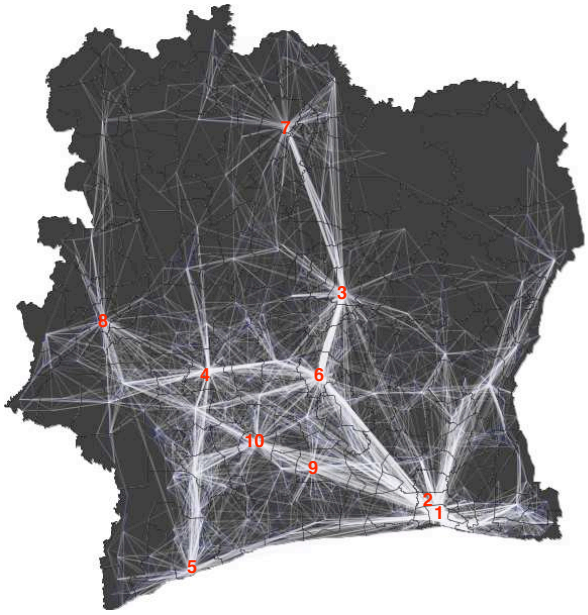
$$W_n = \frac{W_r - W_{min}}{W_{max} - W_{min}} \tag{1}$$

where W_n is the normalized weight, W_r is the raw weight, W_{max} and W_{min} are the maximum and minimum value of all weights. We set up a threshold ξ and filter all edges with $W_n < \xi$. We adjust ξ

and finally set $\xi = 0.001$ for G_a^n, $\xi = 0.01$ for G_a^d and $\xi = 0.0003$ for G_a^t.



(a) Mobile Communication



(b) Human Mobility

Figure 4: Networks of Mobile Phone Communication and Human Mobility.
1: Abidjan; 2: Abobo; 3:Bouaké; 4:Daloa; 5:San-Pédro; 6: Yamoussoukro; 7: Korhogo; 8: Man; 9: Divo; 10: Gagnoa

The flow of G_a^n is shown in Figure 4a, where the brightest part is Abidjan, the economic capital. The southern part is much brighter than the northern part. The North looks dark, except for its capital city,

Regions	*CallRank*	No. Towers
Lagunes	0.2658	463
Bas-Sassandra	0.0996	106
Haut-Sassandra	0.0717	68
Lacs	0.0547	56
Vallée du Bandama	0.0531	49
Savanes	0.0459	46
Moyen-Comoé	0.0430	35
N'zi-Comoé	0.0427	50
Agnéby	0.0366	42
Dix-Huit Montagnes	0.0362	34
Zanzan	0.0352	36
Marahoué	0.0348	37
Sud-Bandama	0.0347	45
Fromager	0.0346	44
Sud-Comoé	0.0334	39
Moyen-Cavally	0.0238	35
Worodougou	0.0231	24
Denguélé	0.0171	12
Bafing	0.0141	9

Table 3: *CallRank* of 19 Regions of Côte d'Ivoire

Communes	*CallRank*	No. Towers
Cocody	0.0827	96
Yopougon	0.0807	78
Abobo	0.0792	45
Adjame	0.0773	20
Plateau	0.0755	35
Attecoube	0.0727	11
Treichville	0.0727	19
Port-Bouët	0.0705	26
Marcory	0.0704	31
Koumassi	0.0698	17
Anyama	0.0658	5
Songon-Agban	0.0589	3
Bingerville	0.0530	1

Table 4: *CallRank* of 13 Communities of Capital City, Abidjan in Côte d'Ivoire

Korhogo. The North West and North-East is the darkest in the map, which may be consistent with their lower level of the economic development, and fewer economic activities being involved[6]. Furthermore, most of the hub nodes (identified as shining points in the map) are major cities given in Table 1. We mark these ten major cities in Figure 4a. In addition, some smaller hub nodes other than ten major cities are identified as well. This finding may tell us that monitoring mobile phone communications is a good way to identify important cities and find cities that play an increasingly important role in communications. The network flow of G_a^d is actually very similar to G_a^n so we don't include the visualization here.

The human mobility flow of G_a^t is shown in Figure 4b. Comparing Figure 4a with Figure 4b, the most evident pattern we observe is that mobile phones facilitate communication between the north and south where geological distance between the two poles is farthest. Specifically, there are a large number of calls between Abidjan/Abobo and Korhogo. Abidjan is the major economic and trade center, while Korhogo is an important producer of agricultural goods. The frequency of mobile communications between the regions may thus be reflective of their economic ties. In addition, mobile communications between the South and the East are stronger than their trajectory connection. Hence, overall human mobility is more localized while mobile phones greatly extend communication between areas that are separated by large geological distances.

3.2 Extract Economic Indicators from Mobile Phone Calls

3.2.1 CallRank and economic activity

PageRank is the most widely used indicator in network analysis to measure the importance of nodes. We introduce an indicator named *CallRank* to measure the importance of a region based on the mobile phone communication graph by running weighted PageRank algorithm. The obtained *CallRank* values for G_a^n on 19 regions and for G_c^n on 13 communes in Abidjan, are listed in Table 3 and Table 4, respectively.

[6] `http://www.fao.org/ag/AGP/AGPC/doc/Counprof/Ivorycoast/figure4.htm`

We hypothesize that the *CallRank* score is reflective of the economic importance of an area. From Table 3, we find that the *CallRank* value (0.2658) of the region Lagunes is higher than that of any other region. Actually, the economic capital Abidjan, and the second largest city, Abobo, are both located in Lagunes. Table 2 shows that average income per capita is highest in Lagunes. The region Bas-Sassandra has the second largest number of towers and *CallRank* score, which is located in the South-West development pole whose annual average per capita income is the second highest after the South. Bas-Sassandra's capital, San-Pédro, is the top 5 largest city in the nation and the second largest port after Abidjan. The region Haut-Sassandra has the third largest *CallRank*. However, according to Section 2, Haut-Sassandra is located in the Centre-West which is one of the poorest areas in the nation. However, it is the main cacao producing area. Its capital Daloa is an important trading center, particularly for cocoa. The western region of Daloa is responsible for a quarter of the Côte d'Ivoire's national output. Therefore, *CallRank* seems to reveal the importance of regions in the national economy rather than the economic development of the regions. The regions with the lowest *CallRank* values, Worodougou, Denguélé and Bafing are located in the North-West, which are all the poorest areas, and do not show characteristic economic functionality.

Similarly, we calculate the *CallRank* scores for Abidjan consisting of 10 communes in the Abidjan-Ville and three sub-prefectures outside. The results of G_c^n and G_c^d are very similar, especially the top three and bottom three regions are the same. Their *CallRank* score matches the real economic situation well, since Cocody is the richest commune and Yopougon is the most populous commune; they are the highest ranked in Table 4. By contrast, the three sub-prefectures that have the lowest *CallRank* score because few mobile connections and economic activities occur there.

3.2.2 Correlation between mobile indicator and social-economic statistics

Although we intuitively observe a qualitative relationship between *CallRank* and economic activity, we seek to find out useful mobile indicators that exhibit quantitative correlations with real socio-economic factors. Such correlations would be especially useful for developing countries who cannot afford the high cost of a national census or large-scale surveys, but nevertheless seek to monitor regional economic development in an timely manner.

There exist severe limitations to collect fine-grained macro-economic indicators in poorer countries. The best economic indicators of Côte d'Ivoire we have obtained pertain only to the ten development poles reported in [9]. In addition to the total average annual per capita income and poverty rate that we provide in Table 2, other economic indicators include annual average per capita income and poverty rate measured in the urban and rural areas, the ratios of former to the latter, as well as Gini index[7]. Correspondingly, the mobile data indicator are extracted from G_p^n, G_p^d and G_p^t and shown in Table 5.

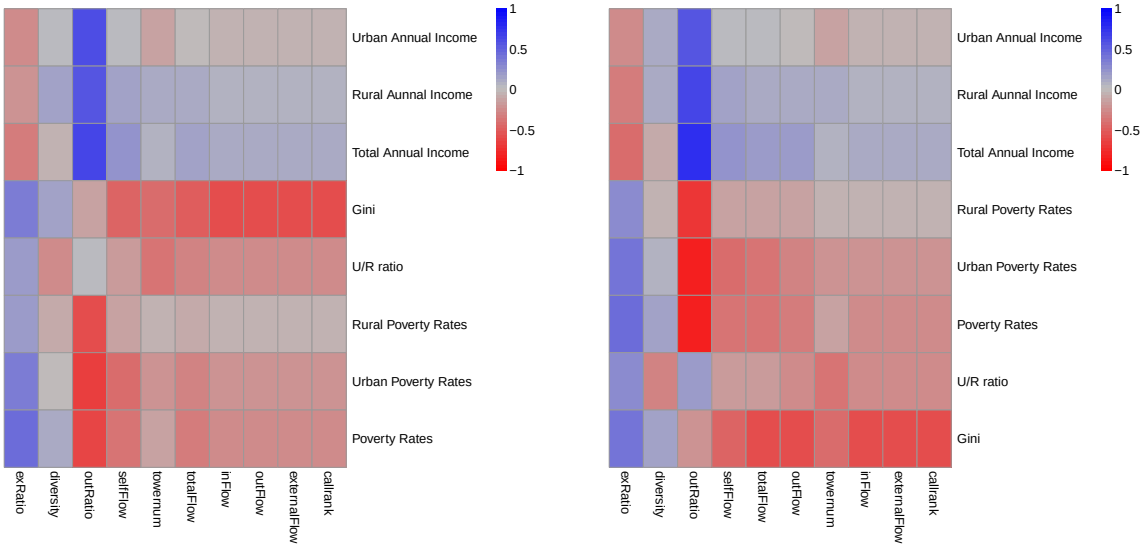
Figure 5 shows that Spearman rank correlation between of mobile network indicators and social-economic statistics. Since we only have ten data points (development poles), most correlations do not show statistical significance. The correlations between G_p^n network measures and economic statistics (Figure 5a) show that only outRatio vs. Annual Income (0.63), and outRatio vs. Urban Poverty Rates (-0.66) exhibit statistically significant correlations with p-values less than 0.05. In addition, outRatio calculated from G_p^d shows an even stronger correlation with Annual Income (0.80) and Poverty Rate (-0.83). It is rather surprising to see that the directionality of calls has a strong power of inference with respect to income level. It may be the case that rich areas have greater opportunities or means to initiate calls to other areas than to receive calls from other areas. This is possibly as results of their leading or commanding economic positions.

Subsequently, we correlate indicators from G_p^t (i.e. trajectory record based network) with the same regional economic statistics, but did not find any significant correlations between these indicators and economic statistics. We thus conclude that human mobility is less shaped by economic factors than

[7]The Gini coefficient measures the economic inequality in terms of income distribution.

Table 5: Mobile network indicators from calling record and trajectory record networks on pole level

Mobile Indicators	Description
inFlow	the weighted in-degree
outFlow	the weighted out-degree
selfFlow	the sum of weights from antenna-to-antenna edges within the pole
externalFlow	inFlow + outFlow
totalFlow	selfFlow + externalFlow
exRatio	externalFlow / totalFlow
outRatio	outFlow / externalFlow
towernum	number of towers
callRank	PageRank
diversity	a measure of normalized entropy of communication adopted from [1]



(a) No. of call weighted network vs. economy (b) Duration of phone call weighted network vs. economy

Figure 5: Correlation between Economic Indicators and Mobile Network Measures at the Pole Level

indicated by mobile communications.

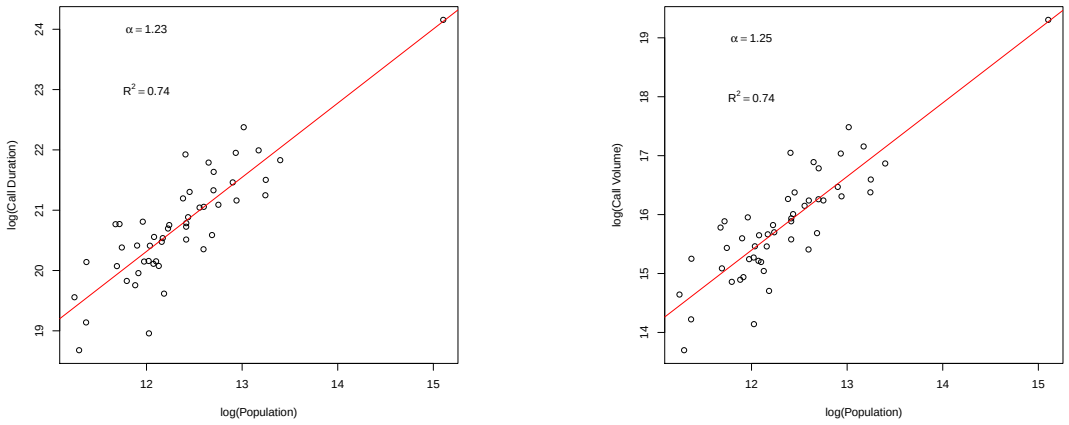
3.3 Population scaling for calling activity

Previous research [11] studied the scaling relationships of a wide range of urban socio-economic indicators with respect to city size as measured by population level. Notably, it is reported that phone call time scales superlinearly with population [6] in US mobile phone calls.

To explore how mobile phone call volume scales with socio-economic indicators, we first examine the relationship of call activity with respect to the population of departments,

$$Y \approx c \times N^{\alpha} \tag{2}$$

where Y is any aggregated measures and N is the population. The power law exponent α is categorized into three categories: linear ($\alpha = 1$), sub-linear ($\alpha < 1$), and super-linear $\alpha > 1$ [11].



(a) Duration of Calls Against Population

(b) Number of Calls Against Population

Figure 6: The Population Scaling of Mobile Phone Calls over 50 Departments of Côte d'Ivoire

Figure 6 shows the linear regression on the log-log scale.

We find that the scaling exponents for call duration (weighted degree of G_d^n) and call frequency (weighted degree of G_d^d) are 1.23 and 1.25, respectively, indicating super-linear scaling with respect to departmental population. Compared with the scaling exponent identified from the United States Call Details Records (CDR) (1.14), the scaling exponent obtained here is higher. This may indicate a stronger "digital divide" effect in developing nations vs. developed ones, since people living in rich areas (i.e. populous areas) use mobile phone much more frequently than poor areas, possibly due to the high cost of technology use in developing country.

We also zoom in and study the deviation of each observation from the regression line. In general, we find those departments with high positive deviations (i.e. those above the regression line) are mainly locate in the rich areas, such as the South and South-West; While the departments that are under the regression lines are more likely to be distributed in the poor areas, such as West, North and North-West.

4 Rich-club and Digital Divide

4.1 Rich-club phenomenon

The rich-club coefficient measures the degree of connectivity among "rich" nodes in a network, quantifying the strength of the "rich club" effect. Here, the "richness" of a node can have various definitions, such as degree of a node, centrality of a node, or other user-defined measure. The classic rich-club coefficient calculation only applies to unweighted network but not suitable in our case, because the call volume is a very important indicator for connectivity between different regions. Here we adopt weighted rich-club coefficient proposed by [12].

Every node has a richness parameter r. For each value of r, a club that consists of all nodes with richness larger than r is formed. For each of these clubs, $E_{>r}$, the number of links connecting the members, and $W_{>r}$, the sum of the weights attached to these links are measured. Then the ratio $w(r)$ between and the sum of the weights attached to the $E_{>r}$ strongest links within the whole network is

calculated as follows

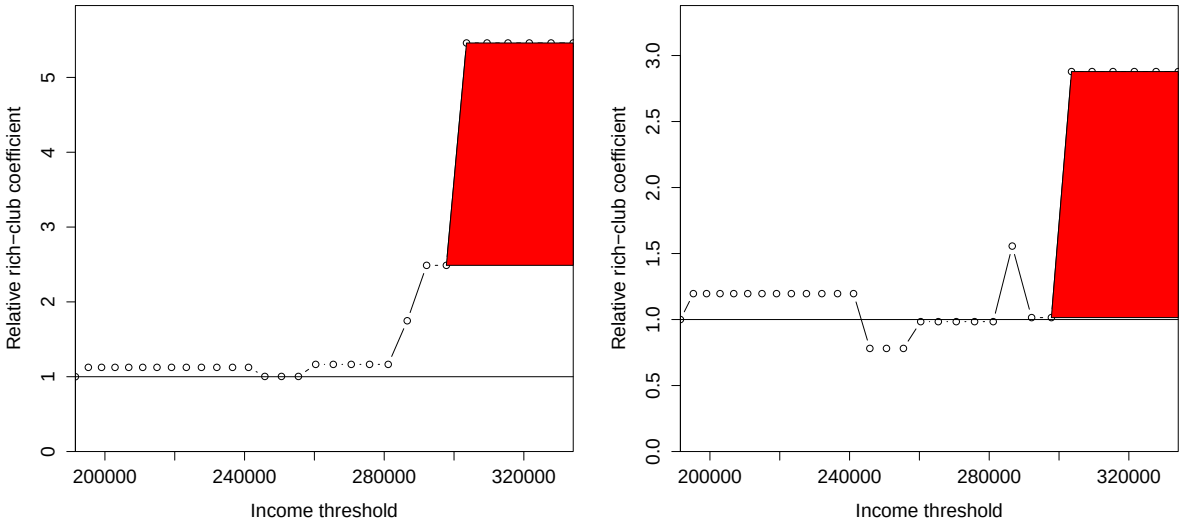
$$\phi^w(r) = \frac{W_{>r}}{\sum_{l=1}^{E_{>r}} w_l^{rank}} \tag{3}$$

where $w_l^{rank} \geqslant w_{l+1}^{rank}$ with $l = 1, 2, ..., E$ are the ranked weights of links on the network and E is the total number of links. To account for the factor that even random networks can exhibit a baseline degree of rich-club effect, the null model is generated by randomizing the original network while preerving its degree distribution. The final define rich-club coefficient is defined as:

$$\rho^w(r) = \frac{\phi^w(r)}{\phi_{null}^w(r)} \tag{4}$$

where $\rho^w(r)$ refers to the weighted rich-club effect as assessed vs. the appropriate null model. When $\rho^w(r)$ is larger than one, the observed rich-club coefficient in original network is larger rather than expected from the random null-model.

We calculate the rich-club coefficient for G_p^n, G_p^d and G_p^t against different rich levels measured by the annual income of ten poles. The results of G_p^n and G_p^d are actually very similar, so we only show G_p^n in Figure 7a and G_p^t in Figure 7b.



(a) Rich-club in Mobile Communication

(b) Rich-club in Human Mobility

Figure 7: Rich-club Effect. The red areas show that the "rich-club" consists of poles with higher than CFAF 300,000 annual average per capita income

We clearly observe a rich-club effect for higher levels of annual income in our calling network and mobility network. Specifically, there are sudden increase in rich-club coefficient when the income level is above CFAF 300,000. Looking at Table 2, we find that only three poles have entered the CFAF 300,000 level "rich-club", i.e. South, South-West and North-East. This implies that the mobile communication or human transition activities among these poles dominates that of the nation.

4.2 Communication-induced communities based on mobile phone data

Communication data allows us to draw new boundaries between communities inferred by communication patterns, rather than those based on historical and administrative boundaries. In line with previous work that re-draw country maps based on the mobile communication data of the United Kingdom [13], Belgium [14], and the United States [15], we adopt the "Louvain method" [16] to perform community detection on G_a^n, G_a^d and G_a^t. We aim to answer the following questions. First, do people have stronger connections (either through mobile phones or travels) with people that are also in a developing country? Second, what are the differences between the communication patterns of rich and poor regions?

From the call network weighted by the number of calls we detected 18 communities in G_a^n which are shown in Figure 8a. Different colors represent different communities. The black borders represent administrative region boundaries. White lines represent sub-prefecture boundaries. Each prefecture area is colored by the community color of its majority antennae. The prefectures without antenna are left blank. Consistent with previous results in [13–15], communication-based communities are well mapped into geographic space. Four largest communities are spotted. The first (in green) covers the four regions in the West; the second (in yellow) mainly covers the two regions in the Centre East; the third is in the North area, and the last one is in the East. They are all poor areas. By contrast, the South and South-West, the richest areas, contain many diverse small communities. Moreover, several communities are formed across the adjacent borders of multiple regions.

Overall, the rich areas tend to further split into smaller inner communities while poor areas tend to merge into a large community. Several explanations can be drawn: (1) rich areas consist of heterogeneous sets of people that separate along gaps in socio-economic status, which form smaller yet stronger "cliques"; (2) poor areas consists of relatively homogeneous sets of people with universal low social status among which communication barriers are not as clearly defined as for rich areas; (3) communities in poor areas are too small to be detected by our algorithm because their communication activity is well below that of rich areas.

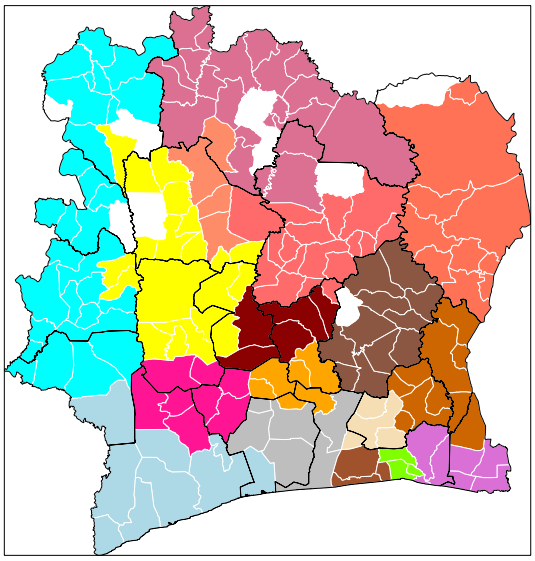
Figure 8b shows that only detect seven communities are found on G_a^d, whose number of communities is much less than G_a^n. Clearly, the nation is overall divided in the four big parts, consisting of three big communities in the West, Center, East, and several small communities in the South. The richest areas in Côte d'Ivoire are in the South-West and South, while other regions are in the poorest or middle level. This result may again suggest the heterogeneous composition or much higher level of communication in the rich areas, and the relative homogeneity or low level of communication in the poorer areas.

The same detection method is applied to G_a^t, i.e. the human trajectory based network. Here, we obtain 28 communities. The cellphone towers in the same community are highly influenced by the geological distance. Due to space limitation, we can not include this figure in our paper.

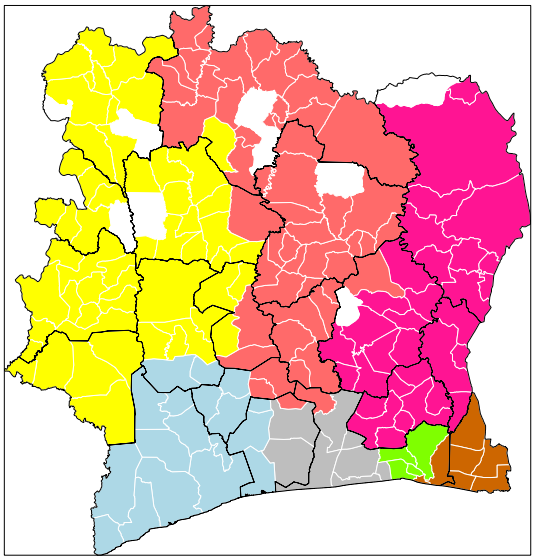
Additionally, given that a large population of approximately 3.6 million individuals live in the capital city, Abidjan, and many disparities exist among its communes, we also conduct community detection based on the call communication and human trajectory data from Abidjan *only*. Three networks are constructed from these data, including two Call Networks weighted by the call frequency (G_c^n) and duration time (G_c^n) and Human Trajectory Network (G_c^t). The community detection results are shown in Figure 9. All the antenna towers (in total 396) are located by their latitude and longitude in the graph. The center of the each commune is geo-located and represented by an ID ranging from 1-10[8]. An antenna with the same color indicates that it is assigned to the same community based on call and trajectory networks.

Based on the call frequency weighted network G_c^n, we detect eight communities as shown in Figure 9a, whereas only five communities are detected from call duration weighted networks with several adjacent communities merged further as shown in Figure 9b. Nine communities are found based on human trajectory records (Figure 9c), which are similar to the community group detected from the call frequency

[8] At the city level, no commune boundaries are found to divide the map further.



(a) Call Frequency

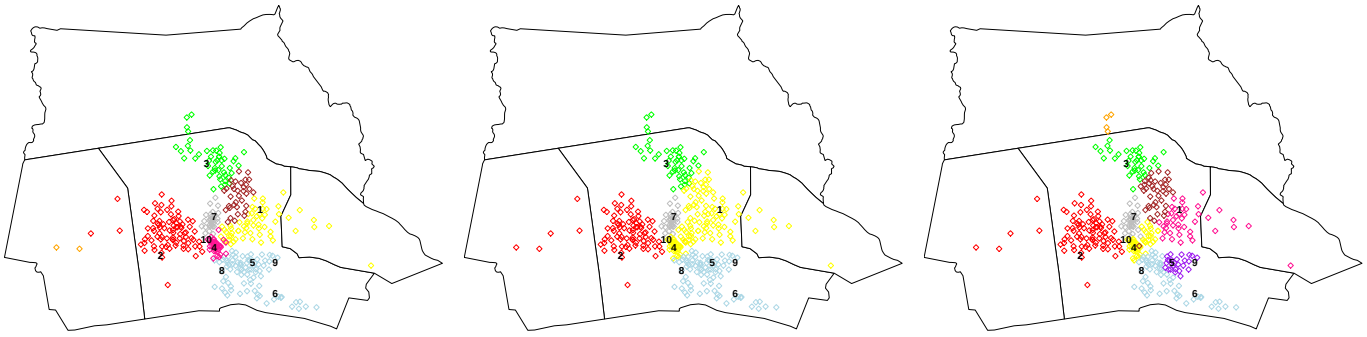


(b) Call Duration

Figure 8: Community Boundaries Drawn from the Mobile Phone Call Frequency and Duration

weighted network, with the exception of certain big communities that are further divided due to the geological distance restrictions.

In sum, from Figure 9, we observe that (1) adjacent areas are more likely to be in the same community, which is consistent with earlier findings at the regional level; (2) the most populous communes, Yopougon and Abobo, form a single large community themselves (in red and green respectively); Cocody (the



(a) Call Frequency (b) Call Duration Time (c) Human Trajectory

Figure 9: Community Detection Based on Call Networks (call frequency, duration) and Human Trajectory Network of Abidjan
1:Cocody; 2: Yopougon; 3: Abobo; 4: Plateau; 5:Marcory; 6: Port-Bouët; 7: Adjamé; 8: Treichville; 9: Koumassi; 10: Attécoubé.

wealthiest commune) dominates the yellow community based on two call networks. The communes with ID, 5, 6, 8, 9, merge into one big community in both Figures. 9a and 9b; (3) finally, the most striking point is that in all these figures, Adjamé (ID: 7), where major slums are located, one small community dominates (in gray at the centre). This finding indicates strong connectivity within the commune, with few connections to other communes. Therefore, consistent with the above findings at the regional level we find that the communities drawn from the Abidjan networks are not only affected by the geological distance, but also reveal different economic development across communes.

5 Conclusion

We analyze the mobile phone traffic and human mobility traces from a large-scale mobile data collected in Côte d'Ivory with the aim to understand the economic development in this country and provide potential suggestions for poverty reduction. First, we develop several network structure indicators from calling record networks weighted by number of calls, duration of calls, and human trajectory network weighted by number of transitions. We find that the relative frequency of initiating calls to other areas significantly correlate (as high as 0.8) with important local social-economic statistics like poverty rate and annual income. It may thus serve as a potential real-time signal to track economic development at the regional level. Second, we found the super-linear scaling properties of the calling activities in terms of call frequency and duration time with respect to the population size. This may indicate mobile phone usage is overly concentrated in populous areas which are generally rich areas. Third, we investigate the mobile communication pattern among rich and poor areas defined by annual income at the development pole level. We found rich areas communicate more frequently with other rich areas than poor areas thus forming a "rich-club". Finally, we run community detection on mobile communication graph and re-defined the borders of the country. We found that rich regions (e.g. South) further split into smaller communities, indicating the heterogeneity and segregation inside these regions. By contrast, several poor regions tend to merge into one large community (e.g. West and North), indicating the relatively less complicated and more homogeneous type of communication among poor people.

Our research may provide some policy implications for the Côte d'Ivory government: (1) use large-scale mobile phone data to monitor the real-time economic development across the country in an efficient and less-costly way, and identify those areas in need of economic and financial aids at an early stage. (2) reduce the poverty due to the unbalanced communications. Mobile technology can serve as a great tool for poverty reduction, but only when it facilitates and deepens the level of communication between people of different socio-economic classes. However, our research findings do not show any strong mobile phone communication between rich and poor regions. Therefore, the government of Côte d'Ivory may seek to engage in efforts that reduce the "digital divide" and support communication between rich and poor areas, and across various socio-economic strate of the Côte d'Ivory population.

References

1. Eagle N, Macy M, Claxton R (2010) Network diversity and economic development. Science 328: 1029-1031.
2. Granovetter M (1973) The Strength of Weak Ties. American Journal of Sociology 78: 1360–1380.
3. Page S (2007) The difference: how the power of diversity creates better groups, firms, schools and societies. Princeton University Press.
4. Newman MEJ (2003) The structure and function of complex networks. SIAM Review 45: 167-256.
5. Krings G, Calabrese F, Ratti C, Blondel V (2009) Scaling behaviors in the communication network between cities. Institute of Electrical and Electronics Engineers.
6. Calabrese F, Dahlem D, Gerber A, Paul D, Chen X, et al. (2011) The connected states of america: Quantifying social radii of influence. In: Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom). pp. 223-230.
7. Vodafone (2005) Africa: The impact of mobile phones. Technical Report 2, Vodafone Police Paper Series.
8. Bhavnani A, Chiu R, Janakiram S, Silarszky P (2008) The role of mobile phones in sustainable rural poverty. Technical report, World Bank, ICT Policy Division, Global Information and Communications Department.
9. Report IC (2009) Cote ivory: Poverty reduction strategy paper. Technical Report 09/156, IMF Country Report.
10. Dubresson A (1997) Abidjan: from the public making of a modern city to urban management of a metropolis, Tokyo, Japan: United Nations University Press. pp. 252-91.
11. Bettencourt L, Lobo J, Helbing D, Kuhnert C, West G (2006) Growth, innovation, scaling and the pace of life in cities. PNAS 104: 7301-7306.
12. Opsahl T, Colizza V, Panzarasa P, Ramasco JJ (2008) Prominence and control: The weighted rich-club effect. Physical Review Letters 101.
13. Ratti C, Sobolevsky S, Calabrese C F Andris, Reades J, Martino M, et al. (2010) Redrawing the map of great britain from a network of human interactions. PLoS ONE 5: e14248.
14. Blondel V, Krings G, Thomas I (2010) Regions and borders of mobile telephony in belgium and in the brussels metropolitan zone. E-journal for Academic Research on Brussels 42.

15. Thiemann C, Theis F, Grady D, Brune R, Brockmann D (2010) The structure of borders in a small world. PLoS ONE 5: e15422.

16. Blondel V, Guillaume J, Lambiotte R, Mech E (2008) Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment 2008: P10008+.

Ubiquitous Sensing for Mapping Poverty in Developing Countries

Christopher Smith
ICRI: Cities
London, UK
chris.smith@ucl.ac.uk

Afra Mashadi
Bell Labs, Alcatel-Lucent
Dublin, Rep. of Ireland
afra.mashadi@alcatel-lucent.com

Licia Capra
University College London
London, UK
l.capra@ucl.ac.uk

ABSTRACT

Social surveys and censuses offer a good indication of poverty and inequality in a country. However, due to the expenses associated with data collection, the granularity and frequency of such information is often limited. In developing countries in particular, census data may be collected very infrequently, thus failing to accurately reflect the changes associated with a growing economy. In this paper, we propose to use ubiquitous sensing as a proxy for estimating socioeconomic indicators and analyse aggregated mobile phone communication data in Côte d'Ivoire. We discover a number of features that show a strong correlation with poverty indicators. We then demonstrate how these features can be used to provide poverty estimates at a spatial resolution finer than previously available.

INTRODUCTION

Social surveys and censuses periodically collected by National Statical Institutes contain valuable information describing the social and economic well being of a country and the relative health of different areas. Such data is used by policymakers and agencies to guide the formulation and implementation of policies and programs that aim to improve the life of the citizens. Poverty maps derived from survey data, and spatial descriptions of the distribution of poverty are most useful when they are finely disaggregated (i.e., when they represent small geographic units, such as cities, towns or villages), and more importantly, when they are most up to date. Spatially rich and temporally accurate knowledge of socioeconomic indicators would help in alleviating poverty by enabling efficient investment in infrastructure and consequently mitigating against the detrimental effects of poverty and inequality. However, this form of data collection is known to be an onerous task due to the cost involved, especially for nations where political instability and a weak economy exacerbate the problem.

Côte d'Ivoire is an example of a developing country which has suffered recent political strife and economic turmoil. Agriculture employs roughly 68% of its total population with Côte d'Ivoire being the world's largest producer and exporter of cocoa beans and a significant producer and exporter of coffee. Although rich in agriculture and natural resources (e.g., diamond), the economy is highly sensitive to fluctuations in international prices for these products. Furthermore, recent events, including civil war, have resulted in a loss of foreign investment and economic contraction. In late 2011, Côte d'Ivoire's economy began to recover from a severe downturn in the first quarter of the year that was caused by widespread post-election conflict. In June 2012 the World Bank announced $4.4 billion in debt relief for Côte d'Ivoire under the Highly Indebted Poor Countries Initiative. Côte d'Ivoire's long term challenges are known to include political instability and degrading infrastructure [1].

Given this state of affairs, it is perhaps not surprising that no data pertaining to a full survey of the country's population appears to have been made available since the late 1990s. To address this problem, we propose the use of ubiquitous sensing as an alternative to the traditional method of collecting sociodemographic data through census and social surveys. Ubiquitous sensing refers to the passive collection of peoples' digital footprints (e.g., location based social networking check-ins, phone calls, etc.) which can provide a detailed picture of human mobility and communication. If we are to provide a viable alternative, or at least a useful complement, to traditional censuses, we need to ensure that data is sourced uniformly from the population, with minimum bias. Online social network and location based service data is likely to be lacking in this regard, as they suffer known demographic biases and uptake of such services tends to be clustered geographically. In this paper, we show how Call Detail Records (CDRs) can be mined in order to derive proxies for poverty indicators, which can then be used to estimate poverty on a continuous basis and at low cost, as opposed to the slow iteration of census survey cycles. Côte d'Ivoire contains more than 17 million mobile phone users (around 77% of the total population) and is well developed by African standards, being ranked 51st in the world [1]. With this high penetration rate there is significant potential for methods exploiting ubiquitous sensing to have a real impact.

[1]CIA World Factbook-`https://www.cia.gov/library/publications/the-world-factbook/geos/iv.html`

Submitted to D4D @ NetMob 2013

We first describe the features we mine from the data and report their correlation with poverty as measured by the Multiple Poverty Index. We then demonstrate how these features could be used to build a regression model that estimates poverty at a finer level of granularity, potentially enabling policymakers and agencies to efficiently allocate limited resources to avert and re-balance inequality and invest in infrastructure where needed most.

RELATED WORK

There is today an unprecedented amount of location based digital footprint data, such as geo-tagged tweets, Foursquare check-ins and CDRs, which has been the subject of much research aiming to understand the dynamics of human mobility and communication on a scale not previously possible. Noulas *et al.* [17] study urban mobility patterns of people in metropolitan cities by analysing the check-ins of a large sample of Foursquare users. They find that human mobility appears to obey a universal law which isolates as a key component the rank-distance, which factors in the number of places between origin and destination, rather than pure physical distance. Cheng *et al.* [8], investigate the effect of contextual factors such as population and social status on the mobility patterns of citizens through their digital footprints obtained from Foursquare check-ins. They observe that in addition to geographic and economic factors, social status is highly coupled with one's mobility in the city. In particular, they show that people in wealthy cities travel more frequently to distant places than people in less rich cities. Kramer [14] found that the difference between the number of positive and negative words used in Facebook status updates covaries with self-reported "satisfaction with life" in the US. Similarly, Quercia *et al.* found that sentiment expressed in tweets [18] and the topic of tweets [19] in London, aggregated by the area associated with the tweet or Twitter profile, correlates with socioeconomic deprivation of that area. A limitation noted in these works, however, is the large demographic bias of online social networking services. The majority of Twitter users are male, under 35 and with a relatively high income. Mislove *et al.* [16] also suggest that the ethnicity of twitter users, in the US at least, is not representative of the general population. Similarly, although Facebook has a more even gender distribution, in the UK around 60% of users are less than 35 years of age.[2]

In order to avoid such a population bias an alternative data source that is more representative of the population is required. One such source are CDRs, which have been extensively studied for a broad range of purposes, from understanding human mobility [6, 11, 7, 26] to land use identification and urban planning [4, 24, 20]. Various ways of characterising geography based on the traffic of mobile phones and their users' trajectories have been examined. Specific to understanding the relation between CDRs and socioeconomic factors, there has been only a handful of works in the literature [25, 10, 5, 9]. Eagle *et al.* [9] measured the communication diversity from fixed line phone call records in England, and found that higher diversity (i.e., the more evenly dispersed a person's communication between people and places) correlates with socioeconomic deprivation aggregated to telephone exchange areas. Blumemstock [5] looked at the relation between users' demographics (collected through personal interviews) and their mobile phone usage from a sample of employees from a company in Rwanda. Observations include that gender and social status of the users had a direct correlation with the volume of their call activity. The closest work to that presented here is research undertaken by Soto *et al.* [25] and Frias-Martinez *et al.* [10], in which the authors have proposed models to infer and predict the socioeconomic indicators of a region. Specifically, [25] proposes a Support Vector Machine model operating on 279 features of individual users' CDR to infer the socioeconomic level of census regions. They used features categorised into 69 *behavioural* (such as total number of calls), 192 *social* (such as number of contacts) and 18 *mobility* features (such as distant travelled). The authors report the performance accuracy rate of around 80%. Finally, [10] has extended [25] to provide forecasts of socioeconomic factors. The drawback to this approach is that by including so many features and their interactions in a complex model policymakers are presented with a 'black-box' predictor, with little hope of understanding how the estimates are reached. Arguably, for such predictions to play a role in the decision making process, it is vital that it can be understood how they were formed. Furthermore, many of the features used in these works require detailed knowledge of individual behaviour, which for privacy reasons may not be readily available.

For these reasons our approach differs from the above works in two important ways: i) we consider only CDRs aggregated by the antennas through which the calls are connected; and ii) our results suggest that far less input variables are needed to infer poverty levels from the aggregated CDRs. We thus avoid the privacy issues associated with individual user data, and allow for a more detailed understanding of how our estimates are formed.

MINING CALL DATA RECORDS

In order to infer poverty levels of areas from human communication patterns, we require a communication dataset representative of the population as well as a ground-truth dataset of poverty information to validate our approach against. We describe each of these datasets next.

Call Records

We obtained a dataset of anonymised voice calls between five million of Orange customers in Côte d'Ivoire between December 1st 2011 and April 28th 2012[3]. Orange is the second main provider of mobile services in Ivory Coast, keeping 48% of market share as well as possessing the largest

[2] **http://www.insidefacebook.com/2010/06/08/whos-using-facebook-around-the-world-the-demographics\-of-facebooks-top-15-country-markets/** - retrieved 29/05/2012

[3] As part of D4D challenge by Orange, see **http://www.d4d.orange.com/**

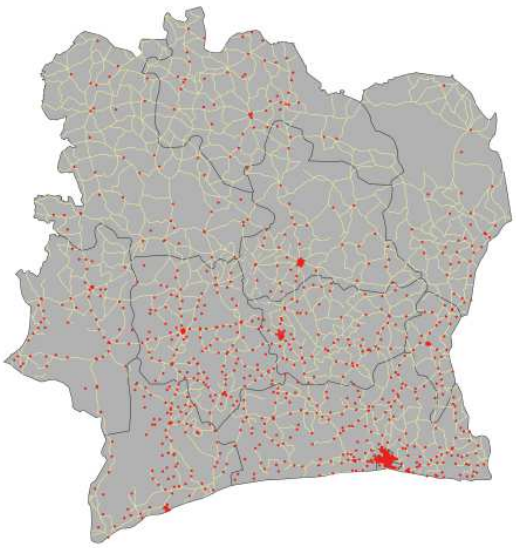


Figure 1. Spatial Distribution of Antenna in Côte d'Ivoire

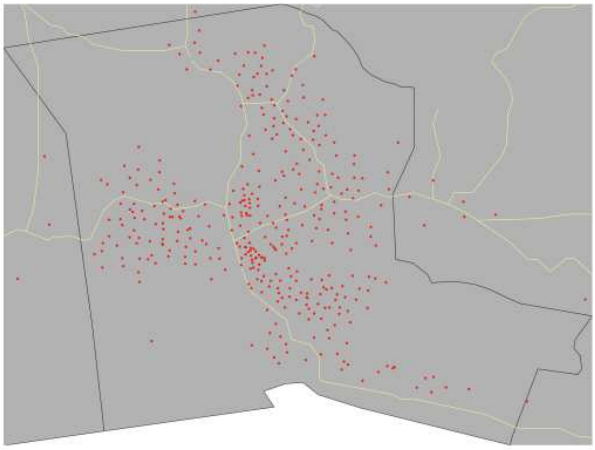


Figure 2. Spatial Distribution of Antenna in the city of Abijdan.

network of retail stores[4]. For the purpose of this study we used the subset data of antenna to antenna traffic, which contains hourly aggregated volume and duration of calls between pairs of antenna. Figure 1 shows the location of the antennas within eleven regional boundaries of Côte d'Ivoire. Notice that in the south there is a dense concentration of antennas in and around the secondary capital, Abijdan, which has a higher population density than the rest of the country and is where most economic activity and trading takes place. We first mapped the antennas to the regions in which they fall before summing volume and duration of calls made between each pair of regions. Some antennas were found to have the same coordinates, and it is unknown to the authors whether these are errors or whether there are genuinely more than one antenna in the same location. However, for our purposes what is important is the location of source and target, therefore we reassigned the identifiers of antennas with duplicated locations.

Multidimensional Poverty Index

In order to study the correlation between the call detail records and poverty, we also require a ground-truth dataset of poverty levels for areas in Côte d'Ivoire. For this purpose, we used the Multidimensional Poverty Index created by the University of Oxford[5]. The index incorporates a range of indicators in order to reflect the complexity of poverty and better inform policies aimed at relieving it. In addition to the single valued index and population estimates, the MPI contains several indicators that aim to capture people's experience of deprivation, such as poor health, lack of education, inadequate living standard, low income, disempowerment and threat of violence. Although detailed in terms of its coverage of the various facets of poverty, the Côte d'Ivoire MPI is derived from survey data from the year 2005. The temporal lag between our ground truth data and the mobile phone data introduces a limitation to our findings and the methods we present. This will be discussed further in the discussion section. Figure 3 depicts the aggregated MPI for eleven regions of Côte d'Ivoire, where the darker colour indicates higher MPI thus higher poverty.

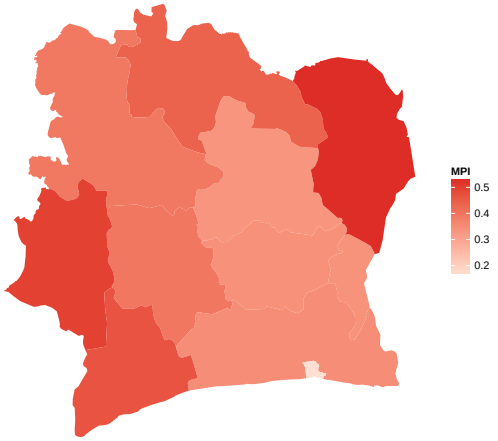


Figure 3. Map of 11 regions showing Multidimensional Poverty Index. Lighter colour indicates higher poverty.

Testing Hypotheses

We next formulate a number of hypotheses and derive features of the flow data to test them.

[4] `http://www.orange.com/en/group/global-footprint/countries/Group-s-activities-in-Ivory-Coast`

[5] `http://www.ophi.org.uk/policy/multidimensional-poverty-index/`

Activity

The first set of features are simple aggregates of the flows between regions. We expect to find that the level of mobile communication activity within a region will reflect its social and economic activity, and thus its level of prosperity [1]. We find strong negative correlations between the total outgoing volume ($r = -.774$, p–value$= .005$) and duration ($r = .791$, p–value$= .004$) of calls within a region and its MPI score, thus confirming that aggregated communication activity provides a simple proxy for poverty level. These aggregates highlight the relationship between communication activity and poverty in a region, however, we also aim to capture the relationship between poverty and the interactions between regions. We therefore investigate further hypotheses involving more complex features. Furthermore, it may not always be the case that these simple aggregates will provide accurate enough estimates of poverty, particularly at finer levels of granularity.

Gravity Residuals

The next set of features involves using a gravity model to estimate the flow between the centroids of the eleven regions. First introduced by Zipf in 1946 [28], gravity models rest on the hypothesis that the size of flow between two areas is proportional to the mass (i.e., population) of those areas, but decays as the distance between them grows. Despite some criticisms (see for example [22]), the model has been successfully used to describe macro scale interactions (e.g., between cities, and across states), using both road and airline networks [3, 12] and its use has extended to other domains, such as the spreading of infectious diseases [2, 27], cargo ship movements [13], and to model intercity phone calls [15]. We hypothesise that the difference between observed and expected flows between areas reflects the level of social and economic activity in those areas, and thus will be related to poverty.

We use the following equation to find the expected flows between regions,

$$F_{u,v}^{est} = g\frac{m_u m_v}{d_{u,v}^2} \tag{1}$$

where m_u is the population of region u taken from 2010 estimates and $d_{u,v}$ is the as-the-crow-flies distance between centroids of regions u and v. We then take two error measures at the areal level: firstly, we find the pairwise residual (i.e., the error between the real and estimated flow on each link), then the *link residual* is the average of a region's links; secondly, we sum the total observed and estimated incoming/outgoing flows for each region and measure the difference, or the *sum residual*. In previous work we modelled the flow of passengers in London's rail system in a similar fashion, and found that the gravity residuals were related to deprivation of neighbourhoods. Results of all correlations are presented in Table 1. We see strong, negative correlations between a region's MPI and both its sum residual and link residual, meaning that when flows between two areas are overestimated (negative residual) the poverty level is likely to be higher, and conversely, when flows are underestimated (positive residual) the poverty level is likely to be lower.

Diversity

Our next set of features aims to capture the opportunity for development afforded by an advantageous position in an information flow network. By studying a the social network represented by a fixed line call dataset, Eagle *et al.* [9] showed that the average *diversity* of the social connections of people living in a neighbourhood correlates strongly with the level of socioeconomic deprivation (a concept closely related to poverty) in that neighbourhood. In this work we are constrained by the aggregation of the call records to antenna and are unable to look directly at the underlying individual social network. Instead, we hypothesise that the diversity of a region's connections to other regions will also reflect the level of poverty in the region. We thus take two measures of a region's diversity: first, the *degree* of the region, and the second, simply termed *diversity*, which is found using the following formula from [9],

$$diversity(i) = \frac{-\sum_j v_{i,j}\log(v_{i,j})}{\log(k_i)} \tag{2}$$

where $v_{i,j}$ is the fraction of antenna i's flow that goes to j, and k_i is the unweighted degree of i. The feature we name *degree* is the number of links connecting each region whose weights are above the 8th decile of the overall weight distribution. We test all deciles and found the 8th to give the strongest correlation with MPI. Thus, *degree* represents the number of heavily weighed connections a region has. Both degree and diversity were first calculated per antenna, then the average of all antenna within a region was taken. As with the gravity model residuals, we see a strong negative correlation with MPI, which shows that the more diverse a region's connections, the lower the poverty level is likely to be.

Introversion

Finally, we hypothesise that a region's level of *introversion* may be a signal of its poverty level. In other words, if an area has relatively fewer connections to other regions compared to the number of connections that exist within it, the less it will be able to benefit from new sources of opportunity arising further afield. In conjunction with our first hypothesis, that higher activity reflects lower poverty, for two regions with equal activity we would expect that with lower introversion to have the lower poverty level. This is similar in spirit to the theory of open economies, albeit on a different scale, which expects nations that close their borders to international trade to fair less well than those that are more open [21]. It is also related to the idea of diversity of connections, except that we now take into account space and consider only a binary relationship, that is, the ratio of self-flow to total flow. We first calculated the introversion of antennas with the following equation, and then found the average introversion of all antennas within each region.

$$introversion(i) = \frac{f_{i,i}}{\sum_{i \neq j} f_{i,j}} \tag{3}$$

where $f_{i,j}$ is the flow between antennas i and j. This measure produces values in the range $[0, 1]$ where 0 means completely extroverted and 1 means completely introverted. Again, the introversion of regions correlates strongly with MPI, this time

Hypothesis	Feature	Pearson's r	95% Confidence Interval	p-value
Activity	total volume	-.777	-.939, -.331	.005
	total duration	-.783	-.941, -.345	.004
Gravity Residuals	link volume residual	-.781	-.940, -.340	.005
	link duration residual	-.525	-.856, .109	.097
	sum volume residual	-.804	-.947, -.393	.003
	sum duration residual	-.822	-.952, -.437	.002
Diversity	diversity volume	-.834	-.956, -.469	.001
	diversity duration	-.848	-.960, -.506	.001
	degree volume	-.787	-.942, -.354	.004
	degree duration	-.750	-.931, -.274	.008
Introversion	volume introversion	.793	.368, .944	.004
	duration introversion	.795	.373, .945	.003

Table 1. Correlations between MPI of 11 regions in Côte d'Ivoire and features derived from mobile phone data.

positively, confirming our hypothesis that areas with higher levels of poverty also tend to be more introverted.

We have seen that a number of different features of communication patterns correlate strongly with MPI at the regional level. However, to be effective in targeting areas most in need of help and aiding the policymaker's decision process, we need to be able to provide estimates at a much finer level of granularity. What follows is a demonstration of the kind of estimates which could be derived from the features described above.

Estimating Poverty

In this section we use the features we derived in the previous section to estimate the level of poverty at a finer granularity. Unfortunately, due to the data collection limitation there is no poverty information available at this level with which we can validate the results, therefore we intend this exercise to be taken as a demonstration of the way in which communication data could be used. At this point it may be objected that at a finer level of granularity we would expect to see weaker correlations between poverty and our flow features. This may well be the case, however, as we would also be working with a larger number of data points which would allow us to combine several features into a more sophisticated model, one can argue that the predictive accuracy would actually increase.

To demonstrate the potential for using communication data to estimate poverty level at a finer level of granularity we use diversity of call duration as this had the strongest correlation in our previous experiments. We first derive a linear model using ordinary least squares regression,

$$MPI_u^{est} = 1.346 - 1.385 \times diversity(u) \qquad (4)$$

Figure 4 depicts the MPI level for the eleven large regions of Côte d'Ivoire as predicted by this model, where the darker areas indicate higher estimated poverty level.

We can then use this model to estimate poverty levels at the sub-prefecture level, of which there are 255 in Côte d'Ivoire. Figure 5 shows the choropleth of the estimates for sub-prefectures. Notice the change in spatial pattern compared to the regional map in Figure 3. The coarser grained map depicts poverty increasing as we radiate out from the city of Abijdan. Instead, our finer grained estimates suggest that the

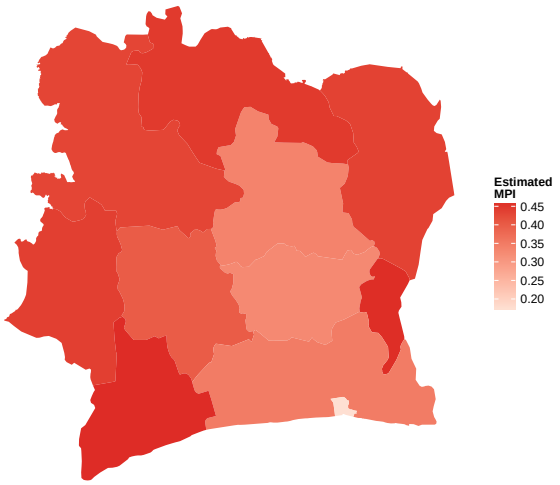


Figure 4. Poverty map estimated based on the link diversity antennas in 11 regions.

South-east of the country may contain areas of high poverty near Abijdan and conversely the North-west may contain areas of low poverty. The grey spaces in the choropleth indicate sub-prefectures for which we cannot obtain estimated poverty levels since they contain no antennas. Estimates could be extended to these regions by borrowing information from neighbouring areas, using tessellation to determine the effect of nearest antennas, or some combination of the two.

DISCUSSION

We have demonstrated the potential of CDR data to provide an invaluable source of poverty estimates, even without knowledge of individual behaviour. We have uncovered several features of communication patterns among mobile phone users in Côte d'Ivoire that track poverty of regions as defined by the Multidimensional Poverty Index. Our results have important implications for policymakers and agencies working in countries which lack the resources to manually collect so-

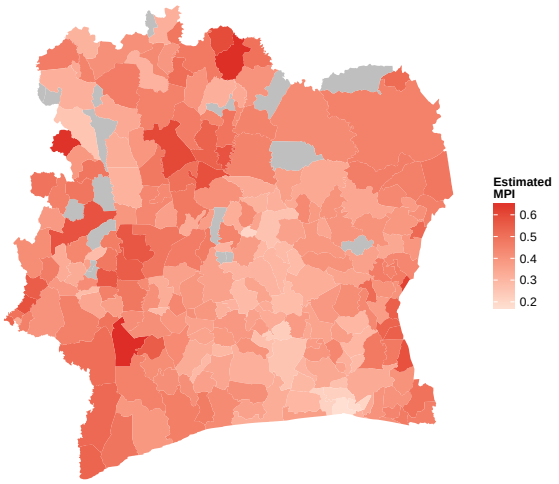


Figure 5. Poverty map in finer granularity estimated based on the diversity of connections between antenna.

cioeconomic data. Indeed, tools built upon the methods we have described would be a useful augmentation to socioeconomic data collection processes in any country. The cost of producing estimates from passively and automatically collected communication data is negligible compared to that of manual surveying, thus the main barrier to obtaining up to date poverty estimates has been removed. Côte d'Ivoire is a perfect example of a country in which timely and accurate information regarding poverty is severely lacking. In cases such as this the ability to obtain estimates of poverty levels on a continuous basis would represent a vast improvement. Limited resources could be allocated in much more efficient manner thereby helping to alleviate some of the detrimental effects of poverty and inequality.

However, we also note that the problem of a lack of up to date and spatially accurate socioeconomic data also represents a limitation to our results. In order to discover proxies for poverty indicators we require knowledge of those indicators, and to be confident that those proxies accurately track poverty level we need the proxy data and ground truth data to be close in space and time. Instead we are forced to work with a lag of 7 years between the poverty data we use as ground truth and the mobile phone data from which we derive our proxies. However, although this temporal lag will undoubtedly affect the accuracy of predictive models based on our proxies, such as the simple linear model we present above, we argue that the legitimacy of the methods we have developed is not compromised. Rather, we would only expect the accuracy and utility of the methods to increase were this lag removed.

Furthermore, it may not be strictly necessary to undertake a new, comprehensive, manual survey in order to build highly accurate predictive models. Previous work has shown that by taking a machine learning approach we may estimate socioeconomic indicators by training on a sample of census data [23]. Indeed, our own more recent experiments on deprivation in London neighbourhoods suggest that by incorporating spatial properties, the size of the required training sample could be as little as 10% of areal units.

A valuable extension to our work would then be to obtain more up to date socioeconomic data, perhaps by working with agencies on the ground to collect data from various locations. This would allow us to build a clearer picture of the relationship between communication patterns and poverty at a higher resolution. As part of our future work, we improve upon the results obtained thus far by exploring variations on the features we have defined. For example, in the gravity model rank-distance or the cost and duration of travel may be more appropriate than straight line distance between two regions, since they better reflect the considerations people make. In addition, we will explore different ways of assigning variables pertaining to antennas to the area around them. At present, antennas near areal borders are treated as if they only relate to the area in which they fall. To overcome this we could use population weighted tessellation for example. Subject to data availability, we also aim to study the generalisation of models built using the methods we have presented by comparing results in other countries, and finally, by obtaining longer term data we can investigate changes in communication patterns as changes occur in the socioeconomic well being of areas, thus helping to tease out causal relationships.

REFERENCES

1. Aker, J., and Mbiti, I. Mobile phones and economic development in africa. *Center for Global Development Working Paper*, 211 (2010).

2. Balcan, D., Colizza, V., Gonalves, B., Hu, H., Ramasco, J. J., and Vespignani, A. Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences 106*, 51 (2009), 21484–21489.

3. Barrat, A., Barthélemy, M., Pastor-Satorras, R., and Vespignani, A. The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America 101*, 11 (Mar. 2004), 3747–52.

4. Becker, R., Caceres, R., Hanson, K., Loh, J., Urbanek, S., Varshavsky, A., and Volinsky, C. A tale of one city: Using cellular network data for urban planning. *Pervasive Computing, IEEE 10*, 4 (2011), 18–26.

5. Blumenstock, J., and Eagle, N. Mobile divides: gender, socioeconomic status, and mobile phone use in rwanda. In *Proceedings of the 4th ACM/IEEE International Conference on Information and Communication Technologies and Development*, ACM (2010), 6.

6. Calabrese, F., Pereira, F., Di Lorenzo, G., Liu, L., and Ratti, C. The geography of taste: analyzing cell-phone mobility and social events. *Pervasive Computing* (2010), 22–37.

7. Candia, J., González, M., Wang, P., Schoenharl, T., Madey, G., and Barabási, A. Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical 41*, 22 (2008), 224015.
8. Cheng, Z., Caverlee, J., Lee, K., and Sui, D. Exploring millions of footprints in location sharing services. In *Proc.of AAAI ICWSM* (2011).
9. Eagle, N., and Macy, M. Network Diversity and Economic Development. *Science 1029* (2010).
10. Frias-Martinez, V., Soguero-Ruiz, C., Josephidou, M., and Frias-Martinez, E. Forecasting socioeconomic trends with cell phone records. In *3rd ACM Symposium on Computing for Development* (2013).
11. Girardin, F., Calabrese, F., Fiore, F., Ratti, C., and Blat, J. Digital footprinting: Uncovering tourists with user-generated content. *Pervasive Computing, IEEE 7*, 4 (2008), 36–43.
12. Jung, W., and Wang, F. Gravity model in the Korean highway. *EPL (Europhysics Letters) 81* (2008).
13. Kaluza, P., Kölzsch, A., Gastner, M. T., and Blasius, B. The complex network of global cargo ship movements. *Journal of the Royal Society, Interface / the Royal Society 7*, 48 (July 2010), 1093–103.
14. Kramer, A. D. I. An Unobtrusive Behavioral Model of Gross National Happiness. In *Proceedings of the 28th ACM CHI* (2010), 287–290.
15. Krings, G., Calabrese, F., Ratti, C., and Blondel, V. D. Urban gravity: a model for inter-city telecommunication flows. *Journal of Statistical Mechanics: Theory and Experiment 2009*, 07 (May 2009), L07003.
16. Mislove, A., Lehmann, S., Ahn, Y., and Onnela, J. Understanding the Demographics of Twitter Users. *Fifth International AAAI* (2011), 554–557.
17. Noulas, A., Scellato, S., Lambiotte, R., Pontil, M., and Mascolo, C. A tale of many cities: universal patterns in human urban mobility. *PloS one 7*, 5 (2012), e37027.
18. Quercia, D., Ellis, J., Capra, L., and Crowcroft, J. Tracking Gross Community Happiness from Tweets. In *Proceedings of ACM CSCW 2012* (2012).
19. Quercia, D., Seaghdha, D. O., and Crowcroft, J. Talk of the City : Our Tweets , Our Community Happiness. In *Proc.of AAAI ICWSM* (2012).
20. Ratti, C., Williams, S., Frenchman, D., and Pulselli, R. Mobile landscapes: using location data from cell phones for urban analysis. *ENVIRONMENT AND PLANNING B PLANNING AND DESIGN 33*, 5 (2006), 727.
21. Sachs, J. D., and Warner, A. M. Sources of slow growth in african economies. *Journal of African Economies 6*, 3 (1997), 335–376.
22. Simini, F., González, M. C., Maritan, A., and Barabási, A.-L. A universal model for mobility and migration patterns. *Nature* (Feb. 2012), 8–12.
23. Smith, C., Quercia, D., and Capra, L. Finger On The Pulse: Identifying Deprivation Using Transit Flow Analysis. *Proceedings of ACM CSCW 2013* (2012).
24. Soto, V., and Frías-Martínez, E. Automated land use identification using cell-phone records. In *Proceedings of the 3rd ACM international workshop on MobiArch*, ACM (2011), 17–22.
25. Soto, V., Frias-Martinez, V., Virseda, J., and Frias-Martinez, E. Prediction of socioeconomic levels using cell phone records. *User Modeling, Adaption and Personalization* (2011), 377–388.
26. Toole, J., Ulm, M., González, M., and Bauer, D. Inferring land use from mobile phone activity. In *Proceedings of the ACM SIGKDD International Workshop on Urban Computing*, ACM (2012), 1–8.
27. Viboud, C., Bjø rnstad, O. N., Smith, D. L., Simonsen, L., Miller, M. A., and Grenfell, B. T. Synchrony, waves, and spatial hierarchies in the spread of influenza. *Science (New York, N.Y.) 312*, 5772 (Apr. 2006), 447–51.
28. Zipf, G. The P 1 P 2/D hypothesis: On the intercity movement of persons. *American sociological review 11*, 6 (1946), 677–686.

15 February 2013

ANALYZING SOCIAL DIVISIONS USING CELL PHONE DATA

Orest Bucicovschi[1], Rex W. Douglass[1,2], David A. Meyer[1,2], Megha Ram[1,2], David Rideout[1] and Dongjin Song[1,2,3]

[1]*Department of Mathematics 0112*
[2]*UC Institute on Global Conflict and Cooperation 0518*
[3]*Department of Electrical and Computer Engineering 0407*
University of California/San Diego, La Jolla, CA 92093

obucicov@math.ucsd.edu, rexdouglass@gmail.com, dmeyer@math.ucsd.edu, megharam09@gmail.com, drideout@math.ucsd.edu, dosong@ucsd.edu

ABSTRACT

Mobile phone use is increasingly common, especially in developing countries. Telecommunications traffic between areas served by individual mobile phone antennae in Côte d'Ivoire, provided by France Telecom-Orange's subsidiary as part of the D4D Challenge, thus provides a finely resolved measure of interactions within the geographically distributed population. We extract a community structure from this communication network, and observe that it is geographically localized. Fitting the intensity of interactions between nodes to a gravity model, we find that it decays with distance, consistent with the geographic locality of the network communities. We develop a novel methodology to compare this network community structure with geographical divisions of the population originating in ethnic, language, religious, or political differences. The same methodology also supports direct comparisons between such social divisions. Since political disputes between factions within a country are a primary obstacle to development, understanding these cleavages and the dependencies between them is a first step towards successful development.

Key Words: cell phone network, communities, spatial statistics, categorical variables, Potts model, gravity model.

1. Introduction

How can the deluge of tracking data now available from personal electronic devices be used to improve our knowledge of economic development and politically-motivated conflict in Côte d'Ivoire? Effective policy planning requires detailed and timely information on social and economic relationships. In developing countries, the necessary infrastructure and bureaucracy for collecting these data through conventional means may be lacking. In particular, data on demographic composition and political preferences are often obsolete or unavailable, with significant barriers to future collection. In countries rife with political discord and violence, such as Côte d'Ivoire, proxies for these types of data can significantly improve conflict resolution programs. Mobile phones, however, have become almost ubiquitous, even the developing world, where as of 2011 there were more than three-quarters as many subscriptions as inhabitants [1, p. 3]. In Côte d'Ivoire this fraction is closer to 85% [1, p. 50]. In this project we point to the possibility of inferring unknown social and demographic characteristics by leveraging prior but possibly outdated knowledge of the spatial distribution of different groups with contemporary cell phone data.

Our project uses one of the mobile phone datasets* from the Data for Development (D4D) Challenge [2]. It consists of aggregate communication between 1216 cell tower antennae from December 2011 to April 2012, along with the geographical locations of the towers. We proceed in four stages. In the first stage we use the call volumes between pairs of cell towers, together with the distances between them, to demonstrate the decrease in communication with distance. In the second stage we partition the call volume weighted network into 'communities' of antennae that have stronger connections to one another than to antennae in other communities. As suggested by the results of the first stage, these communities are geographically localized. In the third stage we create estimates of the distribution of different demographic quantities across space using other data sources. In our first example, we use maps which partition Côte d'Ivoire into areas in which the primary dialect is one of 60, grouped into five major language sets.† In the fourth stage we introduce a new statistical test for independence between pairs of categorical identities (*e.g.*, network community and linguistic group) in the presence of strong spatial dependencies. We conclude by discussing ways in which the approach scales to a wide variety of interesting applications, for development in Côte d'Ivoire and in other countries.

2. Gravity model

There is a long history of quantitative investigation into the effect of distance on social and economic interactions, dating back over a century [3–7]. An important question is at what power of the distance does the strength of the interaction decay? Much research indicates that international trade decays roughly linearly with distance [8], while a number of authors

* These data were made available by France Telecom/Orange Côte d'Ivoire within the framework of the D4D Challenge.

† Four major languages: Gur, Kru, Kwa and Mande, and a region in which Gur and Kwa languages are spoken by about the same fractions of the population.

find a quadratic fall off of communication intensity with distance [9]. Such quantitative decay laws are called "gravity" models.

Set 1 of the D4D Challenge data supports a careful analysis of the intensity of traffic between mobile phone towers, and in particular, a test of the gravity law. In the context of mobile phone communication this model predicts that the intensity of traffic, as measured by the total duration of communication between two geographically localized sites, will be proportional to the product of their populations and inversely proportional to the square of the distance between them [9].

We test this law by computing (1) the total duration of communication, A_{ij}, between each pair of antennae i and j; (2) the total duration communication in which each antenna i is involved, k_i; and (3) the distance d_{ij} between each pair of antennae, in units of the Earth radius. We use the quantities k_i to represent the population of residents near a given antenna, assuming the volume of calls to be proportional to the local population, as has been demonstrated in other settings [9].

Figure 1 shows a histogram of the logarithms of the edge weights, $\log A_{ij}$, with a bin width of .01. It reveals the familiar log normal distribution, as found for a dataset of Belgian mobile telecommunications [9]. The curve in green is the Gaussian

$$He^{(x-\bar{x})^2/(2\sigma)^2},$$

with $H = 534000$, $\bar{x} = 9.49$ and $\sigma = 2.08$.

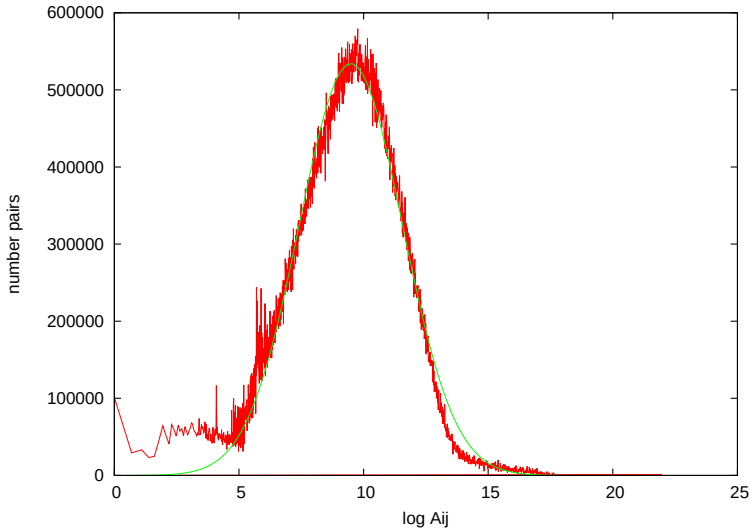


Fig. 1. Histogram of interaction strengths (red), superimposed on a log normal distribution (green).

To test the gravity law we find the best fit

$$\log A_{ij} = \alpha \log(k_i k_j) - \gamma \log d_{ij} + g,$$

with $g = \log G$. It is

$$\log A_{ij} = (0.985 \pm 0.001) \log(k_i k_j) - (0.367 \pm 0.001) \log d_{ij} - (27.67 \pm 0.04).$$

Figure 2 shows a scatter plot of the resulting points, with the line indicating equality. Since it is difficult to deduce by eye the functional form of the data from the scatter plot, we aggregate the points with similar distances to form Figure 3. The red data points plot the mean of $\log A_{ij} - \log(k_i k_j)$ against $\log d_{ij}$, averaged over those data points which fall within the given interval of $\log d_{ij}$ values. (There are 250 such intervals between the largest and smallest $\log d_{ij}$.) An error in each mean is estimated from the data points falling in the same bin, shown by the error bars. The two lines are fits to the full data set: the green to $\log(Gk_ik_j/d_{ij}^2)$, the blue to $\log(Gk_ik_j/d_{ij}^\gamma)$. This analysis effectively isolates the role of the exponent γ on d_{ij}, and makes it clear that $\gamma = 2$ provides a very poor model for the data, while a $\gamma \approx 1/3$ fares much better.

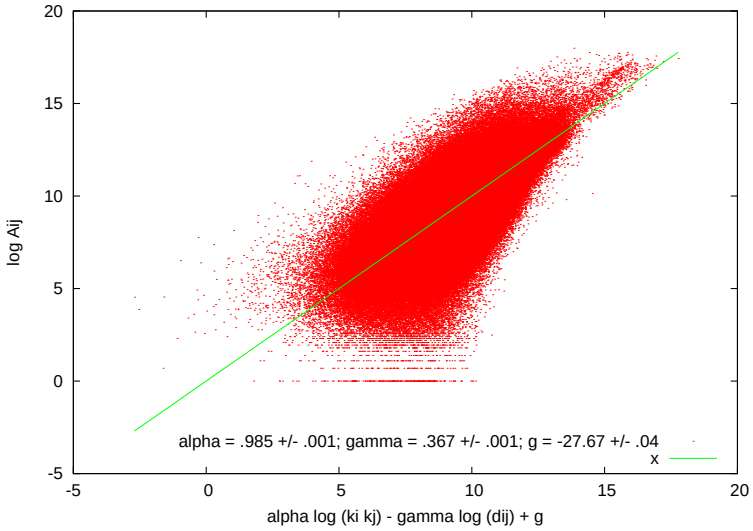


Fig. 2. Linear regression to estimate parameters α, γ, g.

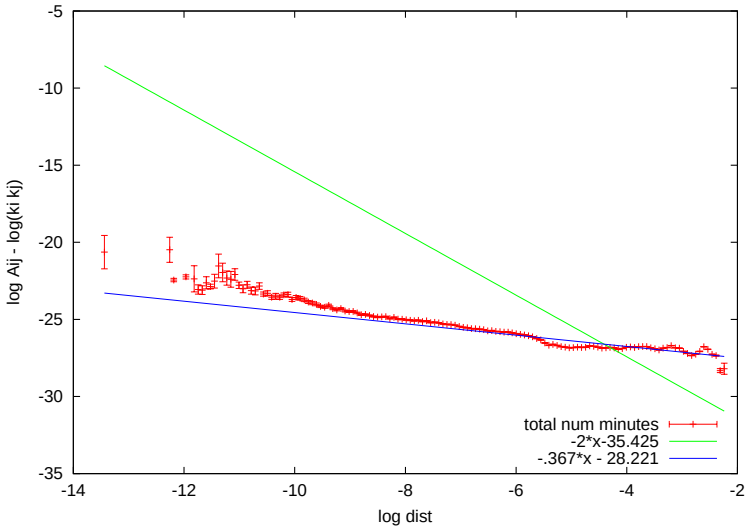


Fig. 3. Visualizing the value of γ by aggregating data with similar distances. The blue line, with slope 0.367, provides a much better fit than the green with slope 2.

From physics we know that in three spatial dimensions quantities such as the intensity of light, strength of the electric field, or magnitude of the gravitational force decay with the square of the distance. This occurs because in each case the field/interaction is diluted into a sphere, which at radius r has area proportional to r^2. In the case of people living on the surface of the Earth, which is only two dimensional, one may expect the strength of the interaction to decay based on the circumference of a circle at radius r, *i.e.*, as $1/r$. In our results above, however, the fact that we find neither of these is not terribly surprising, as people are not uniformly distributed on the Earth's surface. They instead live in communities connected with complex transporation infrastructures such as roads, bridges, railroads, and flight

networks. It is not unreasonable to expect that the resulting geometry is fractal in nature, with a dimension, based upon the above reasoning, somewhere between one (in which the interaction strength would be independent of distance) and two.* In fact, such fractal geometries of populations were first estimated by Appleby, in the context of optimizing telecommunications infrastructure [11]. While it is curious that our estimate of γ for Côte d'Ivoire differs so much from the results obtained using Belgian mobile communications data [9], these heuristics suggest that this may reflect infrastructural differences between a developing and a developed country. In any case, it is important to note that although interactions are stronger at shorter distances, they decay relatively slowly as distance increases, quantifying Tobler's First Law of Geography, "Everything is related to everything else, but near things are more related than distant things." [12].

3. Network communities

In the preceding discussion of gravity models for mobile telecommunications data we completely ignored the network aspect of the latter. The antenna–antenna call volume network is a weighted graph, however, and this can be analyzed on its own. In particular, it is very natural in any social network to look for communities of nodes which are more closely connected to one another than they are to nodes in other communities. Newman showed that the modularity of a partitioning of a network into communities, the amount by which the number of edges between members of the same community is greater than it would be if edges were assigned randomly with the same vertex degrees [13], makes equally good sense for a weighted network [14], in which case it is:

$$\sum_{i,j}\Big(\frac{A_{ij}}{A}-\frac{k_i k_j}{A^2}\Big)\delta(c_i,c_j),$$

where $A=\sum_{i,j}A_{ij}$, $k_i=\sum_j A_{ij}$ and c_i is the community to which node i is assigned. The problem of detecting community structure in a network then, becomes the challenge of maximizing the modularity.

In connection with their work on a Belgian telecommunications dataset, Blondel, Guillaume, Lambiotte and Lefebvre devised an efficient algorithm for finding a community structure with almost maximal modularity in large networks [15]. Applying their algorithm [16] to the antenna–antenna call volume network for Côte d'Ivoire, we find a set C of 11 communities, indicated by distinct symbols at their geographical locations in Figure 4.

The first observation to make about these communities is that they are localized geographically. This is partially explained by the results of §2, namely that the strength of interactions between nodes in this communication network is reasonably well described by a gravity model, and thus diminishes with distance.

* An alternate approach to utilizing insight from gravitational physics to understand the behavior of complex social networks can be found in [10].

The second observation is that these communities may not be arbitrary geographical clusters. We used the Côte d'Ivoire language map provided on the *Ethnologue: Languages of the World* website to see the regional distribution of languages and language groups in the country [17]. Although French is the official language of Côte d'Ivoire and Jula is the most widespread native language, according to Ethnologue's compilation of languages, there are 71 regional languages grouped into four major language families. In Figure 4 the map of Côte d'Ivoire is partitioned according to the primary language spoken in each region. Unsurprisingly, there appears to be some coherence between the communication network communities and the language regions.

Quantifying the dependence between these two sets of categories—network community and primary language—should allow mobile communications network data to contribute to a deeper understanding of the relationships between various language groups. These

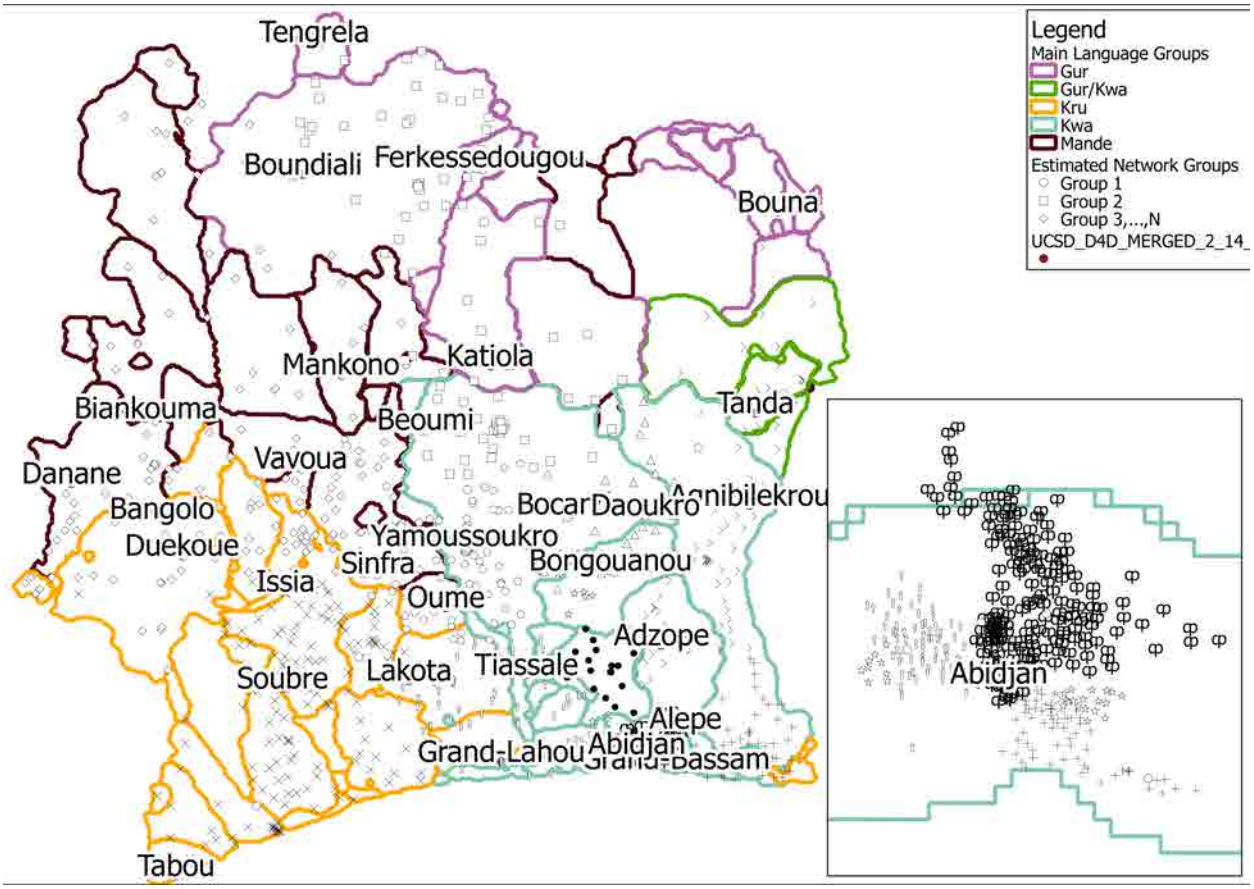


Figure 4. Each cell tower is marked with a symbol representing the network community computed for it. These network communities appear to follow primary language boundaries (for 60 dialects in Côte d'Ivoire, which are grouped into 4 major languages indicated by boundary colors) more closely than political or geographic boundaries (not shown). Further, areas with high population density (Abidjan inset) show clustering at a neighborhood level which exceeds the resolution of existing language data.

relationships are not purely social, but may also be indicative of economic ties between language groups. Moreover, language communities are intimately related to ethnic and religious groups, cleavages between which have become political differences fueling the conflict in Côte d'Ivoire [18,19]. In the next section we develop the tools to more fully analyze these economic and political relationships as they have the potential to significantly impact the development of the state.

4. Language communities refine network communities

4.1. Association between community structures

In the previous section we partitioned (renormalized) the antenna–antenna call duration network into a set of communities C, with $|C| = 11$. And we discussed the geographical partitioning of Côte d'Ivoire into regions by primary language; let L be the set of these 60 languages. Knowing the geographical locations of the antennae, we can assign each an element of L, the local primary language. Comparing the map of the network communities and the primary languages (Figure 4), it appears that there is some association. That is, if we think of the community and language labels of an antenna i as being a sample (c_i, ℓ_i) from a joint probability distribution over $C \times L$, we expect the collection of 1216 samples to indicate that it is not a product distribution.

More precisely, we can construct the $|C| \times |L|$ contingency table with entries $N_{c\ell}$ being the number of antennae that have labels $c \in C$ and $\ell \in L$; the sum of the entries will be $N = 1216$, the total number of antennae. For the labels to be independent, the contingency table should be rank 1, *i.e.*, if the marginal counts are $N_{c\cdot} = \sum_\ell N_{c\ell}$ and $N_{\cdot\ell} = \sum_c N_{c\ell}$, then $N_{c\ell} = N_{c\cdot} N_{\cdot\ell}/N$. The standard test for this is Pearson's χ^2 test [20]: Compute the statistic

$$\hat{\chi}^2 = \sum_{c,\ell} \frac{(\hat{N}_{c\ell} - \hat{N}_{c\cdot}\hat{N}_{\cdot\ell}/N)^2}{\hat{N}_{c\cdot}\hat{N}_{\cdot\ell}/N}, \tag{1}$$

and ask what is the probability of observing a value this far from the mean. Under the hypothesis of independence, χ^2 asymptotically has the distribution of the sum of $d = (|C| - 1)(|L| - 1)$ standard normal random variables, as the number of samples goes to infinity, so this probability can be approximated [21].

For the actual data, $\hat{\chi}^2 \approx 6253$, much greater than $(11-1)(60-1) = 590$. Since $N = 1216$ is small relative to the number of entries in the contingency table, 660, however, the asymptotic approximation is not so good, and it is better to sample the ensemble of closest rank 1 contingency tables, compute the value of the χ^2 statistic for each, and compare with the resulting estimated distribution. The results are shown in Figure 5; the mean of this sample is approximately 657, which is slightly larger than the normal approximation, but so much smaller than $\hat{\chi}^2 \approx 6253$ as to make the latter impossibly unlikely under this null hypothesis.

Since the samples (c_i, ℓ_i) are located geographically, however, and since both the net-

work communities and the language communities are localized in accordance with Tobler's First Law of Geography [12], the corresponding spatial autocorrelations in the c_i and in the ℓ_i mean that there are effectively fewer than N samples. Cerioli has suggested a correction to the χ^2 statistic to account for this [22]:

$$\hat{\chi}^2_c = \frac{d}{d+\hat{\lambda}} \sum_{c,\ell} \frac{(\hat{N}_{c\ell} - \hat{N}_{c\cdot}\hat{N}_{\cdot\ell}/N)^2}{\hat{N}_{c\cdot}\hat{N}_{\cdot\ell}/N},$$

where $\hat{\lambda}$ is a measure of the observed spatial autocorrelations. Again, under the hypothesis of independence (and some conditions on the autocorrelations) χ^2_c asymptotically has the distribution of the sum of d standard normal random variables.

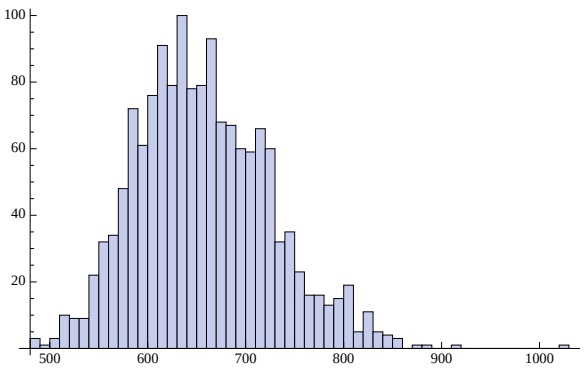


Fig. 5. Histogram of χ^2 for closest rank 1 contingency tables.

Estimating spatial autocorrelations well, however, is a notoriously difficult problem. It is particularly difficult in the presence of long range dependencies [23], which are suggested by the power law decay with distance of the gravity model approximation to the spatial interaction strengths that we computed in §2. Finite size/data and edge effects conspire to bias many estimators that are asymptotically unbiased and efficient. Communications traffic data, however, allows us to take another approach, which we explain next.

4.2. Uncoupled Potts models

The weights on edges in the antenna–antenna call duration network represent the strength of interactions between their endpoints. By the correspondence to the generalized gravity model demonstrated in §2, they also represent the strength of interactions between the geographical areas the antennae cover. These interactions are the origin of the spatial autocorrelations that complicate quantification of the association between different community structures. Inspired by the Potts model in statistical physics [24] (and not unrelated to the modularity of a network community structure [14]) we can thus use these weights as coupling constants for the community label variables at different antennae:

$$E(c) = -\frac{1}{2} \sum_{i \neq j} \frac{A_{ij}}{M} \delta(c_i, c_j), \tag{2a}$$

so that the "energy" is lower when nodes i and j with stronger interaction A_{ij} lie in the same community (have the same label). Here $M = \max A_{ij}$ is a conceptually unimportant but convenient normalization factor. Since we are interested in the association between two community structures, we can consider two uncoupled Potts models on the same network, the second one also having "energy":

$$E(\ell) = -\frac{1}{2} \sum_{i \neq j} \frac{A_{ij}}{M} \delta(\ell_i, \ell_j). \tag{2b}$$

Now let the probabilities of various label assignments be defined by the partition function

$$Z(\beta_c, \beta_\ell) = \sum_{c,l} e^{-\beta_c E(c) - \beta_\ell E(\ell)}, \tag{3}$$

where c and l are restricted to those with $N_{c\cdot} = \hat{N}_{c\cdot}$ and $N_{\cdot\ell} = \hat{N}_{\cdot\ell}$. That is,

$$\Pr(c, \ell \mid \beta_c, \beta_\ell) = \frac{1}{Z(\beta_c, \beta_\ell)} e^{-\beta_c E(c) - \beta_\ell E(\ell)}. \tag{4}$$

Given the edge weights A_{ij}, (2*ab*) and (3) define a two parameter family of probability distributions (4). The probability of any particular label assignment depends only on its energies (2*ab*). Each probability distribution (4) thus implies a probability distribution over energies, $E = E(c) + E(\ell)$. Given the actual data $(\hat{c}, \hat{l})$, the maximum likelihood estimates for β_c and β_ℓ are

$$\begin{aligned}(\beta_c^*, \beta_\ell^*) &= \arg\max_{\beta_c, \beta_\ell} \Pr\big(E(\hat{c}) + E(\hat{\ell}) \mid \beta_c, \beta_\ell\big) \\ &= \arg\max_{\beta_c, \beta_\ell} \Pr\big(E(\hat{c}) \mid \beta_c\big) \Pr\big(E(\hat{\ell}) \mid \beta_\ell\big) \\ &= \Big(\arg\max_{\beta_c} \Pr\big(E(\hat{c}) \mid \beta_c\big), \arg\max_{\beta_\ell} \Pr\big(E(\hat{\ell}) \mid \beta_\ell\big)\Big),\end{aligned} \tag{5}$$

where the second equality follows from the fact that the probability distribution (4) is a product distribution—since the two Potts models are not coupled, the random variables c and ℓ are independent. To compute the β^*s we run Markov Chain Monte Carlo (MCMC) simulations, collecting a large number of sample configurations after equilibration at a fixed β and computing the mean (expectation) value of E. The β_c and β_ℓ that make this equal to $E(\hat{c})$ and $E(\hat{\ell})$ are good approximations to the β^*s. As illustrated in Figure 6, for our network community and language data, $\beta_c^* \approx 11.6$ and $\beta_\ell^* \approx 2.7$.

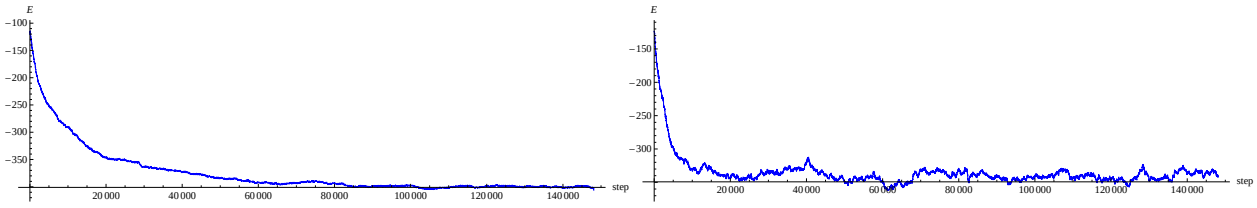


Figure 6. MCMC energies as a function of step, for $\beta_c^* = 11.6$ (left) and $\beta_\ell^* = 2.7$ (right). The horizontal axes are at $E(\hat{c}) \approx -401$ and $E(\hat{\ell}) \approx -349$.

The MCMC simulations start with a random label assignment restricted to have $N_{c\cdot} = \hat{N}_{c\cdot}$ (or $N_{\cdot\ell} = \hat{N}_{\cdot\ell}$). At each step the simulation chooses two random nodes and switches their labels. If the change in energy ΔE is negative, the switch is made. If it is positive, the switch is made with probability $e^{-\beta \Delta E}$. These dynamics satisfy detailed balance, so the

simulation equilibrates asymptotically. Thus after an equilibration period we can (approximately and approximately independently) sample from the equilibrium distribution defined by (2*a*) (or by (2*b*)).

Having estimated the β^*s, we run a new MCMC simulation of the *joint* distribution (4) with these values. (At each step the simulation chooses not only a random pair of vertices, but also either the c or the ℓ labels to switch at random.) Sampling from this simulation after it equilibrates, we compute the matrix-valued statistic that is the contingency table for the labels. The resulting ensemble of contingency tables takes the place of the ensemble of contingency tables introduced in §4.1, arising from N samples from the product probability distribution corresponding to the closest rank 1 contingency table (the one with $N_{c\ell} = \hat{N}_{c\cdot}\hat{N}_{\cdot\ell}/N$). That is, the former is the correct null hypothesis of independent labels, *conditioned on geographical position*. We can now, just as in the standard setting, choose any statistic (*i.e.*, function) of contingency tables, and estimate its distribution by computing it for each sample. Comparing the observed value of this statistic with the distribution tells us the probability that it is at least as large as it is.

The obvious statistic to use is χ^2 as defined by (1). Since this is a measure of distance from being rank 1, we consider also two other such measures: the Frobenius distance from being rank 1,

$$\hat{d}_1 = \sqrt{\sum_{k=2}^{\min\{|C|,|L|\}} \hat{\sigma}_k^2},$$

where σ_k is the k^{th} largest singular value of the contingency table [25], and the relative entropy compared with the closest product distribution,

$$\hat{S} = \sum_{c,\ell} \frac{\hat{N}_{c\ell}}{N} \log \frac{N\hat{N}_{c\ell}}{\hat{N}_{c\cdot}\hat{N}_{\cdot\ell}}.$$

The actual data gives values:

$$\hat{\chi}^2 \approx 6253 \qquad \hat{d}_1 \approx 143.6 \qquad \hat{S} \approx 2.204.$$

Figure 7 shows histograms of these statistics for samples from an MCMC simulation, together with histograms of the same statistics for samples from the ensemble of closest rank 1 contingency tables. They have means and standard deviations:

$$\chi^2 \approx 993 \pm 58 \qquad d_1 \approx 63.1 \pm 2.2 \qquad S \approx 0.498 \pm 0.020.$$

from which the *tiny* probabilities of sampling values as large as the actual values can be estimated. Notice, moreover, that the mean value of χ^2 for samples of the distribution defined by (2*ab*), (3) and (4) is extremely unlikely under the hypothesis of independence of C and L: it has a one-sided p-value of about 10^{-23}. So the spatially dependent model moves the baseline substantially, but the actual values are still very far away.

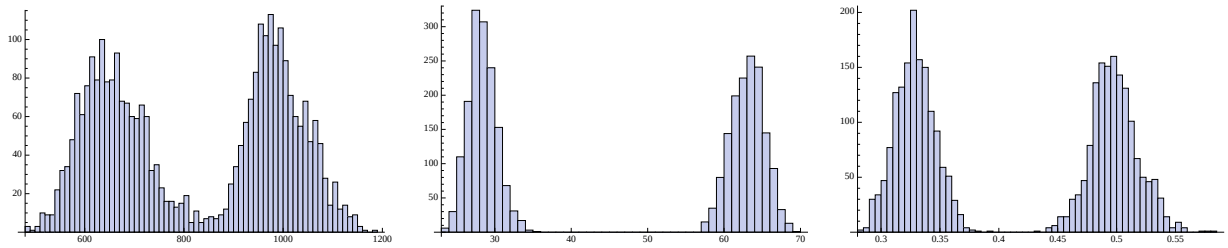


Figure 7. Histograms of χ^2, d_1 and S, respectively, calculated for samples from MCMC simulation of (2ab), (3) and (4) using β_c^* and β_ℓ^* (5), together with histograms of the same statistics sampled from the ensemble of closest rank 1 contingency tables. The latter lie far to the left of the former, while the actual data gives values far to the right. d_1 has the largest correction.

5. Quantifying the association of geographical partitions

The example on which we focussed in §4, namely the association between the network communities and a geographical partition (into languages) is a special case of the more general problem of quantifying the association between two geographical partitions. The communication network edge weights A_{ij} provide a measure of interaction between spatial locations that can be used to define two uncoupled Potts models even when neither takes values in C, the set of network communities. Thus this is a general method with which to handle the notorious "problem of spatial autocorrelation" [26], provided we have access to mobile communications data.

A particularly interesting example of the general problem arose in Côte d'Ivoire's 2010 Presidential election. The first round took place on 31 October 2010 and eliminated Aimé Henri Konan Bedie from the running, resulting in a run-off election between incumbent President Laurent Gbagbo of the Ivorian Popular Front/Front Populaire Ivoirie (FPI) and former Prime Minister Alassane Outtara of the Rally of the Republicans/Rassemblement des Republicains (RDR). During the second round of elections on 28 November 2010, Ouattara won an absolute majority of the vote and was elected to serve a 5-year term.

The first round of voting was perceived as demonstrating a high degree of correlation between the regional distribution of ethnic groups and electoral successes for each candidate:

> For starters, three of the top four vote-getters fully dominated the election in their own ethnic homelands. The central Baoulé people, long Ivory Coast's politically dominant group, voted overwhelmingly for their own candidate, Henri Konan Bédié, a former president of the country (1993-1999). ... Those regions [the northwest and north-center] were dominated by Alassane Ouattara, of Dyula paternal descent, who won more than 70 percent of the vote even in animist Senufo country. Ouattra polled fairly well in most of the rest of the country, but in the Baoulé heartland he received less than ten percent of the vote. Finally, Albert Mabri trounced all other candidates in the relatively small area occupied by the Dan. ... Laurent Gbagbo carried the major districts of his Bété-speaking homeland, but not decisively; he took only half of the

> votes, with Ouattara and Bédié splitting the rest. The Bété region is home to large numbers of immigrants from other parts of the country, most of whom likely cast their ballots their own ethnic "favorite sons". But Gbagbo did crush the other candidates across most of the southeast, the economic heartland of Ivory Coast as well as its most Christian region. He also did extremely well in some of the non-Muslim Mande areas of the west. [27]

Our methods can be used to analyze quantitatively such impacts of ethnic groupings on election outcomes *as well as the impact of election outcomes on the future political and economic development of specific regions*. International confirmation that the 2010 election was free and fair did not prevent incumbent President Laurent Gbagbo from refusing defeat and perpetrating violence against civilians [28]. Not only does this indicate that political success and political defeat can have significant repercussions for civilian safety, but it also suggests that successful candidates will use their power to develop supportive regions of the country economically, at the expense of their opponents' supporters. Thus mobile communications data will support policy-relevant analysis of cultural, political and economic aspects of development in Côte d'Ivoire and other developing countries.

References

[1] International Telecommunications Union, *Measuring the Information Society 2012* (Geneva, Switzerland: ITU 2012)

[2] V. D. Blondel, M. Esch, C. Chan, F. Clerot, P. Deville, E. Huens, F. Morlot, Z. Smoreda and C. Ziemlicki, "Data for development: The D4D challenge on mobile phone data", `arXiv:1210.0137 [cs.CY]`.

[3] E. G. Ravenstein, "The laws of migration", *Journal of the Statistical Society of London* **48** (1885) 167–235.

[4] E. C. Young, "The movement of farm population", *Cornell Agricultural Experimental Station Bulletin* **4269** (Ithaca, New York: 1924).

[5] W. J. Reilly, *Methods for the Study of Retail Relationships*, University of Texas Bureau of Business Research, Research Monograph No. 4, Bulletin No. 2994 (November 1929).

[6] G. K. Zipf, "The $\frac{P_1 P_2}{D}$ hypothesis: on the intercity movement of persons", *American Sociological Review* **11** (1946) 677–686.

[7] J. Tinbergen, *Shaping the World Economy: Suggestions for an International Economic Policy* (New York: The Twentieth Century Fund 1962).

[8] A.-C. Disdier and K. Head, "The puzzling persistence of the distance effect on bilateral trade", *The Review of Economics and Statistics* **90** (2008) 37–48.

[9] G. Krings, F. Calabrese, C. Ratti and V. D. Blondel, "Urban gravity: a model for inter-city telecommunication flows", *Journal of Statistical Mechanics: Theory and Experiment* (2009) L07003/1–8; doi:10.1088/1742-5468/2009/07/L07003.

[10] D. Krioukov, M. Kitsak, R. Sinkovits, D. Rideout, D. Meyer and M. Boguñá, "Network cosmology", *Nature Scientific Reports* **2** Article no. 793 (2012) 1–6, doi:10.1038/srep00793.

[11] S. Appleby, "Multifractal characterization of the distribution pattern of a human

population", *Geographical Analysis* **28** (1996) 147–160.

[12] W. R. Tobler, "A computer movie simulating urban growth in the Detroit region", *Economic Geography* **46** (1970) 234–240.

[13] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks", *Physical Review E* **69** (2004) 026113/1–15.

[14] M. E. J. Newman, "Analysis of weighted networks", *Physical Review E* **70** (2004) 056131/1–9.

[15] V. D. Blondel, J.-L. Guillaume, R. Lambiotte and E. Lefebvre, "Fast unfolding of communities in large networks", *Journal of Statistical Mechanics: Theory and Experiment* (2008) P10008/1–12; doi:10.1088/1742-5468/2008/10/P10008.

[16] "The Louvain method for community detection in large networks", `http://perso.uclouvain.be/vincent.blondel/research/louvain.html`.

[17] M. P. Lewis, ed., *Ethnologue: Languages of the World*, Sixteenth edition (Dallas, TX: SIL International 2009); online version: `http://www.ethnologue.com/`.

[18] N. M. Mulmi, "Ivorian crisis: The intrigues part I", *The African Executive* (2011) `http://www.africanexecutive.com/modules/magazine/articles.php?article=5675&magazine=319#`.

[19] T. Ogwang, "The root causes of the conflict in Ivory Coast", *The Africa Portal Backgrounder*, No. 5 (2011); `http://www.africaportal.org/articles/2011/04/27/root-causes-conflict-ivory-coast`.

[20] K. Pearson, "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling", *Philosophical Magazine*, Series 5 **50** (1900) 157–175.

[21] R. A. Fisher, "On a distribution yielding the error functions of several well known statistics", *Proceedings of the International Congress of Mathematicians*, Vol. 2 (1924) 805–813.

[22] A. Cerioli, "Testing mutual independence between two discrete-valued spatial processes: A correction to Pearson chi-squared", *Biometrics* **58** (2002) 888–897.

[23] J. Theiler, "Estimating fractal dimension", *Journal of the Optical Society of America A* **7** (1990) 1055–1073.

[24] R. B. Potts, "Some generalized order-disorder transformations", *Mathematical Proceedings of the Cambridge Philosophical Society* **48** (1952) 106–109.

[25] C. Eckart and G. Young, "The approximation of one matrix by another of lower rank", *Psychometrika* **1** (1936) 211–218.

[26] A. D. Cliff and J. K. Ord, "The problem of spatial autocorrelation", in A. J. Scott, ed., *Studies in Regional Science* (London: Pion Press 1969) 25–55.

[27] M. W. Lewis, "Ethnic dimensions of the conflict in Ivory Coast", *GeoCurrents* (28 April 2011) `http://geocurrents.info/geopolitics/ethnic-dimensions-of-the-conflict-in-ivory-coast`.

[28] United Nations Operation in Côte d'Ivoire (UNOCI), "Post-election crisis", `http://www.un.org/en/peacekeeping/missions/unoci/elections.shtml`.

Development, Information and Social Connectivity in Cote d'Ivoire

Clio Andris and Luís M. A. Bettencourt

Santa Fe Institute, 1399 Hyde Park Rd, Santa Fe NM 87501, USA

{clio, bettencourt}@santafe.edu

Abstract

Understanding human socioeconomic development has proven to be one of the most difficult and persistent problems in science and policy. Recent developments suggest that the key to progress lies in the consideration of processes where new information is created and embedded in the structure of social networks at a diverse set of scales, from nations to cities and firms. We formalize these ideas in terms of network theory and analyze the D4D Challenge telecommunication in this light to show how incipient socioeconomic connectivity may constitute a general obstacle to development. We also propose a set of further tests of our ideas using telecommunications data and potential measures that we expect would promote socioeconomic development through increases in specific types of social connectivity.

Introduction

The problem of understanding social and economic development in human societies is one of the most fundamental questions in science and policy [1,2]. Despite many studies and general ideas, informed by a history of interventions over the last few decades, the problem has remained stubbornly resistant to useful scientific syntheses [2,3]. As a result, policies and interventions aimed at spurring human development at different levels – from individuals to whole nations – have remained very limited in their successes, and the design of even basic strategies remains the focus of controversy.

Recently, more empirical approaches to problems of development have gained traction, from randomized control trials [3] and census of informal communities [4] on small scales to big(ger) data approaches to the structure and dynamics of developing human societies, e.g. via worldwide remote sensing [5] and the analyses of national mobile communication networks [6,7]**.** Nevertheless, there remains a large gap between what these methods and technologies can measure and a detailed understanding of the processes that underpin socioeconomic development.

A myriad of observations and previous experiences suggest that successful human socioeconomic development requires the simultaneous solution of many problems at once [8,9]. This is not the statement that all problems need to be solved together and once and for all, but rather that incremental progress across many different dimensions of people's lives, from access to better services to security and education, tend to work better than sophisticated solutions to single issues, which are usually not sustainable. This was also the path taken historically by nations that today are considered developed. Thus, solutions targeted at solving one or few aspects of

the problem, e.g. epidemic outbreaks [10], emergent political violence [11], or other humanitarian crises [12,13], though certainly important, are not sufficient or even effective at promoting self-sustaining development. Many nations without these problems display manifest challenges of development, while other societies, where these problem do occur, show clear and sustained growth.
Therefore, there is an acute need for new concepts and methodologies that can inform what development is in terms of specific socioeconomic structures and their dynamics in space and time. Here, we address this problem through a novel synthesis of ideas from economics and other social sciences, grounded on recent developments in network theory and the role of measurements made possible by human interaction networks. Specifically, we address the fundamental nature of development dynamics as a socioeconomic network process, show how it can be measured using telecommunications data from the D4D challenge [14], and suggest how a general path of development in Cote d'Ivoire could be observed and possibly stimulated. We also propose a set of possible additional measurements and policy interventions that provide tests to the concepts developed here.

Together our approach connects important recent theoretical and technical developments in the structure and dynamics of networks to fundamental problems in human societies in a way that can be studied scientifically and successively improved via cycles of policy intervention and empirical observation.

Economic Development Theories

We start by providing some context on general ideas of development and economic growth, and some of their shortcomings when applied to the case of Cote d'Ivoire, as well as of most other developing nations.

The problem of understanding growth lies somewhat outside the main stream of classical economics as it requires that we let go of some concepts of optimality and equilibrium central to microeconomic formulations [15].

Modern theories of endogenous growth, developed through most of the 1990s [15,16], emphasized the creation of new information, in the form of new products, recipes or algorithms, as the source of economic growth. In a nutshell, these theories emphasize the non-rival (non-exhaustible) nature of information, which is qualitatively different from other factors of production, such as labor and capital, as the source of productivity gains in an economy. To put it simply, new money comes from new ideas. Human capital, usually measured in terms of people's education attainment, plays a central role as the producers of this information, while education, R&D, etc must be paid through savings (foregone consumption) to be invested into these activities. While these concepts clearly reflect some important ingredients of economic growth and of development more generally, they provide no specific theory about where information resides in terms of social structures and specific economic agents, such as individuals and firms.

The situation in Cote d'Ivoire provides some examples of this difficulty (though these are by no means unique). At present in Cote d'Ivoire, people with university and other advanced degrees do earn higher wages [17], but also experience some of the highest unemployment rates in the

nation and the longest periods of job search [18]. So, the production of a generally more educated population per se cannot be the answer. A greater emphasis on how innovations are learned and used in human societies is necessary

Recent emphases on the quality of governance and institutions [19] and on the structure of space of capabilities or products [20,21], though surely providing other important ingredients to development, fail to explain how they are integrated in detail in human societies to generate growth.

A generic clue to issues of integration is provided by the fact that informal employment dominates the economy of most developing nations, including Cote d'Ivoire, and is usually associated with low levels of specialization and coordination of labor, and with small value added economic activity [22]. Thus, the crucial question becomes why do developing economies not quickly adopt strategies to promote greater division and interdependence of labor across scales, resulting in both greater value of economic output and on open ended cycles of organizational and technological innovation.

Development, Information and Social Connectivity

A set of interdisciplinary concepts, grounded on the spatial and temporal characteristic of social connectivity and its evolution as a complex adaptive system hold the key to answering this question.

Development is first and foremost the open-ended process of gaining social and economic access to a society at large. Recent ethnographic studies in sociology, especially in Latin America, [9] have emphasized how new urban migrants living in informal settlements have gradually strived for services and citizenship rights and responsibilities. As a result, large, poor and marginalized fractions of the population joined their civic society and formal economies, and gradually raised their socioeconomic status and that of their societies. These processes, that connect the aggregation of information in human societies to scale, innovation and economic growth are the basis of another set of ideas in economics and the social sciences [8,23-25]. However, these dynamics have largely remained untested because the observation of large-scale social structure was technically impossible until recently. We show below, that networks of tele-communications, provide a new window into these fundamental social dynamics.

Developed nations are characterized by high levels of individual and organizational specialization (functional diversity) and their integration at least at 3 fundamental levels: i) the urban system (nation), ii) cities and iii) firms and other social organizations. At each of these scales we should expect network structures that entail the exchange of information, as well as people and goods, involved in socioeconomic processes of growth and innovation. Here, we develop the case that such structures and their dynamics are visible in networks of telecommunication, to which we now turn.

Results

Below we analyze the D4D challenge telecommunication data made available for Cote d'Ivoire. We explore the quantitative characteristics of networks of connectivity at the national and urban levels. More

technical details of the data and of our analysis are given in Materials and Methods, at the end of the paper.

The structure of Cote d'Ivoire's urban system

The creation of urban systems as sets of separate but inter-dependent cities is essential for a nation to develop. It is across places of different sizes that the advantages of large cities, in terms of innovation and organization, pay off. This happens through the incorporation of new ideas and organizational forms into the structure of industrial and primary production, which in turn forms the basis for the economies of smaller cities, and the material basis for nations as a whole.

In theories of central place [26,27], later elaborated by modern economic geography [28], the crucial feature that provides the basis for intercity dependences is a functional hierarchy of economic functions. This means that the largest city in the system contain all economic functions observed in smaller ones, but not the reverse. As such, larger cities supply services (innovation, information, organization) to small cities in their territory, in exchange for food and other material goods. The largest city in an urban system is thought to service, in this way, an entire nation. Despite some urban specialization, such as in large scale manufacturing, that sometimes presents exceptions, these trends are characteristic of most developed nations [29].

Unfortunately, detailed and reliable information about economic functions is hard to come by at the city level, especially in developing economies. Nevertheless, telecommunication data gives us an entry point to investigate to what extent Cote d'Ivoire's urban system is spatially integrated, and to measure the roles of distinct cities in light of the expectation of a national urban hierarchy, expressed in terms of telecommunication call flows.

Figure 1-3 show the structure of the networks of calls between places (prefectures) in Cote d'Ivoire. Figure 1 shows the total number of calls (placed and received) between any two places in the urban system. Immediately apparent, as would be expected from central place theory arguments, is the importance of its largest city, Abidjan (3.8 million people), in terms of the diversity and strength of calls exchanged with many other cities in the nation. However, even at this level we start to observe that links to Northern and Western parts of the country are relatively sparse. The political capital Yamoussoukro also plays no particular role in this network and is in fact strongly connected to the two large population centers, Abidjan and Bouaké (775 thousand people in 2002).

These patterns become clearer when we consider calling patterns between two places on a per capita basis. That is, when we ask what the typical calling patterns of an individual subscriber in a given city may be. Figures 2 and 3 show the calling patterns placed and received per capita, respectively, between two places. We see in Figure 2-3 that larger cities, and Abidjan in particular, are the focus of a large number of calls placed elsewhere. These call patterns show that it is probably not too crude a misrepresentation to say that much of the nation of Cote d'Ivoire is actively listening to what happens in Abidjan. As for Abidjan itself, most of the calls it originates (84%) are to other parts of the city, see below.

The second largest city, Bouaké. plays an analogous role, but over only its immediate neighboring region. Interestingly, the political capital Yamoussoukro, does not escape this pattern: Its strong connections to the two largest cities are more in placing calls rather than receiving them. Other economically important regional centers, such as San Pedro (second largest port, after Abidjan), have more mixed patterns of connectivity, which are to a large extent separate from those of the other larger cities. In this sense the Northern and Western parts of Cote d'Ivoire are largely disconnected, especially on a per capita basis, from the mains economic and political centers of Cote d'Ivoire.

This suggests that the Ivorian urban system is still very much incipient. Most cities display strong communication links regionally, but it is clear that even Abidjan (economic capital) and especially Yamoussoukro (political capital), though displaying a greater reach than smaller cities, fail to maintain a network of communication with most of the nation, especially the North and the West, which are, not coincidentally, the poorest parts of the nation.

This suggests that a path for national development must entail improvements in the integration of the nation as a whole, and should be facilitated and accompanied with the observation of increasing call rates between all places, as especially between the larger cities and the West, South West and North of the country, see below.

Urban agglomeration effects

Large cities are often described as the social, political and economic engines of most (developed) nations. However, it has also been argued that recent urbanization in Africa has failed to deliver on its promise for economic development [30] and has in fact proceeded without much in the way of measurable economic progress. These issues are subtle, however, as we shall discuss below in the context of Cote d'Ivoire, where the socioeconomic gains (and some of the challenges) of its largest city are clear, and yet its consequences on national growth rates remain relatively remote.

Generally, historical and contemporary patterns of national development very much depend on the socioeconomic dynamics that happens inside a nation's largest cities, and in particular on the ability of these places not only to grow but to *realize* increases in social interactivity that can lead to larger and more sophisticated economic specialization and interdependence, organizational and technological innovation and the seizing of latent economies of scale in services and infrastructure [25,31].

Arguments from complex systems theory and from urban economics emphasize the role of *agglomeration* economies in all these processes: the output of socioeconomic processes rises on a per capita basis with the size of cities. This is interpreted in terms of the possibilities for interaction created by spatial and temporal concentration of people in cities [32]. However, even if these conditions are met, the question remains whether cities realize these interactions for good, rather than ill. Negative consequences of increased human interactivity can also occur, in terms of increases in crime rates, the prevalence of infectious diseases and the proliferation of small-scale informal economic agents. Cote d'Ivoire, and Abidjan in particular, manifest all these

potential consequences of urbanization. The good and the bad go very much hand in hand.

We start by showing the parallel analysis of regional connectivity to Figures 1-3, now for parts of Abidjan, see Figures 4-5. Figure 4 show the network of calls between parts of Abidjan and its surrounding region, while Figure 5, show this picture in greater detail for the central city. We observe strong general connectivity between almost every part of the city, despite great differences in their dominant socioeconomic character and relatively large physical distances. At the metropolitan level it is clear from Figure 4 that Abidjan thoroughly integrates its central communes, between each other and with surrounding population centers, such as Anyama, Bingerville, Bonova, Dabou, Grand–Bassam and Songon. Figure 5 shows how central communes in the city exchange information even more frequently than with these other adjacent areas. Especially noteworthy are the almost parallel roles of the two population centers of Yopougon (most populated commune) and Cocody (most affluent residential district), and in a different way the commercial and business centers of Adjamé and Plateau. Interestingly, the commune of Abobo, which also has a larger population, to a large extent of internally displaced migrants, is less connected than Yopougon to the rest of the city, for example. Nevertheless, these figures strongly suggest strong communication patterns between functionally and economically very different parts of the city, which are a sign of functioning urban center as a mixing social network [32].

This suggests that even if Cote d'Ivoire remains relatively disconnected as an urban system, more local socioeconomic connectivity *inside* its cities (in Abidjan, at any rate) seems to already be present and able therefore to facilitate general processes of urban agglomeration.

To test this idea in greater detail beyond Abidjan we performed a simple scaling analysis for the total number of calls received by each prefecture as function of their population. These patterns are well described on average by a power law function [32-34] where the connectivity, C, is a function of population size N, $C=C_0 N^b$. The parameter b-1 measures the on average increase in social connectivity per capita with city population size. We observe, b=1.147 (95% Confidence Interval [1.10,1.20]) implying that connectivity per capita increases by about 15% with each doubling in population size. These scaling effects are in line with patterns recently measured for Portugal and the United Kingdom [35], and predicted by urban scaling theory [32]. They suggest a general acceleration of social economic processes with city size in agreement with many other nations, developing and developed [32-24]. However, in line with the observations of the previous section, we find that the growth in connectivity with population size that is due to *internal* calls within each city is faster, with a b=1.26. This means in practice that as cities grow the fraction of all calls that is internal to the city increases. For Abidjan that is 84% of all calls initiated in the city, illustrating again how most of its information finds uses only locally.

Moreover, the joint signature of urban agglomeration effects and of an urban hierarchy should be visible in patterns of diversity of connection with city size (and therefore of economic productivity, see

below). This has in fact been recently observed for the UK using telecommunication networks [7]. Figures 6 and 7 show that this is indeed the case using two simple measures of diversity, the total number of different places called and the Shannon entropy of the distribution of places called, respectively.

That cities in Cote d'Ivoire generally realize agglomeration effects is good news. However, it is also important to understand what effects of urbanization are enabled by social interactions, benign or malign. The recent history of Cote d'Ivoire, which has many parallels with other examples of urbanization in Africa and Latin America, indicates that its most recent rapid urbanization is largely the result of conflict and political crisis, and less that of well-planned migration to access social and economic opportunity. Abidjan, for example, has grown explosively in population during the last decade of conflict in the region, due to both internally displaced people and refugees from neighboring nations.

Nevertheless, and despite these challenges, rates of poverty in Abidjan are lower than in all other parts of the country (21% compared with 49% for the nation), especially in comparison with rural areas. Urban GDP is largely unknown. However general estimates suggest that Abidjan is responsible for about 40% of national GDP, whereas Bouaké accounts for about 3% and San Pedro for 4% (San Pedro is an important port for Cocoa exports). Accounting for their respective populations results in GDP per capita of 3.677, 1.337 and 4.857 thousand dollars respectively, all considerably larger than Ivory Coast's national GDP per capita of 1.062. Perhaps clearer are the results of the most recent Survey of Living Standards of Households in 2008. Its findings regarding personal income illustrate more fully the relative economic advantages of urban centers and of Abidjan in particular. The findings of this survey estimate that all urban centers in the nation manifest larger incomes than their surrounding rural areas by factors of 1.3 to 1.9 (national average is that urban incomes are 1.82 larger than rural ones). The annual average per capita income in Abidjan is by far the largest in the nation at about 561 thousand CFAF (roughly $1140US) compared with 372 thousand CFAF (roughly $695US) for the national average. In 2007 Mercer Human Resources Consulting, who rank cities around the world in terms of their quality of life, placed Abidjan in 35th place among the most expensive cities in the world!

Thus, economic urban agglomeration effects are at play in Cote d'Ivoire, even if national GDP (and incomes) may have recently decreased in real terms.

Other more sinister urban agglomeration effects are also at play and point to some of the challenges of development in Ivorian cities. It has been estimated [36] that 4.7% of the country's population is infected with HIV/AIDS. In Abidjan however the prevalence of the virus is much higher, at 6.1%. Already in 1997, about 40% of hospital beds in Abidjan were occupied by HIV/AIDS patients.

Urban insecurity is also extremely high, though few reliable numbers exist. A recent report [37] stated that in the first half of 2008, out of a total of 62,424 offences to the penal law registered by the National Police, 75% were registered in the district of Abidjan.

In parallel, economies of scale in urban infrastructure, a general characteristic of

cities worldwide, seem to be at best only incipient in Cote d'Ivoire. While it is in the nations large cities that access to sanitation, treated water, power, and other general services is at all possible [17] these services work mostly intermittently and in small scales, thus squandering some of the possible system wide savings made possible by large population concentrations.

All these results, and many more relating to the concentration of services, education, etc, strongly suggest that urban agglomeration effects are already at play in Cote d'Ivoire. However, many of the negative consequences of increased social connectivity – in terms of violence and infectious diseases for example – seem to trump some of its economic gains. This pattern is historically typical. Development has always happened as large cities, plagued by these problems, develop infrastructural, political and civic institutions that allow them to systematically solve problems of population agglomeration. How mobile communications may play a new role in enabling such solutions will be discussed below.

Entrepreneurship and Informality

Finally, and more speculatively, the somewhat poorly known structure of employment and economic structure in Cote d'Ivoire suggests that most economic activity in the urban centers of Ivory coast are small and unspecialized, with some exceptions in terms of electricity production, cement and the international ports of Abidjan and San Pedro.

The informal sector is in fact been described as 'vibrant' in a recent IMF report [17]. It is estimated to have occupied 4,107,595 people in 2002 vs. about 1,698,300 workers in 1995, an increase of 142% over 7 years. Much of the informal sector is rural, but general estimates put informal employment in Abidjan at about 75% [38].

Given the increases in social connectivity afforded by urbanization and mobile communication, why then are larger and more sophisticated firms the norm, rather than the exception?

In economics, firms are thought to emerge as a result of the minimization transaction costs, which must always be incurred in real markets [39]. These costs become prohibitive when sophisticated production, involving the integration of specialized skills is necessary. Mobile telecommunications play a potentially interesting role in promoting social coordination, minimizing market transaction costs, and in potentially promoting social organization that could result in the growth and sophistication of production. Thus, we hypothesize that in many informal economies cell phones may have encouraged on-demand labor arrangements that may in fact defeat the potential for modern large firms to emerge. Certainly, that seems to be the case of Cote d'Ivoire in recent years. The observation of most cell phone calls during working hors, Figure 8, is consistent with these trends, though it may occur also for other reasons.

An example from the transportation sector illustrates how informal market solutions have been trumping potentially more efficient large-scale services [17]. In Abidjan, mass transportation is mainly operated by the public company SOTRA. In 2000, after a decade of disinvestment, its services translated into one bus for more than 4,500 inhabitants, with

inevitable disaffection for its services. To fill this gap for an essential urban service, informal transport operators using minibuses (called "gbakas" and "504s") rose up to the occasion. Their fleets are estimated at about 6,026 vehicles, while that of regular taxis also swelled up at about 8,000 vehicles, with yet another solution (communal taxis or "wôro-wôrôs") accounting for about another 11,971 units. There are immense opportunities for coordination, integration and formalization of these thriving informal businesses.

The scaling up in scope and quality of informal urban businesses into modern economic sectors is an immense opportunity for mobile communications. New pricing models and innovative uses may provide the means to make these businesses visible, measurable and scalable, while preserving their economic basis and essential services to the public. Such measures may also promote greater trust and satisfaction at the urban level and new economic models that can be exported to other cities across Cote d'Ivoire.

Promoting Social Connectivity for Development

We have noted that some of the challenges of development in Cote d'Ivoire, as in other developing nations, may result from a lack of appropriate socioeconomic integration capable of encouraging individual specialization and interdependence, from the level of the firm to that of the nation.

We now wish to turn the problem around and suggest that the role of mobile communications in developing societies may be changed from diagnostic to cure: If the problem is the promotion of certain kinds of social connectivity, then new uses and subscriber models should help promote desirable functional solutions.

At the national level, greater integration may be obtained by promoting better coverage and product adoption in rural areas and by pricing long distance calling, especially with urban centers sufficiently attractively. Economic activity could be further promoted with plans that emphasize these properties targeted at businesses, see below. Calling distant and rural areas from urban centers, and promoting urban solutions related to issues of modern technological practices in agriculture, services, telemedicine, etc, may help spread urban know-how to poor and remote areas.

Inside cities, problems seem to be less predicated on the lack of overall connectivity, but rather on the promotion of its uses for socioeconomically productive ends. Another question is how to encourage better urban services at larger scales, by exploring latent economies of scale. Calls and messaging that can convey information about public services, and organize the public to demand their improvement could be made very inexpensive, for example. Crowd sourcing models for sharing information about the quality and quantity of public services and make these visible beyond the city and nation may also leverage external influence to promote better organizational, political and technological sustainable solutions. A system for demanding and rating the quality of public services, such as police, and to create e.g. crime hot-spot maps, that are visible not only to urbanites directly involved, but to the world at large will also create a system of incentives for development.

Finally, subscriber pricing models that encourage the formation of visible and formal small firms, by shifting the structure of corporate transaction costs, may help create a culture of small formal entrepreneurship that is the basis of most job creation and innovation in developed societies. This may be achievable by making the costs of communication cheaper for small formal businesses with each other, but more expensive between private individuals. The role of telecommunications in making financial transactions more transparent and formal is also an area that holds much promise in cash fund transfer models using mobile devices, for example.

Discussion

Mobile telecommunications have quickly come to dominates most distant human interactions, playing a fundamental new role in the communication of information and the coordination of socio-economic activities in human societies. In Cote d'Ivoire this market is exploited by four operators, which share more than 9 million subscribers, out of a population of about 20 million. Of these Orange claims a number of about 5 million subscribers. Thus, issues resulting from variable regional market share and biased geographical coverage may affect our results. This information was not provided with the data for the D4D challenge but should be incorporated in future analysis in order to make scientific findings reliable and useful.

Beyond these important empirical issues, we proposed here a framework to formalize socioeconomic development in terms of measurable telecommunications network data at three levels of aggregation: individuals and firms, cities and the urban system. We gave evidence from D4D data for Cote d'Ivoire that lack of socioeconomic integration is likely a general factor impeding development and suggested both diagnostics and solutions to promote the dynamics of connectivity and information exchange that can lead to processes of sustainable development.

A telephone call is a mutual desire to transfer information over geographic space. Studies show that in developing countries, the telephone is used more often for personal reasons than for business [40]. Yet, in the informal economy, many social or business ties can hard to distinguish, as friends and family often provide job opportunities.

Information transfer is important for finding work for day laborers in the informal economy. In this case, calls are evidence that information is needed about job availability, conditions for work (ex. weather, daily pay), availability and price of market goods (ex. crops) for sale, or possibilities for demand for market sale of personal goods. In a formal economy, these calls would not be necessary because there are long-term labor arrangements and organizations that ensure the reduction of transaction costs related to incipient and organization.

A change in communication patterns may be an indicator of the growth of a more stable economy, with continuity in job availability and wages, as well as continuity in services and goods provided. These services can trickle down to more formalized and stable healthcare, education, public services, infrastructure creation and competition in the global economy through activity and employment. Finally, because human communication underlies all development processes, telecommunication services

have the potential for taking an active role in promoting the kind of social interaction that can promote sustainable growth outcomes.

Clearly these hypotheses will require extensive testing and refinement, and access to more data over time and with more careful consideration of important socioeconomic units of analyses, from individuals to businesses and places. When this is done, however, we believe that network analysis of big human interaction data can start delivering on its potential to be a transformative tool to promote the socioeconomic development of all human societies.

Materials and Methods

Telecommunication Data Sources

Call Detail Records (CDR) from Cote d'Ivoire from December 2011 – April 2012 are provided by Orange within the framework of the Data for Development (D4D) Challenge. Origin and destination of cell towers are identified by their longitude and latitude locations within Cote d'Ivoire; 1231 cell tower locations are also provided by the Orange D4D Challenge. Of these, 1094 are associated with either an incoming or outgoing call during the period. Calls that originate from a "-1" tower were eliminated from our analyses

Only pairs of cell towers exchanging 50 or more calls over 5 month are included in our analysis. All calls are directed. Data is processed using the R package (version 2.15.1) (R Studio (2012) The R Foundation for Statistical Computing)

Demographic Assignment

The population of Cote d'Ivoire is assigned to each of 1231 cell tower locations provided by the Orange D4D Challenge. We use population grid data provided by Afripop.org. In this dataset, each grid cell in the raster contains a value. We estimate a 30 kilometer range from each cell tower point (assuming GSM typical ranges), which captures a total of 18,842,000 people, and renders 256,411 out of range.[1] Each person is assigned to the cell tower closest to his or her home location (e.g. through a Voronoi/Thiessen polygon).

Geographic Aggregation and Connectivity Patterns

Lines and self-loops between administrative units are created in the ESRI ArcGIS 10.1 environment. The geographic coordinate system used is GCS-WGS-1984 and projection is GCS-Cote-d-Ivoire.

Calls and population are then aggregated to one of 235 prefectures and one of 50 districts, based on the location of the cell tower. If the cell tower location falls within the boundaries of these administrative units, their call data (number of incoming calls, number of outgoing calls, and population) is summed within the unit. Incoming and outgoing calls involving the Abidjan Prefecture account for the majority of calls.

Similarly, Abidjan is divided into communes and geocoded from a number of maps of the area (provided by Microsoft Bing and Google Maps). The set of geographic units consists of 10 communes, as well as three suburbs, Bingerville, Songon and Anyama. Calls are then aggregated based on cell towers within these spatial units.

Prefecture-to-prefecture calls range from 101 calls (exhibited by many prefecture

pairs, such as M'Bengue to Toumodi) to 105,351,197 calls that both originated from and were received in Abidjan. District-to-district calls range from 101 calls (exhibited by three district pairs such as Danane to Adzope) and are also at a peak with internal Abidjan calls.

References

[1] Sen A. (1999) Development as Freedom, (Alfred A. Knopf, New York NY)

[2] Easterley W. R. (2002) The Elusive Quest For Growth (MIT Pres, Cambridge, MA)

[3] Duflo E., Banerjee A. (2011) Poor Economics (Public Affairs, Jackson TN)

[4] Patel S., Burra S, d'Cruz C., (2001) Slum/Shack Dwellers International (SDI) - foundations to treetops, *Environment and Urbanization* 2001 13: 45

[5] Angel S (2012) Planet of Cities. Cambridge, MA: Lincoln Land Institute. 340 p.

[6] Eagle, N, de Montjoye Y., Bettencourt L. (2009), "Community Computing: Comparisons between Rural and Urban Societies using Mobile Phone Data", IEEE Social Computing, 144-150.

[7] Eagle, N., Macy, M. and Claxton, R. (2010) Network diversity and economic development. Science 328: 1029-1031.

[8] Jacobs J (1970) The Economy of Cities. New York: Jonathan Cape Ltd. 280 p.

[9] Holston J (2008) Insurgent citizenship (Princeton University Press, Princeton NJ)

[10] Flu Phone: http://www.cam.ac.uk/research/features/flu phone-disease-tracking-by-app/

[11] Robertson C, Sawford K, Daniel SLA, Nelson TA, Stephen C. Mobile phone–based infectious disease surveillance system, Sri Lanka. Emerg Infect Dis 16: October 2010.

[12] Sarcevic A., Leysia Palen, Joanne White, Mossaab Bagdouri, Kate Starbird, Kenneth M. Anderson, (2012). "Beacons of Hope" in Decentralized Coordination: Learning from On-the-Ground Medical Twitterers During the 2010 Haiti Earthquake *2012 ACM Conference on Computer Supported Cooperative Work*, Bellevue, WA.

[13] Starbird, Kate and Leysia Palen (2011). "Voluntweeters:" Self-Organizing by Digital Volunteers in Times of Crisis. To appear in the *ACM 2011 Conference on Computer Human Interaction (CHI 2011)*, Vancouver, BC, Canada.

[14] D4D Challenge: http://www.d4d.orange.com/home

[15] Barro R. J., Sala-i-Martin, X. I., (2003) Economic Growth (MIT Press, Cambridge MA)

[16] Romer P. M. (1994) The Origins of Endogenous Growth, The Journal of Economic Perspectives 8:3-22.

[17] International Monetary Fund (2009) Côte d'Ivoire: Poverty Reduction Strategy Paper, IMF Country Report No. 09/156.

[18] Direction des Services Socioculturels et de la Promotion Humaine de la Mairie du Plateau, Rapport sur le contexte socio-économique en Côte d'Ivoire (2009), Available online at http://recap.itcilo.org/fr/documentation/files-eloise/rapport-sur-le-contexte-socioeconomique-en-cote-d-ivoire

[19] Acemoglu D., Johnson S., Robinson J. A. (2005) Handbook of economic growth, Chapter 6 Institutions as a Fundamental Cause of Long-Run Growth.

[20] Leontief, Wassily W. *Input-Output Economics*. 2nd ed., New York: Oxford University Press, 1986.

[21] Hidalgo C.A., Klinger B., Barabasi A.-L., Hausmann R. (2007) The Product Space Conditions the Development of Nations. Science 317: 482-487.

[22] Losby J. L. et al. (2002) Informal Economy Literature Review. Available online at http://www.kingslow-assoc.com/images/Informal_Economy_Lit_Review.pdf

[23] Hayek FA (1945) The use of knowledge in society. American Economic Review 35: 519-530.

[24] Arrow K (1962) The economic implications of learning by doing. Rev. Econ. Stud. 29: 155-173.

[25] Bettencourt LMA, Samaniego H, Youn H (2012) Professional diversity and the productivity of cities. Available at http://arxiv.org/abs/1210.7335

[26] Christaller W (1966) Central Places in Southern Germany. New York: Prentice Hall.
[27] Lösch A (1954) The Economics of Location. (Yale: Yale University Press.)
[28] Fujita M, Krugman P, Venables AJ (2001) The Spatial Economy: Cities, Regions, and International Trade. Cambridge, MA: MIT Press. 384 p.
[29] Mori T, Nishikimi K, Smith TE (2008) The Number-Average Size Rule: A New Empirical Relationship Between Industrial location and City Size. J. Reg. Sci. 48: 165-211.
[30] The Economist (2012) The urbanization trap: http://www.economist.com/blogs/graphicdetail/2012/10/daily-chart
[31] Jones C, Romer P (2010) The New Kaldor Facts: Ideas, Institutions, Population, and Human Capital. American Economic Journal: Macroeconomics 2: 224-245.
[32] Bettencourt LMA (2012) The Origin of Scaling in Cities. Available http://www.santafe.edu/media/workingpapers/12-09-014.pdf
[33] Bettencourt LMA, Lobo J, Helbing D, Kuehnert C, West GB (2007) Growth, innovation, scaling, and the pace of life in cities. Proc. Natl. Acad. Sci. U.S.A. 104: 7301-7306.
[34] Bettencourt LMA, West GB (2010) A unified theory of urban living. Nature 467: 912913.
[35] Schlaepfer M, Bettencourt LMA, Raschke M, Claxton R, Smoreda Z, et al. (2012). The Scaling of Human Interactions with City Size. Available at http://arxiv.org/abs/1210.5215
[36] HIV report
[37] violence report
[38] Programme des Nations Unies pour le Développement (PNUD) 2003, Rapport national sur le développement humain en Côte d'Ivoire 2004 Cohesion Sociale et Reconstruction Nationale, Available online http://www.ci.undp.org/publication/Cote%20Ivoire%20HDR%202004.pdf
[39] Coase R.H. (1937) The nature of the firm, Economica 16: 386-405.
[40] Donner J. (2009) Blurring livelihoods and lives: The social uses of mobile phones and socioeconomic development. Innovations: Technology, Governance, Globalization, 4**:** 91-101. Available at http://ideas.repec.org/a/tpr/inntgg/v4y2009i1p91-101.html

Figure Captions:

Figure 1: Call Network for the Urban System of Cote d'Ivoire.

Figure 2: Network of Calls Placed for the Urban System of Cote d'Ivoire.

Figure 3: Network of Calls Received for the Urban System of Cote d'Ivoire.

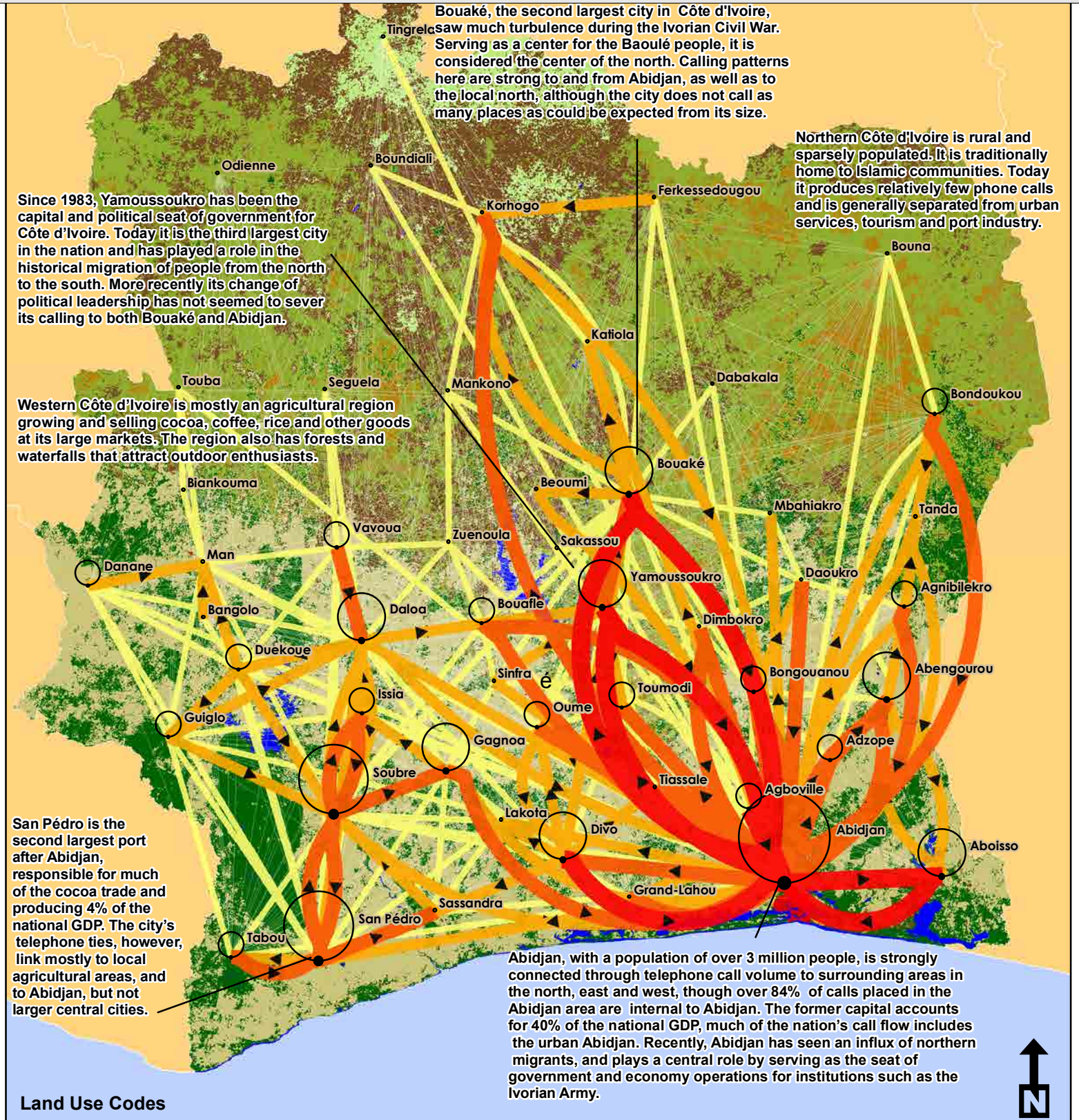
Figure 4: Regional Call Network for Abidjan.

Figure 5: Central Abidjan Call Network. Geographic Units of Analysis are Communes.

Figure 6: The relation between number of cities called and the population size of the origin.

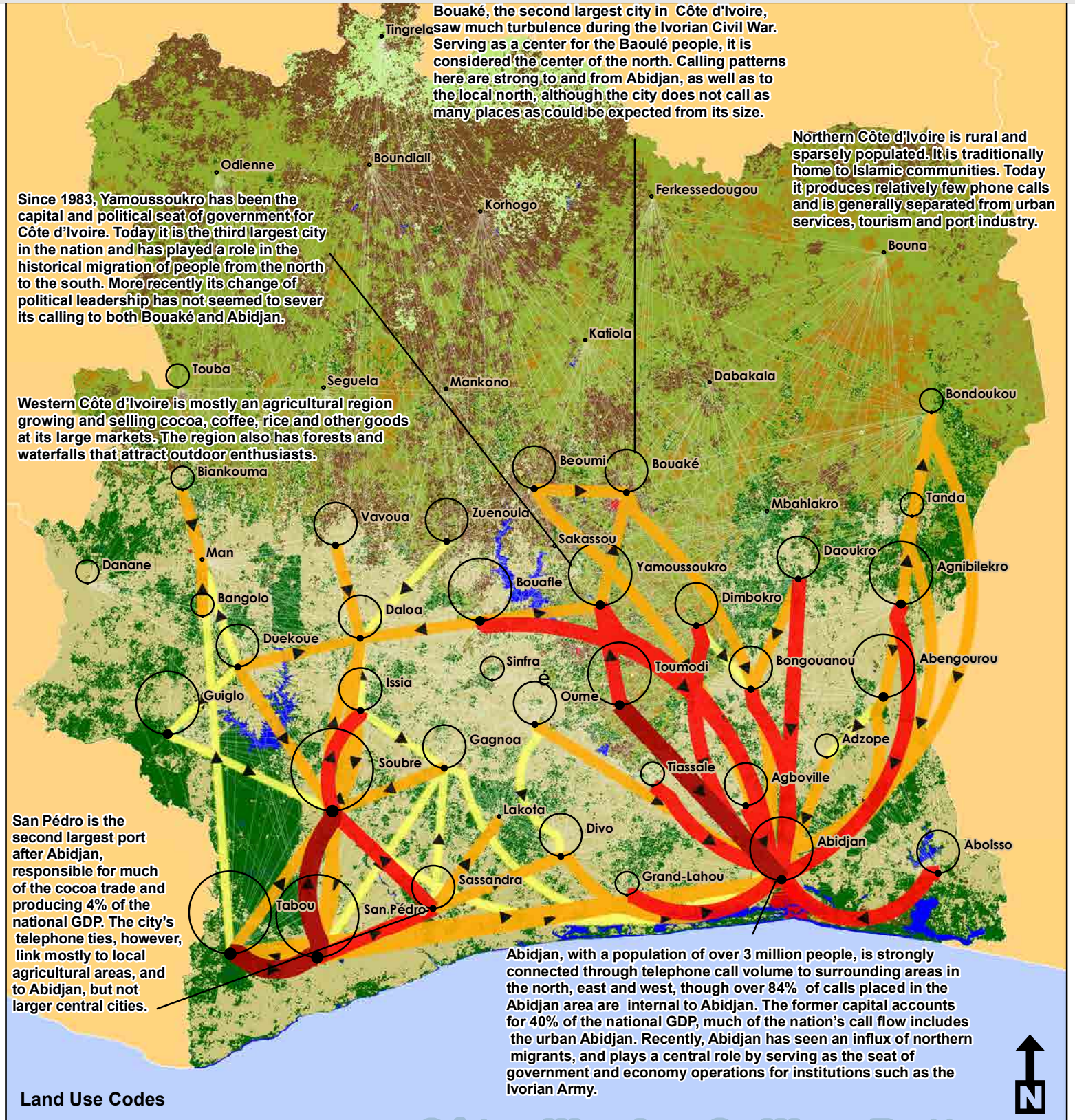
Figure 7: The relation between the diversity (Shannon entropy) of calls placed and population size of cities.

Figure 8: Temporal patterns of calls and call duractions.



Côte d'Ivoire Calling Patterns

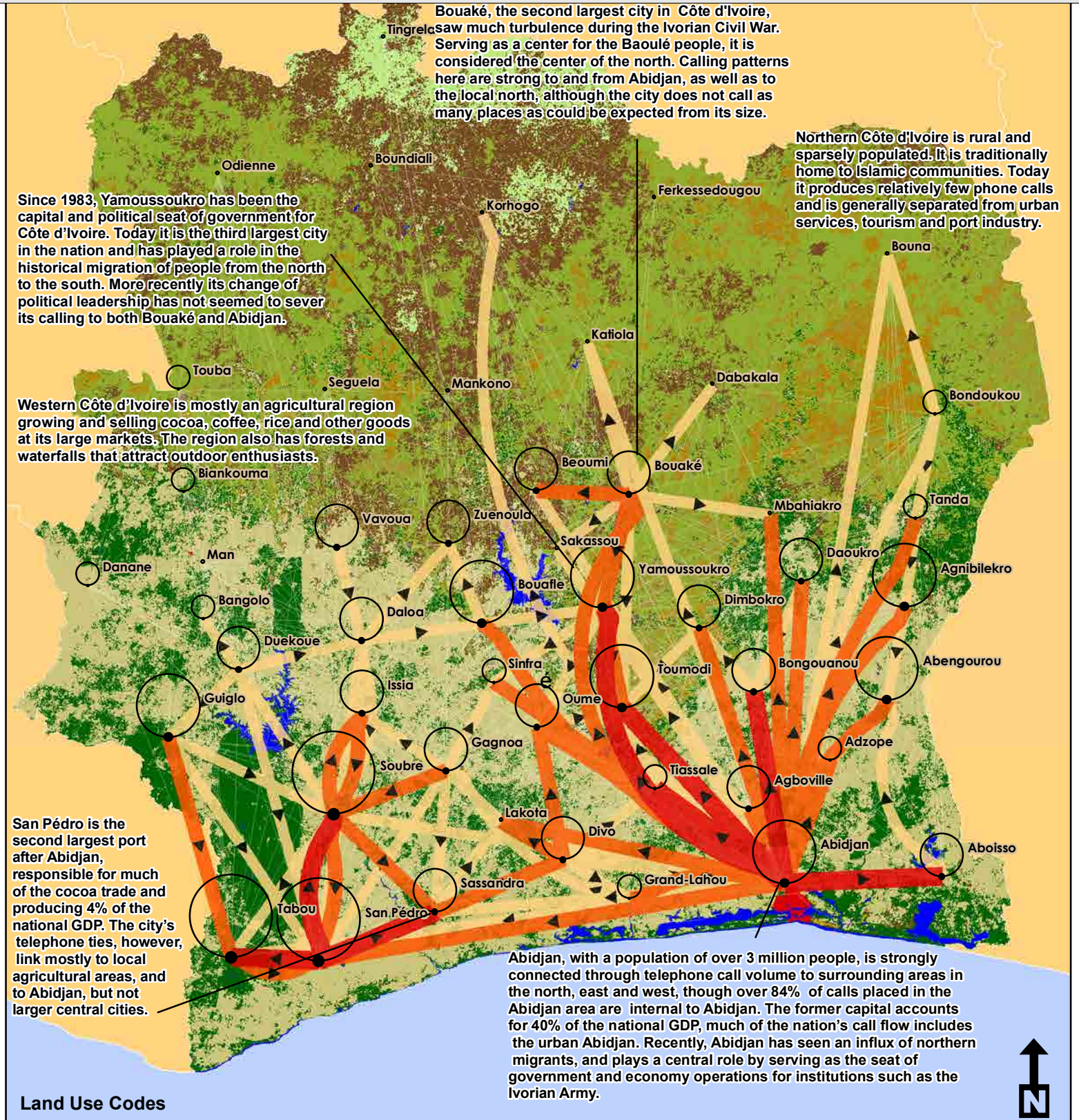
Calls are summed over a five month period, Dec. 2011 - Apr. 2012. Telecom data provided by Orange. Population data from Afripop.org. Hydrology data accessed online from http://psugeo.org/Africa/Tools.htm. Land Use Data from Food and Agriculture Organizations (FAO) of the United Nations http://www.fao.org/geonetwork.



Côte d'Ivoire Calling Patterns

Calls Placed Per Capita

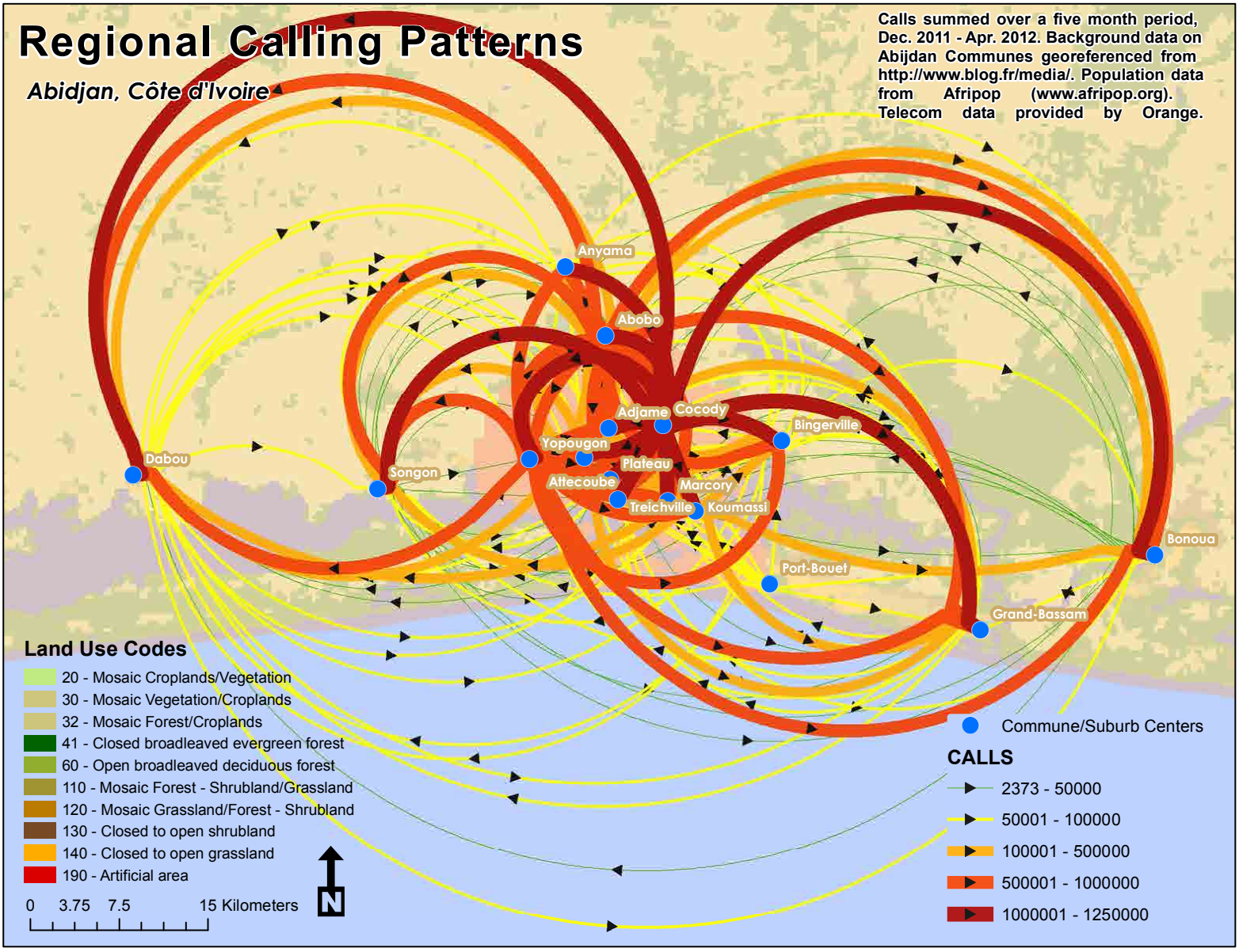
Calls are summed over a five month period, Dec. 2011 - Apr. 2012. Telecom data provided by Orange. Population data from Afripop.org. Hydrology data accessed online from http://psugeo.org/Africa/Tools.htm. Land Use Data from Food and Agriculture Organizations (FAO) of the United Nations http://www.fao.org/geonetwork.

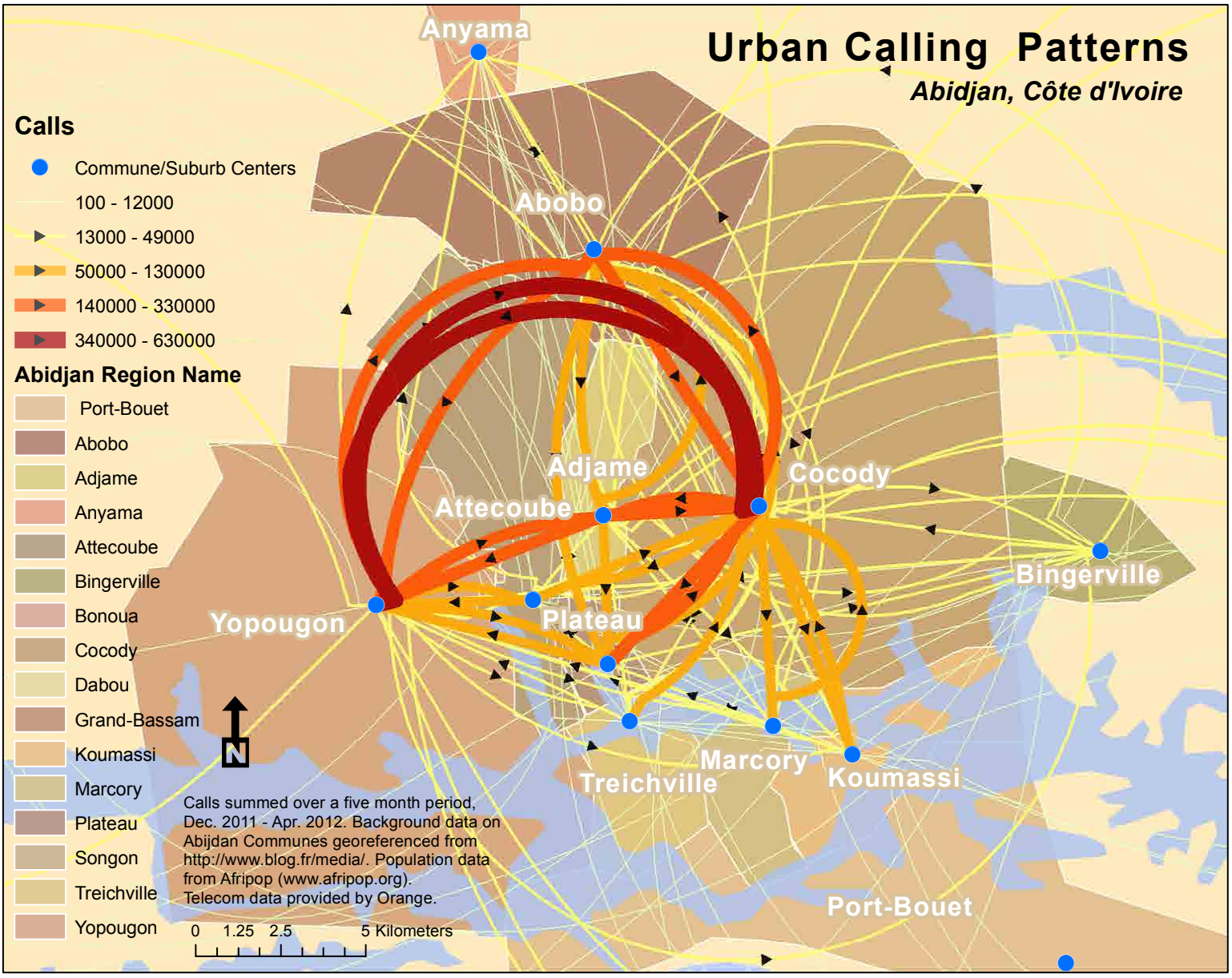


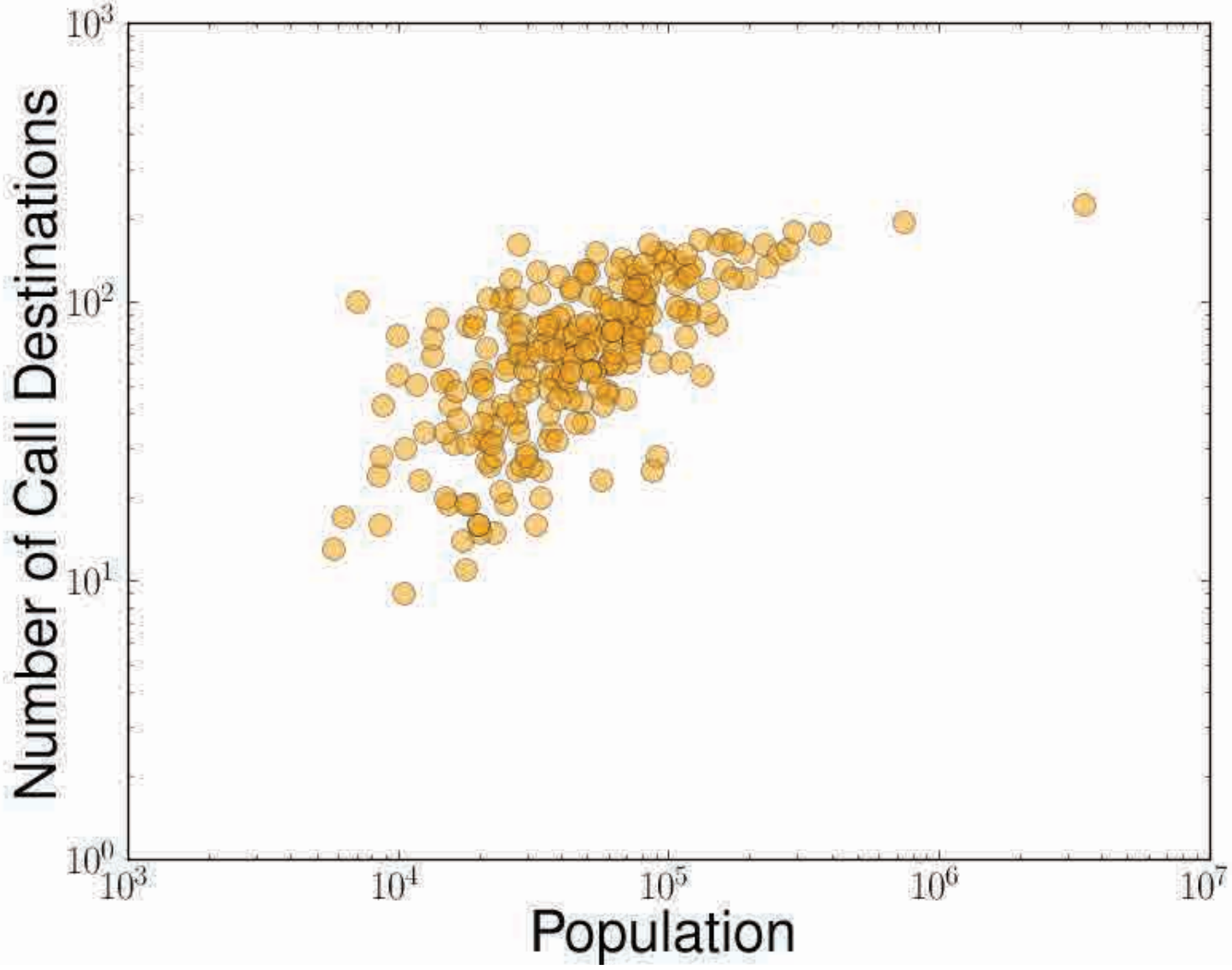
Côte d'Ivoire Calling Patterns

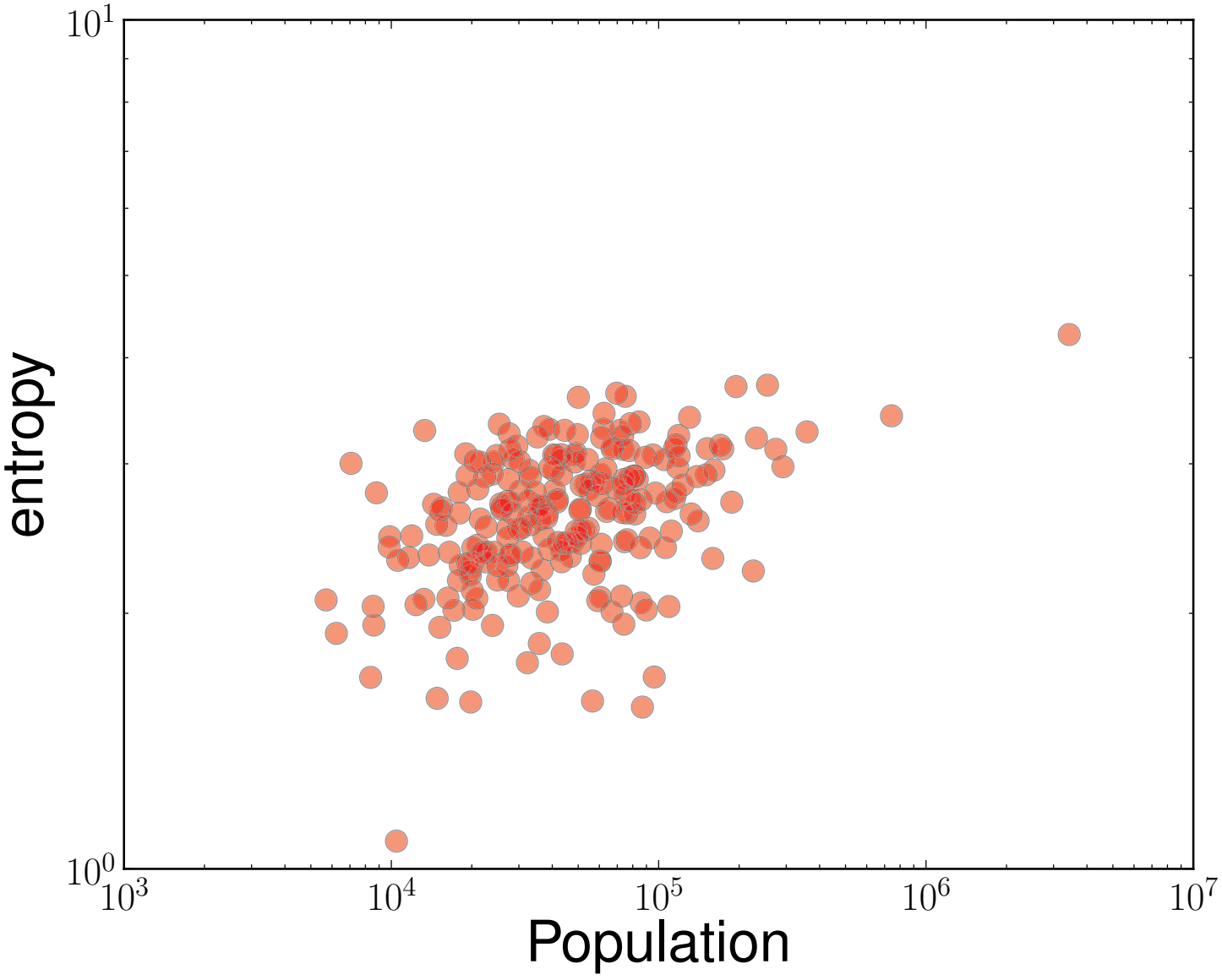
Calls Received Per Capita

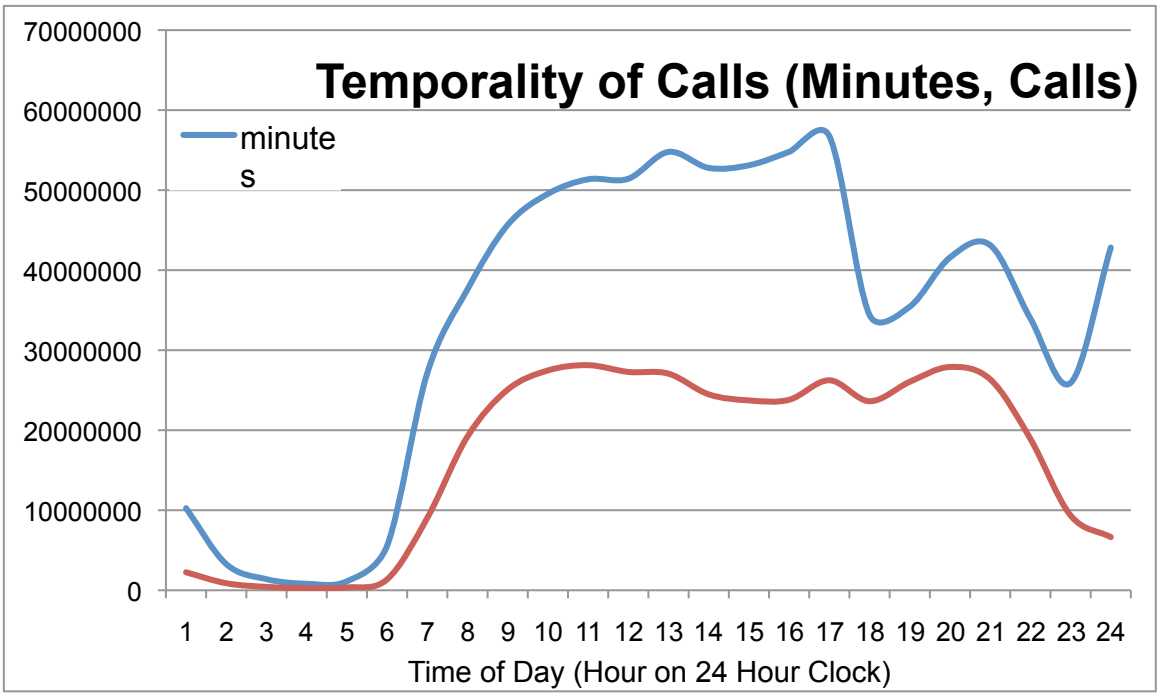
Calls are summed over a five month period, Dec. 2011 - Apr. 2012. Telecom data provided by Orange. Population data from Afripop.org. Hydrology data accessed online from http://psugeo.org/Africa/Tools.htm. Land Use Data from Food and Agriculture Organizations (FAO) of the United Nations http://www.fao.org/geonetwork.


Regional Calling Patterns
Abidjan, Côte d'Ivoire
Calls summed over a five month period, Dec. 2011 - Apr. 2012. Background data on Abijdan Communes georeferenced from http://www.blog.fr/media/. Population data from Afripop (www.afripop.org). Telecom data provided by Orange.
Anyama
Abobo
Adjame
Cocody
Bingerville
Yopougon
Plateau
Attecoube
Marcory
Treichville
Koumassi
Dabou
Songon
Port-Bouet
Grand-Bassam
Bonoua
Land Use Codes
20 - Mosaic Croplands/Vegetation
30 - Mosaic Vegetation/Croplands
32 - Mosaic Forest/Croplands
41 - Closed broadleaved evergreen forest
60 - Open broadleaved deciduous forest
110 - Mosaic Forest - Shrubland/Grassland
120 - Mosaic Grassland/Forest - Shrubland
130 - Closed to open shrubland
140 - Closed to open grassland
190 - Artificial area
0 3.75 7.5 15 Kilometers
N
Commune/Suburb Centers
CALLS
2373 - 50000
50001 - 100000
100001 - 500000
500001 - 1000000
1000001 - 1250000


Urban Calling Patterns
Abidjan, Côte d'Ivoire
Calls
Commune/Suburb Centers
100 - 12000
13000 - 49000
50000 - 130000
140000 - 330000
340000 - 630000
Abidjan Region Name
Port-Bouet
Abobo
Adjame
Anyama
Attecoube
Bingerville
Bonoua
Cocody
Dabou
Grand-Bassam
Koumassi
Marcory
Plateau
Songon
Treichville
Yopougon
Anyama
Abobo
Adjame
Attecoube
Cocody
Bingerville
Yopougon
Plateau
Treichville
Marcory
Koumassi
Port-Bouet
N
Calls summed over a five month period, Dec. 2011 - Apr. 2012. Background data on Abijdan Communes georeferenced from http://www.blog.fr/media/. Population data from Afripop (www.afripop.org). Telecom data provided by Orange.
0 1.25 2.5 5 Kilometers


Number of Call Destinations
Population
10^0
10^1
10^2
10^3
10^3
10^4
10^5
10^6
10^7


10^1
10^0
entropy
10^3
10^4
10^5
10^6
10^7
Population


Temporality of Calls (Minutes, Calls)
minutes
70000000
60000000
50000000
40000000
30000000
20000000
10000000
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
Time of Day (Hour on 24 Hour Clock)

Can Fires, Night Lights, and Mobile Phones reveal behavioral fingerprints useful for Development?

David Pastor-Escuredo[1,2], Thierry Savy[3] and Miguel A. Luengo-Oroz[1,2]

[1] Biomedical Image Technologies, DIE- ETSIT, Universidad Politécnica de Madrid, CEI Moncloa UPM-UCM, Madrid, Spain.

[2] Biomedical Research Networking Center in Bioengineering, Biomaterials and Nanomedicine, CIBER-BBN, Spain

[3] Complex Systems Institute of Paris Ile-de-France, ISC-PIF, Paris, France

Submitted to: D4D Challenge, NetMob 2013

Version: 15 March 2013

ABSTRACT

Fires, lights at night and mobile phone activity have been separately used as proxy indicators of human activity with high potential for measuring human development. In this preliminary report, we develop some tools and methodologies to identify and visualize relations among remote sensing datasets containing fires and night lights information with mobile phone activity in Cote D'Ivoire from December 2011 to April 2012.

1. Introduction

Fires, luminosity from lights at night and mobile phone activity are three ubiquous signals which can serve as proxy indicators of human activity. Recent literature shows that the ability to measure these signals can be used to understand patterns reflecting economic development and modulating factors as violent conflicts.

Fire data

We can identify and quantify fires at a global scale using remote sensing (eg, data from MODIS imaging system). Fire detection data is used for forest and agricultural monitoring, climate change and air quality modeling (Davies 2009). Fire detection might also be used as an input into early warning systems to flag potential human rights violations or humanitarian emergencies. For instance, it has been used to identify burning campaigns in human settlements in Darfur (Sudan) during periods of ethnic violence (Bromley 2010). In another study in Kenya, the United Nations used satellite fire imagery to locate areas where violence had potentially occurred (Anderson 2008).

Luminosity data

Another type of satellite data that can provide useful information of conflict zones is imagery of luminosity from lights at night visible from space (eg. data from DMSP-OLS sensors). Changes in the lights at night signature in cities from Russia and Georgia between 1992 and 2009 were measured to detect the effects of war in the Caucasus region, with the potential application of corroborating reports of unknown quality that emanate from war zones (Witmer 2011)]. Luminosity measurements have also been shown as a way to improve the quality of socioeconomic indicators in developing countries whose standard statistic data sources may be weak, for example countries with no recent population or economic censuses. In particular, luminosity has been correlated against the gross domestic product of several countries during the period 1992-2008 (Chen 2011).

Mobile Phone Data

Call detail records from mobile phone activity have recently been proposed as a potential and promising source of information to understand human behavior that might lead to proxy indicators for populations' well-being. For some information, data extracted from mobile phone activity may provide a less expensive alternative to field surveys and may be useful to helping policy makers monitor existing programs in remote locations. More precisely, recent breakthrough research has explored how calling patterns such as call reciprocity and call diversity can be used to detect the socioeconomic status of populations. This kind of research has been developed using data from UK (Eagle 2010), Rwanda (Blumenstock 2010) and Latin America (Frias-Martinez 2012). Mobile phone data has also been used to track population movements following natural disasters or economic shocks. After Haiti's earthquake, a study in areas of high mobile use showed that mobile data can rapidly provide estimates of population movements during disasters (Bengtsson

2011, Lu 2012). Another recent study in Kenya showed correlations between mobile phone data used to measure populations movements on weekly, monthly, and annual time scales with data on change in residence from the national census conducted during the same time period (Wesolowski 2013). These methods have also been used to understand migration in urban settlements (slums) in Kenya o infer information on informal employment (Wesolowski 2010). In another example, spatially explicit mobile phone data and malaria prevalence information from Kenya have been used to identify the dynamics of human carriers that drive parasite importation between regions (Wesolowski 2012).

In this report, we present our preliminary research devoted to understand how to merge, visualize and analyze information from mobile phone call detail records with remote sensing - fire and light - data in order to discover potential applications for development. First, assuming the hypothesis that mobile phone activity can be used to understand human behaviours, we will explore some relationships between mobile phone activity, lights at night, and locations with significant fires. Then, we present our on-going efforts to develop a visualization software able to integrate data contained in the call detail records with mapping and remote sensing information (in this case, fire and light). The initial purpose of these tools is to better understand what types of questions related to development is possible to ask with these types of data, and which are the right tools that can be used to answer them. Therefore, we close the report with some open questions and ideas that we hope can be used to explore new research ventures.

2. Data

2.1 D4D Data

The mobile phone datasets are based on anonymized Call Detail Records (CDR) of phone calls and SMS exchanges between a subsample of Orange's customers in Ivory Coast. The datasets were made available under the Data for Development Challenge and their description can be found here (Blondel 2012).The data covers a total of 3600 hours between December 1, 2011 and April 28, 2012. Due to technical reasons data is missing for total period of about 100 hours. In particular, in this research, we have used two types of data:

a- Antenna-to-antenna traffic. For 1231 antennas with a precise geographic location, the number of calls as well as the duration of calls between any pair of antennas has been aggregated hour by hour. This data is available for the entire observation period and represents a subset of the communications between Orange customers.

b- Individual Trajectories: High Spatial Resolution Data. Individual movement trajectories can be approximated from the geographic location of the cell phone antennas during calls. This dataset contains high resolution trajectories of 50000 randomly sampled individuals over two-week periods. To protect privacy new random identifiers (individuals) are chosen in every time period. Time stamps are rounded to the minute.

2.2 Fire data

Using the NASA FIRMS resource that provides fire detection at a global scale (earthdata.nasa.gov/firms), we have filtered all the fires occurring in a geographical bounding box that covers Cote d'Ivore. In total, we have found 59469 fires between December 1, 2011 and April 28, 2012 in the selected region (see Fig1.a). The fires are detected using data from the MODIS instrument, on board NASA's Aqua and Terra satellites (Davies 2009). The satellites take images as it passes over the earth, acquiring data continuously and providing global coverage every 1-2 days. Fire detection is performed using an algorithm that exploits the strong emission of mid-infrared radiation from fires. However there is no information about the thermal anomaly that is detected - eg. it can be an agricultural fire, an urban fire or a flare from gas. A MODIS active fire detection represents the center of a 1km (approx.) pixel flagged as containing one or more actively burning hotspots/fires. It is not possible to determine the exact size of a fire represented by 1 pixel, but studies have shown that in good acquisition conditions a fire with a size of $1000m^2$ can be reasonably detected (Giglio 2003).

2.3 Lights at night data

We have downloaded satellite imagery from the Visible Infrared Imager Radiometer Suite (VIIRS) sensor on board the NASA-NOAA Suomi NPP satellite (npp.gsfc.nasa.gov/). The data was acquired in April and October 2012 and is a global composite of cloud-free images utilizing the day-night band of VIIRS. For this research, we have used a subsampled image covering Cote d'Ivoire with a resolution of 3km per pixel (see Fig1.b).

3. Mobile phone data analysis in fire spots

3.1 Characterizing fires locations with night lights information and CDRs

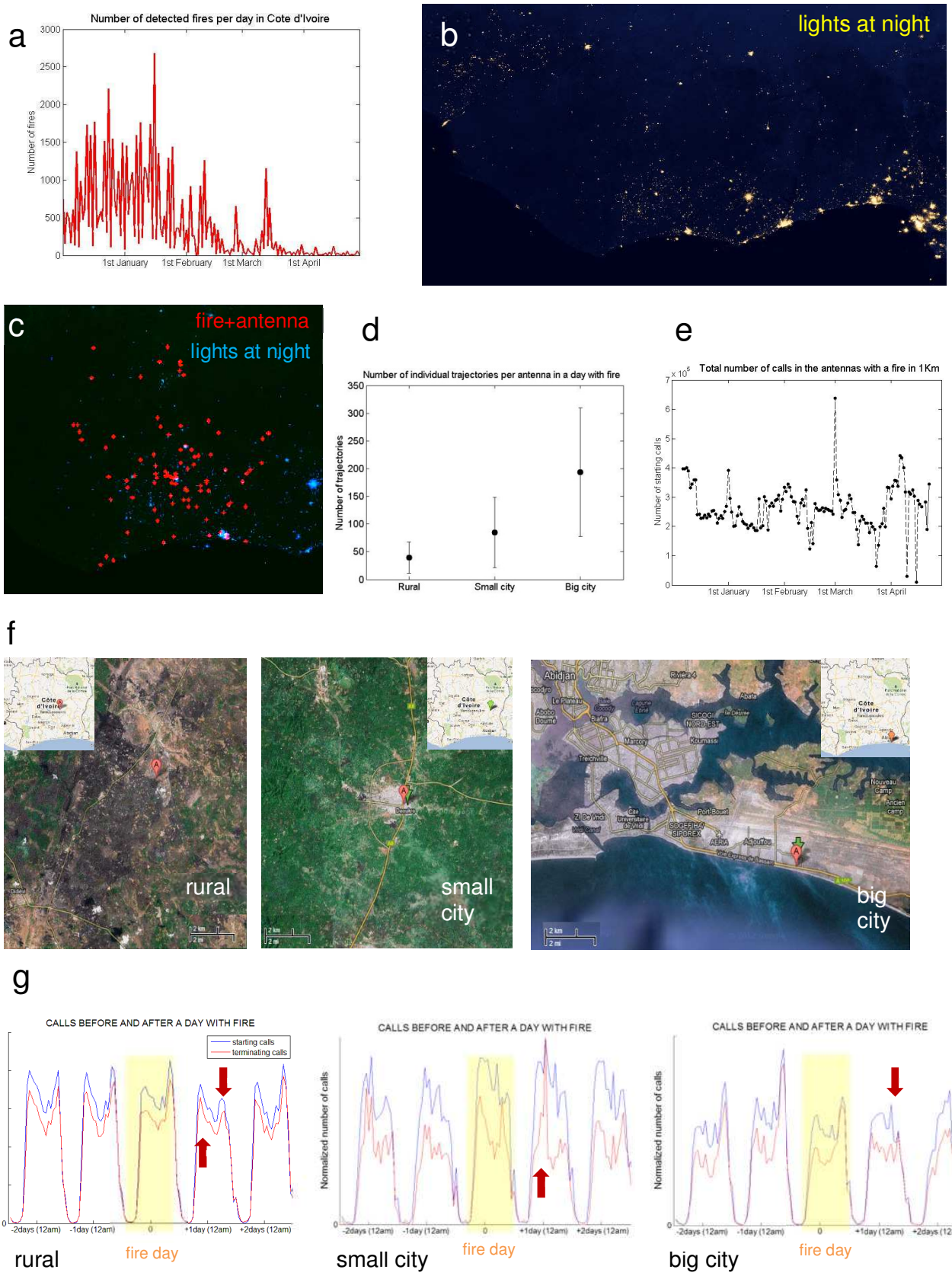
We crossed the geographic location information from the available antennas and the fires. During the studied time interval, we identified 95 antennas at less than 1km distance of a fire, 81 antennas had a unique fire and 14 antennas were affected by two fires - making a total of 109 fires (see Fig1.c).

Based on the hypothesis that luminosity at night is correlated with population and economic activity, we set up a classification system devoted to automatically identifying the urban and rural areas using the lights at night information. With this system we classified the nature of the 95 places with both a mobile phone antenna and a fire during the studied period. For each one of the antenna positions, we integrated the intensity signal of the light at night imagery in an area with a 7.5 km radius. These values were clustered into 3 classes using a k-means algorithm. The resulting classes were composed by 8, 15 and 72 locations. A visual inspection of the elements of each class, both in the lights at night imagery and in Google Maps, revealed that the cluster with 8 locations corresponds to big cities (eg. Abidjan), the cluster with 15 elements signals small cities and the larger class corresponds to remote rural areas and some roads (see Fig1.f). In order to asses this type of classification – from urban to rural, we selected all the individual trajectories (from dataset *b*) that passed by a fire the same day of the fire detection. We found that for 18% of the antennas, no trajectory was logged that day. We do not know if it is because of the impact of the fire –ie. the network went down -, or if it is due to the low population sampling in the mobile phone data. From the remaining antennas with at least one person logging the day of the fire, we measured the number of individual trajectories for each cluster. The result shows a direct relationship between the number of trajectories and the urban-rural classification obtained from the lights data (see Fig.1d). This suggests that both mobile traffic and lights at night are proxies for the urban-rural classification of the fire locations.

In a second experiment, we explored the potential impact of fires at the antenna level (starting and terminating calls) at a short time scale of a few days. We added the number of calls per hour of all the antennas in each of the clusters after a temporal alignment with the fire date (t=0 is 12h of the day of the fire). The mobile phone activity is given by hour, while for the fire temporal resolution it is daily. When adding the activity of all the antennas of a category, in order to avoid over representation of a single antenna with many calls, we normalized the number of calls per hour of each antenna by the maximum number of calls in the same antenna during five days centered in the fire date. That is, we obtained the mean number of calls per hour from two days before a fire until two days after that fire (see Fig1.g). Interestingly, we observed that the day after the fire, there are more calls in the morning and less in the evening. In rural areas, the typical pattern of daily calls has a couple of peaks, one in the morning and one in the evening – being the latter of slightly higher intensity. We noticed that after a day with a fire event, the intensity of the two peaks was inverted, that is, more calls in the morning than in the evening. This might be because people are seeking information on what happened. Two days after the fire, the typical pattern is recovered, suggesting a rapid resumption of normal activity (at least in most of the cases). In the small city category, we observed a large increase of terminating calls in the morning of the day after the fire, which again might be associated to the requests for information attention-worthy events. In the big cites, our data indicate that fires result in a reduced number of calls.

We also tried to explore potential longer term effects that might be caused by devastating fires. However, we obtained very unstable results when examining patterns at a weekly-monthly time scale. We suspect that it is due to non-uniform calling patterns in the analyzed antennas, which brings an additional level of complexity. When examining the data at daily scale, we found three patterns that make any kind of temporal alignment around fires a difficult task (see Fig1.e). First there is an inherent pattern in the mobile phone activity of Orange customers in Cote d'Ivoire: there are many calls at the beginning of the month, with a decreasing drift until the end of the month. We think that this is pre-paid phone typical pattern – people have money at the beginning of the month so they top-up their sim cards. It would be interesting to know the proportion of pre-paid and contract subscribers for a deeper study of this effect. Second, during the period analyzed -from December to April 2012- we observed that for antennas in several regions there are fewer calls around Christmas. Third, there are some punctual spikes of activity, as the 1st of January or the 1st of March. All these factors point to an open question on how to aggregate the information from different antennas occurring at different times points -in our case when fire is detected – at a temporal scale in the order of weeks or months.

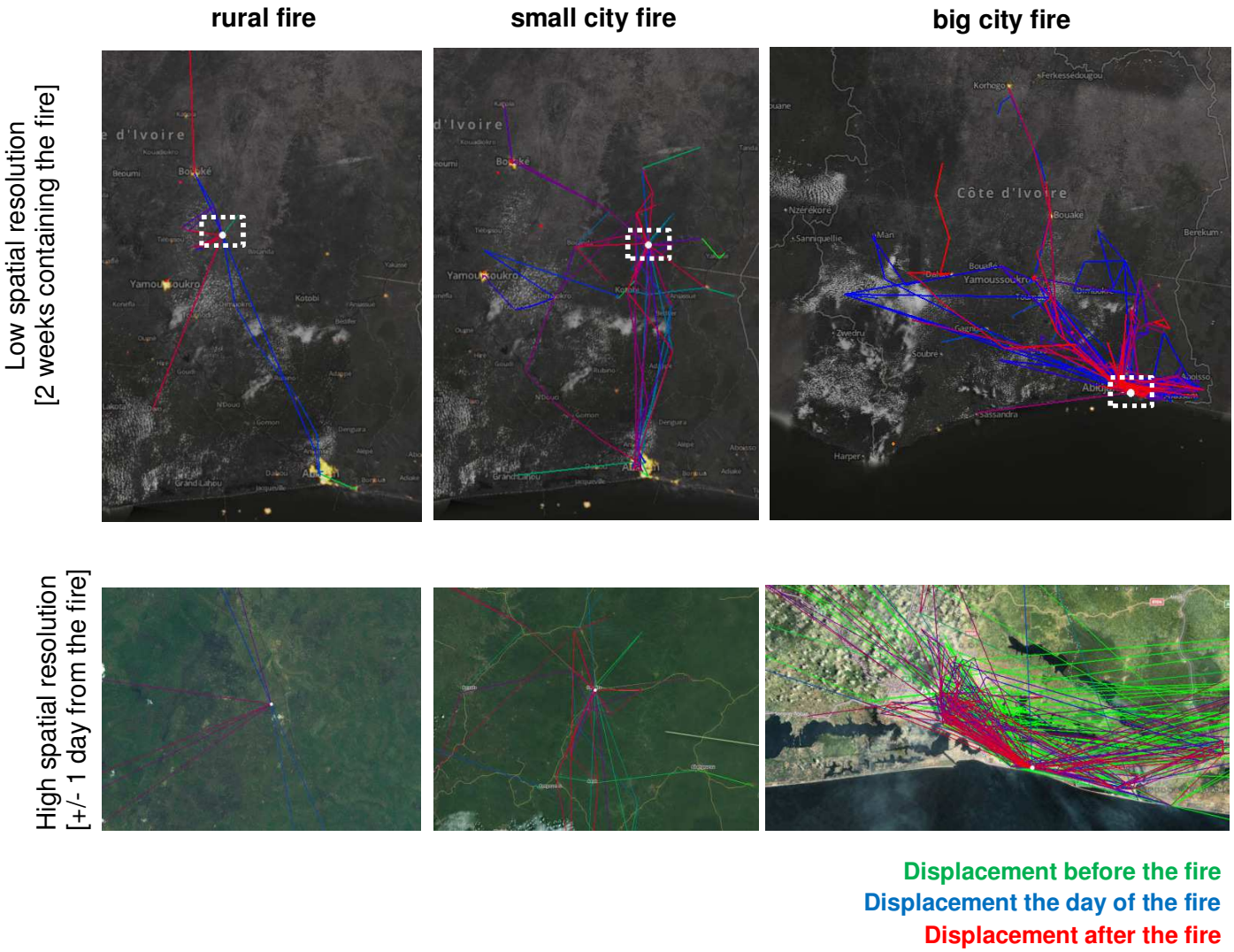
Figure 1


a
Number of detected fires per day in Cote d'Ivoire
Number of fires
3000
2500
2000
1500
1000
500
0
1st January
1st February
1st March
1st April
b
lights at night
c
fire+antenna
lights at night
d
Number of individual trajectories per antenna in a day with fire
Number of trajectories
350
300
250
200
150
100
50
0
Rural
Small city
Big city
e
Total number of calls in the antennas with a fire in 1Km
x 10^6
Number of starting calls
7
6
5
4
3
2
1
0
1st January
1st February
1st March
1st April
f
rural
small city
big city
Abidjan
g
CALLS BEFORE AND AFTER A DAY WITH FIRE
starting calls
terminating calls
Normalized number of calls
-2days (12am)
-1day (12am)
0
+1day (12am)
+2days (12am)
rural
fire day
small city
fire day
big city
fire day

3.2 Visualizing Dynamic Call Detail Records

The previous analyses reinforced the idea that handling complex geo-localized information as mobile phone activity and fire data requires the development of dedicated visual analytics tools capable of showing the data in a comprehensive manner. There exist available ad-hoc visualization platforms that allow seeing general mobile traffic statistics at high scale (eg. www.geofast.net) and some non-interactive demonstrative videos (eg. www.villevivante.ch/). However, there is not a standard interactive visual analytics platform that allows to visually understanding individual trajectories of this nature with high spatio-temporal resolution. Because of this, we have developed the basic architecture of an interactive visualization platform focused on providing a multiscale and dynamic view of the individual trajectories and their spatial relations together with heterogeneous databases such as the detected fire locations. This platform - so-called MOBILOMICS (www.die.upm.es/im/archives/mobilomics/)- has been developed using *Processing* language (processing.org/) and the *Unfolding Maps* library (unfoldingmaps.org/) and allows us to load several maps from OpenStreetMap (licensed under the Open Data Commons Open Database License), mobile phone trackings and specific spatio-temporal events. With this platform, we can browse and explore the data in space and time at different scales, searching and selecting specific subsets of individual trajectories that share similar characteristics. As a visualization example, in Fig2, we see a capture that shows all the trajectories of people that logged the same day that a fire occurred in three different antennas corresponding to each of the 3 categories: rural – small city – big city.

Figure 2



4. Discussion and open questions

In this report we have presented the idea of merging fire, night lights and mobile phone information to explore behavioral fingerprints that might be useful for development. We have introduced some analytical methodologies, shown preliminary data analysis and developed an interactive visualization platform that allows integrating and exploring individual trajectories and fire data with high spatiotemporal resolution. While this research is in progress, and the preliminary discoveries neither are concluding nor statistically significant, we have better understood the kind of questions related to development that this data might answer, and we have found some useful clues encouraging further investigation. The relation between phone calls and light at night data suggests that the previous works relating light data with macroeconomic indicators as GDP might be improved by adding mobile data analytics to the mix, so that a finer scale both in time and scale becomes feasible to not only measure economic activity, but also to characterize urban and rural areas. Regarding the potential effect of fires in mobile phone activity – and potential proxy indicators of human behavior - , we have identified some changes in the patterns of calls at hourly scale just after a fire occurred. We hypothesize that this information – and the time needed to recover a normal calling pattern - might be used to track the recovery of a zone after a fire emergency. Further research should be done to understand the particular cases depending on the nature of the fire (eg. urban fire, agricultural related fire or conflict fire) that might be approached by crossing this output with emergency records or field surveys, and with potential use to optimize emergency resources and protocols. When analyzing the effect of fires at longer time scales – i.e weeks, we have identified some data biases that imply normalization challenges that should be investigated. However, these long term effects could be of high interest. For instance, in an urban environment, we might imagine some people forced to change their home or job location because of a fire. In a rural environment, we might imagine this impact translated into a decreased agricultural activity. At a broader scale we could infer possible economic deceleration due to devastating fires thanks to phone activity records -other official resources should be used to verify and extract robust characterization and classification of these social dynamics. For further research, it would be extremely useful to dispose of a longer dataset of mobile phone activity, as we suspect that there will be many antennas close to the almost 60000 fires identified during the 5 months studied. The effects of the subsampling of the mobile dataset remain unknown and it would be interesting to asses which subsample is representative and optimal for this kind of research. In future developments, we expect to include the data analysis methods and different ground truth data from the off-line world into the visualization interface, with special emphasis in the dynamic properties of the individual trajectories data in order to better understand the particular flows in and out the fire locations depending of the nature and consequences of the fire.

Acknowledgments

The authors would like to thank Eva Kaplan and Constanza Blanco for their fruitful discussions and feedback on the study. This research was partially funded by the Picata program from the Moncloa Campus of International Excellence, Universidad Politécnica de Madrid and Universidad Complutense de Madrid, Spain.

Bibliography

Davies, D. K., Ilavajhala, S., Wong, M. M., & Justice, C. O. (2009). Fire information for resource management system: archiving and distributing MODIS active fire data. Geoscience and Remote Sensing, IEEE Transactions on, 47(1), 72-79.

Anderson, D., & Lochery, E. (2008). Violence and Exodus in Kenya's Rift Valley, 2008: Predictable and Preventable?. Journal of Eastern African Studies, 2(2), 328-343.

Bromley, L. (2010). Relating violence to MODIS fire detections in Darfur, Sudan. International Journal of Remote Sensing, 31(9), 2277-2292.

Witmer, F. D., & O'Loughlin, J. (2011). Detecting the Effects of Wars in the Caucasus Regions of Russia and Georgia Using Radiometrically Normalized DMSP-OLS Nighttime Lights Imagery. GIScience & Remote Sensing, 48(4), 478-500.

Chen, X., & Nordhaus, W. D. (2011). Using luminosity data as a proxy for economic statistics. Proceedings of the National Academy of Sciences, 108(21), 8589-8594.

Eagle, N., Macy, M., & Claxton, R. (2010). Network diversity and economic development. Science, 328(5981), 1029-1031.

Blumenstock, J., & Eagle, N. (2010). Mobile divides: gender, socioeconomic status, and mobile phone use in Rwanda. In Proceedings of the 4th ACM/IEEE International Conference on Information and Communication Technologies and Development (p. 6). ACM.

Frias-Martinez, V., & Virseda, J. (2012). On the relationship between socio-economic factors and cell phone usage. In Proceedings of the Fifth International Conference on Information and Communication Technologies and Development (pp. 76-84). ACM.

Bengtsson, L., Lu, X., Thorson, A., Garfield, R., & von Schreeb, J. (2011). Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: a post-earthquake geospatial study in Haiti. PLoS medicine, 8(8), e1001083.

Lu, X., Bengtsson, L., & Holme, P. (2012). Predictability of population displacement after the 2010 Haiti earthquake. Proceedings of the National Academy of Sciences, 109(29), 11576-11581.

Wesolowski, A., Buckee, C. O., Pindolia, D. K., Eagle, N., Smith, D. L., Garcia, A. J., & Tatem, A. J. (2013). The Use of Census Migration Data to Approximate Human Movement Patterns across Temporal Scales. PloS one, 8(1), e52971.

Wesolowski, A., Eagle, N., Tatem, A. J., Smith, D. L., Noor, A. M., Snow, R. W., & Buckee, C. O. (2012). Quantifying the impact of human mobility on malaria. Science, 338(6104), 267-270.

Wesolowski, A., & Eagle, N. (2010). Parameterizing the dynamics of slums. In AAAI Symposium on Aritificial Intelligence and Development.

Blondel, V. D., Esch, M., Chan, C., Clerot, F., Deville, P., Huens, E., ... & Ziemlicki, C. (2012). Data for Development: the D4D Challenge on Mobile Phone Data. arXiv preprint arXiv:1210.0137.

Giglio, L., Descloitres, J., Justice, C. O., & Kaufman, Y. J. (2003). An enhanced contextual fire detection algorithm for MODIS. Remote sensing of environment, 87(2), 273-282.

Exploration and Analysis of Massive Mobile Phone Data: A Layered Visual Analytics approach

Stef van den Elzen, Jorik Blaas, Danny Holten, Jan-Kees Buenen, Jarke J. van Wijk,
Robert Spousta, Anna Miao, Simone Sala, Steve Chan

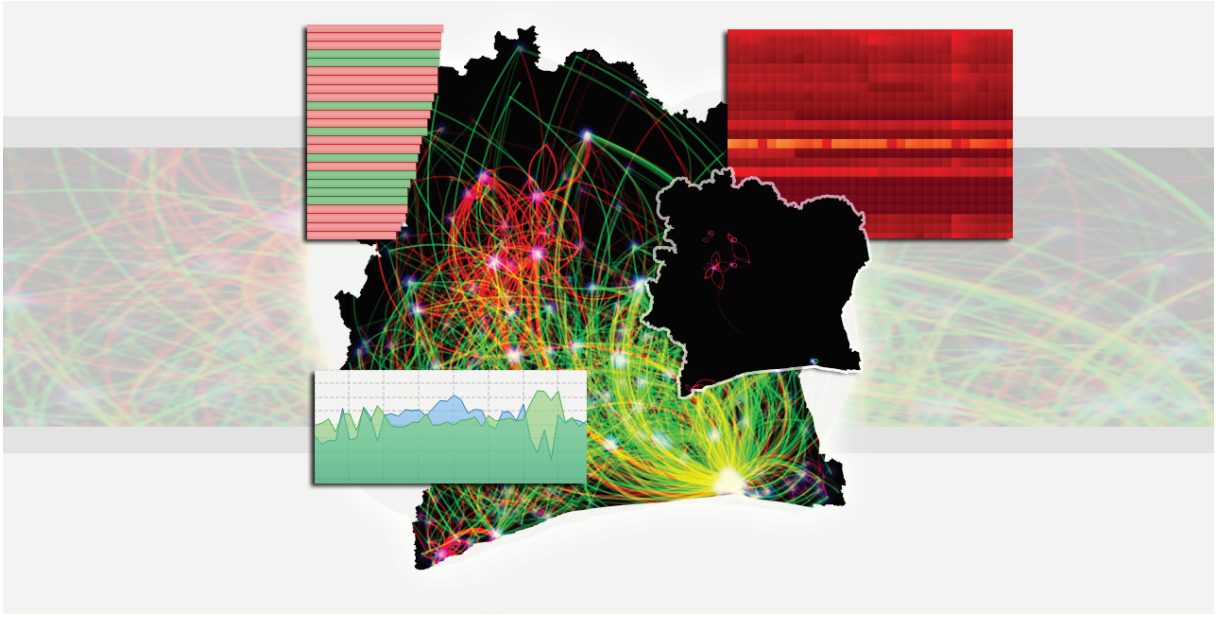


Fig. 1. Different coherent components of the visual analytics solution for the exploration and analysis of Massive Mobile Phone Data in the context of the Orange Data for Development D4D-challenge.

Abstract— We present a system for the exploration and analysis of massive mobile phone data that enables users to gain insight. First we identify user tasks and develop a system following a visual analytics approach by tightly integrating visualization, interaction and algorithmic support. The system is then evaluated by exploring a massive mobile phone data set containing 2.5 billion calls and SMS exchange between around 5 million users located in Ivory Coast over a period of 5 months. From the use cases a number of findings are gathered, such as localized increase and decrease of calls due to major events.

Index Terms—Mobile Phone Data, Visual Analytics

1 Introduction

Four datasets on mobile phone communication were released in the context of the Orange *Data for Development* (D4D) challenge. The datasets are based on 2.5 billion anonymized Call Detail Records (CDR) of phone calls and SMS exchanges between five million of Orange's customers in Ivory Coast between December 1, 2011 and April 28, 2012 [10]. In this paper we focus on the first dataset: tower-to-tower traffic. Data is provided to us in tab-separated-value (TSV) file format. For each hour in the timespan we are given the number of calls and duration (aggregated) between any pair of towers. Additionally, we are provided with the geographic location (latitude and longitude) of each cell tower. Initial data cleaning was performed by Orange Labs in Paris, such as removing double entries, new subscribers, and, communication to other providers. We further cleaned the data by removing entries that had missing tower identifiers.

In this paper we describe how we support the analysis and exploration of massive mobile phone data sets by identification of user tasks and according requirements for a visual analytics prototype. Next, this is implemented and applied to the provided real-world mobile phone data. We present a system for the exploration and analysis of massive mobile phone data that enables users to gain insight on different levels of abstraction both in time and space. The prototype provides a smooth user experience despite the massive amount of data. We implement a visual analytics approach adhering to the visual analytics mantra: *analyse first, show the important, zoom, filter and analyse further, details on demand* [21].

Visual Analytics is the science of analytical reasoning facilitated by interactive visual interfaces [32]. It aims at an integrated approach combining visualization, human factors and automated data analysis using methodologies from information analytics, geospatial analytics and scientific analytics to effectively support the decision making process [22].

- *Stef van den Elzen is with Eindhoven University of Technology and SynerScope BV. E-mail: s.j.v.d.elzen@tue.nl.*
- *Jorik Blaas, Danny Holten, Jan-Kees Buenen are with SynerScope BV. E-mail: {jorik.blaas, danny.holten, jan-kees.buenen}@synerscope.com*
- *Jarke J. van Wijk is with Eindhoven University of Technology. E-mail: vanwijk@win.tue.nl*
- *Robert Spousta, Anna Miao, Simone Sala, Steve Chan are with Prince of Wales Fellowship at Massachusetts Institute of Technology. E-mail: {spousta, annamiao, salas, s_chan}@mit.edu*

Correlations of call change behavior and local events are successfully identified using the prototype. We believe *call change* occurs due to *major events*. We found both an increase and decrease in the number of calls over locally concentrated communication channels strongly correlated with events.

The paper is organized as follows. First, user tasks and according requirements to the visual analytics approach are discussed in Section 2. Next, we discuss related work in Section 3. A layered visual analytics approach is presented in Section 4, where the different components of the system are discussed in detail. Section 5 portrays typical use cases utilizing our approach. Finally, conclusions and directions for future work are given in Section 6.

2 Design Principles

In this section we identify different user tasks, derive requirements and discuss design decisions following from this. We believe that change in call behavior occurs due to major events. Therefore, to detect events, we focus on call change derived from the first D4D dataset (tower-to-tower communication).

The user tasks can be categorized into higher level tasks: *exploration*, *analysis* and *presentation* of massive mobile phone data. The main goal of **exploration** is to *gain insight* and to *form hypotheses*. The main goal of **analysis** is to *confirm or reject hypotheses*. While performing analysis, visualization is not only supportive but can also raise new questions therefore, users typically switch often between exploration and analysis during data exploration. **Presentation** is needed to *convey findings* to both expert users and a broader audience. In order to support this, familiar visualizations are needed. Table 1 provides and overview of more detailed tasks and requirements. In addition to these requirements we aim for effortless switching between the exploration, analysis and presentation process. In summary:

- data has to be shown at various levels of aggregation, both temporally and spatially;
- data has a temporal, a geospatial, and a network character; all have to be shown;
- we use multiple linked views for this;
- details have to be shown on demand;
- where possible, use automated methods to simplify analysis and reduce the amount of data;
- features like overall call behavior, call change and communities have to be clearly visible;
- where possible, use familiar mappings and metaphors for easy understanding.

In Section 4 we discuss how the system implements all requirements by detailed discussions of each of the individual components and their integration and coherence.

3 Related Work

We briefly discuss related work as means of placing our work in context, and to motivate the development of a new visual analytics prototype: no tool exists that fulfills our requirements.

Many approaches are explored using only automated methods offering no interaction and visualization e.g., [8, 15, 23, 33].

A visual analytics system, developed by Andrienko et al. [6, 7], for extracting place histories from mobile data, combining geovisualizations, geocomputations and statistical methods allows for the exploration of spatial, temporal and thematic components of the data. However, the main focus here is on social aspects and place extraction and does not allow for the exploration of call change inferred from events. Similar approaches, not exploring call change or event detection are discussed by Kwan and Lee [24] and Sagl et al. [25]. A system developed by Correa et al. [13, 31] also mainly focuses on social behavior patterns. Also, the geographical component is not taken into account and no algorithmic support is offered. Temporal communication patters in mobile call graphs focusing on structural network change is discussed by Ye et al. [37]. Höferlin et al. [16] focus on individual trajectory movement exploration extracted from mobile data. A more general visual analytics system for the exploration of spatio-temporal data not tailored towards mobile data is presented by Von Landesberger et al. [34].

Egocentric temporal exploration of CDR data is performed by Qi et al. [36]. However, the geographical content is not taken into account and cannot be visualized. Furthermore, the method focuses on egocentric exploration and does not allow for an overview of the (higher level) call patterns. Egocentric exploration of temporal call behavior focusing on regions rather than individual towers is explored by Blondel et al. [9] in their web-based Geofast tool.

4 Visual Analytics Approach

In this section the developed prototype enabling the effective analysis and exploration of massive mobile phone data is presented. The prototype application follows a visual analytics approach using multiple coordinated views that tightly integrates visualization, interaction and automatic computation methods. A combination of visualization and automated methods is used, because purely visual methods fall short due to scalability issues; the data provided is large and screen space is limited. This can partially be overcome by interaction methods such as zoom, pan and filter techniques. However, this leaves less apparent patterns in the data hidden. Also, purely automatic methods fall short due to aggregation of results and loss of context. Furthermore, automatic methods are often highly focused and designed for one specific task, not allowing for the exploration and discovery of unexpected patterns. A system effectively integrating visualization, interaction and algorithmic support leverages the benefits of the individual parts.

Initial data transformation steps were taken to be able to implement the requirements as defined in Section 2. These steps are discussed next, followed by a detailed description and discussion of design decisions of the individual coordinated views and their integration into the system (see Figure 2 for a screen-shot of the graphical user interface).

Table 1. Requirements to support users in the exploration and analysis of massive mobile phone data. We acknowledge this list is not exhaustive and can be extended further, however, we believe the system should at least support these user tasks.

	Task *User wants*	**Requirement** *the system should*
Exploration	identification of:	
	higher level communication channels;	provide an overview of communication channels both in space and time;
	changes in call behavior, this because we believe that a change in call behavior occurs due to major events; and	provide an overview of call change behavior, again, both temporal and spatial aspects should allow for exploration;
	communities (cell towers with similar behavior over time).	provide (algorithmic and visualization) support for community identification.
Analysis	perform comparison of:	
	multiple levels of abstraction for time, space, and, visualization;	enable effortless switching between different abstraction levels;
	multiple points in time; and	enable for simultaneous comparison of multiple time points;
	multiple visualizations of similar or different data dimensions.	enable for simultaneous comparison of multiple views and data dimensions.
General	interactively browse through different portions of the data; being guided by complex patterns and correlations; and do all of this in real time.	provide appropriate visualizations for each abstraction level that; emphasize clues for further navigation to; provide a real-time smooth exploration user experience.

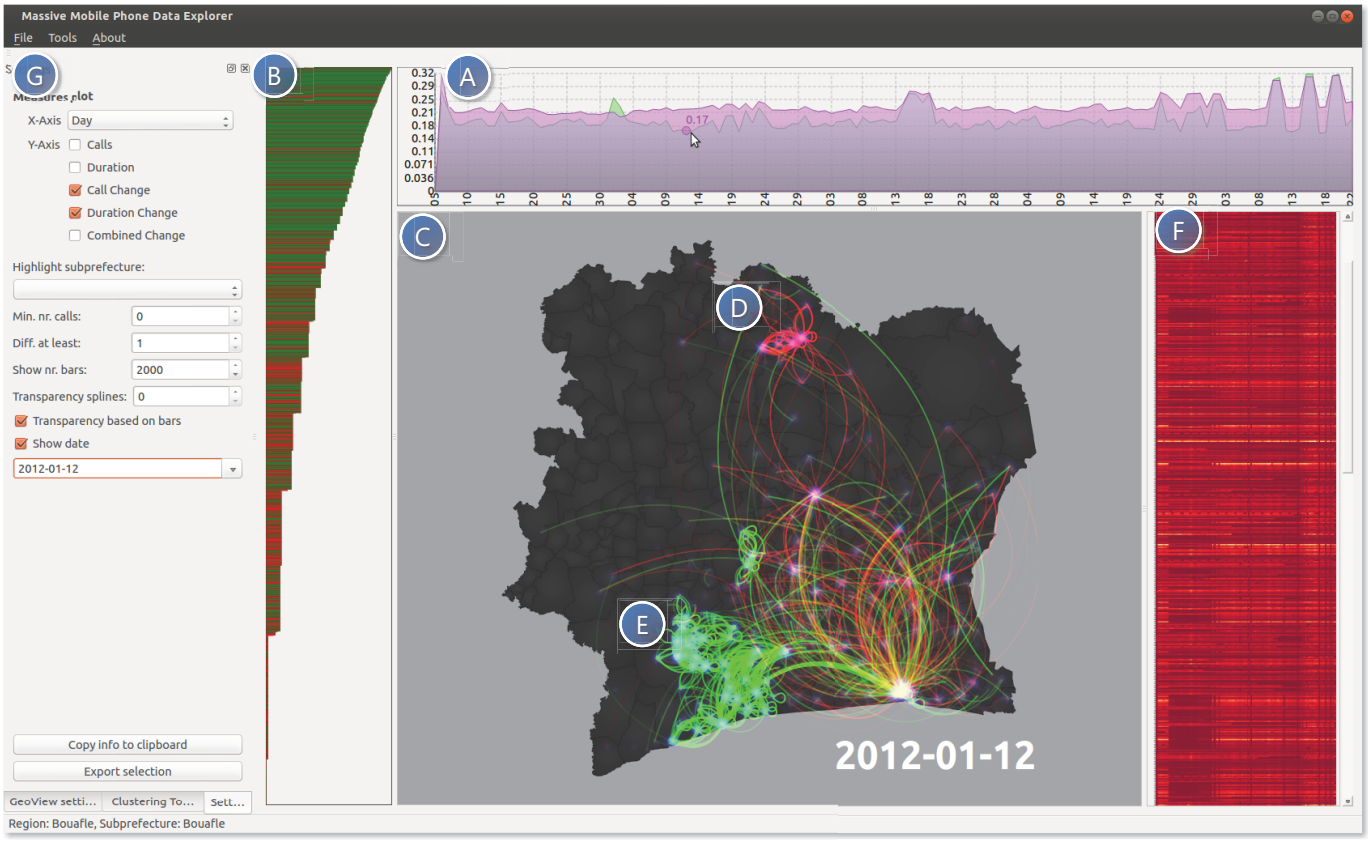


Fig. 2. Graphical user interface: providing high level overview of different measures such as the number of calls and call change over time in the measure overview (A). Individual contributions of communication channels to the measure at selected time interval are shown in the measure contribution view (B) and in the geospatial view (C) revealing localized call change behavior: large decrease (D) and large increase (E) of the number of calls. The matrix view (F) provides a high level overview of call behavior over time for the individual towers. Different settings and algorithmic support are offered by according controls (G).

4.1 Analyze first, show the important

The data provided consists of 3.9GB (zipped) and 33.5GB (unzipped) TSV files. Clearly, this does not fit into memory. In order to provide a real-time exploration experience to users different techniques are employed. Here we choose for a combination of *pre-computation*, *divide-and-conquer*, and, *load-on-demand* strategy. The tower-to-tower communications, mainly focused on in this paper, were provided in ten separate TSV files each spanning a period of two weeks. One of the requirements is to provide an overview, therefore all files need to be loaded into memory, however due to size constraints this is not possible. In order to acquire manageable data, we first processed this data using scripts taking an advantage of a line-by-line streaming approach that *divided* the large files into smaller files each containing the data for one day. Furthermore, the data lines in the smaller files are sorted descending on number of calls between any pair of towers. Finally, separate files were created for calls and duration. This process is repeated to create files on different abstraction levels such as weeks and months. Instead of loading everything into memory at once, smaller chunks can now be loaded on demand if detailed information is requested. Due to the relative small file sizes this allows for real-time exploration. In addition, data can also partially be loaded, requesting only the most important data because the files are internally sorted.

In addition to splitting and organizing each file, different measures are identified for which an overview needs to be provided. These measures were then pre-computed to be shown as a line graph in the measure overview (more on this in Section 4.2). The pre-computed measures are *number of calls, duration, call-change, duration-change*, and, *combined-change*. Each of the measures are pre-computed for the different abstraction levels (*days, weeks* and *months*). The number of calls measure and duration measure aggregates per abstraction level the total number of calls and total duration respectively. The call-change, duration change and combined change are computed based on the extended Jaccard index, rather than taking the absolute difference of the number of calls for two points in time. This has the advantage that a large change is reported when the number of calls goes from 100 to 1000 but also if the number of calls goes from 1 to 10 for a certain communication channel. Let E be the set of all cell tower pairs having communication on one or more points in time. For each pair of cell towers involved $(e \in E)$, at two points in time t_x and t_y we compute the *individual* call change ICC_e:

$$ICC_e(t_x,t_y) = \frac{M_e(t_x) \times M_e(t_y)}{M_e(t_x)^2 + M_e(t_y)^2 - M_e(t_x) \times M_e(t_y)}, \quad (1)$$

where $M_e(t_x)$ gives the value of the according measure (here number of calls) at time point t_x for cell tower pair (communication channel) e. If $M(t) = 0$, 1 is used to prevent a final value of 0. Next, all individual call changes ICC are summed and divided by the number of involved tower pairs E to provide a final call change value CC for two points in time:

$$CC(t_x,t_y) = \frac{1}{|E|} \times \sum_{e \in E} ICC_e(t_x,t_y) \quad (2)$$

Duration change and combined change (calls + duration) are computed in a similar fashion. In addition to the change values CC we also store the individual change values ICC and again sort these descending. These files are later loaded-on-demand if additional information is required on the aggregated measures e.g., provide information on the

contribution of each communication channel to the aggregated measure. By pre-computation of these measures we can provide multiple overviews that allow for a smooth real-time exploration of the data.

4.2 Measure overview

In the measure overview line graphs are shown. Users can select what to show on the x-axis and what to show on the y-axis. On the x-axis different aggregation levels of time can be set; users are able to choose one from *days*, *weeks*, and, *months*. On the y-axis (a combination of) different measures can be conveyed such as number of *calls*, *duration*, *call change*, *duration change*, and, *combined change*. Showing multiple measures in the overview enables users to explore correlation. For example, in Figure 3, we see that the number of calls in the D4D-data highly correlates with duration.

The measure overview provides a high-level overview of interesting points in time that require further investigation. Depending on the chosen measure one can focus on (any combination of) curves, peaks and dips. For example, if call change is selected users can identify points in time where change is high by focusing on peaks. The identified points can then be further explored in detail using the different linked views.

On mouse hovering the according measure value is highlighted and the actual value is shown. Furthermore, the aggregated measure for the selected time point is broken down into individual values that are shown in the measure contribution view.

4.3 Measure contribution view

The measure contribution views shows the individual contributions to the aggregated measure of the selected time point in the measure overview. The individual contributions of communication channels (antenna-to-antenna) are shown as horizontal bars (see Figure 4). The horizontal bars are sorted from highest contribution (most important) to lowest contribution to the aggregated measure value. Each bar shows the region, department, sub-prefecture and identifier of both the sending and receiving cell towers. Furthermore, the number of calls (or a different selected measure) over this communication channel is shown along with the measure of the previous day for comparison purposes. This difference is also encoded as bar color; red or green indicates that on this point in time there were less or more calls compared to the previous point in time. By default only the fifty most important contributors are shown. Users are enabled to adjust this value to their likings. In addition we provide users with filtering options. Filtering is possible on the minimum number of calls required or the minimum difference of the number of calls between the previous and current point in time. This enables users to focus on communication channels with low, average or high activity.

All communication channels shown in the measure contribution view are also shown in the geospatial view for spatial identification. By hovering over an individual communication channel in the contribution view, the according channel is also highlighted in the geospatial view.

4.4 Geospatial view

The geospatial view displays the map of Ivory Coast. On top of this map all communication channels and involved cell towers are rendered that are currently shown in the contribution view. The communication channels are rendered as arcs. The direction of the communication is encoded clockwise (see Figure 5). Also here, color depicts whether the number of calls (or a different measure) is lower (red) or higher (green) compared to the previous point in time. The opacity of the arcs depend on the contribution value, similar to the length of the bars in the contribution view; the more important a link is, the higher its opacity. This emphasizes the more important communication channels for easy identification.

On mouse hovering, the according region, department and sub-prefecture are shown. Zoom and panning mechanisms can be used to navigate the map and focus on specific regions.

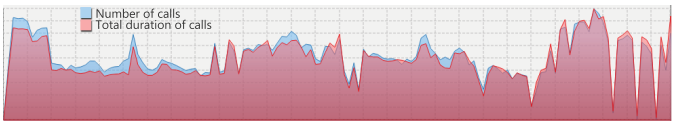


Fig. 3. Measure overview simultaneously rendering multiple measures enabling high-level correlation exploration.

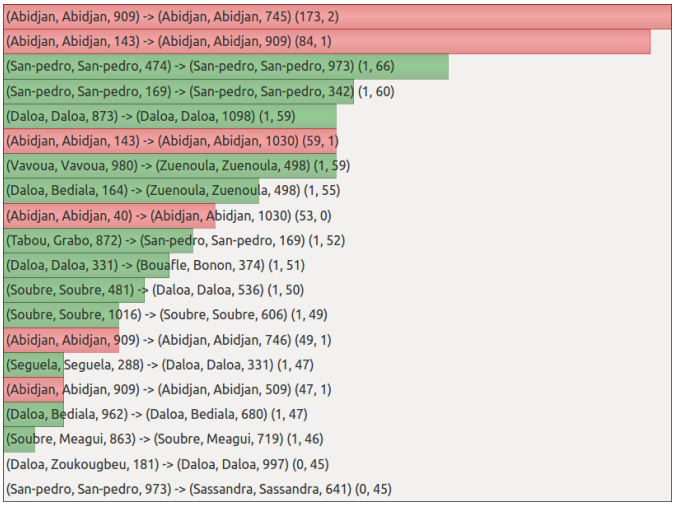


Fig. 4. Measure contribution view, providing information on most important contributors (communication channels) to the according measure and selected point in time.

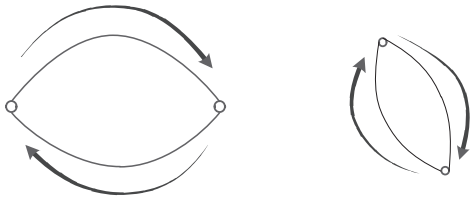


Fig. 5. Clockwise encoding of communication channel direction such that the visual variable color can be employed to encode a different aspect.

Technical details The arcs are rendered as quadratic Bezier curves. First the vector from source to destination point is determined. Next we compute the vector orthogonal to this vector with half the length and position it halfway between the source and destination point. Now we take the endpoint of this orthogonal vector as the control point for the quadratic Bezier curve. Due to the computation of the orthogonal vector, taking source and destination point into account, the clockwise direction is automatically inferred. The towers are rendered as white dots with a radial gradient from white opaque (innermost) to full transparent blue (outermost). Finally, additive blending techniques are used to render the towers and arcs on the map to create a subtle aesthetically pleasing glow effect in dense areas. In addition, this allows users to differentiate between increasing and decreasing traffic; precisely overlapping arcs where one is increasing (green) and the other decreasing (red) are rendered as yellow.

4.5 Matrix view

The matrix view provides a holistic overview of the behavior over time for each individual cell tower. On the vertical axis all cell towers are shown. The horizontal axis denotes time. Each row represents the behavior of one cell tower over time. On the intersection of a tower and point in time a rectangle is drawn. This rectangle represents a measure value, e.g., number of calls for one antenna at one time interval. The

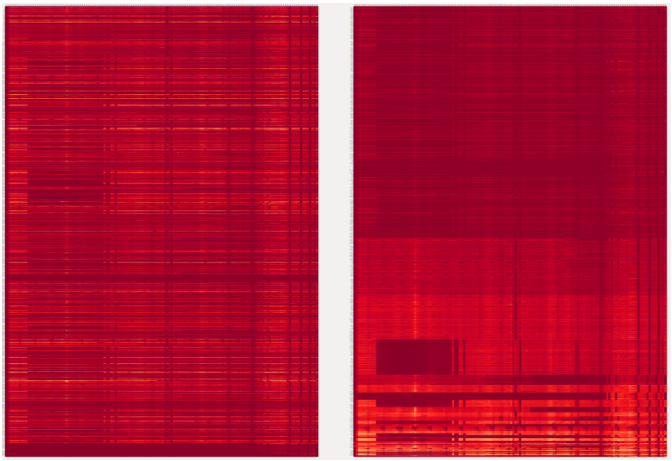


Fig. 6. Non-clustered (left) and clustered (right) matrix view grouping towers with similar communication behavior over time.

rectangles are rendered using a heat-map technique; each rectangle is colored based on the according value, here we use a dark-red to yellow to white colormap; dark-red represents the lowest value, white the highest. Zoom and panning techniques are provided to navigate and explore the matrix.

The matrix enables users to identify towers of similar behavior over time. However, because of limited screen resolution this poses serious difficulty on the task. Therefore, users are enabled to interactively apply clustering methods to the rows shown in the matrix. Once clustered, the rows of the matrix are re-ordered (see Figure 6). Towers that belong to the same cluster (all having similar behavior) are grouped together. In addition, the clusters themselves are also sorted based on cluster-size. We offer cluster parameters to users, which can interactively be adjusted. The result of a parameter change is directly reflected in the matrix view by reordering. The parameters available to users are cluster method (e.g., hierarchical, k-means, k-medians), distance metric (e.g., Euclidean, Manhattan, Pearson, Spearman, Kendall), number of desired clusters, and, time period to take into account while clustering.

If a clustering is applied on the matrix view, users are enabled to only show clusters of interest. Furthermore, the geographical location of the towers belonging to a cluster can be shown or hidden in the geospatial view.

4.6 Linking and integration

From each of the initial views additional views can be opened for the inspection of details. For example, if one communication channel is identified in the contribution view, an additional overview can be created showing the number of calls (or different measure) for the entire timespan to verify outlier behavior. Similar, the number of calls for a specific cell tower can be shown from the matrix view. The creation of new views enables comparison both in time and space to verify hypotheses. Finally, data for a combination of point (or period) in time, range of cell towers, and, range of regions, can be exported to a file for further investigation in external tools such as SynerScope [1] (see Figure 7). Also, facilities for easily searching the internet for events on a specific date and region are built-in. On double clicking in the geospatial view the platform-specific default browser is opened with according constructed search strings.

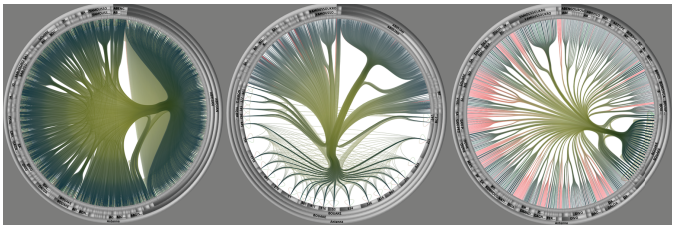


Fig. 7. Part of the D4D-data exported and loaded into SynerScope for further investigation revealing towers with unilateral behavior (only sending or only receiving at certain days).

5 Use cases

In the following sections, typical use cases are presented that demonstrate the power of the the visual analytics approach to the exploration and analysis of massive mobile data in the context of the D4D challenge. First some background knowledge on Ivory Coast is discussed to provide a context. This background knowledge is assembled based on United Nations reports [26, 27, 28, 29, 30]. Next we provide general findings and interesting correlations are extracted from complex patterns found while browsing through the data using the prototype.

5.1 Background knowledge

On November 28, 2010 elections were held to choose a new president for Côte d'Ivoire. There were two candidates to be chosen from, the current president Mr. Gbagbo (leader of the FPI party) and the opponent Mr. Ouattara (leader of the RHDP party). On December 2, 2010 the Independent Electoral Commission announced that Alassane Ouattara garnered 54.1 per cent of the votes while Laurent Gbagbo received only 45.9 per cent of the votes. That same day, the constitutional council declared the electoral results to be invalid, due to missing the deadline for announcing the provisional results. The next day, December 3, 2010 the constitutional council proclaimed the final results of the presidential elections. This time, however, Laurent Gbagbo received 51.45 per cent of the votes while Alassane Ouattara received 48.55 per cent of the votes. These results of the first announcement were later certified as the rightful outcome of the elections by the UNOCI [28]. However, Laurent Gbagbo did not step down. This started the post-electoral crisis, resulting in violent attacks, killing of civilians, rape, torture, displacements, and, inhumane and degrading treatment. While human rights abuses have been committed by both sides, most of the killings have been carried out by elements of the forces loyal to Mr. Gbagbo [28]. The situation continued to deteriorate until former President Gbagbo was apprehended on April 11, 2011 [26]. However, pro-Gbagbo militias, mercenaries and FDS (former army) elements continued fighting. Some 50 of those elements surrendered to FRCI (new army) on April 29, 2011, the rest fled towards the Liberian border area, where they continued to kill civilians and loot property in south-western Côte d'Ivoire. Clashes between the FRCI and pro-Gbagbo militias and mercenaries continued to be reported there, as well as violence against civilians in the west and south-west [26]. On December 11, 2011 legislative elections were held in a generally calm and peaceful manner, however, the country is still struggling to recover from the devastating crisis [27]. Here our data starts. We are provided with CDR-data covering the period December 5th, 2011 until April 22nd, 2012. Because media is government controlled and many reports are made that journalists are oppressed and newspapers are banned [26, 27, 28, 30], we solely rely on U.N. reports and reports of the International Crisis Group [18, 19, 20] as our source of major events that occurred during the data-period. During this period, the situation remains particularly fragile in western Côte d'Ivoire, where large numbers of weapons, armed elements, former combatants, militias and dozos, as well as competition over the control of resources are significant sources of insecurity [29]. Most of the incidents occurred in the west, although insecurity has increased in other parts of the country. Law enforcement, while present throughout the country, remains ineffective, and some areas are still under the protection of the dozos, which increases insecurity [29].

5.2 General findings

In the measure overview events that generate a peak in the number of calls and also in call change behavior are directly visible, such as the celebration of new year (see Figure 8).

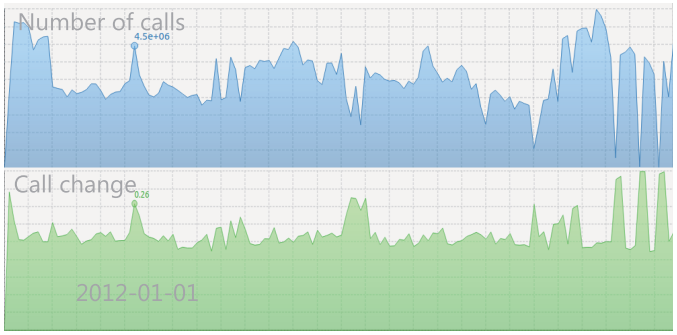


Fig. 8. Peak in the number of calls and call change due to new year.

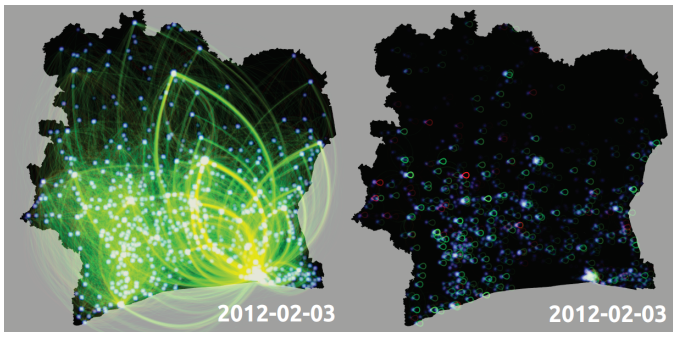


Fig. 9. Predominantly local communication (right) and higher level communication patterns revealed (left).

From the inspection of the highest contributors to the number of calls on any day it becomes clear that the highest number of calls is very local. This appears in the geospatial view as many predominantly self-loops (see Figure 9(right)). By plotting all contributors using a high transparency value for the individual edges, higher level communication is revealed (see Figure 9(left)). We see for example, that there is a strong link of communication between Bouake and Abidjan, but significantly less strong between Bouake and Yamoussoukro.

From the clusters in the matrix view, some clusters can directly be explained. For example, some clusters show a high week-weekend pattern, with less traffic in the weekends. These towers are based in Abidjan, more specific in the Plateau and Adjame region where most companies are located, thus indeed less traffic in the weekends (see Figure 10).

5.3 Local event increased call correlation patterns

During the following events there is a local increase of cell phone traffic. There is a clear correlation of call change and events that are directly visible when exploring the data. Below, these correlations are discussed in chronological order.

On January, 21, 2012, there is a meeting of the FPI (pro-Gbagbo) in Abidjan. This meeting is violently disrupted by supporters of the RHDP (pro-Ouattara). One person was killed, several were injured and property was damaged. Also, national police officers were assaulted. On this day we see, Figure 11, that there is an heavy increase in telephone calls in the west (pro-Gbagbo), also noticeable is the increase in traffic from Abidjan to the western region (probably supporters calling their friends and family, informing them of the disruption).

On February, 11, 12 and 13, 2012 clashes between communities are reported in Arrah. During these days (especially 12 and 13) there is indeed a local increase in the number of calls to the Arrah region, directly visible in the geospatial view (see Figure 12). If we bring up detailed information on the number of calls from the specific communication channels (antenna-to-antenna) there is indeed a remarkably high spike at these days, confirming something is going on.

On February, 21, 2012, the village of Zriglo is attacked, killing six

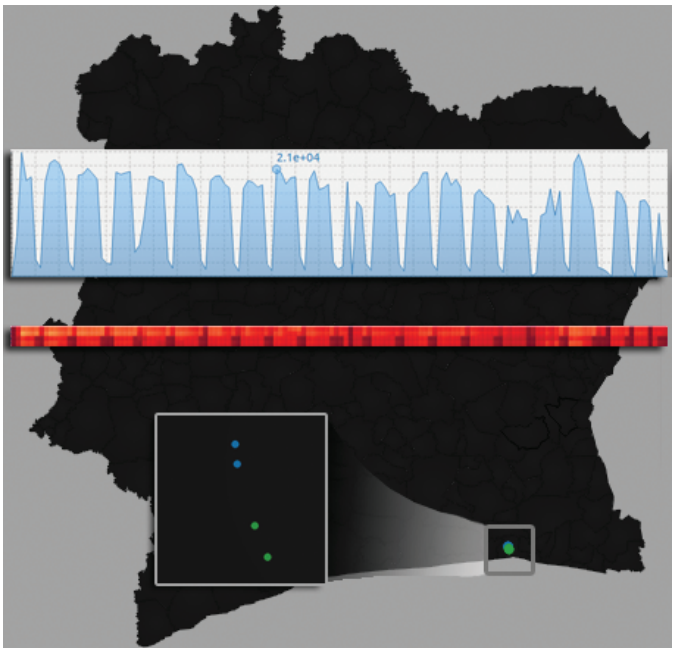


Fig. 10. Cluster of towers having a strong week-weekend pattern are located in Abidjan.

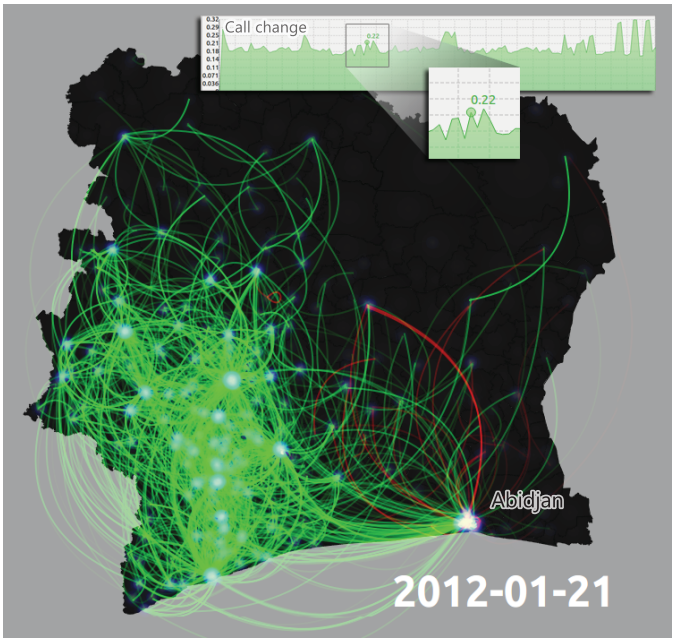


Fig. 11. FPI meeting in Abidjan disrupted by supporters of the RHDP.

persons and wounding many more. There is indeed an increase in the number of calls to this village for this day. This also becomes apparent if the individual antenna-to-antenna communication is inspected, showing an unusual peak (see Figure 13).

Cocoa is a key commodity in Ivory Coast. The country is the world's largest producer of cocoa beans, accounting in 2010/2011 for a total 35% of the world's total production [17]. Because of such a high economic value cocoa has already a driving factor for conflict in the country [12, 35].

Cocoa trees in Ivory Coast are harvested twice a year: a main harvest happens between September to December, while a second minor harvest happens from April to June [11]. April/May is a particularly important time for cocoa farmers, as two important events other than

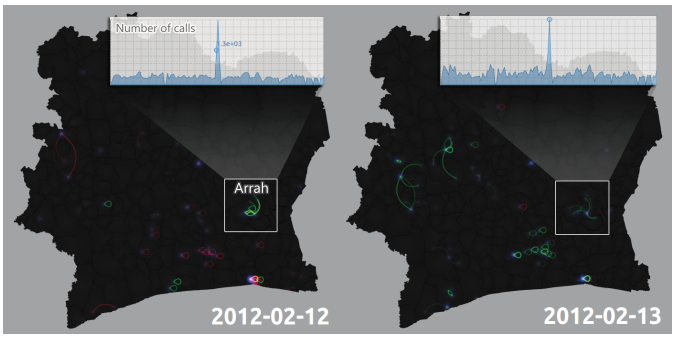


Fig. 12. Clashes between communities in Arrah.

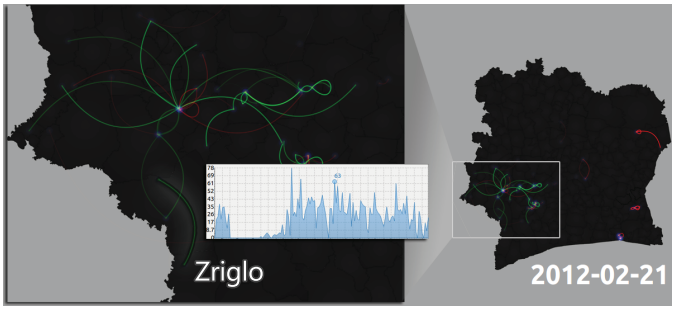


Fig. 13. Attack on the village of Zriglo.

the aforementioned harvest happen: (a) it is one of the two times when pesticides are applied on cocoa plants; (b) if precipitation has been abundant farmers can establish new cocoa plantations or expand existing ones (starting field operations in May). Moreover, yam varieties growing in forest areas are planted in April and May [11]. Two main hypotheses can be thus made:

- Events happening in April/May (i.e. harvesting, marketing and input supply) are likely to produce an increase in telephone traffic from the western cocoa-growing regions towards urban areas and other (market, logistics) hubs.
- Correlation with abundant (i.e. above normal) rainfall in March/April is likely to produce additional increase in telephone traffic from the western cocoa-growing regions towards agricultural inputs supply hubs.

Two positive rainfall anomalies that may have impacted agricultural activities and may be linked with increased phone calls have been identified in two regions: Bas-Sassandra and Dix-Huits Montagnes (see Figure 14). In the Bas-Sassandra region it was reported a positive rainfall anomaly during the first decade of April (see Figure 15) and an increase in telephone calls on April 8, 2012, was noticed in two different areas: (a) the area of Sassandra and San Pedro subprefectures; and (b) the area of Tabou, Grand-Bereby and San Pedro subprefectures.

Higher data granularity was available only for the area of Sassandra. Rainfall was absent during the first decade of April, with the exception of an highly-anomalous storm on April 3 reported in Sassandra [5]. Still this event does not explain the increase of telephone calls on 8 April. It is important to highlight that April 8 2012 matched with Easter. Some sort of correlation with religious events may hence be assumed.

In the Dix-Huits Montagnes region it was reported a positive rainfall anomaly during the second decade of April (see Figure 16) and an increase in telephone calls on April 13 and 16, 2012. Particularly the increase in phone calls was localized in the Man sub-prefecture, which is an important center for both cocoa and coffee production national and is the most important production area of coffee in the whole country.

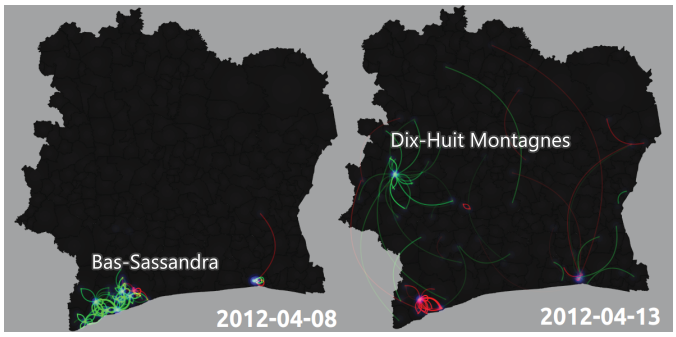


Fig. 14. Increased phone calls correlated with rainfall anomalies.

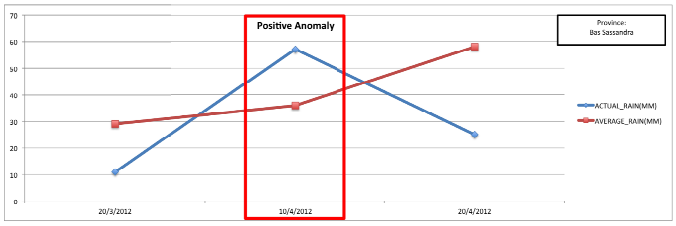


Fig. 15. Rainfall activities in Bas-Sassandra region between 20 March and 20 April 2012 (10-day cumulated estimates) [4].

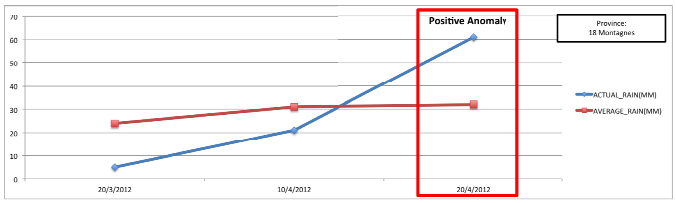


Fig. 16. Rainfall activities in Dix-Huit Montagnes region between 20 March and 20 April 2012 (10-day cumulated estimates) [4].

5.4 Local event decreased call correlation patterns

In the events described below we found a strong local decrease of call activity (majority of complete shutdown of call activity). Again, these local decreases were clearly visible in the visualizations while browsing the data. Once more, these events are discussed in a chronological fashion.

On December 15th, 2011 all cell towers in the western region appear to be shut down (see Figure 17). A large number of antennas is connected to the electric grid of Ivory Coast. On this day half the country was shut down due to electric failure. Next, we identify additional towers by applying a hierarchical clustering on call behavior over time (see Figure 18). These cell towers have similar call behavior over time. If we inspect one of these towers (typically they all have this pattern) we see that these towers are not entirely shut down, but they remain to have an unusual low number of calls. Then, this low activity remains until the 4th of January, when they appear to be turned on again.

On January 15th, 2012, there were confrontations between communities in Gagnoa resulting in the deaths of 16 people, injuries to many more and the burning of several houses. On this day there is indeed a regional call change. The calls in this region drop to 0 (see Figure 19). A possible cause is the fleeing of locals and damaging of the cell towers.

In early February, 2012 (no date mentioned) there were reports of confrontations between farmers and cattle breeders in Odienne. This led to injuries to several persons and the displacement of some 200 people. On February 5th, 2012 a shutdown of towers in the region around Odienne immediately show up in the call change graph (see

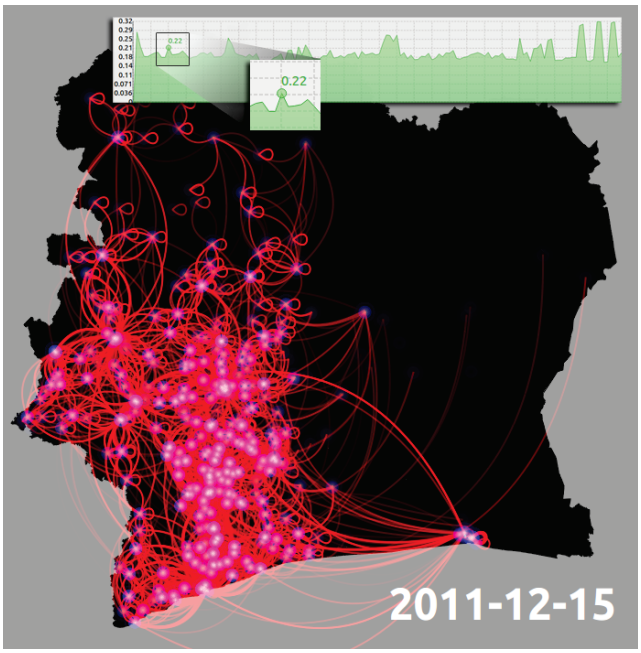


Fig. 17. Electric failure results in shutdown of cell towers connected to the power grid.

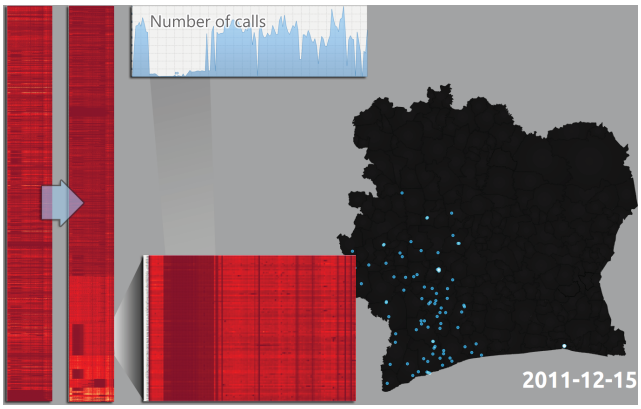


Fig. 18. Cluster of towers identified in western region with unusual low activity for a significant period.

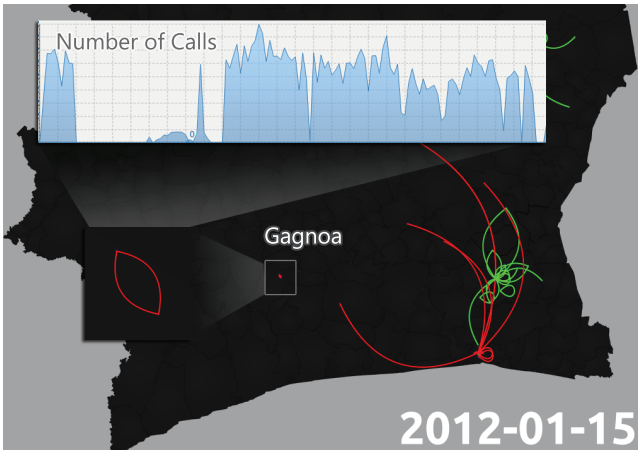


Fig. 19. Confrontations between communities in Gagnoa.

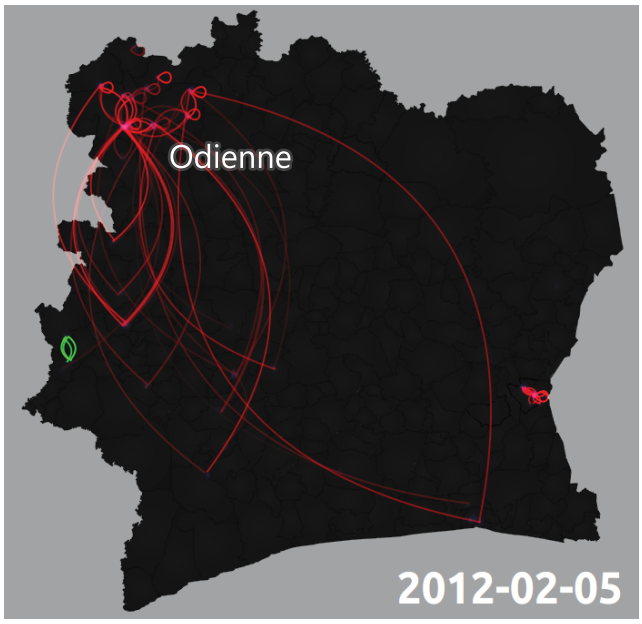


Fig. 20. Confrontations between farmers and cattle breeders in Odienne.

Figure 20). It should be underlined that farmer-pastoralists conflicts in the area of Odienne have been already reported by scholars [14], and the livelihoods of local farmers suffered additional significant stress in the recent two years. Particularly, in March 2012 FAO [2] classified Ivory Coast as a country in need of external assistance because of severe localized food insecurity, citing the northern regions as a food insecurity hotspot because of lacking support services and conflict-related damages to agricultural activities. We performed an assessment of disasters and weather for the time covered to test for correlation. However, no disasters were reported during the whole period according to the International Disaster Database [3]. Also, no significant weather events in terms of rainfall that may have disrupted telecommunications were recorded both in Gagnoa [5] and Odienne, based on Interpolated Estimated Dekadal Rainfall provided by NOAA/FEWSNet[1] [4] for the whole Denguele region (see Figure 25).

On March 3rd, 2012, near Daloa and neighboring big cities Man and Duekoue there was a drop in the number of calls. The number of calls on these communication lines dropped from the normal level of 100-200 to 10-20 on this day (see Figure 21). On March 5th, 2012, in Agboville, we again see a decreased call activity (no calls) of at least two towers in the area, that directly pop-up in the call change behavior (see Figure 22). A thunderstorm was recorded on March 3rd, 2012 in Daloa [5]. This event may have seriously impacted telecommunications activity in the neighboring areas. The same weather conditions were reported in Abidjan on March 5, 2012 and we can fairly assume that Agboville (65 km distant from Abidjan) was affected as well by the thunderstorm.

On March 13th, 2012 cell towers in the western region are shut down (see Figure 23). This again might be the result of electric failure.

Finally, if we cluster towers on the number of calls over time, several more interesting clusters are revealed, all having a different shutdown period (see Figure 24). To our opinion the shut down of towers can have a number of reasons: (a) this is missing data, Orange could or did not register calls, (b) external factors making communications more difficult, like weather and disasters, (c) sabotage on the antennas, (d) electric failure or diesel replenishment problems for off-grid cell towers, or (e) other technical problems.

[1]Quantitative estimate of rainfall combining METEOSAT derived Cold Cloud Duration imagery and data on observed rainfall (GTS-Global Telecommunication System by the NOAA Climate Prediction Centre)

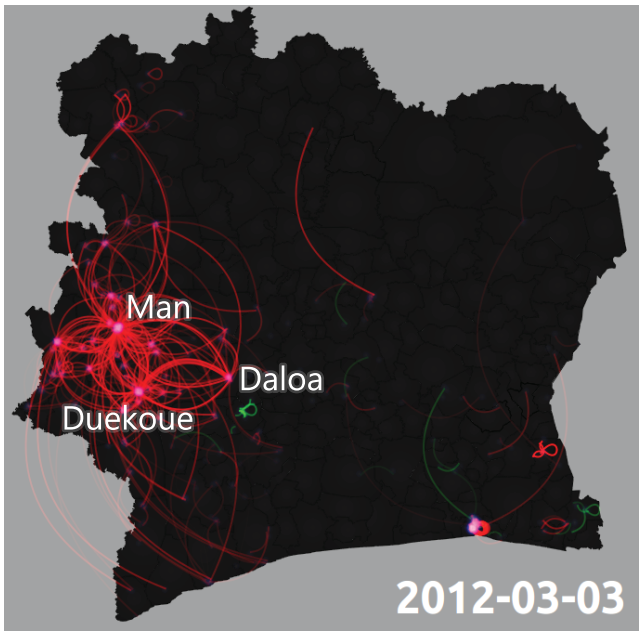


Fig. 21. Thunderstorm reports near Daloa, Man and Duekoue, presumably impacting telecommunication activity.

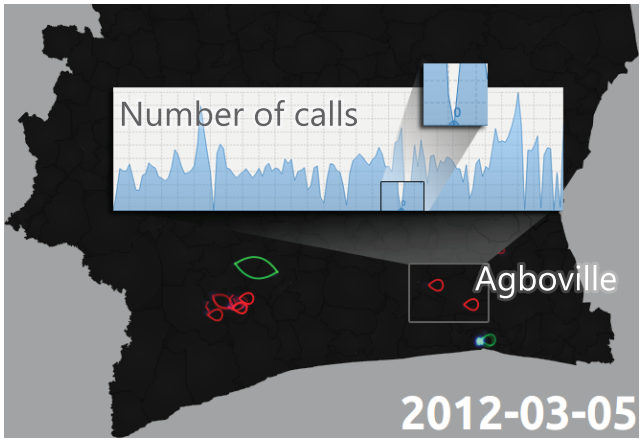


Fig. 22. Bad weather conditions influencing local call activity around Agboville.

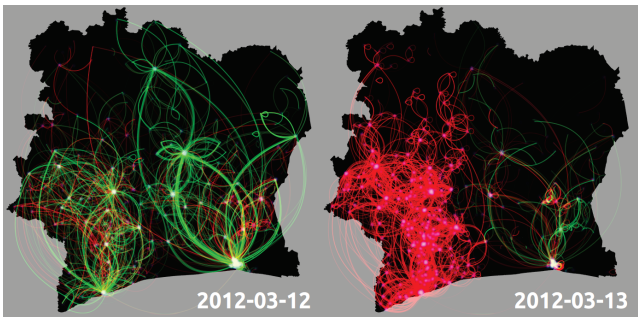


Fig. 23. Day before (left) supposedly electric failure in the western region (right).

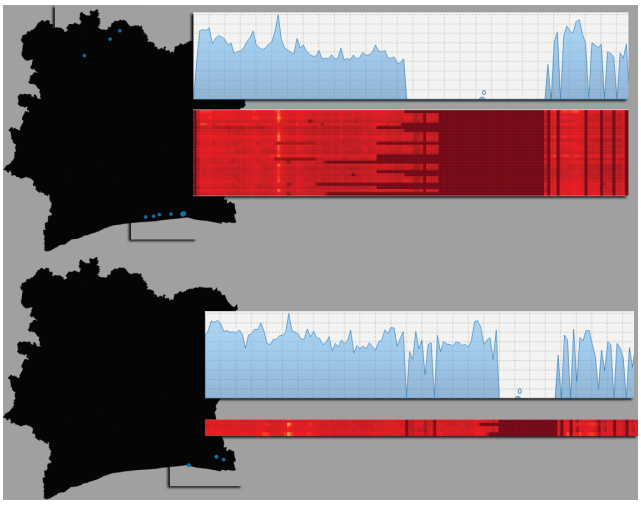


Fig. 24. Clusters of towers with different shut down periods.

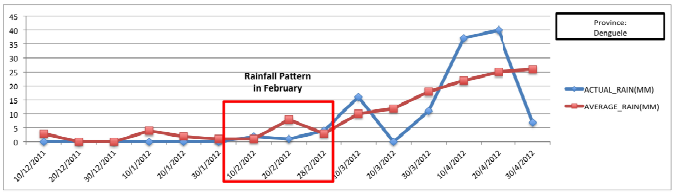


Fig. 25. Highlight of rainfall pattern in Denguele region in February 2012. No rainfall activity was recorded on the whole region [4].

The first explanation can be ruled out as it is stated in the D4D data information report that indeed there is missing data but this only covers a period of about 100 hours ($\pm$ 4 days) [10]. However, we notice shut downs (or significantly lowered communication) for periods more than 15 continuous days. As we have seen weather conditions explain some of the local decreased cell tower activity. However, due to time constraints no explanation is found for the clusters of towers that have a significant period of lowered (or none at all) activity. These cases are currently being investigated in a collaborative effort with Orange.

5.5 Socio-economic development

We believe that by focusing on call change, local events can be detected as we have shown in the use cases. Both the local increase and decrease of call behavior provides insight in complex patterns. By focusing on clusters of cell towers having similar call behavior, events can be detected. These events are of different nature, such as weather (heavy rainfall in important cocoa area), social (new years eve), political (party meetings), disorder (clashes between communities) etc. and can only be explained by domain experts by enriching the visual analysis process with external data to gain insight in complex correlations, anomalies and communities both in time and space. This in turn enables event detection which is an important first step towards prediction, improving early intervention through development, aid and other civil initiatives.

6 Conclusions

We aimed at developing a tool for the exploration and analysis of massive mobile data supporting all aspects of the process. We identified user tasks and requirements from which appropriate visualization, interaction and automated support techniques are selected. Next, we implemented these in a highly interactive prototype. We next showed the effectiveness of our visual analytics approach by applying the prototype on massive mobile phone data containing 2.5 billion calls and SMS exchange between around 5 million users located in Ivory Coast over a period of 5 months, provided by France Telecom within the context of the Orange D4D challenge. From the typical use cases obtained while browsing the data, we extracted significant and interesting events by cross-correlating these using UN reports and weather information.

6.1 Future Work

In the context of the D4D challenge we mainly focused on the first dataset, containing detailed information of tower-to-tower communication due to time constraints. We believe it would be valuable to incorporate additional visualizations and automated techniques that enable also the exploration and analysis of the remaining datasets. Also, we believe exploration and analysis of non-aggregated data provides even more insight in complex patterns.

REFERENCES

[1] SynerScope connecting the dots. http://www.synerscope.com. Accessed: 13/02/2013.

[2] Crop prospects and food situation. *Food and Agriculture Organization of United Nations*, (1), March 2012.

[3] Centre for research on the epidemiology of disasters, em-dat, the international disaster database. http://www.emdat.be/database, 2013. Accessed on 1/02/2013.

[4] Noaa/fewsnet. famine early warning system network. http://www.cpc.ncep.noaa.gov/products/fews/africa/, 2013. Accessed on 1/01/2013.

[5] Weather underground. historical weather. http://www.wunderground.com/history/, 2013. Accessed on 1/02/2013.

[6] G. Adrienko, N. Adrienko, M. Mladenov, M. Mock, and C. Politz. Identifying place histories from activity traces with an eye to parameter impact. *Visualization and Computer Graphics, IEEE Transactions on*, 18(5):675 –688, may 2012.

[7] G. Andrienko, N. Andrienko, M. Mladenov, M. Mock, and C. Politz. Discovering bits of place histories from people's activity traces. In *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on*, pages 59 –66, oct. 2010.

[8] R. Baruah and P. Angelov. Evolving social network analysis: A case study on mobile phone data. In *Evolving and Adaptive Intelligent Systems (EAIS), 2012 IEEE Conference on*, pages 114 –120, may 2012.

[9] V. Blondel, M. Esch, and C. Chan. Geofast. http://www.geofast.net. Accessed on 14/02/2013.

[10] V. D. Blondel, M. Esch, C. Chan, F. Clérot, P. Deville, E. Huens, F. Morlot, Z. Smoreda, and C. Ziemlicki. Data for development: the d4d challenge on mobile phone data. *CoRR*, abs/1210.0137, 2012.

[11] M. Boas and A. Huser. Child labour and cocoa production in west africa. the case of côte d'ivoire and ghana. research program on trafficking and child labour. fafo-report 522. web edition. 2006.

[12] R. Carroll. Chocolate war erupts in ivory coast, the guardian (14/05/2004), 2004.

[13] C. Correa, T. Crnovrsanin, C. Muelder, Z. Shen, R. Armstrong, J. Shearer, and K.-L. Ma. Cell phone mini challenge award: Intuitive social network graphs visual analytics of cell phone data using mobivis and ontovis. In *Visual Analytics Science and Technology, 2008. VAST '08. IEEE Symposium on*, pages 211 –212, oct. 2008.

[14] Y. Diallo and M.-P.-I. für ethnologische Forschung. *Conflict, Cooperation and Integration: A West African Example (Côte D'Ivoire)*. Max Planck Institute for Social Anthropology working papers. Max Planck Inst. for Social Anthropology, 2001.

[15] N. Eagle and A. Pentland. Eigenbehaviors: identifying structure in routine. *Behavioral Ecology and Sociobiology*, 63:1057–1066, May 2009.

[16] B. Höferlin, M. Höferlin, and J. Räuchle. Visual analytics of mobile data. In *Proceedings of the Nokia Mobile Data Challenge 2012 Workshop*, 2012.

[17] I. C. C. O. International Cocoa Organization. Annual report 2010/2011. *International Cocoa Organization*, 2012.

[18] I. C. G. International Crisis Group. Côte d'ivoire: Is war the only option? *Africa Report*, 171, 3 March 2011.

[19] I. C. G. International Crisis Group. A critical period for ensuring stability in côte d'ivoire. *Africa Report*, 176, 1 August 2011.

[20] I. C. G. International Crisis Group. Côte d'ivoire: Defusing tensions. *Africa Report*, 193, 26 November 2012.

[21] D. A. Keim, F. Mansmann, J. Schneidewind, J. Thomas, and H. Ziegler. Visual data mining. chapter Visual Analytics: Scope and Challenges, pages 76–90. Springer-Verlag, Berlin, Heidelberg, 2008.

[22] D. A. Keim, F. Mansmann, J. Schneidewind, and H. Ziegler. Challenges in visual data analysis. In *Proceedings of the conference on Information Visualization*, IV '06, pages 9–16, Washington, DC, USA, 2006. IEEE Computer Society.

[23] G. Krings, F. Calabrese, C. Ratti, and V. D. Blondel. Urban gravity: a model for inter-city telecommunication flows. *Journal of Statistical Mechanics: Theory and Experiment*, 2009(07):L07003, 2009.

[24] M.-p. Kwan and J. Lee. Geovisualization of human activity patterns using 3d gis : A time-geographic approach in michael f . goodchild and donald g . janelle . eds . 2003 . *Spatially integrated social science*, 27:48–66, 2003.

[25] G. Sagl, M. Loidl, and E. Beinat. A visual analytics approach for extracting spatio-temporal urban mobility information from mobile network traffic. *ISPRS International Journal of Geo-Information*, 1(3):256–271, 2012.

[26] U. N. Security Council. Twenty-eighth report of the secretary-general on the united nations operation in côte d'ivoire. *S/2011/387*, 24 June 2011.

[27] U. N. Security Council. Twenty-ninth progress report of the secretary-general on the united nations operation in côte d'ivoire. *S/2011/807**, 30 December 2011.

[28] U. N. Security Council. Twenty-seventh progress report of the secretary-general on the united nations operation in côte d'ivoire. *S/2011/211*, 30 March 2011.

[29] U. N. Security Council. Special report of the secretary-general on the united nations operation in côte d'ivoire. *S/2012/186*, 29 March 2012.

[30] U. N. Security Council. Thirtieth progress report of the secretary-general on the united nations operation in côte d'ivoire. *S/2012/506*, 29 June 2012.

[31] Z. Shen and K.-L. Ma. Mobivis: A visualization system for exploring mobile data. In *Visualization Symposium, 2008. PacificVIS '08. IEEE Pacific*, pages 175 –182, march 2008.

[32] J. J. Thomas and K. A. Cook. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. National Visualization and Analytics Ctr, 2005.

[33] V. Traag, A. Browet, F. Calabrese, and F. Morlot. Social event detection in massive mobile phone data using probabilistic location inference. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pages 625 –628, oct. 2011.

[34] T. von Landesberger, S. Bremm, N. Andrienko, G. Andrienko, and M. Tekusova. Visual analytics methods for categoric spatio-temporal data. In *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on*, pages 183 –192, oct. 2012.

[35] D. Woods. The tragedy of the cocoa pod: rent-seeking, land and ethnic conflict in ivory coast. *The Journal of Modern African Studies*, 41:641–655, 11 2003.

[36] Q. Ye, B. Wu, D. Hu, and B. Wang. Exploring temporal egocentric networks in mobile call graphs. In *Fuzzy Systems and Knowledge Discovery, 2009. FSKD '09. Sixth International Conference on*, volume 2, pages 413 –417, aug. 2009.

[37] Q. Ye, T. Zhu, D. Hu, B. Wu, N. Du, and B. Wang. Cell phone mini challenge award: Social network accuracy; exploring temporal communication in mobile call graphs. In *Visual Analytics Science and Technology, 2008. VAST '08. IEEE Symposium on*, pages 207 –208, oct. 2008.

Exploring the Multilevel Community Structure in D4D Dataset

Xin Liu [1,2,3]
tsinllew@ai.cs.titech.ac.jp

Tsuyoshi Murata [1]
murata@cs.titech.ac.jp

Ken Wakita [1]
wakita@is.titech.ac.jp

[1]Tokyo Institute of Technology
2-12-1 Ookayama
Meguro, Tokyo
152-8552 Japan

[2]CREST, JST
K's Gobancho, 7, Gobancho
Chiyoda, Tokyo
102-0076 Japan

[3]Wuhan University of Technology
122 Luoshi Road
Wuhan, Hubei
430070 China

1. INTRODUCTION

Community detection in spatial networks is recently introduced by Expert et al. [2]. In spatial networks long-distance edges are often restricted, thus space can influence the network connectivity, either directly or indirectly. This is the so-called "space effect". Therefore, the key problem is to take out the space effect and reveal the hidden community structure that are not due to the space factor.

To handle this problem, Expert et al. proposed a Spa-Modularity. As the standard NG-modularity proposed by Newman and Girvan [6], Spa-Modularity involves a comparison between the observed network and a null model. The difference is that Expert et al. simulate the space effect in the null model using Newton's gravity law, with the hope that the space effect of the observed network can be taken out when make a comparison.

Recently, we created a family of Dist-Modularity for detecting community structure in various networks [5]. In this paper we applies Dist-Modularity to the antenna-to-antenna spatial network of D4D dataset [1]. The advantage of Dist-Modularity is that we can control the magnitude of space effect in the null model, and reveal multilevel community structure.

2. DIST-MODULARITY

Dist-Modularity is a variant of modularity. The difference is that the null model involved in Dist-Modularity captures the similarity attraction feature of many real-world networks, i.e., edges tend to link to nodes that are similar to each other.

We use d_{ij} to denote the similarity distance between v_i and v_j — the smaller of d_{ij}, the more similar of the two nodes. d_{ij} can be estimated by a distance function that takes some available information about v_i and v_j as input, e.g., the network structural information (such as the neighbors of v_i and v_j), additional nonstructural information (such as attributes and contents of v_i and v_j), or a mixture of them. Generally, d_{ij} should satisfy the following constraints

$$d_{ij} \geq 0 \text{ with equality iff } i = j, \tag{1}$$

$$d_{ij} = d_{ji}, \tag{2}$$

$$d_{ij} \leq d_{it} + d_{tj}. \tag{3}$$

Eq. (1) is self-evident — a node is the most similar to itself and hence the distance is zero. Eqs. (2) and (3) indicate the triangle inequality and the symmetry constraints, respectively.

Dist-Modularity is defined as

$$\mathrm{Q}^{\mathrm{Dist}}(\mathscr{C}) = \frac{1}{2m} \sum_{i,j=1}^{n} (A_{ij} - P_{ij}^{\mathrm{Dist}})\, \delta(l_i, l_j). \tag{4}$$

where $\mathscr{C}$ is a partition represented as a community assignment vector on the right-hand side of the equation, with element l_i indicating the community membership of the ith node v_i, n is the number of nodes, m is the number of edges, δ is the Kronecker's delta, A_{ij} is the element of the adjacency matrix $\mathbf{A}$ representing the number of edges between v_i and v_j in the observed network, and P_{ij}^{Dist} is the expected value of that number in the corresponding null model. For an undirected network, we define

$$P_{ij}^{\mathrm{Dist}} = (\tilde{P}_{ij} + \tilde{P}_{ji})/2, \tag{5}$$

$$\tilde{P}_{ij} = \frac{N_i N_j f(d_{ij})}{\sum_{t=1}^{n} N_t f(d_{ti})}. \tag{6}$$

N_i is a notion representing the importance of v_i, and satisfies the normalization condition

$$\sum_{i=1}^{n} N_i = 2m. \tag{7}$$

$f(d)$ is a deterrence function which will be explained later. Note that this null model preserves the number of edges of the observed network, as we can derive that $\sum_{i,j=1}^{n} P_{ij}^{\mathrm{Dist}} = \sum_{i=1}^{n} N_i = 2m$.

From Eq. (5) and (6), we can find that

- P_{ij}^{Dist} is positively related to N_i and N_j. That is, in this null model edges tend to link to important nodes;
- P_{ij}^{Dist} is negatively related to d_{ij} (Suppose $f(d)$ is a strictly monotonic function). That is, in this null model edges tend to link to nodes that are similar to each other, an evidence of the similarity attraction feature.

We have a freedom in choosing N_i and $f(d)$. With different choices, Dist-Modularity can apply to various networks. For example, if we specify $f(d) = 1$ the similarity attraction feature actually vanishes. At the same time if we specify $N_i = k_i = \sum_{j=1}^{n} A_{ij}$, Dist-Modularity reduces to NG-Modularity. The mechanism behind Eq. (5)-(7) is driven by the field theory in Physics. For more information please see Ref. [5].

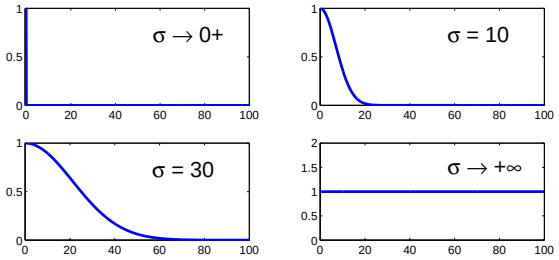


Figure 1: The plot of function $f(d) = e^{-(d/\sigma)^2}$ with different values of σ.

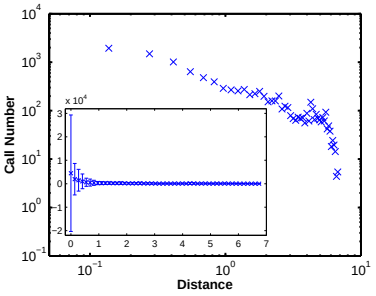


Figure 2: The number of calls between antennas decreases with their spatial distance.

3. DIST-MODULARITY FOR SPATIAL NETWORKS

In spatial networks, we use d_{ij} to denote the spatial distance between v_i and v_j. Then, the similarity attraction feature becomes that "edges tend to link to nodes that have a shorter spatial distance". This is just the space effect. Hence, Dist-Modularity applies to spatial networks and can reveal the hidden community structure beyond space effect.

In general the space factor influences network connectivity in a complex way, and it is not easy to simulate the space effect in a single manner. Thus, we give a particular interest to the following deterrence function

$$f(d) = e^{-(d/\sigma)^r}, \tag{8}$$

where $\sigma \in (0, +\infty)$ is a parameter. Note that $f(d)$ is a monotonically decreasing function falling in $(0, 1]$. As Fig. 1 shows, when σ is small $f(d)$ decreases sharply, indicating a strong space effect; when σ is large $f(d)$ decreases slowly, indicating a mild space effect. As $\sigma \to +\infty$, $f(d)$ decays to a constant function and the space effect vanishes. Therefore, the advantage of this deterrence function is that we can control the magnitude of space effect by tuning σ.

4. EXPERIMENT

The D4D dataset are based on anonymous records of mobile phone calls between five million of Orange's customers in Ivory Coast during the period from Dec 1, 2011 to Apr 28, 2012 [1]. We focus on the antenna-to-antenna network of this dataset. The nodes represent 1216 antennas which are associated with spatial coordinate information. The edges are placed between antennas that have communications, with edge weight representing the numbers of calls *. Note that this network is temporal [3] — it has ten consecutive slices and each slice represents a two-week period (all together accounts for the five-month period from Dec 1, 2011 to Apr 28, 2012). We find that this network has space effect. That is, the number of calls between antennas decreases with their spatial distance, as shown in Fig. 2.

To reveal the community structure, we optimize Dist-Modularity (specifying $N_i = k_i$) by the modularity-specialized label propagation algorithm [4]. As shown in Fig. 3, we can explore communities at different levels along the σ axis, and the community evolution along the time slot.

*The edge weight representing duration of calls is also available, but as an example here we take the one representing the numbers of calls.

Let $\bar{d} = \sum_{i,j=1}^{n} d_{ij}/n^2$ be the average spatial distance over all node pairs. Fig. 4 shows the community structure in one of the network slice when σ equals to $0.1\bar{d}$, $0.5\bar{d}$, $1\bar{d}$, $5\bar{d}$, and $10\bar{d}$, respectively. Fig. 5 shows the community evolution at $\sigma = 1\bar{d}$. From Fig. 4 we can find that as σ increases, the community structure gradually correlates with the geography. At $\sigma = 1\bar{d}$, the border lines of the 19 regions of the country becomes a perfect classifier. Later, some of the communities are combined, resulting in coarser communities at $\sigma = 10\bar{d}$. Note that as $\sigma = 10\bar{d}$, it is so large that the space effect in the null model can be ignored. Thus we can expect that community structure at this point is quite similar to that obtained by optimizing NG-Modularity.

In spatial networks the space factor influences network connectivity in a complex way, and it is not easy to simulate the space effect accurately. Dist-Modularity enables us to simulate the whole landscape of possible space effects, and reveal the multilevel community structure.

Acknowledgments: Our study and research activity were performed using mobile communication data made available by France Telecom and Orange Cote d'Ivoire within the D4D Challenge. This research was partly supported by the Japan Science and Technology Agency (JST), the Core Research of Evolutionary Science and Technology (CREST) research project.

5. REFERENCES

[1] V. D. Blondel, M. Esch, C. Chan, F. Clerot, P. Deville, E. Huens, F. Morlot, Z. Smoreda, and C. Ziemlicki. Data for development: the d4d challenge on mobile phone data. *arXiv:1210.0137*, 2012.

[2] P. Expert, T. S. Evans, V. D. Blondel, and R. Lambiotte. Uncovering space-independent communities in spatial networks. *Proc. Natl. Acad. Sci. USA*, 108(19):7663–7668, 2011.

[3] P. Holme and J. Saramäki. Temporal networks. *Physics Reports*, 519:97–125, 2012.

[4] X. Liu and T. Murata. Advanced modularity-specialized label propagation algorithm for detecting communities in networks. *Physica A*, 389(7):1493–1500, 2010.

[5] X. Liu, T. Murata, and K. Wakita. Extending modularity by capturing the similarity attraction feature in the null model. *arXiv:1210.4007*, 2012.

[6] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69:026113, 2004.

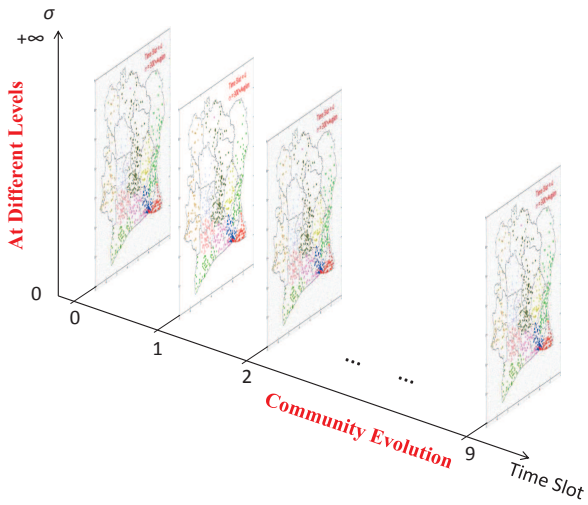


Figure 3: Explore multilevel community structure in the antenna-to-antenna network.

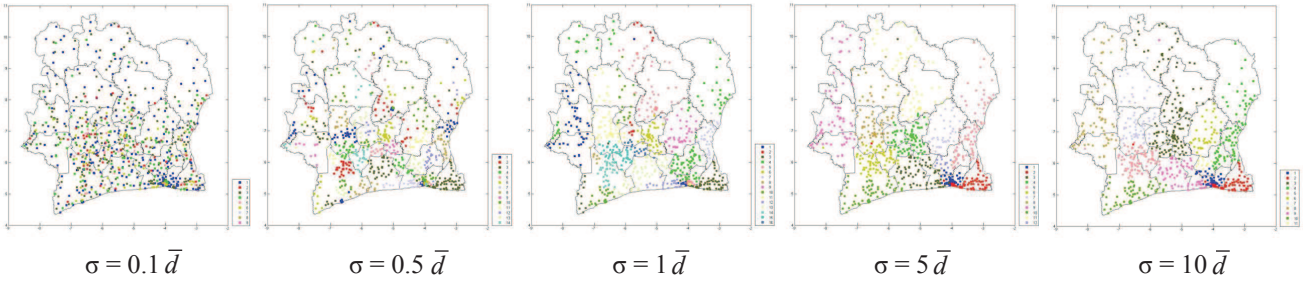


Figure 4: Explore communities at different levels as σ increases from 0 to $+\infty$.

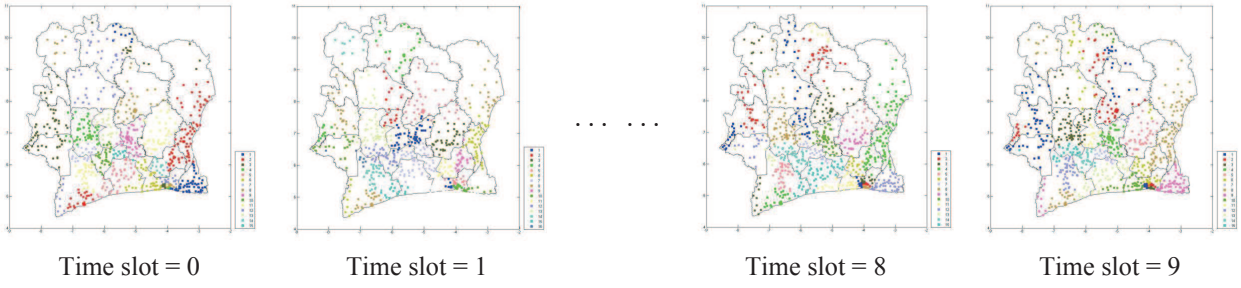


Figure 5: Explore the community evolution along the time slot.

Social Capital for Economic Development: Application of Time Series Cluster Analysis on Personal Network Structures

Benedict Lim[1], Derek Doran[2], Veena Mendiratta[3], Maria Rodriguez[1], Diego Klabjan[1]
[1]Northwestern University, [2] University of Connecticut, [3]Bell Labs, Alcatel-Lucent
Contact Author: veena.mendiratta@alcatel-lucent.com

Abstract

The availability of cell phone usage data opens up possibilities for using the insights from analysis of these datasets to guide policies and in turn boost economic development. In this paper, we propose the use of time series clustering to social network attributes and metrics as a way to quantify social capital. We demonstrate how our approach can allow policymakers to use the results from clustering to identify particular groups in the population and potentially provide assistance to boost social capital levels, and in turn economic growth.

1. Introduction and Motivation

Social capital, though widely agreed to share the core idea that social networks have value and is closely related to the structure of a social network, has taken on numerous meanings in different contexts and sociological analysis [Krishna and Shrader (1999)]. We define social capital as: social network attributes or metrics that have potential benefits towards economic development. Numerous studies have been conducted on the relationship between economic development in developing countries and social capital [Morris (1998), Knack and Keefer (1996)]. Research has shown evidence for communities with higher social capital to be associated with higher individual income levels [Narayan and Pritchett (1997)], and that social capital is beneficial for the economic development of developing countries. We discuss how certain social network attributes can directly benefit developing countries, especially in agrarian communities, and how this benefit could be used to quantify social capital.

Historically, data related to social capital and personal networks was collected via surveys where there is a limitation on the level of detail to which a network can be described, and where the research question is specific to individual entities in different environments and not about patterns of interaction within the individual's social group. Furthermore, the bulk of social capital and personal network analyses have been conducted on small scales and longitudinal network studies are rare [Lubbers et el (2010)]. With digital connectivity the collection of network data and social network analysis has been eased considerably and the evolution of large networks over time can be analyzed. At the same time, personal network analysis is still relevant as the influence of a social network on an individual actor, and vice versa, is largely determined by the actor's social environment at a distance of a small number of hops [Spreen and Zwaagstra (1994)]. The evolution of complete social networks has been well covered in the research community with Markov chain models and stochastic actor based models [Snijders (1996)].

While some work uses typical metrics taken from the social network analysis literature for forensic analysis of mobile phone data [Catanese (2012)], numerous kinds of use clustering strategies have also been proposed. Kurucz et al. introduce a spectral clustering algorithm to classify users according to the structural attributes of a mobile call graph [Kurucz (2009)], while Wu et al. propose a way to consider the evolution of clusters of users over time by correlations in their calling feature vectors [Wu (2009)]. Motahari et al. categorize users into sub-networks according calls to families and utilities over cell data records from two different cities [Motahari (2012)].

Our contributions in the paper are as follows. We consider a structural analysis of social networks embedded within cell phone datasets in order to quantify social capital. We then propose the use of clustering to complement social network analysis to allow for targeted policy implementation to specific groups. We perform two types of analysis: the first focuses on how time series clustering can provide additional insight into personal network analysis, as well as how to group similar users based on the evolution of their personal network attributes; the second analysis uses the temporal data in aggregate to cluster users. Our analysis finds that groups of users with the same evolution of social capital levels can be identified. Such information can be used by governments of developing countries to maintain optimal levels of connectivity and social capital amongst the citizens to act as a catalyst for strong economic and social development.

This analysis used the dataset provided by Orange in conjunction with the "Data for Development" (D4D) open data challenge. We consider communication sub-graphs of 5,000 random users over 10 two-week time periods from December 2011 to April 2012 (SET4TSV). We define 'alters' as users who the main user contacted during a given time period. Undirected communication graphs for each user were created by considering all communications between the user and the user's alters at up to two degrees of separation from the user. The actors in each personal network are given unique identifiers and remain unchanged throughout the twenty weeks.

This paper is structured as follows: Section 2 gives an overview of the metrics that were extracted from the dataset, Section 3 explains the methodology and algorithms used, Section 4 details key results from the clustering exercise, and Section 5 shows how the results could be for social development in the Ivory Coast. Finally, we give conclusions in Section 6 and discuss future research directions.

2. Network Attributes and Metrics

2.1 Longitudinal Personal Network Attributes

For the first step of the analysis, we extracted the number alters with one degree of separation in the user's personal network. We then extracted two other personal network attributes: density and betweeness centrality. We also extracted three time aggregated attributes: persistence of ties, cliques, and second degree influence. Other measures such as Burt's constraints [Burt (1996)] were considered, but were not included as these attributes were highly correlated to the basic measures of degree, density and betweenness [McCarty (2002)].

We define below the personal network attributes that are analyzed in the paper.

Density: A metric that measures how well connected are the nodes in a real network relative to the theoretical number of connections possible. It is calculated by taking the ratio of the number of edges and the number of possible edge.

$$Graph\ Density,\ d = \frac{Actual\ Connections}{Maximum\ Possible\ Connections} = \frac{2e}{n(n-1)}$$

Where e is the number of edges and n is the number of nodes in the graph.

Density was selected as a network attribute because of its close relationship with the concept of structural holes, and the analysis of structural holes is directly linked to the analysis of social capital [Burt (1996)]. In fact, Burt's measure of redundancy is a scaled version of ego network density [Borgatti (1997)]. Burt's theory of structural holes postulates that there is a link between good ideas and structural holes, and that brokerage across these holes provides social capital. Thus, we can infer that the lower the density of a two degree personal network, the higher the probability that the user can help close the structural holes and link two alters in the network.

This act of brokerage [Burt (2004)]through mobile devices is a source of social capital that is especially important in a developing country like Ivory Coast. Moreover, in a primarily agrarian and less technologically advanced

country like Ivory Coast [CIA World Factbook], information dissemination is extremely important to improve the efficiency and competitiveness of businesses [Jensen (2009)].

Betweenness Centrality: A measure of the user's centrality in a network calculated from the number of shortest paths from all alters to all others that pass through the user. The betweenness centrality of a node *v* is given by the expression:

$$g=s\neq v\neq t\sigma st(v)\sigma st$$

Where *v* is the user, σ_{st} is the total number of shortest paths from user s to user t, and $\sigma_{st}(v)$ is the number of those paths that pass through v. In this paper we use the normalized betweenness centrality, which limits the range to [0,1] without loss of precision. We use the normalized value so that this measure would be on the same scale as density.

$$normalgv=gv-\min(g)\max g-\min(g)$$

The betweenness centrality measure is used as a personal attribute for analysis because we believe that it is an important determinant of social capital - it quantifies how important the user is in the his network as an intermediary and his influence, which can be translated into social capital. Moreover, for a group of users in a particular region, high levels of betweenness centrality for all users would indicate strong community ties.

We observe a positive correlation (R^2=0.3839) between the number of alters and the betweenness centrality of the main actor because the larger the network, there is less probability that the first and second degree alters are connected, and thus most of the shortest paths would pass through the main alter. We observe a negative correlation (R^2=0.31) between number of alters and the density of a network, as expected, because the larger the number of first degree friends, the lower the probability those friends as well as the second degree friends are connected.

2.2 Time Aggregated Network Metrics

This section describes the new metrics we created to describe characteristics of personal networks across the 10 time periods.

Persistence of Ties: A measure of how persistent the ties are between a user and alters across the 10 periods.

$$Persistence\ of\ Ties=110i=110xiy$$

Where x_i is the number of alters the user calls in period *i*, and *y* is the total number of distinct alters the user has over all 10 periods. If a user calls different alters every period, this measure would be low; alternatively, if a user calls the same alters in every time period, this measure would be high. Thus, higher persistence value implies that the user has stronger connections with the user's first degree network, translating into higher social capital.

Cliques: We total the number of cliques of sizes 3 and larger across all 10 periods, which gives us an indication of the level of community within each of the user's personal network. The measure of cliques differs from persistence in the sense that it is a measure of how tightly knit the community (within two degrees from the user) is as opposed to analyzing only the first degree of alters. The trust generated from such tightly knit communities is a crucial source of social capital [Glaeser, Laibson, Scheinkman, Soutter (1999)].

Second Degree Influence: This ratio is a measure of the potential influence outside his own direct community a user can have within a two degree network.

$$Second\ Degree\ Influence=110i=110\theta i\beta i$$

Where θ_i is the total number of second degree friends in period *i*, and β_i is the total number of first degree friends in period *i*. For example, if in period 1 a user has 4 first degree friends, and each of the user's first degree friends has another 4 friends, the ratio that period will be 16/4=4. Taking the average over the 10 periods will give the second degree influence metric described.

Having a high level of second degree influence is a source of social capital as it demonstrates the user's sphere of potential influence in the user's community- The higher the number the more people the user could potentially spread information to through his first degree alters.

3. Clustering Methodology

3.1 Longitudinal Personal Network Clustering

As seen from the correlations in the previous section, personal network structural attributes are very sensitive to the number of alters in the network. To address this issue, we propose a two-step clustering methodology. First, we perform time-series clustering by the number of first degree alters each user interacted with over the ten periods. In the second step, we perform another clustering within each of the clusters based on the personal network attributes described in Section 2.1. Subsequently, we define a distance metric so that users in the same cluster exhibit similar density and betweenness over time, the distance metric used for each pair of users is the sum of Euclidean distance across all the periods, and is given by:

$$Distance=sqrt(i=110(d1,i-d2,i)2+(g1,i-g2,i)2)$$

Where $d_{1,i}$ and $d_{2,i}$ denotes the density measure as defined previously for a pair of users (denoted as user 1 and 2) for period i. $g_{1,i}$ and $g_{2,i}$ denotes the betweenness measure for a pair of users for period i.

Two step clustering will allow us to compare users based on density and betweenness because the users who are in the same cluster after the first step of clustering will have relatively similar number of alters in each of the 10 periods. If the first step of clustering was not performed, two users could potentially have the same evolution of betweenness centrality, but very different number of alters and therefore could not be grouped together.

The number of clusters for the first step was selected based on checking the silhouette coefficient for up to 30 clusters, as well as an examination of how meaningful each cluster was for a certain number of clusters [Tan, Steinbach, Kumar (2006)]. Figure 1 graphs the values of the silhouette coefficients against k, the number of clusters. We find that 8 clusters is the optimal number of clusters for our purposes, and the distribution of clusters was conducive for our second step of clustering.

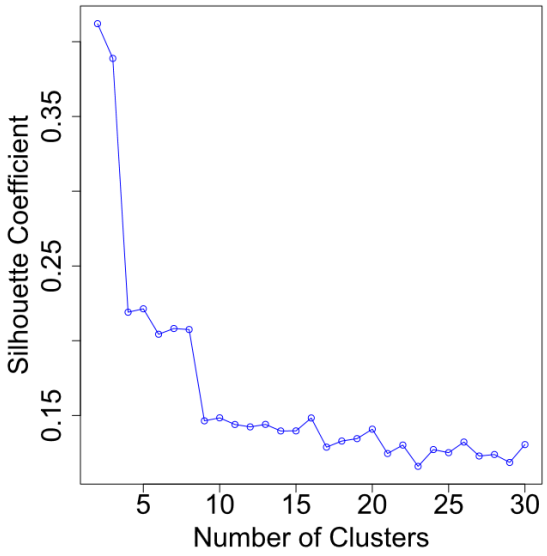


Figure 1. Silhouette coefficients for each k-clustering

Before clustering the network attributes, analysis showed that density and betweenness are relatively uncorrelated across the 10 time periods. For the second step of clustering, the optimal number of clusters was selected by maximizing the silhouette coefficients for number of clusters k from 2 to 10. Based on the silhouette coefficient method, the optimal number of sub-clusters for each of the 7 main clusters was between 2 and 3.

Two Step Clustering Psuedocode
Step 1: For each user for each period, calculate
K = number of alters
D = Density
G = Betweenness
Step 2: Perform time series k-means clustering over all 10 periods on K, the number of alters, into 8 clusters
Step 3: For each of the 8 clusters –
i) Initialize disparity matrix D[n_{users} x n_{users}]
ii) Compute distances for every user pair and populate disparity matrix
□□□□□□□□□=□□□□(□=110(□1, □−□2, □)2+(□1, □−□2, □)2)
Step 4: Cluster
Step 5: Output
i) Cluster Assignments
ii) Plots

3.2 Time Aggregated Network Metrics Clustering

The clustering exercise for the aggregated network attributes differ from the methodology in Section 3.1. First, we extracted the three attributes outlined in Section 2.3, we then mean normalized the attributes to remove skewness and to have them all on a similar scale. Lastly, we determined the optimal number of clusters using the silhouette coefficient maximization method to select 7 clusters and then ran k-means clustering.

Aggregated Attributes Clustering Psuedocode
Step 1: For each user, calculate
P = Persistence of ties
C = Average number of cliques
S = Second degree influence
Step 2: K-means clustering on all three metrics
Step 3: Output
iii) Cluster Assignments
iv) Plots

4. Results and Discussion

In this section, we discuss the two major results from the two clustering exercises.

4.1 Longitudinal Personal Network Clustering Results

After performing the first round of time series clustering on number of first degree alters, we observe that we can group users that have similar alter fluctuation trends - each cluster has its own distinctive average number of alters across each of the ten periods

Figure 2 shows the evolution of the means of all 8 clusters. In one example, cluster number 5 could be characterized by users who saw a rising trend in the number of first degree alters throughout the 10 periods, a group of users who became increasingly social. In the simplest context of social capital, the higher the number of first degree friends, the higher the increase in the size of their network and concurrently their social capital. Thus we can infer that users in this cluster saw an increase in social capital by virtue of interacting with more alters over time.

Figure 3 gives a more detailed look into cluster 5 showing that users indeed were clustered based on having similar evolution of the number of first degree alters.

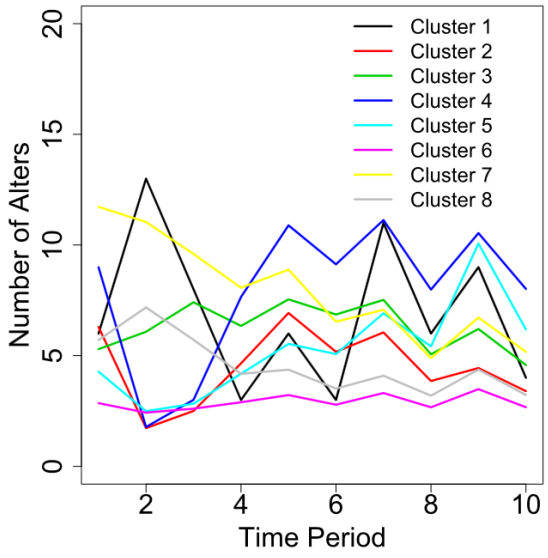


Figure 2. Evolution of Average Number of Alters for Each Cluster

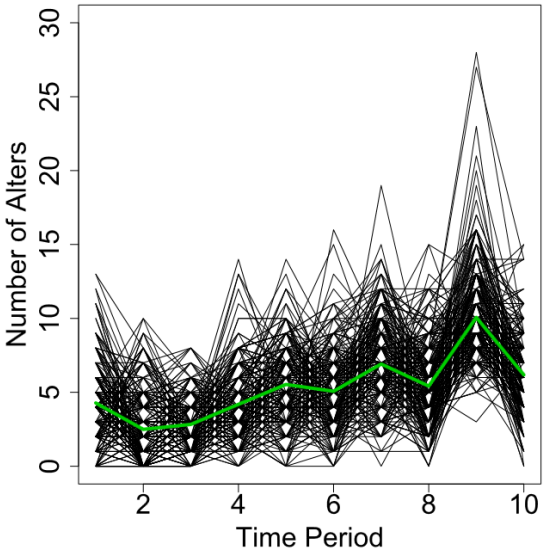


Figure 3. Plot of All the Users in Cluster 5

4.2 Longitudinal Personal Network Second Stage Clustering Results

At the second stage of clustering, we gain even more insight into the evolution of personal network attributes of different groups of users, painting a more comprehensive picture of the social capital of users in these groups. Continuing with our in-depth analysis of users in Cluster 5, we split the cluster into 3 clusters (number of clusters decided by silhouette coefficient) based on the density and betweenness attributes.

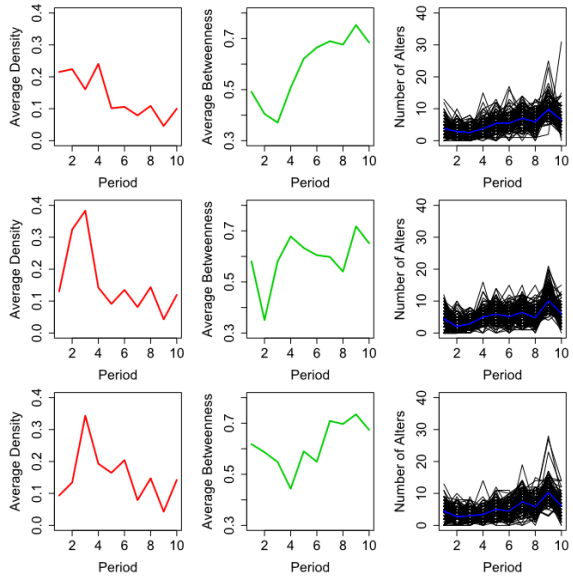


Figure 4. Second Stage Clustering of Users in Cluster 5

Each of the rows in Figure 4 shows the evolution of network attributes for each of the three sub-clusters in Cluster 5. The first, second and third columns corresponds to the evolution of density, betweenness and number of alters respectively. Though the curves for density and betweenness show some correlation with the evolution of number of alters, we can see distinct 'personalities' of each cluster: For example, users in sub-cluster 2 saw a sudden rise in density and fall in betweenness in period 2, before sharply reverting close to previous levels in period 3. A logical inference from this trend would be that of a sharp decrease in social capital during this period Government policies could be targeted at this group of users to boost their levels of social capital.

4.3 Aggregated Network Attributes Clustering

Figure 5 shows the cluster centers for each of the attributes for each cluster from the clustering exercise on the aggregated network metrics. From this figure we can easily interpret and classify users in each cluster: For example, users in Cluster 1

could be labeled as users who are in small tight knit units such as families due to the high persistence of ties but small 'size' of their network. Whereas users in Cluster 7 would have the opposite characteristics of calling different groups of people every period, but having a large network with many people interacting with each other. Users in Cluster 7 might contain users such as farmers or businessmen who depend on a wide but interconnected business network.

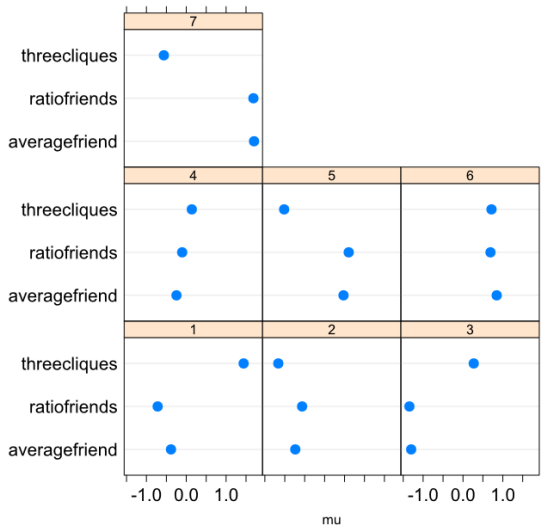


Figure 5. Centroid Positions for Each Cluster

5. Uses of Cluster Analysis for Social Development in Ivory Coast

We have described above the use of social network attributes as proxies for social capital and how longitudinal clustering can provide information to describe the evolution of social capital. We believe that the ease of collection of cellphone network data, coupled with demographic data on each user, can yield extremely powerful tools for the government. The evolution of network attributes can be used for relative comparisons of social capital levels between users. Ultimately, scores could be assigned based on each of the measures of network attributes, to quantify the level of social capital each user has in each cluster.

Considering time series clustering results combined with geographic and demographic data, there could potentially be two practical usages of this analysis: The first would be that of finding identifying groups of users who have falling social capital trends and reaching out to these users to improve their social capital levels, to spur economic and social development. Another usage of this analysis could be to analyze the impact of major events such as civil wars or sporting events on the evolution of social capital amongst the citizens of Ivory Coast. Such an analysis can be used to evaluate the effectiveness of state sponsored events or to decide when to intervene to prevent social capital levels from falling to levels that would hinder development.

6. Future Research and Conclusions

In this paper, we proposed two different methods of using clustering to extract information on social capital from cellphone usage data. One direction of further research could involve refining the network attributes to link them more closely to measures of social capital, as well as combining different data sets to better analyze the underlying reasons for certain evolutions of network attributes. In conclusion, powerful insights can be drawn from cellphone data sets to promote socio-economic growth in developing countries.

7. Bibliography

Borgatti, Stephen. "Structural Holes: Unpacking Burt's Redundancy Measures." *Analytic Tech Teaching Corner*. Connections, 1997. 15 Feb. 2013.

Burt, Ronald. "The Social Capital of Entrepreneurial Managers." *Financial Times* (1996) .

Burt, Ronald. "Structural Holes and Good Ideas." *The American Journal of Sociology* 110.2 (2004).

"Central Intelligence Agency." CIA. N.p., n.d. Web. 15 Feb. 2013.

Catanese, Salvatore and Ferrara, Emilio and Fiumara, Giacomo. "Forensic Analysis of Phone Call Networks" *Social Network Analysis and Mining* (2012).

Glaeser, Edward, David Laibson, Jose Scheinkman, and Christine Soutter. "What Is Social Capital? The Determinants of Trust and Trustworthiness." Quarterly Journal of Economics 65 (1999): 811-46.

Jensen, Robert. "Information, Efficiency and Welfare in Agricultural Markets." 27th International Association of Agricultural Economists Conference (2009).

Knack, Stephen, and Philip Keefer. "Does Social Capital Have an Economic Payoff? A Cross-Country Investigation." *The Quarterly Journal of Economics* 112.4 (1997): 1251-288.

Krishna, Anirudh, and Elizabeth Shrader. "Social Capital Assessment Tool." *Prepared for the Conference on Social Capital and Poverty Reduction* (1999).

Kurucz, Miklos and Benczur, Andras and Csalogany, Karoly and Lukacs, Laszlo. "Spectral Clustering in Telephone Call Graphs", Proc. 2007 SNA-KDD Workshop, 2007.

Lubbers, Miranda, Jose Molina, Jurgen Lerner, Ulrik Brandes, Javier Avila, and Christopher McCarty. "Longitudinal Analysis of Personal Networks. The Case of Argentinean Migrants in Spain." *Social Networks* 32 (2010): 91-104.

Motahari, Sara and Mengshoel, Ole and Reuther, Phyllis and Appala, Sandeep and Zoia, Luca and Shah, Jay. "The Impact of Social Affinity on Phone Calling Patterns: Categorizing Social Ties from Call Data Records". *Proc of 2012 SNA-KDD Workshop*, 2012.

McCarty, Christopher. "Structure in Personal Networks." *Structure in Personal Networks*. Journal of Social Structure, 15 Feb. 2013.

Morris, Matthew. "Social Capital and Poverty in India." *IDS Working Paper 61* (1998): n. pag. Print.

Narayan, Deepa, and Lant Pritchett. "Cents and Sociability: Household Income and Social Capital in Rural Tanzania." *Economic Development and Cultural Change* 47.4 (1999): 871-97.

Snijders, Tom A.B. "Stochastic Actor-Oriented Models for Network Change." *Journal of Mathematical Sociology* 21 (1996): 149-72.

Spreen, M., and R. Zwaagstra. "Personal Network Sampling, Outdegree Analysis and Multilevel Analysis: Introducing the Network Concept in Studies of Hidden Populations." *International Sociology* 9 (1994): 475-91.

Tan, Ping-Ning, Steinbach, Michael, and Kumar, Vipin. *"Introduction to Data Mining"*. AddisonWesley, 2006.

Wu, Bin and Ye, Qi and Yang, Shengqi and Wang, Bai. "Group CRM: A New Telecom CRM Framework from Social Network Perspective*". Proc. of CNIKM, 2009.* pp. 3-10.

Estimating Human Dynamics in Cote d'Ivoire Through D4D Call Detail Records

Ken Wakita[1,2] and Ryo Kawasaki[1]

[1]Department of Mathematical and Computing Science, Tokyo Institute of Technology,
{wakita,kawasak0}@is.titech.ac.jp
[2]JST/CREST

Abstract

The article outlines how we have identified the cities, city population, strength of social ties among cities, urban mobility in the largest city of Abidjan, and locations of residential and work areas. To deal with otherwise inadequate information contained in the D4D call detail records we introduced some daring assumption and invented approximation techniques. Although further investigation is and assessment is required, these techniques let the dataset talk about the societies found in the nation.

Keywords:

Population estimation, Human dynamics, Urban mobility, Spatio-temporal reasoning

1 Introduction

Cote d'Ivoire is a country in West coast of Africa whose population is estimated to be about 20 million. It is a multiethnic society that is formed from six major ethnic groups that differ linguistically, culturally, and religiously. Growth of immigrants during 70's and 80's added ethnic variety to the country and it is reported that the number of minor ethnic groups counts about 60. The country is respected for the higher rate of economic growth in 80's and 90's. However, it suffered from domestic conflict that caused Ivorian civil war, which brought a decade of political unstabileness.

The research was at first motivated by interest in the structure of the multi-linguistic, multiethnic culture and started from attempt to identify ethnic groups and weak ties that connect different ethnic groups, expecting social network clustering techniques give us insight into the social structure of the country. Later, we were more interested in the structure and relations of cities and urban mobility.

The article briefly reports our estimation and approximation techniques that we have used to compliment the otherwise inadequate D4D dataset. The techniques have discovered boundaries of cities, estimated population of the cities which are in line with the population statistics, urban mobility that is used to segment the largest city into residential and workplace areas.

2 Characterization of Major Cities

This section describes the techniques that we have used to estimate city boundaries, population of the area covered by each antenna, and social tie strength between cities. These techniques are based on inter-antenna geographical proximity as well as the amount of communication between antennas.

Firstly we have tried to estimate the strength of social ties ($w_{i,j}$) between coverage areas of two antennas using the following formula:

$$w_{i,j} = c_{i,j}/(p_i \times p_j)$$

where $c_{i,j}$ is the number of calls made between the antennas and p_i is the estimated population of the area. It reflects our assumption that the number of calls made from an antenna should be proportional to the population of its coverage area.

As a special case of the formula, where $i = j$, we have the following intra-antenna estimation:

$$w_i = \sqrt{w_{i,i}} = \sqrt{c_{i,i}}/p_i$$

In this formula, the new parameter w_i can be interpreted as digital fluency (or digital addictedness) of the people in the coverage area i.

We assumed that w_i's are equal nation-wide and therefore $w_j = w_i$ for every pairs of i and j.

Figure 1 presents the result of this analysis. To avoid visual clutter, we have trimmed edges with weaker tie with regard to $w_{i,j}$. We have also reduced the number of nodes by segmenting the whole map into 30×30 grids and replacing antennas in the same grid by their barycenter. Grids are depicted by small circles and strong ties between grids, namely pairs of grids whose accumulated $w_{i,j}$ are high, are connected by lines. In this figure, we can find many long-distance lines span out from grids in large cities such as Abidjan and Bouakè to grids all over the country. It should also be noted that neighboring grids

Figure 1: Strength of social ties between antennas

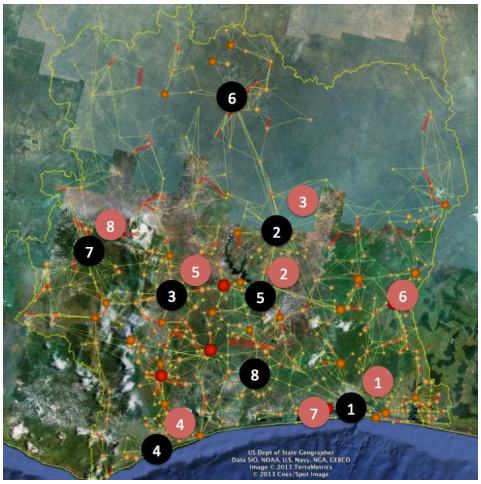


Figure 2: Identification of major cities. Eight major cities are in black and the estimated cities are in red.

are connected with short lines. When we further remove inter-grid edges by raising the threshold of tie-strength, long most of the distance edges disappears and we can observe paths formed short lines, revealing a highway net.

Secondly, we attempted to identify the city boundaries by means of geographical proximity between antennas. However, this technique, being too aggressively merging neighboring cities, produces poor result. Therefore, we then considered the strength of ties between antennas. The process of finding cities is essentially segmentation of antennas by calculating the closures of geographical proximity and the strength of social tie.

As we have already estimated the population in the coverage areas of antennas p_i, we can easily accumulate them and determine the population of the estimated city. Figure 2 presents the result. The black circles are locations of eight largest cities documented in [1] and the red circles denote the locations of eight largest estimated cities that are found by our analysis. The circles are labeled by numbers in the descending order of population. Our method seems to overestimate the population of Dabou (Red-7), which is ranked at 14th position in [1]. Korhogo in north (Black-6) which is 6th the largest city, was ranked 11th in our method. Nevertheless, our method gives surprisingly good estimate of the population despite difficulties such as the young generation being not found in the dataset, economic discrepancies, differences of education level, divergence of antenna density, and etc.

3 Urban Human Dynamics in Abidjan

The previous section discussed the inter-city structure of Cote d'Ivoire, whereas, this section examines human dynamics in the largest cities of Abidjan. Abidjan is a large city that has 3.6 million population. We would like to study the patterns of human mobility in Abidjan, and partition this million city into residential and work areas.

For this analysis, we used the second D4D dataset (Set 2), which gives fine grained trajectory records of the mobile users. We have done statistical analysis of the dataset to estimate the temporal population of the coverage area for each antenna, following the technique proposed by [2] for call detail records. In contrast to this previous work which took advantage of their fine-grained records including per-person records of radio beacons, the analysis of D4D Set2 is much more difficult because it is given for sampled mobile users and tracks only phone calls and text message exchanges and does not give the trajectory of the user when one is not communicating over the cellar network. To deal with this problem, we have marked the mobile user as *missing* when we can not find one's records in the trajectory database.

The upper image of Figure 3 is the illustration of the result of our modified analysis. Three antennas were taken from the city center (Red line) and northeast border (Light blue line), and in between of them (Gray line). These lines presents average estimated population over the period of the dataset (two weeks). It can be seen that during the night populations of these antennas are almost emptied. It is due to the fact, during the night very little communication takes place and our method fails to identify the location of the mobile users.

To overcome this problem, we normalized such estimated population by the maximum estimated population among antennas for each time and obtained the lower image of the figure. After this modification, we can see an interesting temporal patterns of human mobility among the three antenna areas. The light blue line has higher estimated population in the night and lower number during the daytime, whereas the red line presents the opposite character. It seems that the red and light blue antennas are located at

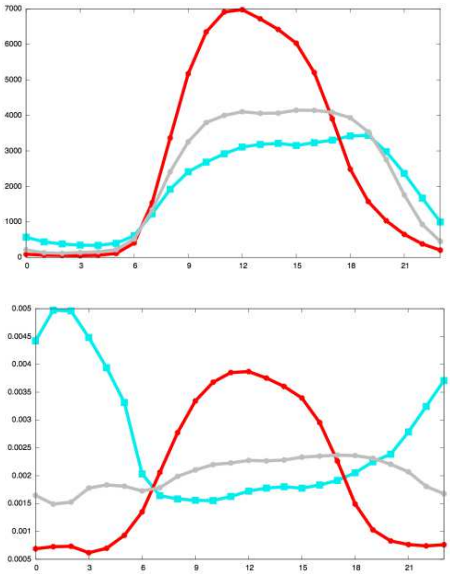


Figure 3: Temporal population of the estimated population (upper image) and normalized estimated population (lower image) of the coverage area of antennas #916 (Red), #1022 (Blue), and #772 (Gray).

a workplace and residential area. The gray antenna can be located in a area in between these two areas or it can be a mixture of workplace and residential area.

We have applied this area classification technique to all the antennas in Abidjan and obtained Figure 4. Our estimation suggests that workplaces are located in the city center and residential areas surround the center. We have also applied this technique for other cities identified in the previous section. Our classification technique found work places in Daloa (third the largest city) and residential areas in Yamoussoukro (fifth the largest), and three other smaller cities (see Figure 5). It would be possible to draw a conclusion that the economy of the cities in this country is largely dependent on agriculture and

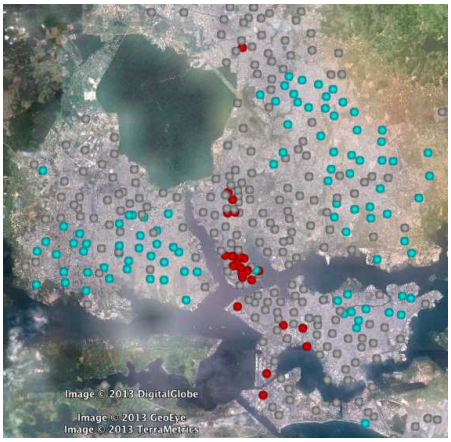


Figure 4: Classification of Abidjan areas into workplaces and residential areas

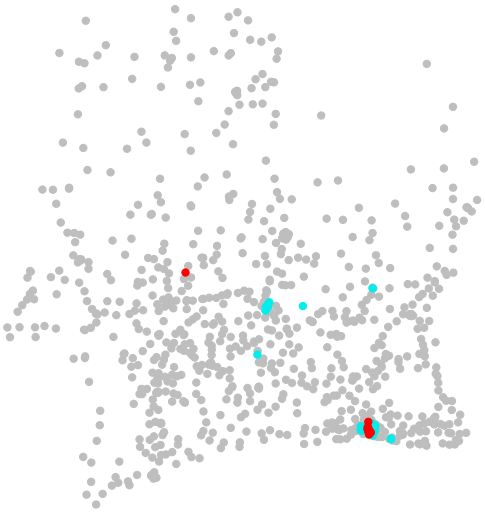


Figure 5: Nation-wide classification of antenna areas

industrialization is yet to come, except of Abidjan but it can be caused from premature of our approach.

4 Summary

In this work, we proposed spatiotemporal analysis technique for call detail records that identify the location and border of cities, estimate population of the coverage areas of antennas, and classified the coverage areas into residential and workplace areas. Our technique is to some extent useful for finding interesting human dynamics in Cote d'Ivoire.

This research result has been obtained from Set 1 and 2 only. We would like to continue working on other datasets included in the Set 3 and Set 4 and other third party datasets.

We are currently working on investigation of linguistic diversity and ethnic groups of Cote d'Ivoire.

Acknowledgment Our study and research activity were performed using mobile communication data made available by France Telecom and Orange Cote d'Ivoire within the D4D Challenge. This research was partly supported by the Japan Science and Technology Agency (JST), the Core Research of Evolutionary Science and Technology (CREST) research project.

References

[1] Canty and Associates LLC. Cote d'Ivoire - top 100+ cities by population, 2011.

[2] Masayuki Terada, Tomohiro Nagata, and Motonari Kobayashi. Population estimation techniques for the mobile spatio statistics (in Japanese). *NTT DOCOMO Technical Journal*, 20(3):11–16, 2012.

"Calling Abidjan" – Improving Population Estimations with Mobile Communication Data (IPEMCODA)

Harald Sterly[1], Benjamin Hennig[2], Kouassi Dongo[3]

February 16, 2013

[1]: Department of Geography, University of Cologne (h.sterly@uni-koeln.de)

[2]: Department of Geography, Department of Geography, University of Sheffield (B.Hennig@Sheffield.ac.uk)

[3]: Université de Cocody-Abidjan / Centre Suisse de Recherches Scientifiques en Côte d'Ivoire, CSRS (kouassi.dongo@csrs.ci)

Abstract

Analyzing call data records of Ivoirian mobile service provider Orange Telecom, we assess the potential of mobile phone data for the improvement of population estimations. Especially in countries with a lack of reliable and spatially disaggregate census data, the combination of existing population data with satellite imagery, land use data and population modelling yields good results in rural and periurban areas. In larger agglomerations these methods have limitations. It can be shown that, while less advantageous in rural areas, the use of mobile call data can lead to an improved understanding of the population distribution in larger urban centres.

1 Introduction

There is a lot of evidence that the 21st century will be both an urban and as well an urbanizing century. The United Nations expect the cities in the low and middle income countries to accommodate 89% of the global population increase between 2011 and 2050 (UNDESA 2012). In Côte d'Ivoire this value would even be higher (92%), where the urbanization rate has been declining over the past years but with over 3,35% (estimation for 2005-2010) still is only slightly lower than the Sub-Saharan average (3,67%) and much higher than world average (2,14% for the same period of time, UNDESA 2012).

This means a huge challenge for urban governance, planning and administration, especially as urban poverty rates in Côte d'Ivoire remain high, with an even growing number of people living in low-income housing conditions (57% in 2009, UNSTATS 2013) increasing income inequality in urban areas (UN HABITAT 2010b). However, much of this data that is or could be guiding policy making regarding more sustainable and equitable urban development, is based on sparse national population accounts, the UN estimations for Côte

d'Ivoire for example being extrapolated on the basis of the last national census data from 1998 and official national estimates dating to 2008. Recent methodical improvements in population data estimations like the gridded population data provided by AfriPop (Afripop 2012) or LandScan contribute to substantial information improvement in less populated areas, however their information content for densely populated areas such as Abidjan or Yamoussoukro is limited as well. Approaches as the participatory slum upgrading programme (PSUP) launched in Ivorian cities are countered by the limited knowledge about population distribution and dynamics, which is of key importance for planning and targeting service and infrastructure improvement with limited resources.

2 Objectives and Methodology

We are therefore exploring a novel approach for improving the knowledge on intra urban population distribution and dynamics through the analysis of mobile call data. Each mobile phone communication activity is transmitted and received through GSM base stations / antennas, thus each call or message can be attributed to the respective base stations involved in sending and receiving. Through a relatively simple analysis of a large sample of call data records (CDR) we attribute each caller to a base station considered to represent his or her place of residence. The population distribution is then extrapolated from the callers at each base station referring to the total number of callers considered in the analysis and the total population on a national level.

This approach is based on the following assumptions. First, as we don't have any detailed information on spatial and social differences in the subscriber rates and usage of the mobile service provider's service, we have to assume a uniformity of both variables. Second, we assume the estimation of the total population on national level as more reliable that the estimations for the dynamic and densely populated urban areas. Third, the approach relies on the randomness of the sample of CDRs drawn from the total number of CDRs. The first two assumptions pose serious limitations to the approach, however as we regard the procedure as a proof of concept, we take this into account and will discuss them later.

Data

We were granted access to a dataset of call detail records of the Ivoirian mobile service provider Orange Telecom. Access was granted in the context of a call for projects, with the objective to explore the potential of mobile call data analysis for the field of socio-economic development. CRD's of 2.5 billion calls and text messages of approx. five million Orange users were collected over a time period of 150 days between 01.12.2011 and 28.04.2012. The caller data have been preprocessed by Orange to ease data handling and to ensure privacy protection of the callers and data consistency (Blondel et al. 2012).

The data provided consists of four different data sets, out of which we used two for our analysis. The first set ('high spatial resolution') covers the calling activities of 50.000 randomly sampled users, consisting of 10 subsets each covering 14 days. For every call the following information is provided: Caller ID, Date/Time, Basestation ID. For privacy reasons, in each of the 10 subsets the caller IDs are newly assigned. The second data set ('long term data') covers the calling activity of 500.000 randomly sampled users, also consisting of 10 subsets of 14 days. For privacy reasons in this data set the data is spatially aggregated on the level of the 255 Ivorian Subprefectures, thus the information provided for each call is: Caller ID, Date/Time, Subprefecture ID. All steps of data analysis were

done using the open source statistical software R (2.15.2), on a standard desktop computer (Apple Mac Mini). Visualization was done in Quantum GIS and Adobe Illustrator.

For an assessment of the quality of population estimation, a raster dataset of population in Côte d'Ivoire was used (AfriPop 2012), which was

Analysis of the Data

In a first step the long term data set was used to make an approximation of the population on the level of the administrative unit of the 255 subprefectures. The call data of all 10 subsets were combined and aggregated by subprefectures, thereby assigning the 500.000 callers to those subprefectures from where they placed the highest number of calls. The population of each subprefecture was then extrapolated assuming a population of 21 million for the national level. We used this figure as an approximate value, as population estimations for Côte d'Ivoire range considerably (e.g. 20.11 million for 2011, UNDESA 2012; 21.5 million for 2011, CIA World Fact Book).

For the calculation of higher resolution population figures the 'high resolution data set' was used. In order to assign callers to places of residence, the daily mobility of mobile phone users had to be taken into account. From an analysis of the daily calling activity and locations of a small sample of callers the time of "being home" for most of the population (especially in urban areas) was derived as from 7pm until 5am. The base station from which the respective callers placed the highest number of calls during the overall time interval of 14 days was considered as associated with his or her place of residence. After aggregation of the number of callers for each base station the population was extrapolated for each of the 10 data subsets. A weighted average for these population figures was calculated as in each CDR subset the entry for the base station was missing for a significant number of calls. For spatial analysis and visualization, Voronoï polygons were calculated for the base stations. For a number of urban areas, these polygons were clipped by the limits of the urban areas, derived from manual digitalisation of Landsat ETM+ data from 2007.

3 Results

Figure 1 shows the results of the population estimation based on the analysis of the long term data set, compared with data derived directly from the AfriPop dataset, aggregated on subprefecture level. Partly the results of the CDR analysis and the raster data from AfriPop are consistent. They reproduce a number of similar patterns, namely the higher population densities in the Southeast and the Centre, the lower densities in the Northeast and Northwest, and particularly the urban areas of Abidjan, Yamoussoukro, Bouke, Man and others. However, there are also considerable differences, especially in the Western parts of the country and, most notable in the Southwest. Fig. 1d shows the relation of population densities ($\frac{Popdens_{CRD}-Popdens_{AfriPop}}{Popden_{AfriPop}}$), indicating that generally the the population estimation with CDR analysis leads to higher population figures in the urban areas and lower ones in rural areas/areas with little population density. The reasons for this are probably differentiated mobile phone subscription rates in urban and rural areas, but might also be attributable to inaccuracies in the underlying census data and population models of the AfriPop dataset. There is one other notable difference: in the Southwest, in the region of Bas-Sassandra the figures derived from mobile usage is significantly higher. This can probably be explained by the positive economic development of the region (to a large extent related to the port of San-Pédro), resulting in both population growth as well as higher mobile subscription rates.

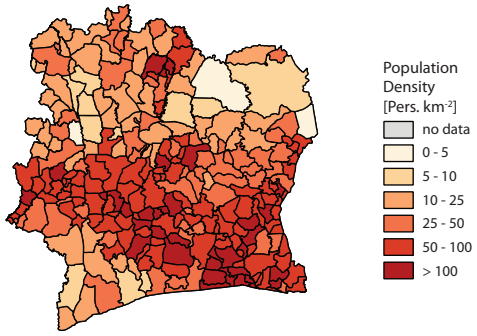


Figure 1a: Population density derived from AfriPop datset, on subprefecture level

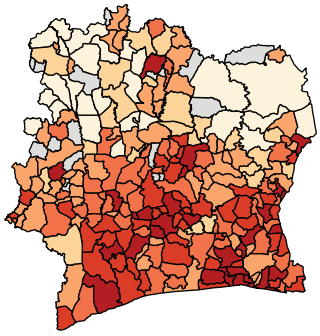


Figure 1b: Population density derived from CDR analysis, on subprefecture level

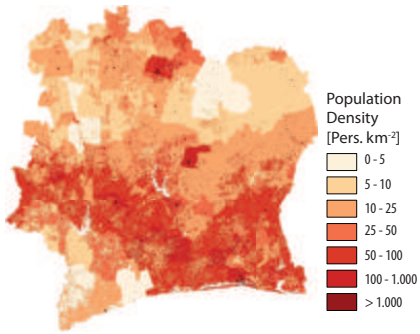


Figure 1c: Population density (AfriPop dataset), raster image with cellsize approx. 90m

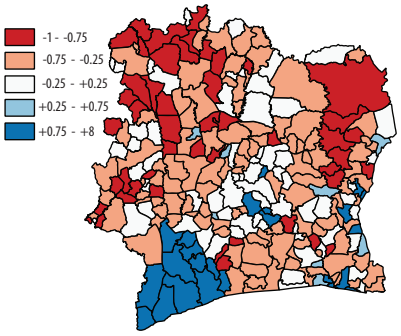


Figure 1d: Population density differences between data derived from CDR and AfriPop dataset, normalized by the population density of the AfriPop datset

Figure 2 shows the results of the analysis of the 'high spatial resolution data' for the urban agglomeration of Abidjan. The gridded population data (at least not in this version of the AfriPop dataset) does not reflect the differentials in intra-urban population distribution (Fig. 2a). The high number of base stations in the urban area leads to a very differentiated picture of the population within the agglomeration. The densely populated areas of Yopougon (in the West), Adjamé and the Plateau (Centre), Treichville and Marcory (South-Centre) are clearly represented in the data (Fig. 2b).

Summarizing the CDR derived population estimates from the analysis of the high resolution dataset for the agglomeration area of Abidjan results in a population figure of 7.25 million, which is 1.69 times the estimates of UNDESA for 2011 (4.3 million). Taking the figures derived from the long term dataset still gives a population of around 5.4 million for the subprefectures of Abidjan, Bingerville, Songon and Anyama. The large difference in the results of the CDR analysis are probably attributed to the high number of callers not attributable to any base station in the high spatial resolution dataset, leading to a higher weight attributted to the data included in the calculation.

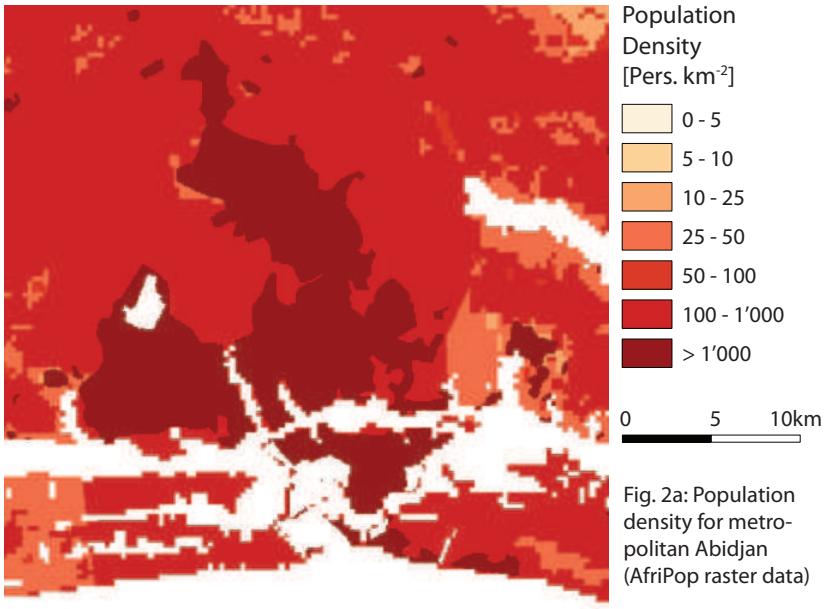


Fig. 2a: Population density for metropolitan Abidjan (AfriPop raster data)

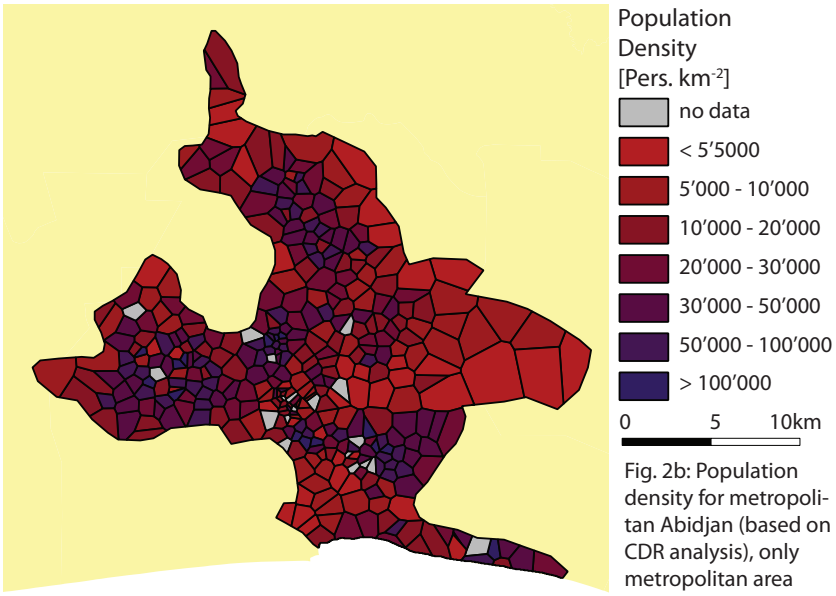


Fig. 2b: Population density for metropolitan Abidjan (based on CDR analysis), only metropolitan area

4 Discussion and Conclusion

As noted before, we took as one basic assumptions the homogeneity of subscription and usage rates of Orange. The results of the analysis fundamentally challenges this assumption, even if we presume a limited accuracy of the raster image from AfriPop as well. Operationalizing this approach for population estimation would require substantial work here and the systematic triangulation with other existing data sources.

Our second central assumption relates to the existing population estimations for the national level, which we used as a basis for our analysis. Although the models used for calculating these figures are partially fuelled with data from recent surveys, there is still some level of uncertainty involved. A better approach would require the availability of information on spatially differentiated mobile service subcription.

Existing approaches of advanced population modelling using satellite imagery, land use data and sociocultural information (Tatem et al. 2007, Linard et al. 2010) yield good results where existing data is of sufficient quality and are particularly useful in delineating urban areas. However, especially in assessing intra-urban population differences, they are still at their limits. Due to the increasing use of mobile phones in most parts of the world, the use of mobile caller data seems to be an opportunity for the improvement of population estimations.

As next steps we are planning to develop correction factors for a) the weight of data from missing base stations, b) different spatial (like e.g. Calabrese et al. 2011) and c) social groups, based on different calling patterns related to socio-economic status (Soto et al. 2011). This would imply in the case of b) knowledge of regional differences in subscription rates, which would have to come from the mobile service provider, and in the case of c) the use of additional socioeconomic data which would probably have to be specifically sampled. Addionally we intend to derive information about urban population dynamics and migration from the data.

References

[1] AfriPop (2012): High resolution, contemporary data on human population. Access at 13.02.2013 on http://www.clas.ufl.edu/users/atatem/index_files/AfriPop.htm.

[2] Blondel, V., Esch, M., Chan, C., Clerot, F., Deville, P., Huens, E., Morlot, F., Smoreda, Z., Ziemlicki, C. (2012): DATA FOR DEVELOPMENT: THE D4D CHALLENGE ON MOBILE PHONE DATA. 29 Sep 2012. Accessed online on 09.12.2012 at arXiv:1210.0137 [cs.CY].

[3] Calabrese, F., Dahlem, D., Gerber, A., Paul, D., Chen, X., Rowland, J., Rath, C., Ratti, C. (2011): The Connected States of America: Quantifying Social Radii of Influence. In: Proceedings of SocialCom/PASSAT. 2011, pp 223-230.

[4] Linard, C., Gilbert, M. and Tatem, A.J. (2011): Assessing the use of global land cover data for guiding large area population distribution modelling. In: Geojournal 76/5, pp 525-538, doi: 10.1007/s10708-010-9364-8.

[5] Soto, V., Frias-Martinez, V., Virseda, J., Frias-Martinez, E. (2011): Prediction of socioeconomic levels using cell phone records. In: UMAP'11 Proceedings of the 19th international conference on User modeling, adaption, and personalization. Heidelberg, pp 377-388.

[6] Tatem, A., Noor, A., von Hagen, C., Di Gregorio, A., Hay, S. I. (2007): High Resolution Population Maps for Low Income Nations: Combining Land Cover and Census in East Africa. In: PLoS ONE December 2007, Issue 12.

[7] UN HABITAT (2010a): The State of African Cities 2010. Governance, Inequality and Urban Land Markets. Nairobi. Report, accessed on 12.02.2013 at http://www.unhabitat.org/pmss/getElectronicVersion.aspx?nr=3034&alt=1

[8] UN HABITAT (2010b): The State of the World's Cities 2010/2011. Bridging The Urban Divide. Accessed online on 12.02.2013 on http://www.unhabitat.org/pmss/getElectronicVersion.aspx?nr=2917&alt=1

[9] UNDESA (2012): World Urbanization Prospects, the 2011 Revision. Data Tables, accessed on 12.02.2013 at http://esa.un.org/unpd/wup/index.html

[10] UNSTATS (2013): Millennium Development Goals Indicators. Online Database accessed on 12.02.2013 at http://data.un.org/Data.aspx?q=slum&d=MDG&f=seriesRowID:710; countryID:384&c=2,3,4&s=countryEnglishName:asc,year:desc&v=1

Understanding ethnical interactions on Ivory Coast

A. J. Morales [2], W. Creixell [1,2,3], J. Borondo [2], J. C. Losada [2], and R. M. Benito [2]
[1] *Departamento de Electrónica. Universidad Técnica Federico Santa María. Chile.*
[2] *Grupo de Sistemas Complejos, Universidad Politécnica de Madrid, ETSI Agrónomos, 28040, Madrid, Spain and*
[3] *Visiting researcher at Center for Spatial Information Science (CSIS), University of Tokyo, Japan.*

Towards the consolidation of peace and national development, the Ivorian society must overcome the lack of cohesion reported by several cooperation agents. In this sense, the present work provides insights on the regional interactions, so efforts may be planned and deployed with more intelligence. We characterize the communication patterns from a social perspective, in order to understand the factors that influence its emergence. We found that in a subregional scale, the ethnical identity plays an important role in the communication patterns, while at the interregional scale, other factors arise like economical interests and available infrastructure.

I. INTRODUCTION

The EU cooperation program for 2008-2013 on Ivory Coast [1], establishes the promotion of social cohesion as indispensable for the strengthening of peace, governance and national stability. In order to promote this cohesion, several improvements on the social services and infrastructures were deployed during this period, in which violence erupted once again among the social groups of the country. Therefore, in order for future actions to be planned and deployed with more intelligence, we aim to provide useful information on the way that the diverse society of Ivory Coast is structured, according to their interactions.

In particular, we intend to characterize and quantify interactions among the geographical and ethnical regions on Ivory Coast to understand the factors that influence the emergent social structure. We accomplish our scope by means of the tools provided by the complex networks theory [2]. For this matter, we used data provided by the *D4D Challenge on Mobile Phone Data*, as well as metadata like ethno-linguistic families, population or infrastructures, to get a better characterization of the analyzed regions.

In summary, on a local and regional scale, the Ivorian communication patterns seem highly influenced by social facts like the ethno-linguistic identity of the habitants, who tend to relate mostly to their own locality and others with similar cultural features. However, between the regions on a wider scale, the underlying infrastructure and the economical interests, seem to play a major influence in the social interacting structure, that end up rupturing the country into two large regions, located at the east and west side of the map.

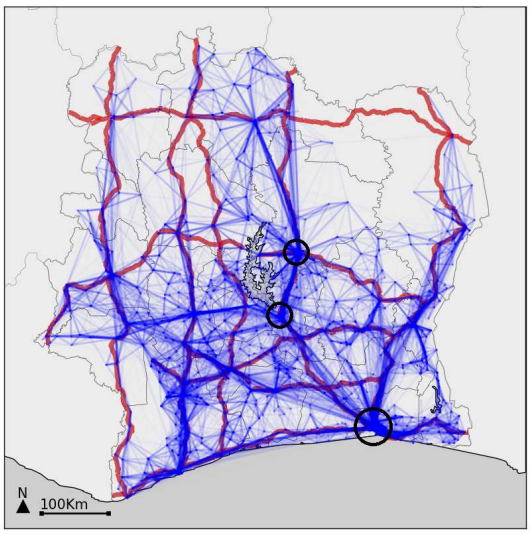


Figure 1. Trajectories network. Snapshot of the trajectories network at the end of the observation period. The blue lines represent the edges of the network and their intensity is proportional to the edge weight. The red lines represent the main roads of the country. Black circles indicate the location of the major cities.

II. DATA SET

The present work is based on the data provided by the *D4D Challenge on Mobile Phone Data*. These datasets were preprocessed in order to construct the complex networks necessary to represent the recorded interactions. We specifically used the antenna-to-antenna traffic dataset (*SET1*), as well as the individual trajectories for 50,000 customers dataset (*SET2*). At both cases, we aggregated the data available from all the observation period. Besides we used the geographical location of the antennas, to find spatial correlation on the Ivory Coast map.

Moreover, in the spirit of the project and to get a better understanding of the analyzed regions, we also used other sources of data available on Internet, like the language map from Lewis, M. Paul [3], and other spatial information files like the main roads location, taken from the African Development Bank [4].

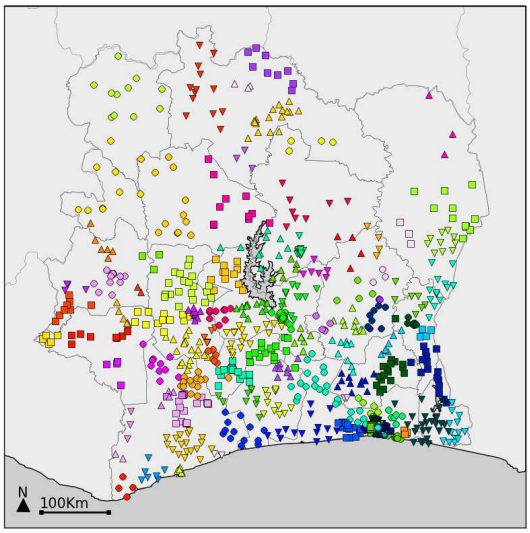


Figure 2. Trajectories network communities. The antennas have been represented by color and shape according to the community detection algorithm.

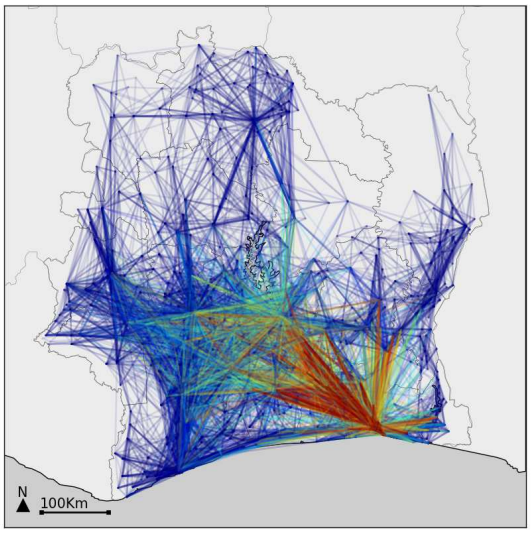


Figure 3. Trajectories network centrality. The edges have been colored according to the closeness centrality mean value of the connected nodes. The red regions indicate higher closeness-centrality, the yellow and pale blue regions indicate medium centrality, and the dark blue regions indicate lower closeness-centrality.

III. ETHINCAL INTERACTIONS

To begin to understand the Ivorian Society geographical structure and interactions, we first analyzed the mobility patterns of the country. For this matter, we built the Trajectories Network from the aggregation of all the individual trajectories found in the dataset *SET2*. An individual trajectory is defined as the sequential set of antennas that served a given user through time, and provides information on the mobility path of the user. In this network, antennas represent nodes, and an edge is created between two antennas, i and j, if a user makes two consecutive calls: first from the antenna i and after from the antenna j. The edges are directed, from i to j, and weighted, according to the number of times that all users performed the trajectory from i to j. The resulting network has 1215 nodes and 187102 edges. A visualization about the dynamical growth of this graph during an arbitrary day is presented on an animation in the supplementary video *VS1* (http://www.gsc.upm.es/materiales/videos/).

A snapshot of the trajectories network is presented in Fig. 1, where the graph has been plotted it in the map of Ivory Coast, using the location coordinates of the antennas. The edges are colored in blue, and the intensity of the edge is proportional to its weight. This network unveils the collective mobility pattern nationwide. The main city (largest black circle), is easily located in the southeast coast where a large amount of edges are concentrated, as well as other major cities in the center of the country (smaller black circles). In the figure we have also superimposed in color red the main roads of the country, to show the impact that the infrastructure has on the mobility patterns, since a large part of the trajectories keep correspondence to them. This network provides information about the flux of people along the roads. Some roads seem to be more frequently used, like the ones linking the north with the south of the country, in contrast to others less frequently used, like the transversal road up in the north. In addition, there are other edges, between the roads, that display the populated areas of the country, and possible roads of lower order. These populated areas are identified after using the modularity optimization method [5]. We found 100 communities, as can be seen in Figure 2, where each antenna has been plotted according to the community it belongs to. It may be noticed that communities comprehend a limited territorial area, which supports that communities emerged from the people displacements around villages and urban areas.

Therefore the patterns found in the trajectories network are a reflection of the country infrastructure and demography, but also of other social factors, like the economical activity, due to underlying reasons that justify the efforts for such displacements. To infer the significance of country regions in the economical activities we studied the closeness-centrality of the antennas in the

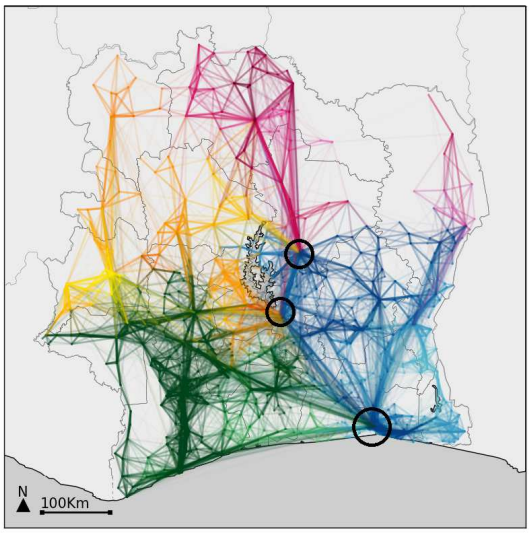


Figure 4. Trajectories network. The edges have been colored according to the linguistic group to which the most connected antenna at each community belongs to. There are four major linguistic families represented in yellow (northwest), purple (norteast), green (southwest) and blue (southeast). Black circles indicate the location of the major cities.

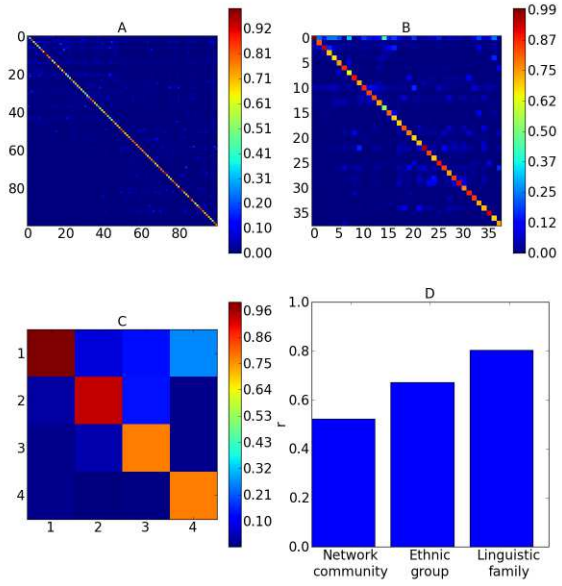


Figure 5. Row normalized adjacency matrices by (A) Network community, (B) Ethnic group and (C) Linguistic family. (D) Assortativity by characteristics coefficient on local scale (community), subregional scale (ethnic group) and regional scale (linguistic family).

trajectories network. This network property is inversely proportional to the average distance, in terms of network connections, that a node presents respect to the rest of the network, providing insights of how central or peripheral a node may be. In Figure 3, we present the trajectories network, coloring the edges according to the mean value of closeness-centrality of the antennas it connect. The red links connect antennas with a high closeness-centrality value, the yellow and pale blue links connect antennas with medium centrality value, and the dark blue links connect peripheral antennas with low closeness-centrality. It can be noticed, that the most central area of the graph (red) corresponds to the main city and the regions it adjoins, and as we get further from it, antennas get more peripheral. However, a medium centrality is noticed around the other major cities in the center of the country. This is in correspondence to the EU cooperation program [1], which identifies the north and the west of the country as the less developed areas.

To further understand the social composition of this graph, we have also taken into account the ethnical and linguistic component of each community. As communities are identified with given localities on the country, we used the ethnical map from [3], to identify the ethnical group that compounds them. To do this, we first identified the antennas with the highest degree at each community and them mapped them to the geographically closest ethnical group, using the location coordinates. The resulting ethnical assignment of the rest of antennas in the community, corresponded quite well with the four large linguistic families to which the ethnical groups belong (see Appendix A). This is shown in Fig. 4, where the trajectories network is presented by coloring the edges according to the linguistic family the network community belongs to. It can be seen that as the most densely connected areas, like the capital city or the cities in the center of the country (black circles), concentrate links from different linguistic areas, while other regions mainly present trajectories from their own linguistic family. Besides it shows that the trajectories in the northern families, occur more frequently with the southern families, than between them.

Although, the trajectories network provides some information to characterize the social structure on Ivory Coast, it does not provide a clear vision of the interactions taking place among the groups found. To further understand it, we have constructed a second network, taking also into account the mobile calls information provided in the dataset *SET1*. In this network, the nodes are the 100 communities found in the trajectories network from the dataset *SET2*, and the edges correspond to the number of calls made from one community to the other, extracted from the dataset *SET1*. The edge direction goes from the emitter community to the receiver community and the weight is equal to the number of oc-

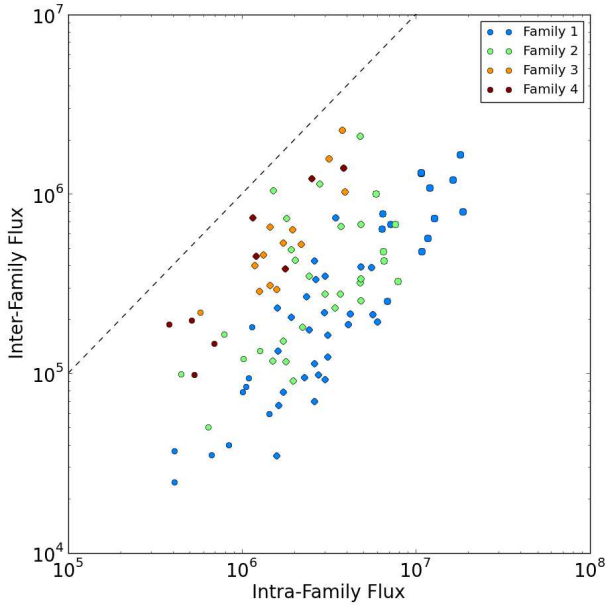


Figure 6. Intra linguistic family flux (calls directed to an antenna in the same linguistic family as the emitter antenna) versus Inter linguistic family flux (calls directed to an antenna in a different linguistic family than the emitter antenna).

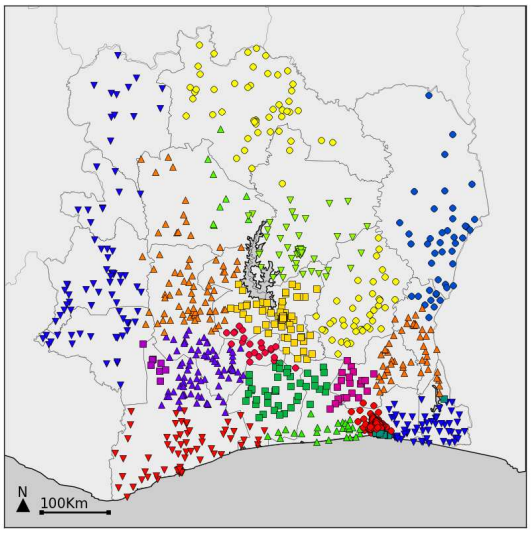


Figure 7. Calls network. The antennas have been represented by color and shape according to the community detection algorithm.

currences found in the data sets. This network represents a second level of interaction among the antennas.

To get a clearer view of the way that these trajectory communities called each other, in Fig. 5A we present the weighted adjacency matrix of the constructed network, normalized by row for a better visual comprehension. This normalization avoids the masking effect that the larger groups have over the small ones, and provides relative information about the destination of the outgoing calls and origin of the incoming ones. It can be noticed, that the diagonal is quite strong, indicating that most of outgoing calls of a community remains in the same community. This effect is also noticeable, when we group these communities using the ethnical metadata, like the ethnical group in Fig. 5B, and the linguistic family in Fig. 5C.

In fact, the preference of people to communicate to similar ones gets stronger as we increase the scale of the network, in terms of the same community, ethnical group or linguistic families. This is shown in Fig. 5D, where we plot the resulting assortative coefficient by characteristic [6] of each matrix. It can be noticed that it increases from 0.5 up to 0.8, when evaluating from network community to language family, which means that at the lower level, the people of a community call more often people in other communities, yet these other mostly belong to the same family language.

Moreover, not all families behave the same way. The south linguistic families (number 1 and 2 in Fig. 5C) present a larger proportion of calls directed to the same linguistic family, in comparison to the north linguistic families (number 3 and 4 in Fig. 5C), since the diagonal values of the north families are redder in the figure than the south families, which are more yellowy. This difference between the northern and southern regions are noticed in Fig. 6, where we represent the communities studied according to the intra-family traffic (calls directed to the same linguistic family) versus the inter-family traffic (calls directed to a different linguistic family). In the figure the symbols represent the communities found in the trajectory network and the color corresponds to the linguistic family they belong to. The further the community is located below the dashed line of slope 1, the higher is the intra-family traffic in comparison to the inter-family traffic. We see that the northern families are more keen to call different families, than the southern ones.

However, the communication between the northern families to the southern families are found to be quite selective. In Fig. 5C, we see that the family 1 (southeast) has a stronger connection with family 4 (northeast), than with family 3 (northwest), which resulted to have a stronger connection to the family 4 (southwest). This means, that each northern family tends to communicate more with their adjoin southern family, resulting in a larger density of calls from the northeast to the southeast regions, as well as from the northwest to the southwest regions. This observation is in good agreement with

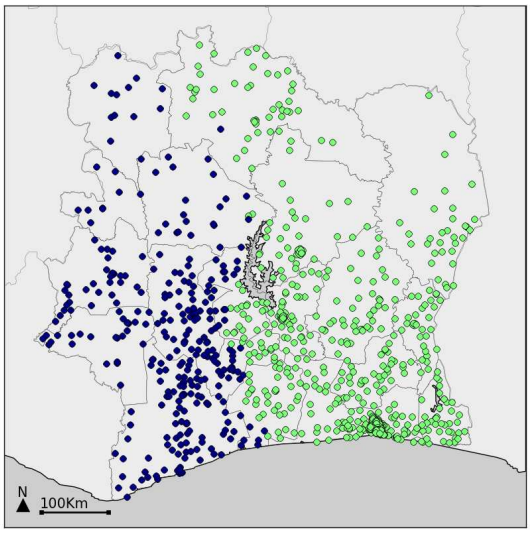


Figure 8. Antennas classification results by the way the calls network communities are related, using k-means clustering classifier.

the mobility patterns shown in Fig. 1, where the vertical roads seam to have a higher significance than the horizontal ones, and in Fig. 4, where we showed that the mobility of the northern families to the south are stronger with the adjoin regions.

To further understand this rupture on the communication patterns between the east and west side of the country, we built a third network, to analyze the calling behavior at the microscale, extracting only information from the dataset *SET1*. The nodes in this network also represent antennas, and an edge is created from the antenna i to the antenna j, when a user that is being served by the antenna i makes a call to another user who is served by the antenna j. It is a directed and weighted network, where the weight of the edges represents the number of calls made from i to j. This calls network, presented 19 communities, according to the modularity optimization algorithm [5], distributed along the geography of Ivory Coast as may be seen in Fig. 7. In the supplementary video *VS2* (http://www.gsc.upm.es/materiales/videos/), we present an animation with the growth of this network and a visualization of the influence that each of the 19 communities have in the network.

To capture how communities influence the rest of the network, we analyzed the density of calls directed to the given communities from the rest of antennas. To quantify such preference, we have classified these communities using a k-means clustering algorithm, according to the density of calls to the rest of communities. The results are presented in Fig. 8, where we have plotted the antennas with different colors, according to the classifier results. A clear division between the east side and west side of the country is appreciated, which corroborates the influence that the underlying infrastructure and other human factors, like cultural bonds or economical interest, may have in the structure of the social interactions, within a country.

IV. CONCLUSIONS

By means of the analysis of the resulting patterns of the trajectories and calling network, we have characterized the interactions and resulting structure of the diverse geographical and social areas of Ivory Coast.

From a social and ethnic perspective, we found that the linguistic identity plays a fundamental role in the communication patterns of this country. The Ivorian people, seem to preferentially communicate to those that belong to the same local community, but more drastically to those of the same linguistic family. Yet this preferences is not equal to all linguistic families, since the peripheral regions of the north, seem to communicate with their adjoin southern regions significantly, which ruptures the map into two interacting regions located at the east and west side of the country. This division of the country patterns, seems to be highly influenced by the underlying infrastructure and economical factors.

On this basis, we conclude that the geographical and social factors, whether cultural or economical, determine the structural features of the social interchange. In the sense that on a local and subregional scale, the ethno-linguistic factor determines the interaction patterns, while on a wider scale, the available infrastructure and economic facts play a major influence in the social dynamics.

ACKNOWLEDGMENTS

This research was supported by the Ministry of Economy and Competitiveness-Spain under Grand No. MTM2012-39101.

Special thanks to the "Beca Iberoamérica Santander Universidades 2012" for its travel financing.

Appendix A: Ethno-linguistic groups in Ivory Coast

Ivory Coast presents a complex society compound by more than 60 different ethnic groups. These ethnic groups are classified into four large linguistic families, as can be seen in Fig. 9, where we present the ethnic map realized by Lewis, M. Paul [3]. Although each ethnic group has its own language, French is the official language and it is broadly spoken along the country.

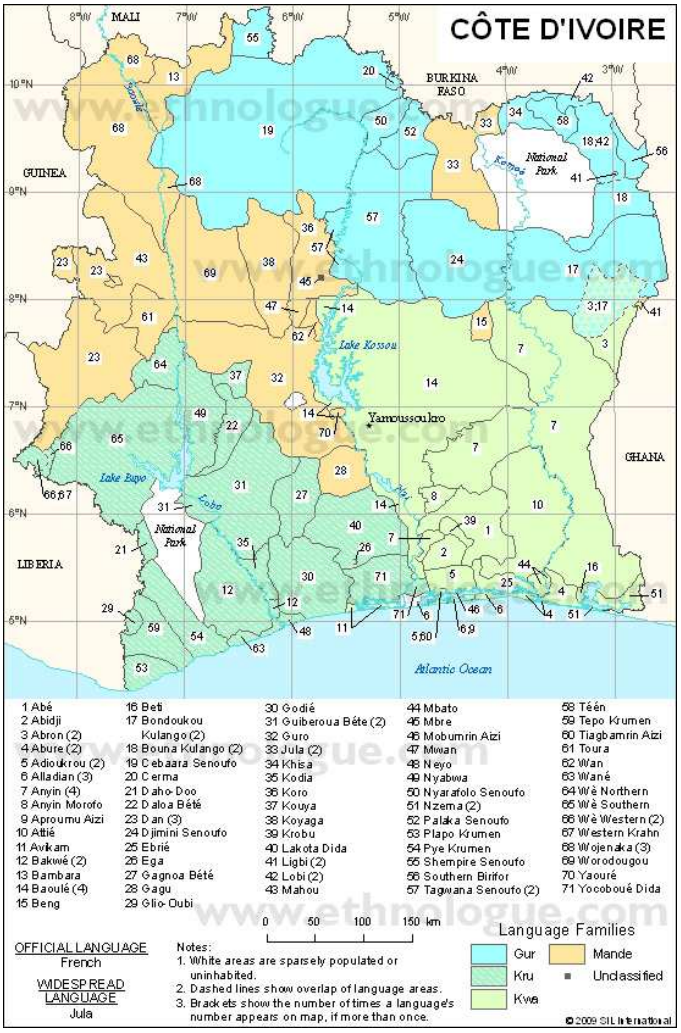


Figure 9. Ethno-linguistic groups in Ivory Coast. Image realized by Lewis, M. Paul [3]

In summary, the Kwa group is located in the southeast side of the country, where the capital city and other major cities are situated, as well as the main Ivorian Airport and Port. The Kru group is located in the southwest side, also in the Atlantic coast. The Mande group is found in the northeast side of the country, and the northwest region is occupied by the Gur family. These last are the least populated regions of the country.

[1] European Commission (2008) Communaute europeenne - Republique de Cote d'Ivore. Document de strategie pays et programe indicatif national pour la periode 2008-2013 http://ec.europa.eu/development/icenter/repository/scanned_ci_csp10_fr.pdf

[2] S. Bocaletti, V. Latora, Y. Moreno, M. Chavez and D.-U. Hwang Complex networks: Structure and dynamics *Phys. Rep.* 424, 175 (2006)

[3] Lewis, M. Paul (ed.), 2009. Ethnologue: Languages of the World, Sixteenth edition. Dallas, Tex.: SIL International. Online version: http://www.ethnologue.com/.

[4] African Development Bank Cte d'Ivoire Roads http://www.infrastructureafrica.org/library/doc/986/cote-divoire-roads

[5] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, (2008) Fast unfolding of communities in large networks, *J.STAT.MECH.*

[6] Newman M.E.J., 2003. Mixing patterns in networks. *Physical Review E* 67, 026126.

Impacts of External Shocks in Commodity-Dependent Low-Income Countries: Insights from mobile phone call detail records from Cote D'Ivoire

Ayodeji Fajebe[1], Peter Brecke[2]
[1]afajebe@gatech.edu, [2]peter.brecke@inta.gatech.edu
The Sam Nunn School of International Affairs
Georgia Institute of Technology
Atlanta, Georgia

Abstract: In this paper, we use a large mobile phone call detail record dataset of France Telecom-Orange's customers in the Ivory Coast (Cote D'Ivoire) to examine the relationship between short-term price fluctuations in Ivory Coast's top agricultural exports and manifestations of those fluctuations identified in the phone call dataset. We use Ivory Coast's primary agricultural export products—cocoa, coffee, and palm oil to investigate this relationship. Using the data provided by Orange, we extract and analyze patterns of socioeconomic ties from the dataset as a measure of short-term, domestic economic concern because of price fluctuations. We estimate a series of regression models and find that fluctuations in commodity prices are statistically significant predictors of short-term domestic economic concern. Changes in the price of cocoa and coffee have a negative effect on the level of concern while palm oil has a positive effect. We also include weather variables—temperature and rainfall—to the model but find their role insignificant. We conclude by discussing the limitations and implications of these findings.

INTRODUCTION

Mobile phones have become one of the most pervasive technologies of our times, penetrating all socioeconomic strata and offering a slew of communication services that are changing the way we live, work, and play. While the mobile phone is now virtually ubiquitous, and thus valued, their inherent ability to also function as sensors of human behavior presents social scientists the opportunity to gain insight into the patterns of behavior of a population in ways not previously available. Behavioral data has generally been difficult (and expensive) to capture on a large-scale. Mobile phone data typically collected by telecommunications companies—call detail record (CDR) data, traditionally used for billing purposes, are now increasingly analyzed for insights into behavioral and socioeconomic trends. For example, researchers have delved into areas such as social ties (Dasgupta et al., 2008; Eagle, Pentland, & Lazer, 2009; Miritello, Moro, & Lara, 2011; Onnela et al., 2007), relationship management (Rygielski, Wang, & Yen, 2002; Yan, Wolniewicz, & Dodier, 2004), transportation (Järv, Ahas,

Saluveer, Derudder, & Witlox, 2012; H. Wang, Calabrese, Di Lorenzo, & Ratti, 2010), human mobility (González, Hidalgo, & Barabási, 2008; D. Wang, Pedreschi, Song, Giannotti, & Barabasi, 2011), urban planning (Vieira, Frias-Martinez, Oliver, & Frias-Martinez, 2010), socioeconomic status (J. E. Blumenstock, Gillick, & Eagle, 2010; J. Blumenstock & Eagle, 2010), and trends (Frias-Martinez, Soguero-Ruiz, Josephidou, & Frias-Martinez, 2013). However, CDR data have not been used to address economic impacts such as the effect of price fluctuations on the level of concern within the society because of those price changes as identified in the patterns of cellphone calls. So we ask: "Can CDR data be useful in examining the relationship between variations in specific, international commodity prices and domestic economic concerns in a commodity-dependent exporting country?"

We ask this because we are interested in finding novel ways to detect or predict the impact of short-term instabilities on low-income, agricultural or other primary product exporting countries. Although it is widely believed that external fluctuations are transmitted to the economies of these countries, the linkages or mechanism of transmission and the impacts are not well understood. Of particular concern are the impacts on countries in Sub-Saharan Africa, many of which are poor but resource-rich and heavily dependent on exports of a small number of primary products. Moreover, the countries in this region are often fragile, both politically and economically, and lack good institutions to help them buffer the shocks. Thus, excessive volatility of prices in international commodities markets could pose serious consequences for their social, political, and economic stability.

There exists a literature concerning price volatility, linkages to domestic economies, and subsequent impacts. Some researchers claim that volatility has a moderately deleterious effect on economic growth (C. Glezakos, 1973; Constantine Glezakos, 1983; Moran, 1983; Voivodas, 1974), while others disagree, arguing that the net effect is not detrimental but somewhat positive because it is followed by a series of investments that lead to higher economic growth (Dawe, 1996; Knudsen & Parnes, 1975; Lam, 1980; MacBean, 1967). Yet other researchers counter both perspectives by claiming that the empirical evidence on the relationship is inconclusive and difficulty to specify because the transmission mechanism is unclear at the domestic level and the impact difficult to measure reliably on a wide scale (Behrman, 1987; Dawe, 1996; Lim, 1976; Savvides, 1984; Tan, 1983). CDR network data offers an opportunity to measure the impact more reliably and cost-effectively using the network of socioeconomic ties that are triggered as a result of a price change.

In this paper, we take a step towards answering the question: Can we measure the impact of external shocks to an economy by identifying, in the call detail record (CDR) data, changes in the amount, direction, and character of cellphone calls? Using the CDR data we received for the Orange "Data For Development" (D4D) Challenge in the Ivory Coast, we estimate a series of fitted regression models and find that variations in the international commodity prices of Ivory Coast's top agricultural exports—cocoa, coffee, and palm oil generate a change in the level of domestic economic concern measured using the patterns of socioeconomic ties extracted from the data. We found the relationship most significant when the model is based on call patterns

(that represent socioeconomic ties) between people in the region surrounding Daloa (an important cocoa producing area) and Abidjan (the economic and political capital). Although we include weather variables (temperature and rainfall) as part of the explanatory variables, they were not statistically significant, probably because of the timing—the data was collected during the dry season when variations are lowest. Changes in the prices of cocoa and coffee have a negative effect on the level of economic concern. When prices rise, the level of concern declines. Changes in the price of palm oil had a positive relationship with the extent of economic concern. Thus we show that large-scale CDR data of a low-income, commodity-dependent, primary product exporting country possess novel information about the impacts of variations in international commodity prices of the country's primary export products.

BACKGROUND

Ivory Coast[1] is a market-based economy that relies heavily on agriculture and related services. Although it has a considerable industrial sector, and more recently crude oil, most of the population (about 70%) is engaged in agriculture (CIA, 2013; FAO, 2009). It is the world's largest producer and exporter of cocoa beans (contributing about 40% of global market share), and a considerable producer and exporter of coffee and palm oil. According to a World Bank (1999, p. 1) report, "cocoa has usually contributed some 35 to 40 percent of exports, 14 percent of GDP, and more than 20 percent of government income." Coffee is also an important export product for the country although it has seen a decrease in production in recent years. Primarily it produces the lower grade Robusta coffee beans. To a lesser extent, palm oil is also an important cash crop for the country even though it is a declining export. It still plays an important role in the economic lives of the people as a recent Bloomberg news report attests: "palm oil production provides the livelihood of 2 million people, that is 10 percent of the population" (Ruitenberg, 2012). Overall, cocoa and coffee contribute some 50 percent of exports. Smallholder cash crop production is the dominant means of farming. These smallholder farmers grow the crops on their farms (either family-owned or community-owned) and sell their products to agents to the market. The whole process is a labor-intensive process from when it is planted, harvested, and sold to agents, before finally sold to a major exporter and exported.

We believe that a significant number of the population are economic actors involved in the value-chain of the commodity products—that is, they all play different roles in the different stages: from farming to export of the products. As such, they will be concerned with the fluctuations in the international prices. Intuitively, sellers believe that the market price of a product ripples through socioeconomic strata irrespective of the market structure (Mincer, 1963; Monroe, 1973; Reardon & Barrett, 2000), especially commodity sellers in Sub-Sahara Africa (Akiyama, Baffes, Larson, & Varangis, 2003). Even if commodity products are sold in the futures market, as is the case here, it is should be expected that individuals engaged at any level

[1] The World Bank classifies the country as a lower-middle income (WorldBank, 2013).

of the value chain will be concerned about the market prices because it inevitably affects them and will track the information, at least emotionally. The price of the commodity in the international market differs from the actual export prices, or much more, differs from the farmgate prices. In some cases, the state sets a government-guaranteed farmgate price—for example, coffee and cocoa (Reuters, 2011, 2012).

While we are unclear on how the Ivorian agricultural export market is coupled to the international commodity markets, using cocoa as a case-in-point, we know that the liberalization of the local market has left it exposed to fluctuations of prices in the world market (Agritrade, 2012). Farmers can sell directly to private operators as a price set by the state, and if not favorable, they can sell to the black market—for example, they can and do smuggle the products across boarders to neighboring countries like Ghana to sell for a higher price (Agritrade, 2012). So price (as well as weather for obvious reasons) are legitimate economic concerns for Ivorians in the industry (AfricaTimesNews, 2012).

Mobile phone

Ivory Coast has one of the highest mobile penetration rates in West Africa—over 17 million mobile phone subscribers in the country (approximately 80% of the population or 3 mobile phones for every 4 people) (ITU, 2012a). According to the ITU (2012b, p. 64), 92% of the population is covered by mobile cellular network. Based on this high teledensity, we assume that mobile phones are homogenously distributed in the population.

DATA

The CDR Dataset

The data for this study was provided by Orange Cote D'Ivoire through the Orange "Data for Development" (D4D) challenge—an open data challenge designed to use anonymized call detail records to help address socioeconomic development questions in the Ivory Coast. The data captures records of calls and text messages exchanged between five million of Orange's customers in Ivory Coast between December 1, 2011 and April 28, 2012 (150 days). Orange Labs in Paris preprocessed the data. Orange removed call information related to spurious customers such as those that subscribed to and cancelled their subscription during the observation period as well as those deemed to have come from "public" phone call centers—that is, private citizens that operate their mobile phone lines as public pay phones, thus removing noise.

We use two datasets from the data provided for our analysis: i) the datasets containing mobile phone base station antenna-to-antenna call traffic data information (SET1), and ii) the dataset containing the antenna position information (ANT_POS). All the datasets were provided in tabulation separated values (TSV) plain text format. The SET1 datasets (SET1TSV0 – SET1TSV9) contained 175,645,538 call events (observations). The data column header

contained: date hour, originating antenna identifier, terminating antenna identifier, number of voice calls, and duration of voice calls.

As part of the preprocessing, Orange paired incoming and outgoing calls together to eliminate double counting. However, about a quarter of the calls were not identified to protect its commercial interest. A non-unique antenna identifier "-1" in the dataset masked these calls. Random identifiers from 1 to 1238 uniquely identified the rest of the antennas in the dataset. In all, the datasets cover a total of 3600 hours. However, due to unspecified technical reasons by Orange, data was sometimes missing in the datasets. The missing data covered a period of about 100 hours. The ANT_POS dataset contained the corresponding geographic location information (longitude and latitude) of the 1238 antennas identified in the SET1 datasets.

Commodity prices data

We created our datasets for the commodity products from publicly available sources to match the period of observation in the data. For cocoa beans, we used the international cocoa organization (ICCO) daily prices of cocoa beans priced in US$ per tonne[2]. For coffee beans, we used the group daily indicator prices for Robusta beans from the international coffee organization (ICO) priced in US cents per pound[3]. For palm oil, we used daily prices derived from the crude palm oil chart produced by PalmOilHq, a market intelligence and news and prices organization[4]. The price chart is marked in Malaysian Ringgit (RM) per metric tonne, and linked to the Malaysian Palm Oil Board prices, the de facto industry price board. Expectedly, price data is not available for special days (weekends and holidays), so we constrained our CDR dataset by this information. Specifically, we used the days for which price data was available for cocoa as the criteria for trimming our call data. We had very few missing price data points—we had seven for palm oil and three for coffee. For these missing data points, we extrapolated the next day's price as the missing price data. We felt this was a reasonable approximation to do because the leading and trailing prices were significantly close together. Since the prices are daily prices and the resolution of our call data was hourly, we 'recreated' our price data as hourly also so that we can maintain the same resolution with our call data. Thus we merely recast the day's price as the price for every hour in that specific day. Hence the price data is constant for all the hours of the day we have call data.

Temperature and rainfall data

Similarly, we created our temperature and weather datasets from publicly available sources. We extracted the data from the historical daily temperature and rainfall records from Weather Underground[5] (an internet weather service). The temperature data was recorded in degrees Celsius while the rainfall was in millimeter (mm). The period of observation in the

[2] Source: http://www.icco.org/statistics/cocoa-prices/daily-prices.html.
[3] Source: http://www.ico.org/coffee_prices.asp.
[4] Source: http://www.palmoilhq.com/palm-oil-prices/.
[5] Source: http://www.wunderground.com/.

dataset largely corresponds to the dry season in the country so temperature and rainfall were relatively constant for this period. The temperature averaged between 24 – 28 degrees Celsius and rainfall was almost 0 mm. We had very few missing temperature and rainfall data points also. For those missing days, we simply extrapolated the next day's temperature for the missing data since they were fairly constant. We also up-scaled (that is, repeated the data hourly) the daily temperature and rainfall data to hourly data like we did for the price data to match the hourly call data.

METHODS

In Ivory Coast, agriculture is most suitable in the South— hence; cocoa, coffee, and palm oil are grown in these regions (south east and south west). Expectedly, economic activities are more prevalent in the South.

To operationalize economic concern, we use the count of the number of minutes that a group of originating and terminating antennas participating in a call are connected within the recorded hour. The connections are representative of the patterns of socioeconomic ties embedded in the network, thus the population. We assume that call patterns related to regions in the South will contain more information about economic concerns by virtue of the fact that more people live here.

We created different call pattern scenarios to better understand the dynamics and use the scenarios as the bases for the models we tested. To do this, we selected the following four cities with a high population —Abidjan, Bouake, Daloa, and Yamoussoukro, for analyses:

Abidjan: the largest city in the country with an estimated population of 3 million is the commercial and financial capital of the country (FAO, 2009). We expect call traffic to be the busiest in this region of the country. More than a third of the antennas in the dataset are located in the proximity of Abidjan. We are thus interested in call patterns between residents in the proximity of Abidjan and the entire country (Model 1), and call patterns between residents in the proximities of Abidjan and Daloa (Model 5).

Yamoussoukro: the political capital of the country with an estimated population of 250,000 inhabitants. We analyzed communication patterns strictly between antennas in the proximities of Yamoussoukro and the whole country (Model 6).

Bouake: located in the central region, with a population of 850,000 is the second largest city in the country, thus important. Here, we are interested the call patterns strictly between antennas located in the region of Bouake and the whole country (Model 2). **Daloa**: is an important trading center, particularly for cocoa with a population of 300,000. It produces about a quarter of the national cocoa output (AfricaTimesNews, 2012). We are interested in the call patterns between antennas in the region of Daloa and the whole country (Model 4) as well as patterns between Daloa and Abidjan (Model 5). Model 3 captured the patterns between all callers on the network regardless of location. Table 1.

We aggregate and group the call records into hourly blocks of data using an SQL database tool. The geolocation of all the antennas in the dataset is shown on a shape map[6] of the country to aid understanding in Figure 1. We also show the geolocation of the antennas we considered as located in the proximities of Abidjan, Bouake, Daloa, and Yamoussoukro in Figure 2.

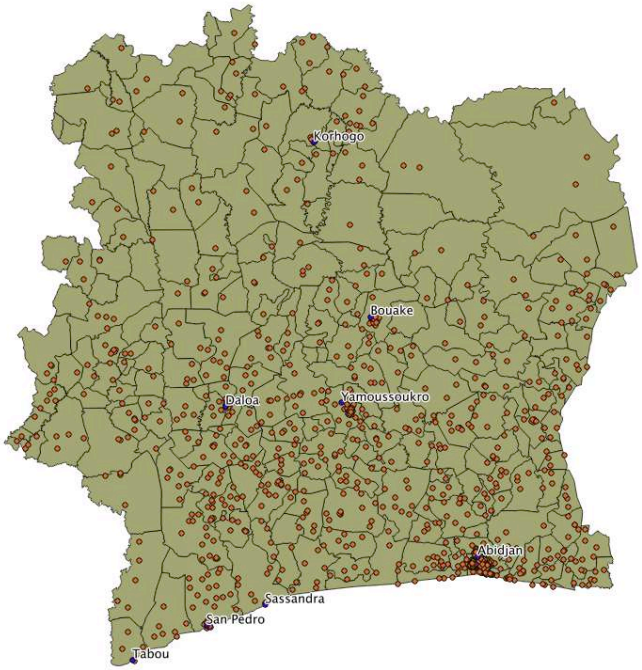


Figure 1: Map showing the geolocation of all the antennas in the dataset (1238 antennas).

[6] The file contained boundaries of the 255 subprefectures of Ivory Coast as used in D4D dataset n°3 - reduced spatial resolution dataset and formatted as an ESRI Shapefile. Source: http://sodexo.orange-labs.fr/GEOM_SUB_PREFECTURE.zip.

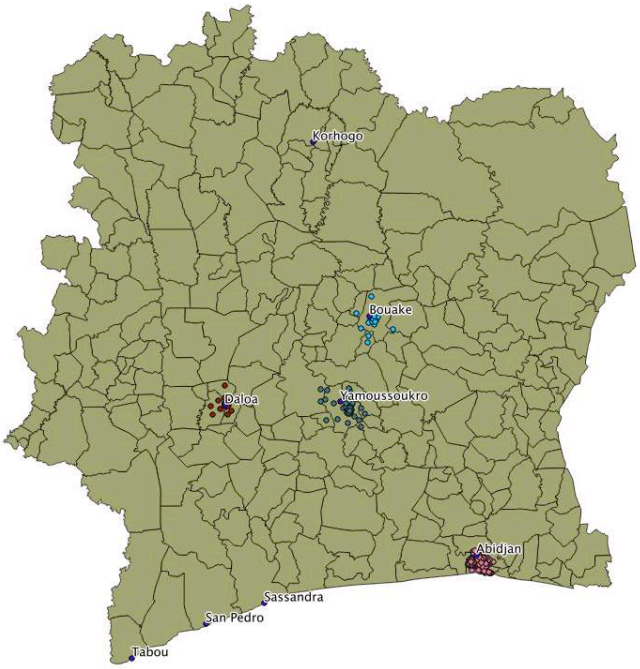


Figure 2: Map showing the geolocation of the antennas selected as located in the proximities of Abidjan, Bouake, Daloa, and Yamoussoukro.

Model	Description	Call pattern captured	N
Model 1	Abidjan model	All conversations that originate from any antenna anywhere in the country but terminates in Abidjan.	2284
Model 2	Bouake model	All conversations that originate from any antenna in the country but terminate in Bouake.	2164
Model 3	CIV model	All conversations that originate and terminate anywhere in the country i.e. the entire SET1 dataset.	2284
Model 4	Daloa model	All conversations that originate from any antenna in the country but terminate in Daloa.	1822
Model 5	Daloa-Abidjan model	All conversations that originate and terminate exclusively between the Daloa and Abidjan.	2284
Model 6	Yamoussoukro model	All conversations that originate anywhere from any antenna in the country but terminate in Yamoussoukro.	2164

Table 1: Summary description of the Models analyzed.

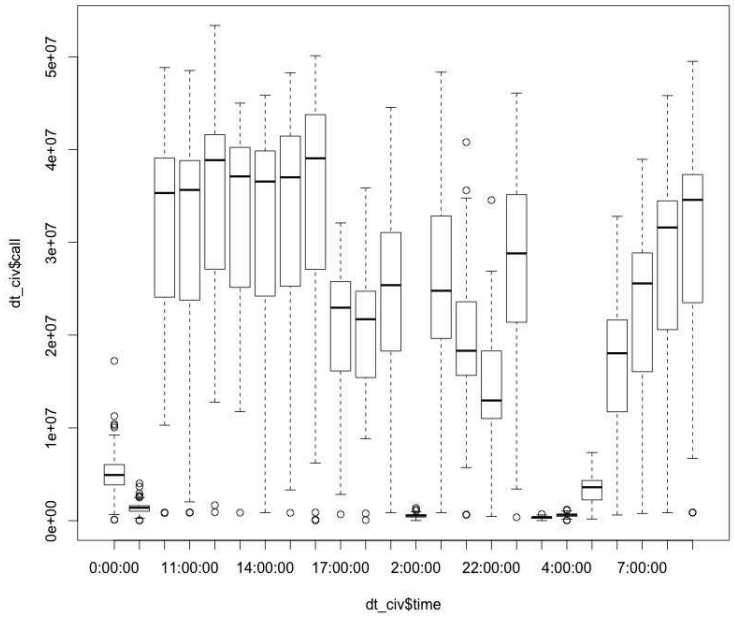


Figure 3: Box plot summarizing the circadian pattern of the all the calls in the dataset (Model 3).

Calls around mid-day have reasonably close medians with about the same interquartile range (IQR), thus suggests same call behavior. Possible calls that may be viewed as outliers exist, but they are close to the lower quartile and thus should not pose a problem to regression result. The median is highest (and about the same) for calls at noon and 4 pm but with slightly different variability. For other hours, the IQR decreases exhibiting reducing variability of calls within those hours as should be expected. Points at a greater distance from the median than 1.5 times the IQR are plotted individually and represent potential outliers.

ANALYSIS

Our analytical strategy involves the estimation of fitted regression models on our count of the total calls in minutes per hour (call volume) by the antennas in the models. In all the six models we present, our dependent variable exhibited a pronounced bimodal distribution. On the average, call volume is highest around noon and 4 pm. Figure 3 shows a boxplot conveying this information (summarizes the calls in the dataset in a 24-hour frame). But the impact is minimized in the model as the count regression method implemented takes a log of the dependent variable. We include the weather variables (temperature and rainfall) into the model for robustness as agricultural communities are typically concerned about the weather (AfricaTimesNews, 2012).

One of the challenges of using call volume between connected antennas as a measure of economic concern in an ethnically diverse population is that it may measure the diversity in the population instead of economic worries. This is because different ethnic groups in the country may have calling patterns (behaviors) that differ from other groups; thus, an aggregate hourly call volume may not sufficiently reflect their responses. Also, it is

known that other local events drive call volume so it may be possible that other events such as the political situation in the country or region drive the call volume instead.

We address the first concern by taking sub-samples of the dataset restricted to certain regions. So we did not just use the entire dataset (that is, Model 3) for the analysis, but focused patterns of communication between specific regions, for example, between Daloa and Abidjan exclusively (Model 5). This should help reduce the variations in responses. Second, even though political events may drive call volume, we assume the influence to be minimal.

Our formal model is shown below:

$$call = \beta_1 cocoa + \beta_2 coffee + \beta_3 palmoil + \beta_4 temperature + \beta_5 rainfall.$$

Where,

$call = hourly\ count\ of\ calls\ in\ minutes\ (call\ volume)$
$cocoa = daily\ price\ of\ cocoa\ beans\ in\ US\ \$/tonne$
$coffee = daily\ price\ of\ Robusta\ coffee\ beans\ in\ US\ cents/lb$
$palmoil = daily\ price\ of\ crude\ palm\ oil\ in\ RM/tonne$
$temperature = daily\ temperature\ in\ Celsius$
$rainfall = daily\ rainfall\ precipitation\ in\ mm$
$\beta_1, \beta_2, \beta_3 = estimated\ regression\ coefficients.$

Typically, count data are analyzed using Poisson-distribution models. However, empirical datasets such as this usually exhibit unique problems such as heterogeneity and over-dispersion—a problem present in our dataset. To address these related problems we use negative binomial regression to estimate all the models.

RESULTS

Our results are shown in Table 2. We check for model significance at the 5% level using the p chi-squared value (model p-value). Model 6 fails at this level. Further, we examine the rest of the models for predictor significance and eliminate Models 2 and 3 for lack of significance on our key predictor variables. We are thus left with Models 1, 4 and 5. We then use a combination of the predictor significance and model p-value to make our final selection. We find that Model 5 has a better significance level on cocoa than Models 1 and 4, and also has a much better p-value. Thus we select Model 5 as our best estimator.

In all the models except Model 3, the principal effect of cocoa is negative and statistically significant. In other words, an increase in cocoa price is associated with a decrease in call volume (keeping all other variables constant). This finding is intuitive because it suggests more calls are made when prices are low in the commodity market. This suggests that people talk more when they are concerned about prices but less when

prices are higher probably because it represents a potential for profit for them. Coffee has a negative significant effect also, except in Model 4. Coffee is almost as important as cocoa in the country so it is reasonable to assume a similar response. Model 4 may be an exception. Palm oil has a positive effect and is statistically significant in three of the models—Models 1, 4 and 5. Also, the coefficient estimates are close. Care should be taken in the interpretation of the estimated coefficients because they are interaction terms.

To answer our research question, we examine the analysis of the deviance table for Model 5 to check for the effect of the predictors as they are added sequentially to the null model. The result indicates that the individual price predictor variables are statistically significant. We show this result in Table 3.

In summary: we find evidence of a statistically significant relationship between commodity prices and call volume used as a measure of economic concern. While it is not surprising that cocoa and coffee co-vary since they are the two main agricultural exports and share similarities in production and export processes, it is somewhat surprising that palm oil has a positive relationship. It may be that the palm oil market is different because there are so many more effective alternatives that purchasers can shift to. In this case a rise in prices is a bad thing because then palm oil loses market share compared to other plant oils.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
(Intercept)	19.2846013***	13.1258294***	14.4684824***	7.5097872***	18.9057256***	1.371e+01***
cocoa	-0.0008264**	-0.0007376**	0.0002877	-0.0007375*	-0.0007830**	-6.068e-04*
	(-1.86065e-08)	(-1.728420e-07)	(2.050029e-09)	(-1.688856e-07)	(-2.622204e-08)	(-1.374175e-07)
coffee	-0.0324960 ***	0.0132427	0.0095415	0.0820444***	-0.0389377***	5.674e-03
	(-2.64101e-08)	(1.092398e-07)	(2.454112e-09)	(6.953663e-07)	(-4.707241e-08)	(4.528478e-08)
palmoil	0.0005694 ***	0.0001259	0.0002488	-0.0006800***	0.0007992***	6.637e-05
	(1.96351e-08)	(4.480961e-08)	(2.715004e-09)	(-2.736557e-07)	(4.099434e-08)	(2.284646e-08)
temperature	-0.0122510	0.0094261	-0.0014221	0.0457814**	-0.0214007	1.179e-02
	(-3.20464e-09)	(4.298984e-08)	(-1.177296e-10)	(2.043820e-07)	(-8.327033e-09)	(5.196790e-08)
rainfall	0.0023831	0.0069140	-0.0047635	0.0020923	0.0024771	6.287e-03
	(2.842603e-09)	(5.944023e-08)	(-1.798164e-09)	(2.941160e-08)	(4.395026e-09)	(5.225176e-08)
$\boldsymbol{\theta}$	0.8766	0.9902	0.9121	0.8137	0.8642	1.0243
N	2284	2164	2284	1822	2284	2164
AIC	75989	62459	81542	51394	74011	62430
2 *Log L*	-75975.0520	-62445.1200	-81528.4630	-51380.0630	-73997.0380	-62415.8700
pchisq	4.466005e-10	0.007438897	0.0132586	6.949996e-13	3.915757e-13	0.1051414

Robust standard error in parentheses; Significance levels: *** $p < 0.001$ ** $p < 0.01$ * $p < 0.05$

Table 2: Table of fitted negative binomial regression models estimating a count of the number of total calls per hour (call volume).

	Df	Deviance Residual	Df. Residual	Dev	Pr(>Chi)
NULL			2283	2751.3	
cocoa	1	33.725	2282	2717.6	6.349e-09 ***
coffee	1	6.577	2281	2711.0	0.01033 *
palmoil	1	25.157	2280	2685.9	5.286e-07 ***
temperature	1	1.387	2279	2684.5	0.23896
rainfall	1	0.342	2278	2684.1	0.55843

Significance levels: *** $p < 0.001$ ** $p < 0.01$ * $p < 0.05$

Table 3: Analyses of Deviance table for Model 5

LIMITATIONS

A number of important limitations and threats may affect the validity of our findings. One important concern is the issue of our measure of economic concern. Economic concern is a composite measure of welfare that is subjective and may be difficult to capture by just call volume alone. Most works that measure such related social indicator point out the difficulty of obtaining an aggregate measure (Diener & Suh, 1997; Shin, 1980). There are also issues concerning the complexity of the social structure in Sub-Sahara Africa that may affect our measurements even if we can adequately measure concern by call volume. Other valid measures of concern at the individual level—for example, measuring public support, might more appropriately reflect concern.

Another important concern that we have is that econometrically assessing the responses of economic concern of agricultural actors to variations in international prices is very still very problematic. Fundamentally, the poor quality of data available for such modeling is usually very difficult to get. This anchors Behrman's argument that the data is not just available to do a good job. Other dimension of data are needed to enable the model achieve an improved resolution. Even if the dimension were improved, a correlation between commodity price and economic concern indicates very little about the causal relationship underlying the choices economic agents make between production and leisure, cash crops and food crops, and wage work and non-wage work (Bond, 1983). The relationship between commodity prices and domestic economic concerns is a deep topic beyond the scope of this analysis. Our findings should be seen as limited in this regard.

Additionally, our analytical strategy assumes that Ivorians are relatively sedentary and make most of their calls in the same vicinity for the duration of study. This assumption enabled us to randomly select the group of antennas in the regions of Abidjan, Bouake, Daloa, and Yamoussoukro as targets of interests. It may be that calls are placed from different antennas in different regions. If this is so, it hinders our assumption that the same people make calls from the same area most of the time, and thus the validity of our results.

Another important limitation of our study is the generalizability of our results. For example, Model 5 that we selected models the relationship between callers in the region of Daloa

and Abidjan exclusively. Thus, the validity of our model for the entire country may be questioned. It may also be difficult to generalize this result across other Sub-Saharan African countries because they differ greatly in their geographical and physical conditions, ethnic conditions, cultural heritage, and weather patterns, even if they have similarities in their economic structures.

In spite of these limitations, this study provides a unique empirical approach of examining domestic economic concern issues in a network data. This data could be made richer and deeper by using the actual information of the callers instead of the base station antennas for identification. We hope that future work on this issue will explore and test our findings in similar contexts.

DISCUSSION & CONCLUSION

Our paper contributes to the literature on social network analysis, and economic welfare and development in two ways. First, we provide empirical evidence using phone call network data from a commodity-dependent lower-income country in Sub-Sahara Africa that shows that people in regions of the country specialized in the primary commodity products the country exports are concerned about the variations in its international prices. This finding is not surprising because some scholars as well as policymakers in developing countries have long believed that the variations, if excessive, has a deleterious effect on the economies of the exporting countries. We provide what we believe is evidence of the first empirical study that uses this kind of "Big Data" to find a relationship between domestic economic concerns and international price fluctuations.

Our metric of economic concern is based on network call data, but the people making the calls, and their reasons, are unclear. It is also possible that the metric, due to its complexity, may be responding to unobservable effect that we have not taken into account. For example, the metric may respond to current political dynamics in the country at the time of observation.

According to the Prebish-Singer hypothesis, commodity prices decline in the long-run (Prebisch, 1950; Singer, 1950), so we expect the prices of these commodities to decline. However, how they decline is the issue of concern because of the implications for the commodity-dependent exporting countries. The volatility concerns are deemed more important because price movements can have profound consequences for the achievement of economic development goals of those countries—impacting areas such as their politics, real incomes, fiscal positions, and terms of trade (Blattman, Hwang, & Williamson, 2007; Deaton & Miller, 1995; Drucker, 1986; Grilli & Yang, 1988).

This study is important because it gives further support to the debate that the linkage between the world markets and exporting countries in the developing world are more tightly integrated and responsive than previously thought. Most importantly, it shows that the impact of short-term price variations in the world markets can now be seen on economic actors in the exporting country using mobile phone CDR data. The resolution provided by this data now

allows us to overcome the "measurement problems" rightly identified by Behrman (1987, p. 559) as the tough obstacle because developing countries lack the structures or effective institutions that can capture the impact and effects on their economic attainment. According to Behrman (1987, p. 565), it is "costly" to get this kind of data for an individual country and even "very costly" to get this kind of data for a large number of primary commodity exporting developing countries.

We hope our study in spite of its limitations is viewed as a step toward illuminating the many ways in which fluctuations in world commodity prices ripple to the domestic economy of exporting countries. As social scientists, it is important that we find interesting ways to examine the plethora of data that the new technologies we are increasingly embracing can be used to help address socioeconomic concerns.

AKNOWLEDGEMENTS

We thank Orange Cote D'Ivoire for providing us the datasets and the opportunity to study this problem.

REFERENCES

AfricaTimesNews. (2012, August 13). Weeks of cool weather worry Ivory Coast cocoa farmers. Retrieved from http://www.africa-times-news.com/2012/08/weeks-of-cool-weather-worry-ivory-coast-cocoa-farmers/

Agritrade. (2012, December 16). Special report: Côte d'Ivoire's cocoa sector reforms 2011–2012. Retrieved February 10, 2013, from http://agritrade.cta.int/en/Agriculture/Commodities/Cocoa/Special-report-Cote-d-Ivoire-s-cocoa-sector-reforms-2011-2012

Akiyama, T., Baffes, J., Larson, D. F., & Varangis, P. (2003). Commodity market reform in Africa: some recent experience. *Economic Systems*, *27*(1), 83–115. doi:10.1016/S0939-3625(03)00018-9

Behrman, J. R. (1987). Commodity price instability and economic goal attainment in developing countries. *World Development*, *15*(5), 559–573.

Blumenstock, J. E., Gillick, D., & Eagle, N. (2010). Who's calling? Demographics of mobile phone use in Rwanda. *Transportation*, *32*, 2–5.

Blumenstock, J., & Eagle, N. (2010). Mobile divides: gender, socioeconomic status, and mobile phone use in Rwanda. In *Proceedings of the 4th ACM/IEEE International Conference on Information and Communication Technologies and Development* (pp. 6:1–6:10). New York, NY, USA: ACM. doi:10.1145/2369220.2369225

Bond, M. E. (1983). Agricultural Responses to Prices in Sub-Saharan African Countries (Réactions du secteur agricole aux prix en Afrique au sud du Sahara) (Reacciones de la agricultura ante los precios en los países del Africa al sur del Sahara). *Staff Papers - International Monetary Fund*, *30*(4), 703–726. doi:10.2307/3866783

Burchardi, K. B., & Hassan, T. A. (2011). *The Economic Impact of Social Ties: Evidence from German Reunification* (Working Paper No. 17186). National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w17186

CIA. (2013, February 15). The World Factbook: Cote D'Ivoire. Retrieved from https://www.cia.gov/library/publications/the-world-factbook/geos/iv.html

Dasgupta, K., Singh, R., Viswanathan, B., Chakraborty, D., Mukherjea, S., Nanavati, A. A., & Joshi, A. (2008). Social ties and their relevance to churn in mobile telecom networks. In *Proceedings of the 11th International Conference on Extending Database Technology: Advances in Database Technology* (pp. 668–677). Retrieved from http://dl.acm.org/ft_gateway.cfm?id=1353424&type=pdf

Dawe, D. (1996). A new look at the effects of export instability on investment and growth. *World development*, *24*(12), 1905–1914.

Diener, E., & Suh, E. (1997). Measuring quality of life: Economic, social, and subjective indicators. *Social indicators research*, *40*(1), 189–216.

Eagle, N., Pentland, A. S., & Lazer, D. (2009). Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, *106*(36), 15274–15278.

FAO. (2009). Country Pasture/Forage Resource Profiles-Cote D'Ivoire. Retrieved December 12, 2012, from http://www.fao.org/ag/AGP/AGPC/doc/Counprof/Ivorycoast/IvoryCoast.htm

Frias-Martinez, V., Soguero-Ruiz, C., Josephidou, M., & Frias-Martinez, E. (2013). Forecasting Socioeconomic Trends With Cell Phone Records. Retrieved from http://www.vanessafriasmartinez.org/uploads/dev13.pdf

Glezakos, C. (1973). Export instability and economic growth: A statistical verification. *Economic Development and Cultural Change*, 670–678.

Glezakos, Constantine. (1983). Instability and the growth of exports: A misinterpretation of the evidence from the western pacific countries. *Journal of Development Economics*, *12*(1–2), 229–236. doi:10.1016/0304-3878(83)90041-X

González, M. C., Hidalgo, C. A., & Barabási, A.-L. (2008). Understanding individual human mobility patterns. *Nature*, *453*(7196), 779–782. doi:10.1038/nature06958

Goyal, S. (2005). Strong and weak links. *Journal of the European Economic Association*, *3*(2-3), 608–616.

Goyal, S. (2009). *Connections: an introduction to the economics of networks*. Princeton University Press. Retrieved from http://books.google.com/books?hl=en&lr=&id=h4V250eVEKgC&oi=fnd&pg=PR7&ots=K26TTcAX39&sig=-RqXtWts0ZEALDdnxQidIdr7aMQ

Granovetter, M. (1983). The strength of weak ties: A network theory revisited. *Sociological theory*, *1*(1), 201–233.

Granovetter, M. (2005). The impact of social structure on economic outcomes. *The Journal of Economic Perspectives*, *19*(1), 33–50.

Granovetter, M. S. (1973). The strength of weak ties. *American journal of sociology*, 1360–1380.

Jackson, M. O. (2009). Networks and economic behavior. *Annu. Rev. Econ.*, *1*(1), 489–511.

Järv, O., Ahas, R., Saluveer, E., Derudder, B., & Witlox, F. (2012). Mobile phones in a traffic flow: a geographical perspective to evening rush hour traffic analysis using call detail records. *PloS one*, *7*(11), e49171.

Jenny, D. (2009). Spatial organization of leopards Panthera pardus in Taï National Park, Ivory Coast: is rainforest habitat a "tropical haven"? *Journal of Zoology*, *240*(3), 427–440.

Kaldor, N. (1976). Inflation and recession in the world economy. *The Economic Journal*, *86*(344), 703–714.

Knack, S., & Keefer, P. (1997). Does social capital have an economic payoff? A cross-country investigation. *The Quarterly journal of economics*, *112*(4), 1251–1288.

Knudsen, O., & Parnes, A. (1975). *Trade instability and economic development: an empirical study*. Lexington Books. Retrieved from http://www.getcited.org/pub/101508117

Krackhardt, D. (1992). The strength of strong ties: The importance of philos in organizations. *Networks and organizations: Structure, form, and action*, *216*, 239.

Labys, W. C., & Maizels, A. (1993). Commodity price fluctuations and macroeconomic adjustments in the developed economies. *Journal of Policy Modeling*, *15*(3), 335–352. doi:10.1016/0161-8938(93)90037-Q

Lam, N. V. (1980). Export instability, expansion and market concentration: A methodological interpretation. *Journal of Development Economics*, *7*(1), 99–115.

Lim, D. (1976). Export instability and economic growth: a return to fundamentals. *Oxford Bulletin of Economics and Statistics*, *38*(4), 311–322.

MacBean, A. I. (1967). Export instability and economic development. Retrieved from http://www.getcited.org/pub/101231335

Maizels, A. (1987). Commodities in crisis: An overview of the main issues. *World Development*, *15*(5), 537–549. doi:10.1016/0305-750X(87)90001-5

McIntire, J., & Varangis, P. (1999). Reforming Cote d'Ivoire's cocoa marketing and pricing system. *World Bank Policy Research Working Paper*, (2081). Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=615012

Mincer, J. (1963). Market prices, opportunity costs, and income effects. *Measurement in economics*, 67–82.

Miritello, G., Moro, E., & Lara, R. (2011). Dynamical strength of social ties in information spreading. *Physical Review E*, *83*(4), 045102. doi:10.1103/PhysRevE.83.045102

Monroe, K. B. (1973). Buyers' subjective perceptions of price. *Journal of Marketing Research*, 70–80.

Moran, C. (1983). Export fluctuations and economic growth: an empirical analysis. *Journal of Development Economics*, *12*(1), 195–218.

Onnela, J.-P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., … Barabási, A.-L. (2007). Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences*, *104*(18), 7332–7336. doi:10.1073/pnas.0610245104

Prebisch, R. (1950). *The economic development of Latin America and its principal problems*. United Nations Dept. of Economic Affairs.

Reardon, T., & Barrett, C. B. (2000). Agroindustrialization, globalization, and international development: an overview of issues, patterns, and determinants. *Agricultural Economics*, *23*(3), 195–205.

Reuters. (2011, April 16). Ivory Coast cocoa exporters seek speedier process. Retrieved January 15, 2013, from http://www.forexyard.com/en/news/Ivory-Coast-cocoa-exporters-seek-speedier-process-2011-04-14T153752Z

Reuters. (2012, December 19). Ivory Coast fixes government-guaranteed coffee price. *Yahoo! News*. Retrieved February 12, 2013, from http://news.yahoo.com/ivory-coast-fixes-government-guaranteed-coffee-price-155300924--sector.html

Ross, M. L. (1999). The political economy of the resource curse. *World politics*, *51*, 297–322.

Ruitenberg, R. (2012, November 12). Ivory Coast Palm Oil Growers Defend Crop Amid France Tax Plan. *Bloomberg*. Retrieved February 10, 2013, from http://www.bloomberg.com/news/2012-11-12/ivory-coast-palm-oil-growers-defend-crop-amid-france-tax-plan.html

Rygielski, C., Wang, J. C., & Yen, D. C. (2002). Data mining techniques for customer relationship management. *Technology in Society*, *24*(4), 483–502.

Savvides, A. (1984). Export instability and economic growth: some new evidence. *Economic Development and Cultural Change*, *32*(3), 607–614.

Shin, D. C. (1980). Does rapid economic growth improve the human lot? Some empirical evidence. *Social Indicators Research*, *8*(2), 199–221.

Singer, H. W. (1950). The distribution of gains between investing and borrowing countries. *The American Economic Review*, 473–485.

Tan, G. (1983). Export instability, export growth and GDP growth. *Journal of Development Economics*, *12*(1), 219–227.

Vieira, M. R., Frias-Martinez, V., Oliver, N., & Frias-Martinez, E. (2010). Characterizing dense urban areas from mobile phone-call data: Discovery and social dynamics. In *Social Computing (SocialCom), 2010 IEEE Second International Conference On* (pp. 241–248). Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5590404

Voivodas, C. S. (1974). The effect of foreign exchange instability on growth. *The Review of Economics and Statistics*, *56*(3), 410–412.

Wang, D., Pedreschi, D., Song, C., Giannotti, F., & Barabasi, A. L. (2011). Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1100–1108). Retrieved from http://dl.acm.org/citation.cfm?id=2020581

Wang, H., Calabrese, F., Di Lorenzo, G., & Ratti, C. (2010). Transportation mode inference from anonymized and aggregated mobile phone call detail records. In *Proceedings of the IEEE International Conference on Intelligent Transportation Systems*. Retrieved from http://people.csail.mit.edu/huayongw/papers/2010_Wang_ITSC.pdf

WorldBank. (2013). Cote d'Ivoire. Retrieved January 20, 2013, from http://data.worldbank.org/country/cote-divoire

Wuchty, S. (2009). What is a social tie? *Proceedings of the National Academy of Sciences*, *106*(36), 15099–15100. doi:10.1073/pnas.0907905106

Yan, L., Wolniewicz, R. H., & Dodier, R. (2004). Predicting customer behavior in telecommunications. *Intelligent Systems, IEEE*, *19*(2), 50–58.

Regional patterns of socio-economic activity in Côte d'Ivoire

Maurizio Dusi, Mohamed Ahmed, Roberto Caporicci
NEC Laboratories Europe
<first.last>@neclab.eu

Nicholas Cheeseman
African Politics at the African Studies Centre,
University of Oxford
nicholas.cheeseman@politics.ox.ac.uk

The number of active cell phone connections in Africa in 2008 was approximately 260M, by 2012 this figure was more than 500M and is expected to rise to more than 700M by the year 2016. This is a 2.7× increase in market penetration in under a decade and clearly represents a significant change to the ease of access to mass-communication.

Both academic literature and industry practice show that personal cell phones in Africa are not limited to private use. Cell phones have proven to provide a reliable and affordable tool for conducting business and facilitating social, economic and political interaction, helping to facilitate activities that span from collecting daily commodity prices [2] to purchasing common services [1]. This all signifies that today, cell phones in Africa are immersed into the daily activity of ordinary citizens, and may therefore offer us a window into how they manage their daily affairs. This is particularly interesting when looking at the social and economic recovery (or lack thereof) of post conflict countries such as Côte d'Ivoire, which has seen a very recent history of political turmoil.

In this brief abstract, we present our first results towards understanding to what extent the traffic patters of ordinary citizens can help us to uncover how the citizens of a post conflict country organise their economic and social ties. Specifically, using the cell phone antenna data collected in Côte d'Ivoire between December 2011 and April 2012 [4], we look at; when people communicate, and who they communicate with. For the analysis, we consider the outgoing traffic that each antenna generates towards the other antennae.

In answering the first question, we are interested whether cell phone usage is more common during the day or in the evening; under the assumption that people use cell phones in the day for business and in the evening for leisure. While the second question looks at the geographic distribution of the set of interactions; with the assumption that a regionally segmented pattern of interaction is indicative of a regionally segregated population.

Starting at when people use their cell phones, from Figure 1, we see that user activity is diurnal, but significantly, it varies little over the week. The peek traffic hours for outgoing calls are bi-modal with modes at 10hrs in the morning and 20hrs at night. Further, we find no significant difference in the observed traffic volume between the standard working day and in the evening, suggesting that users are utilising their phones for business and leisure equally.

Looking at where users call, Figure 2a gives the cumulative distribution of the distance between pairs of antennae and shows that ≈ 30% of calls are between antennae that are within 5Km of each other, while ≈ 50% of calls are between pairs of antennae that are within 20Km. Given that 63 out of 65 population centres in Côte d'Ivoire have less than 250K inhabitants, with an average population of 48K[1], this result is easily biased by calls originating from highly populated areas.

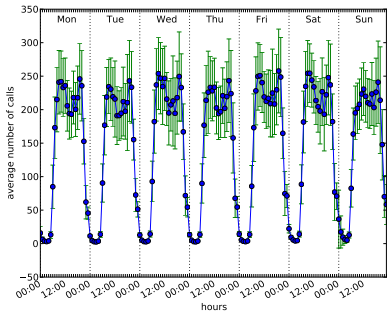
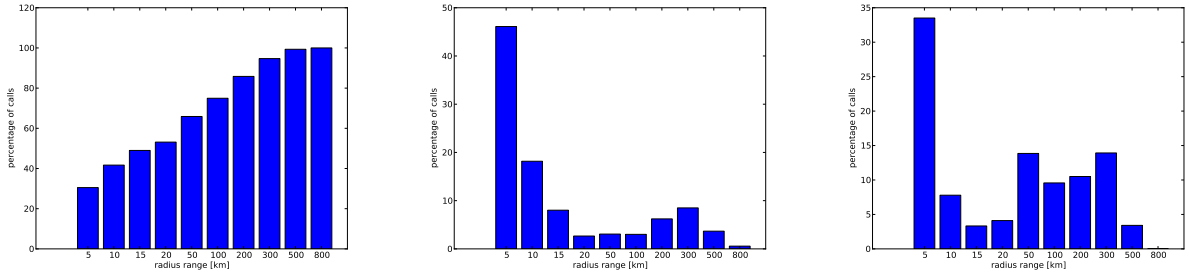


Figure 1: The weekly (mean) volume of outgoing calls between each pair of antennae across the observation period of 150 days.

From the probability distribution functions of the calls, conditional on the nearest population of each antenna (Figures 2b-2f), we see that users in the few highly populated areas ($N \geq 100$K) tend to call very short distances: for $N \geq 150$K, approximately 40% calls are within 5Km, with nearly 30% of calls ranging distances ≥ 50Km. This is in contrast to users in areas with small populations ($N \leq 100$K) as show in Figures 2e and 2f where approximately 70% of calls have a range ≥ 70Km. This suggests that though short distance calls are important to both groups, long distance calls are more significant for people living in rural/small communities. While the evidence in the data is not enough to derive direct answers from the numbers alone, there is significant evidence in literature suggesting that cell phones are an enabling technology for rural populations, allowing people to keep upto date with of economic, political and social events and duties [3].

The large geographical locality in the calls within highly populated regions ($N \geq 100$K) suggests that both business and social contact is normally conducted within a very close range. Though calls are placed evenly throughout the week, only a marginal number of calls that take place are directed from urban to rural areas. Conversely, for more rural areas ($N < 100$K) contact with distant location is more impor-

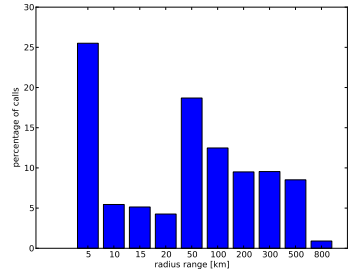
[1]Only two cities have a population above 250K inhabitants: Bouakè with 461K inhabitants, and Abidjan being the biggest city with almost 3M people



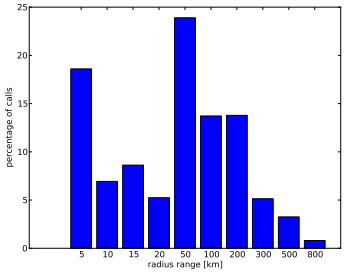
(a) All traffic.

(b) $\geq$ 200K

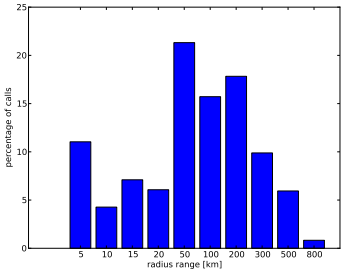
(c) 200-150K



(d) 150-100K



(e) 100-50K



(f) $<$ 50K.

Figure 2: The distribution of call distances. Figure 2a gives the cumulative distribution of all outgoing calls, while Figures 2b-2b gives the PDF of the distance of outgoing calls for regions with N inhabitants

tant.

Figure 3 looks into this result to understand the geographic distribution of traffic. First, we see that in absolute terms (Figure 3a), Abidjan – the commercial capital – is the major focal point. However, if we start to filter for the (geographical) cumulative distribution of outgoing calls, subject to the population size, we find that regional hubs maintain a significant contact to the rest of the county. When looking at the number of outgoing calls from areas with a population $100K \leq N \leq 200$K (Figures 3c and 3d), the cities of Man, Yamoussoukro (the political capital) and Bouake generate the vast majority of calls to the more rural parts of the country and for $N \leq 100K$ (Figures 3e and 3f) we see that Yamoussoukro becomes as the main hub for generating activity.

Interestingly (Figures 3c - 3f) show the important role that regional hubs play in facilitating interaction between themselves, Abidjan and the many smaller communities dotted throughout the country. Though vast majority of traffic is directed between population centres it is not surprising that the region hubs are significantly more likely to be initiators of communication to much smaller communities, than Abidjan. While Abidjan tends to initiate traffic to the regional hubs.

Finally, given that users do not seem to display a significantly different call-rate pattern over the entire week, look at whether the destination of their calls varies. From Figure 4 we see that there is a significant difference in where users call. We find that during the working week, aggregate traffic is concentrated to/from the economic and political hubs, while regional centres appear to be relatively more significant during weekends. Regional hubs such as Man, Bouake, Korhgo, San Pedro appear to become strong focal points for contact during the weekend.

In summary, we find that most communication is directed towards regional economic hubs, and these regional centres attract the vast majority of communication from the many small communities. We find that region hubs seem to play two important roles; first, during the working week they act as economic gateways to small communities, both for other hubs and Abidjan, and second as focal points for distant (presumably social) contact between regional hubs during weekends.

In further work, we are interested in understanding the role that regional hubs play in facilitating interaction between distant parts of the country, and the properties of the interactions between inter-regional hubs. Specifically we would like to identify the commercial and social clustering of inter-hub interactions and to what extent this reflects the political or economic organisation of country.

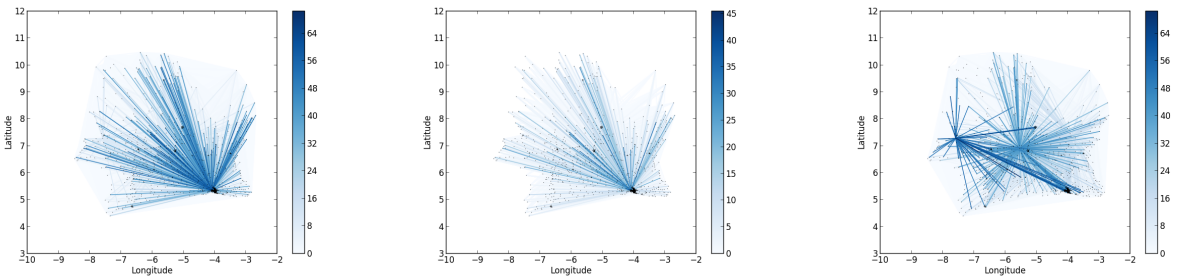
1. REFERENCES

[1] M. Bratton. Briefing citizens and cell phones in Africa. Afr Aff (Lond) (2013) doi: 10.1093/afraf/adt004.

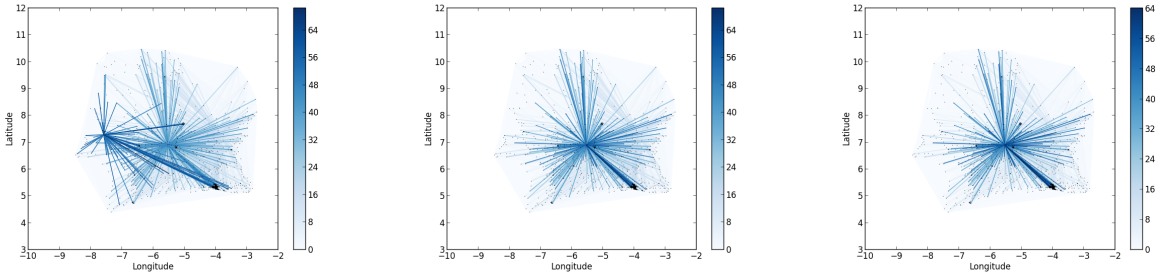
[2] M. Rao. Mobile Africa report 2012: Sustainable Innovation Ecosystems. http://www.mobilemonday.net/04/2012/mobile-monday-fourth-annual-mobile-africa-research-report.html (accessed 18 Feb 2013).

[3] K. Fox. Africas mobile economic revolution www.guardian.co.uk/technology, 2011 (accessed 18 Feb 2013).

[4] V. D. Blondel and M. Esch and C. Chan and F. Clérot and P. Deville and E. Huens and F. Morlot and Z. Smoreda and C. Ziemlicki Data for Development: the D4D Challenge on Mobile Phone Data. arXiv:1210.0137.

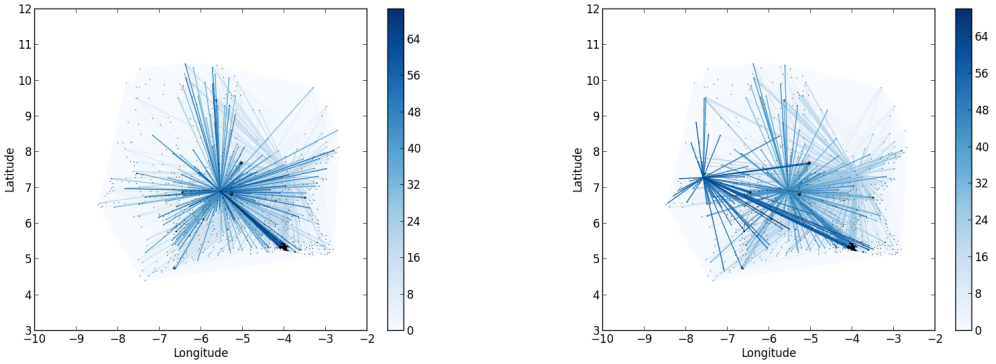


(a) All traffic (b) ≥ 200 (c) < 200



(d) < 150 (e) < 100 (f) < 50

Figure 3: Inter-antennae communication distribution. Figure 3a plots that distribution of outgoing calls during the period (150 days). Figures 3c-3f give percentage of the inter-antennae communication for regions with N inhabitants.



(a) Working week ($N \leq 200$K) (b) Weekend ($N \leq 200$K)

Figure 4: Working week / weekend inter-antennae communication distribution.

Regional Development - Capturing a nation's sporting interest through call detail analysis

Donal Mc Gowan, Neil Hurley

Clique Research Group, University College Dublin, Ireland
donal.mcgowan@ucd.ie

Keywords: D4D, Call Data Records, Cup of Nations, Time Series

Abstract. Although the Ivorian Nation is recovering from its second civil war, there is still one dominating force which brings the nation together, that being their love for football. Using mobile call records acquired during the African Cup of Nations tournament, we look to extract a quantitative metric conveying each departments emotional and social interest for the event. This we achieve through detecting irregular call activity around match times over the entire tournament period. Using these interest scores we target specific areas we feel could benefit from the creation of sporting facilities which we additionally hope would help develop the surrounding area.

1 Introduction

We currently live in an age where the mobile phone has become omnipresent and the idea of being apart from it for a day let alone an hour is absurd. This indispensable device has become essential to our daily lives, keeping us in contact with our social circles, providing real-time information on weather, sports, markets and anything else we wish to google. Throughout developed countries, we see children as young as ten and upwards owning or having access to a phone, with some having multiple devices in their possession. In contrast to this, several developing countries have only recently seen an uptake in mobile users. An executive vice president of a global telecommunications company has been quoted as saying *"In 2000, you had about five million mobile phones in Africa, today, we have about 500 million, In 2015, we expect it to be 800 million"* This exponential growth in users provides us with some insight into how quickly a developing country can adapt to technology within the space of a few years.

The ever increasing problem with current technology lies in the amount of excess data produced. Either from mobile phone, or other forms of information-sensing devices, this rich source of data hold the key to answering vital questions in many disciplines. The scope of this mobile data can be very extensive, with many mobile service providers routinely collecting information on phone usage such as call logs, geo-spatial location and volume of communication. Over a single month a network provider, depending on the size of its customer base,

can accumulate tens of millions of call detail records (CDRs). Analysis of the mobile phone records can be a long and arduous procedure. Whether you are simply processing or carrying out more sophisticated data mining using these data records, it is important to have a specific question or goal in mind, otherwise as the sheer magnitude of data makes casual observations from the data impossible. Past research involving call record analysis has commonly been split between two lines of enquiry, these being scientific research and industrial applications. While the former may look at sociological questions, such as mobile user behaviour and disease or information spread analysis, the later focuses on profit driven questions like churn prediction, link analysis and community detection. Apart from these applications, phone data could potentially be utilised in a humanitarian manner to assist in the development of a nation. Whether we are looking to identify the early signs of epidemics, measure the threat and resultant impact of droughts or look to optimise the usage of certain infrastructures, call data analysis could help to improve reaction times in moments of crisis. Possibly the most famous example of exploiting data in this manner is the Google Flu Trends [8] that uses search query logs to predict flu epidemics. This has been followed by similar work using Twitter data [5].

In this paper, utilising anonymised Call Detail Records extracted from Orange's customer base within the Ivory Coast, we look to extract a quantitative metric conveying each department's interest for the African Cup of Nations. We also propose the concept of two types of interest, which we term emotional and social interest. These metrics will be calculated through detecting irregular call activity around match times over the entire tournament period. Using these interest scores we target specific areas within the Ivory Coast we feel could benefit from the creation of sporting facilities and which we hope would additionally help develop the surrounding area. An important aspect of this research is that it can be immediately verified, as it was recently decided that the Cup of Nations be offset by one year so that it never occurs the same year as the World Cup. As a result of this, the cup was scheduled to take place again this year.

This rest of this paper is organised as follows, Section 2 presents some of the related work in the area of mobile data analysis. We will then give a brief description of the Ivory Coast in Section 3. Following on, Section 4 will provide a brief description of the data to be used throughout our analysis. We then begin our analysis in Section 5, starting with a country-wide overview to detect any emerging trends. Next in Section 6 we compute interest measures on the individual departments within the Ivory Coast. In Section 7 we discuss how this analysis can be exploited to target sports development in areas that would socially profit from development. We conclude our work in Section 8 by providing future areas in which this work could be explored.

2 Related Work

The mobile phone, in particular the smart-phone, generates vast amounts of data on a regular basis. This data, which is collected by the mobile operators, can

contain call detail information, network data usage, phone handset information, customer details and even geo-location data. Such data is commonly accumulated for an entire subscriber base thereby making it almost impossible to manually analyse. To overcome this issue, the telecommunications sector turned to data mining techniques.

Specifically developed to discover patterns in large data sets, data mining methods look to extract information utilising machine learning, statistics and artificial intelligence. Common data mining tasks include pattern analysis, visualisation, and anomaly detection, all of which can be done on static or in real-time data, depending on the requirement. Due to the sheer volume of raw data available, it is generally necessary to pre-process the data, for example, by summarising the data, extracting certain features advantageous to the research being carried out or aggregating the data over certain time periods.

Previous research involving mobile data records has commonly been split across various scientific fields and the two that stand out are that of network analysis and sociology. These areas have had a long standing relationship with mobile data, as seen through their various papers covering topics such as churn prediction [10], link analysis [12], community detection [3] and social network analysis [6]. These topics rely on extracting informative features from call data records or constructing complex network structures using call information so as to accurately predict or model the data. Additionally service providers have shown interest in customer behavioural analysis [4], mobility pattern detection and classification of subscribers [9, 11], along with methods of improving service and characterising workload dynamics of a mobile network [13]. Such in depth analysis of network workloads can provide substantial information into user experience and into how a network handles communication traffic.

Regarding research surrounding social development, much focus has been directed towards anomalous event detection [13, 1] and analysing human behaviour in response to an emergency [2]. Both topics provide invaluable information in first detecting irregular events and subsequently how information spreads both spatially and temporally throughout a network. In addition to this, further research has been carried out on contagion analysis by looking at how to model the spread of disease [7].

In our work, we specifically look to detect irregular call behaviour, for multiple departments, around multiple sporting events. This is similar to the work already carried out in [13]. However instead of focusing on the network workload surrounding an event, we are more concerned with discerning the amount of interest a city shows towards an event. Although the research done in [13] is somewhat different to our, is does provide us with a framework for the calling behaviour around football matches. This will prove helpful to our analysis in the later sections.

3 Ivory Coast

In this section we will provide a short description of the Ivory Coast to aid us in our analysis of the call data in later sections. To begin with, the country is situated in West Africa, bordering the countries of Liberia, Guinea, Mali, Burkina Faso and Ghana. In 2009 it was estimated to have a population of 20 million, with its official language being French. Geographically the country is divided into 19 regions and 81 departments with the capital, Yamoussoukro, situated in the Lacs region. Below in table 1, we provide a list of the top 10 departments ranked by population, with Abidjan the largest city clearly surpassing the rest.

City	Region	Population
Abidjan	Lagunes	3,677,115
Abobo	Lagunes	900,000
Bouak	Valle du Bandama	567,481
Daloa	Haut-Sassandra	215,652
San-Pdro	Bas-Sassandra	196,751
Yamoussoukro	Lacs	194,530
Korhogo	Savanes	167,359
Man	Dix-Huit Montagnes	139,341
Divo	Sud-Bandama	127,867
Gagnoa	Fromager	123,184

Table 1: Top ten cities of the Ivory Coast ranked by population

Within the Ivory Coast there are distinct religions scattered throughout the country, Islam (which dominates the North), Christianity (which dominate the South) and various other indigenous religions. It must be noted that the country itself is currently recovering from its second civil war, which took place after the 2010/11 presidential elections. This war which spanned several months, resulted in many deaths left many cities and regions torn apart.

Concerning sport within the Ivory Coast, the most popular sporting activity is football, with the national team having played in the World Cup twice (Germany 2006, South Africa 2010). Although the World Cup only occurs every four years, the African Cup of Nations takes place every two years. This competition, whose first edition in 1957 only contained three nations, has grown in stature over the years and currently allows 16 qualifying teams. While the location changes for each tournament, based on bidding wars between nations, the period during when the competition is held, that is, the weeks leading up to the end of January and the start of February, rarely changes. On an important note, it was recently decided that the tournament be offset by one year so that it never occurs the same year as the World Cup. As such, it was held again this year making it two years in a row and furthermore allowing the analysis carried out in this paper to be immediately replicable.

4 Call Data

As of the 30^{th} of June 2011, an annual report listed Orange as having over 5.5 million subscribers within the Ivory Coast. Considering the 20 million citizens already estimated to be living within the Ivory Coast, this is a clear indication Orange has a strong position within the market. Our mobile phone datasets have been provided by Orange, to be used in a research context, for the sole purpose of addressing society development questions pertaining to the Ivory Coast. These data sets are based on anonymised Call Detail Records extracted from Orange's customer base covering 150 days, from December 2011 to April 2012, for a total of 3600 hours. This data, which excludes any communication between Orange customers and non Orange customers, additionally has over 100 hours of missing data scattered throughout. In total there are four separate data sets supplied for the Data 4 Development challenge, they are as follows:

- Aggregated communication between cell towers
- Mobility traces: fine resolution dataset
- Mobility traces: coarse resolution dataset
- Communication sub-graphs

For the purpose of our research, we are only interested in the aggregated communication between cell towers. This data contains the number of calls as well as the duration of calls between any pair of antennas, aggregated on an hourly basis. In our analysis of this data we aim to detect irregular patterns that may appear within different regions of the Ivory Coast, specifically during the time of the African Cup of Nations. Given the coinciding nature of this event and our data, our goal is to measure anomalous call patterns throughout the regions of the Ivory Coast. As an added bonus to this analysis, during this specific year the Ivory Coast managed to progress to the final of the tournament where they marginally lost in extra time penalties, so interest in the tournament would have extended right to the final match.

Our analysis is conducted using call data outside of the tournament period as a baseline from which we compare against the tournament data. Within the tournament period there are 28 match events, some containing 2 matches occurring at once, spaced out over a month. With the exact kick off times for each match known, we can focus on the activity around each match from beginning to end.

5 Initial Data Overview

An initial step in any analysis of call data is to acquire a general understanding of data available. One simple method for this is to visualise the entire data, thereby allowing any trends that may be present to be discerned. Our aggregated cell tower data, in its raw form, displays the number of calls and duration of calls

for active regions on an hourly basis. An initial visualisation of these two data sets shows a very similar trend, albeit on different scales.

Taking this into account, we only provide the daily aggregated call duration, as seen in Figure 1. Within the data there is a lull period throughout the beginning of the data covering Dec 15^{th} to Jan 18^{th}. Within this calm period there is a notable spike which, as you would expect, occurs around New Year's Eve. Furthermore the base of the figure indicates the presence of three distinct troughs which, on further review, are as a result of missing data points. For the sake of our research, we are particularly interested in the peaks that appear in the midway period of Figure 1, as these cover the days during which the tournament took place. The impact of the tournament on call behaviour can be immediately seen from the highest peak centrally located in Figure 1. At this specific hour of the day, the final match of the tournament had just ended, where the Ivory Coast lost in sudden death penalties. The notable global spike is a result of the nation clamouring over the ill-fated outcome of the match. Over the entire data period, the average call length is approximately 170 seconds, but varies quite substantially depending on the hour of the day.

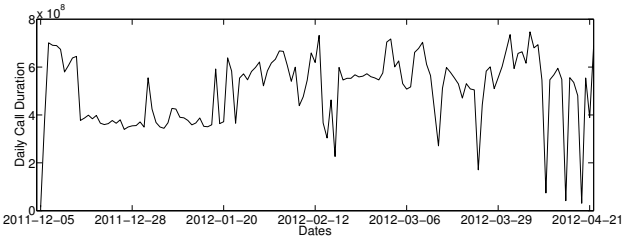


Fig. 1: Daily Call Duration

Focusing in on our period of interest, we reduce the data-set down to around 2,000 hours, from the 18^{th} of January to the 9^{th} of April. Moving forward, we wish to identify interest in the sporting event through detection of irregular activity in and around the times of football matches, not specifically Ivory Coast matches. The reason for this is because several of the bordering countries were themselves competing in the competition and as such there is a possibility of non-Ivorian inhabitants supporting their country of birth, which could produce activity around non-Ivorian matches. Furthermore, as the competition progresses there was likely to be an increased concern over who the Ivory Coast may have to compete against in later rounds, therefore we may possibly see activity around matches involving potential opponents.

Our first step is to split our reduced data, both into the tournament t_T and non-tournament periods t_{NT}. We extend t_T to three additional days either side of the tournament, as we wish to cover any growing and fading interest for the sporting event. Regarding t_{NT}, we take the hour directly after the end of t_T up until the 9^{th} of April. In order to discover the presence of irregular call behaviour,

we contrast data from within t_T with that of the daily hourly averaged data from t_{NT}.

5.1 Group Stage

Initially we focus on the group stages of the tournament, during which 24 matches took place. Figure 2 displays the duration of calls during the group stage (dotted line) against the daily hourly averages during non tournament period (grey continuous line). Additionally we encircle points at which a match begins, with red marks indicating matches involving the Ivory Coast. In the work carried out by [13], they noticed an increase in call activity leading up to and after several Brazilian football matches and a clear decrease in calls during matches. With this in mind we expect to detect similar patterns here, which would indicate interest in certain matches.

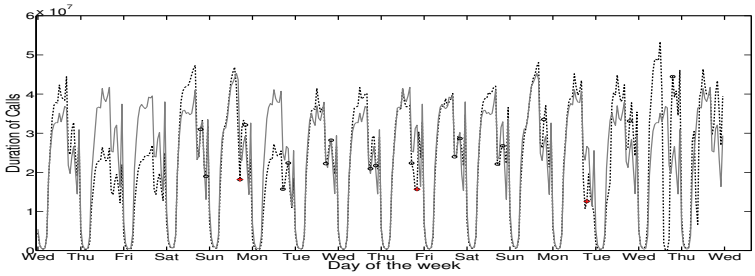


Fig. 2: Daily Average vs Group Stage Call Duration

Looking at Figure 2, we notice that there are several occurrences of anomalous activity through the period, yet activity surrounding the beginning and end of a day seem to somewhat match up well. Firstly, there appears to be a noticeable pattern occurring around matches involving the Ivory coast, which can be seen as a small spike hours before the game immediately followed by a severe drop in activity, far below average, as the game commences. One would assume this is due to fans initially conversing over the game prior to kick off, where they then cease any communication throughout the match, after which they initiate some bursty communication at the end. In contrast for non-Ivorian matches we do see small reductions in activity around game time but nothing significant. Furthermore in some cases we see an actual increase in activity during a match, which would lead us to believe there exists an alternate form of interest regarding these games.

An area which is of particular interest appears at the end of the period involving the final set of group stage matches, as seen in Figure 3. On this day we witness an early morning rise in activity, peaking at mid day, where it then drops to near zero over a span of four hours. A second peak then emerges

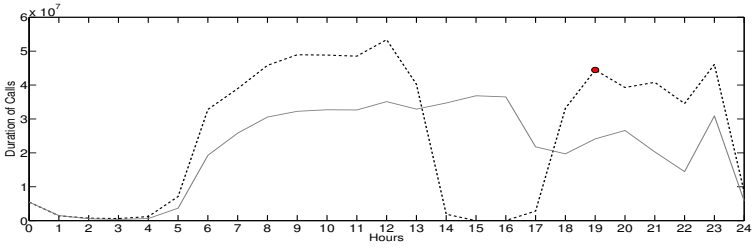


Fig. 3: Extreme Behaviour at Final Group Matches

coinciding with the kickoff of the final game. This game is of obvious concern to football fans, as three of the four teams involved in these games are neighbouring countries of the Ivory Coast, so one would expect a fan base for those countries living within the Ivory Coast. In addition to this, one of the winning teams could possibly be competing against the Ivory Coast in the next set of games.

5.2 Finals

Next we move onto the quarter-finals, semi-finals and grand final which account for 8 matches. Similar to what we had seen in the group stages, Figure 4 sees matches involving the Ivory Coast exhibit distinct below average activity prior to kick off. Furthermore, due to the importance of these knock-out matches, we notice increasingly prominent peaks and troughs around match periods as the final stage progresses, particularly at the final. In addition to this we again witness peaks in activity occurring at kick-off times for some matches. Regarding the final, if we consider the late hour at which it took place, the presence of an uncharacteristic peak occurring post match clearly depicts significant interest in the outcome. On a small note, there appears a clear spike in activity on the day following the final which can be explained by a parade held in Abidjan to celebrate the return of the national football team. As you would expect this would account for the above average activity spanning several hours.

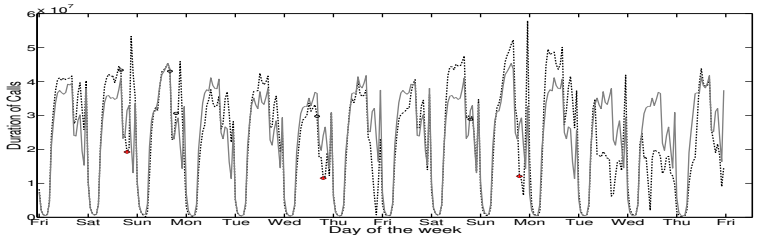


Fig. 4: Daily Average vs Finals Call Duration Contrast

Given these noticeable and reoccurring patterns which appear around match times, we can confirm a strong *interest* in the outcome of the tournament. Another interesting observation is the existence of two opposite forms of interest, which occur at match kick-offs. These two-forms, one a decrease in activity, the other an increase, can be described as *emotional interest* and *social interest* respectively. With the *emotional interest* we see a large decrease in communication during the event, as people don't want to be disturbed. Whereas with our *social interest* we see the opposite: an increase in activity is apparent during the match as everyone is in constant communication, presumably, conversing over the events of the match.

We will look to track the existance of this two in our following regional analysis.

6 Department Behaviour

Taking our analysis one step further, we narrow our focus onto the individual departments within the Ivory Coast. As mentioned in Section 3, the Ivory coast is divided up into 19 regions which are themselves further divided into 81 departments. This analysis will ascertain which department, if any, displays irregular activity during t_T which could possibly indicate an interest in the proceedings of the tournament. In addition to this, we would like to determine what form of interest, social or emotional, each department predominantly exhibits. Moreover, from a network perspective we are interested in clustering the departments together based on their call activity. In doing so, we wish to shed some light on similar behavioural traits displayed between departments.

Fig. 5: Departmental activity

Prior to any analysis we need to create the individual department data, as seen in Figure 5, by mapping the antenna locations within the Ivory Coast, separating each by department borders, and accumulating the communication within each over an hourly basis. Doing this not only provides us with individual department activity, but also inter-department activity, which may be of some importance later. Taking just the total call duration, we apply this conversion and are left with the call duration activity of 50 individual departments which, for the sake of this section, we visualise as time series data. Our next step

is to normalise each department's time series, as the magnitude of each will differ given the varying populations within each department. To do this we take each departments time series and subtract its mean and divide by its standard deviation.

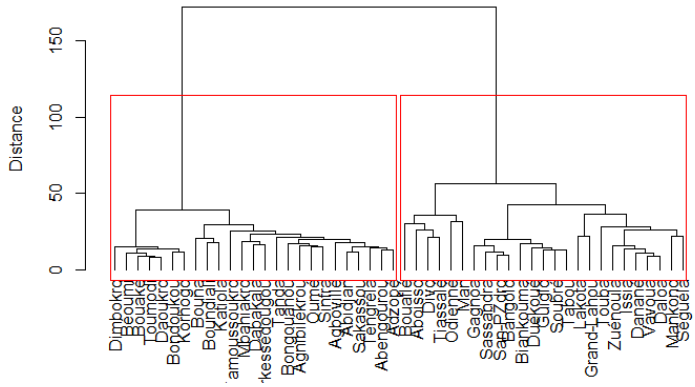


Fig. 6: Hierarchical clustering of regions based on normalised duration of calls

We begin with our networking perspective, where we look to cluster the regions based on the similarity of their time series. This is done by applying a standard Euclidean distance measure so as to acquire a similarity matrix which we pass into a hierarchical clustering method using a Ward's minimum variance linkage method. The resultant dendrogram, seen in figure 6, displays two clear clusters separating the departments. We confirm the appropriate number of clusters (see Figure 7) using the sum squared distance between each member of a cluster and its cluster centroid. We choose the cluster number at which the reduction in SSE slows significantly, which in our case confirms our two cluster split.

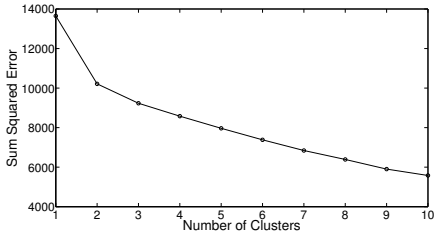


Fig. 7: SSE for a number of cluster's

If we plot these two clusters onto a map of the Ivory Coast (see Figure 8), we find a clear split of the country into its east and west halves. This geographical divide correlates with the ethnic divide throughout the entire country as seen in Figure 9 [1]. The divide between the Beoule / Senoufo and Malinke / Bete ethnic groups follows a strikingly similar outline, excluding Aboisso in the south east.

Fig. 8: Similarity clustering departments from call duration activity

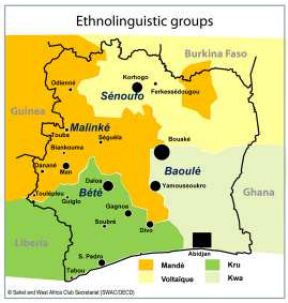


Fig. 9: Ethnic groups within the Ivory Coast

Moving on, our specific interest in this section is to acquire a quantitative score regarding the *interest* each department shows towards the tournament. To achieve this we take a measure of how each department's call activity deviates from their daily average around game time. As we alluded to at the end of Section 5.2, there exists two forms of interest which are apparent at kick-off time, one characterised by a decrease in call activity (*emotional interest*), the other by an increase (*social interest*). Before we go about calculating each of these measures there are a few issues that need to be addressed, in particular:

- how long a period do we take around a game to measure interest?
- how do we account for the lengthly low activity period?

Regarding the length of period around a match, we can use our analysis from Section 5 to ascertain the general activity around these hours. As we have previously mentioned there are cases where a spike can be seen hours leading into and/or out of a match. This also includes peaks and troughs appearing at kick off times. We define $M_i(b,d,a)$ to represent the interest period around the i^{th} match, taking into account the length in hours, before b, during d and after a the match. We already know the length a match takes, so that we can safely set $d = 2$. This takes into account each half of football, injury time and half time team talks. As for b and a, some caution must be taken when choosing these values, to avoid overlapping interest periods. We know from the tournament layout that match times are set out in such a way that there is only a three hour gap between the starting points of two successive games.

[1] http://www.oecd.org/swac/someelementsoftheivoriancrisis.htm

Fig. 10

Using an example (see Figure 10), we can see an instance of two successive games. The green line here represents the interest periods M_i and M_{i+1}. In order to avoid overlap and to assign equal interest periods to every match, we set $b = 1$ and $a = 1$. A low-activity period in the data corresponds to a period of time in which the call volume is significantly lower than the average. Low activity periods are present in the data for a number of different departments on multiple occasions and they can typically span large portions of a day. At the time of writing, we have not been able to explain the reason behind these low activity periods but are confident that they are independent of the tournament activity that forms our primary interest. As we wish to detect interest, through a comparison of call activity with the average activity in the period around match events, to avoid false positives, we ignore any period in which the total sum of data within a period is below the total sum of the average data with n the period.

Our interest scores, which measure the difference in activity between tournament and non tournament periods, are accumulated over the 28 match event within the tournament, as such we set $n_m = 28$. We define $\mathbf{v_{m,c}}$ to be a vector holding the hourly call duration contained within the match interest period M_m of match m for city c. We also define $\bar{\mathbf{v}}_{\mathbf{m,c}}$ to be the average hourly call duration expected, outside the tournament period, for the same hour range and weekday as the match interest period. As an example, if we take a match period on a Monday from 4-8pm, we set $\mathbf{v_{m,c}}$ to take the hourly call duration for those four hours. Accordingly we will set $\bar{\mathbf{v}}_{\mathbf{m,c}}$ to be the average hourly call duration expected on a Monday from 4-8pm during an average day.

Before any interest score is calculated we first obtain a vector holding the difference in call duration $\mathbf{d}$ for all match events during a chosen match interest period M_m. All interest measure calculations are conducted using the original individual department call duration data created from figure 5. The reason for this is because normalising it prior to our calculations would disrupt our difference measurements between $\mathbf{v_{m,c}}$ and $\bar{\mathbf{v}}_{\mathbf{m,c}}$. Taking this into account we let,

$$\mathbf{d} = \sum |(\mathbf{v_m} - \bar{\mathbf{v}}_{\mathbf{m}})| \qquad \forall m = 1, .., n_m$$

To calculate the overall interest score for city i we first calculate $\mathbf{d_i}$ using our match interest period $M_m = M_m(1, 2, 1)$. We then normalise our result by summing $\mathbf{d_i}$ and removing the mean and dividing the standard deviation,

$$total_i = \frac{\sum \mathbf{d_i} - \bar{\mathbf{d}}_\mathbf{i}}{std(\mathbf{d_i})}$$

To calculate the emotional and social interest scores we now take $M_m = M_m(0, 2, 0)$ i.e. we focus specifically on the two-hour period during which the match is taking place. For the emotional score, we look for call levels that are smaller than normal, whilst for the social score, we look for call levels that are larger than normal. Specifically, for each city i, we again calculate $\mathbf{d_i}$ using our new match interest period and define

$$emotional_i = \frac{\sum \mathbf{d_i} - \bar{\mathbf{d}}_\mathbf{i}}{std(\mathbf{d_i})}, \qquad \mathbf{d} = \sum \max(\bar{\mathbf{v}}_\mathbf{m} - \mathbf{v_m}, 0) \qquad \forall m = 1, .., n_m$$

$$social_i = \frac{\sum \mathbf{d_i} - \bar{\mathbf{d}}_\mathbf{i}}{std(\mathbf{d_i})}, \qquad \mathbf{d} = \sum \max(\mathbf{v_m} - \bar{\mathbf{v}}_\mathbf{m}, 0) \qquad \forall m = 1, .., n_m$$

We now provide results for social, emotional and total overall interest using the process described above. It should be noted that throughout this analysis we are not particularly interested in what city shows the best overall interest, but more on how close are the scores and what similarities are there between the top cities.

6.1 Social Interest

In relation to social interest which, as we explained earlier, is concerned with events for which peaks in activity during a match are observed, we highlight the top 10 cities as referenced in table 2. From this we notice that apart from Duekoue and Abidjan at the top, the interest scores are somewhat close and when we visualise them on a map there are two small clusterings of departments in the south and west of the map beside Liberia. Besides this, there are also two outliers Tengrela near Mali and Daoukro near the border to Ghana.

City	Interest
Duekoue	49.842677
Abidjan	38.877667
Zuenoula	31.143590
Daloa	30.595430
Sinfra	29.733157
Divo	28.852031
Guiglo	28.142372
Daoukro	27.145546
Lakota	26.500503
Tengrela	25.485879

Table 2: Top 10 cities for social interest

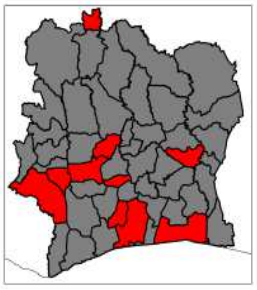


Fig. 11: Top 10

6.2 Emotional Interest

Now moving onto emotional interest results, table 3 demonstrates a close proximity of scores for the top 10. In this case, Figure 12 displays a tight clustering of departments in the centre of the country around the capital. Again we see two outliers in the form of Korhogo and Man. This compact clustering is situated around highly populated departments, that display a keen interest in this tournament.

City	Interest
Bouafle	19.174448
Toumodi	19.141005
Man	19.052178
Korhogo	18.445232
Dimbokro	17.832876
Daoukro	17.814426
Oume	16.946766
Bouake	16.892782
Divo	16.796226
Bongouanou	16.084102

Table 3: Bottom 10 cities for emotional interest

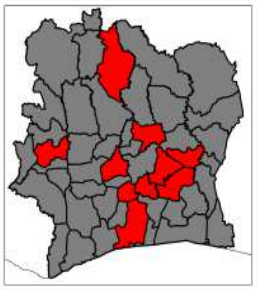


Fig. 12: Bottom 10

6.3 Total Interest

Following on, we now present our scores for the total overall interest during the tournament stage. We do this by supplying a ranked list of both the top and bottom 10 cities. First, when looking at the top 10 departments, we notice an overlap with our results from table 4. There exists two clusterings of interest groups, one in the south around Abidjan, the largest and most heavily populated city, and the other in the north around Mali and Burkina Faso. The cluster in the north can be explained by the highly populated departments and from the tournament perspective in that Mali successfully progressed to compete against the Ivory Coast in the semi finals.

City	Interest
Korhogo	65.033635
Agboville	61.522172
Odienne	61.006107
Man	57.054510
Sinfra	53.414109
Ferkessedougou	52.999377
Abidjan	52.069859
Zuenoula	48.838793
Tengrela	48.055909
Divo	43.934015

Table 4: Top 10 Cities total interest

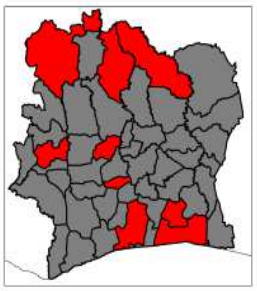


Fig. 13: Top 10

With the clear overlap between the total and socially interested departments, this would lead us to relate social interest with high overall interest. Or to

describe it differently, departments socially interested in matches or teams are more inclined to have a greater interest over the total period of the tournament. This seems realistic considering that areas emotionally interested are primarily focused on their own matches and possible show little interest in other matches. In contrast to this, the bottom 10 ranked cities display a tighter knit cluster located across the middle portion of the country. It is noticeable how several of these departments border the capital, yet are ranked so low in total interest. Lastly we notice Aboisso, showing the lowest score, is clearly seen as an outlier.

City	Interest
Bouafle	25.417656
Touba	23.768517
Seguela	23.334414
Sassandra	22.388688
Mbahiakro	21.057811
Yamoussoukro	19.550397
Dabakala	16.934076
Biankouma	16.710743
Mankono	10.456541
Aboisso	9.941129

Table 5: Bottom 10 Cities total interest

Fig. 14: Bottom 10

7 Regional Development

Using our results from Section 6 we have discovered specific cities which shown significant interest in matches throughout the African Cup of Nations. Moving forward with these findings we now wish develop as strategy as to how one can utilise these results to implement a plan to develop the sporting infrastructure within certain regions. To be specific, the results we are taking away from our department behaviour analysis are those pertaining to the total and social interest scores. Considering these results, we focus on the two clusters found in the figure 13 which, as we mentioned earlier, have a clear overlap to those found in figure 2. For reference purposes, we assign labels to these clusters, with the first residing in the south around Abidjan, $clust_s$, and the other to the north bordering on Mali and Burkina Faso $clust_n$.

Although these two clusters contain departments ranked in the top 10, they share very different characteristics. The first clear distinction is the amount of land mass taken up by each cluster, with $clust_s$ half the size of $clust_n$. The second notable distinction relates to the difference in population density between the two clusters. As we know Abidjan is the highest populated city in the Ivory Coast, and as such there is a drop off in population density per square kilometre as we radiate outwards from it. The lowest of these values can be seen in the north of the country, where general infrastructure is poor.

Considering these issues, our proposal as detailed below, for one cluster may not prove equally viable for the other. Starting with $clust_n$, centred around Abidjan, we find this area is somewhat more developed than other places, having

seen more financial investment in the area to develop and improve transportation and flood protection [2]. We would propose the construction of several sport grounds with multiple functions, to cater for other sports activities within the regions. These grounds can be used as a venue for football skill schools for the younger generation that would hopefully increase their interest in the sport and be utilised to increase their interest in the sport and as locations for friendly matches, local regional and national. Consideration in this regard should be given to the creation of sports scholarships to further promote the popularity and love for football. Similar methods can be put forward towards $clust_n$ and if we consider the countries in close proximity, Mali, Guinea and Barkina Faso, who themselves share the same love for football, there is no reason why this sport would not flourish to the extent as it does in the southern region of the Ivory Coast.

Our hope is that through development of these clusters and potentially other less interested departments, we wish to see the creation of tournaments between cities or regions which would help alleviate hostilities which have arisen as a result of past and recent civil wars.

8 Conclusion

We propose a methodology to identify and extract specific interest metrics demonstrated by cities during the African Cup of Nations football tournament. This process was applied to call detail records extracted from the Orange customer base within the Ivory Coast. Our results show that using information from these records, specifically call duration, we are able to split the nation into two clusters which show strong similarity to the ethnic divides that exist within the nation. In terms of interest measures, using our own proposed definitions of interest, both social and emotional, we have shown that certain clusters exist in separate parts of the country, either around highly populated cities or bordering on other footballing nations. Finally, looking at the total interest measured over the tournament, we discovered two clusters, one located around the largest populated city within the Ivory Coast and the other bordering on nations to the north. These specific clusters have been targeted as areas suitable for sporting development.

References

1. L. Akoglu and C. Faloutsos. Event detection in time series of mobile communication graphs. In *Proc. of Army Science Conference*, 2010.
2. J. P. Bagrow, D. Wang, and A.-L. Barabási. Collective response of human populations to large-scale emergencies. *CoRR*, abs/1106.0560, 2011.
3. V. D. Blondel, J. loup Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks, 2008.

[2] maps.worldbank/afr/cote-divoire

4. J. Candia, M. C. Gonzlez, P. Wang, T. Schoenharl, G. Madey, and A.-L. Barabsi. Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical*, 41(22):224015, 2008.
5. A. Culotta. Towards detecting influenza epidemics by analyzing twitter messages. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pages 115–122, New York, NY, USA, 2010. ACM.
6. N. Eagle, A. S. Pentland, and D. Lazer. Inferring social network structure using mobile phone data, 2008.
7. E. Frias-Martinez, G. Williamson, and V. Frias-Martinez. An agent-based model of epidemic spread using human mobility and social network information. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pages 57 –64, oct. 2011.
8. J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, and L. Smolinski, Mark S. Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, 2009.
9. M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, June 2008.
10. H. N. C. B. McGowan D, Brew A. Churn prediction in mobile tellecommunications. page 18, 2011.
11. C. Song, Z. Qu, N. Blumm, and A.-L. Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, Feb. 2010.
12. D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A.-L. Barabasi. Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 1100–1108, New York, NY, USA, 2011. ACM.
13. F. H. Z. Xavier, L. M. Silveira, J. M. d. Almeida, A. Ziviani, C. H. S. Malab, and H. T. Marques-Neto. Analyzing the workload dynamics of a mobile phone network in large scale events. In *Proceedings of the first workshop on Urban networking*, UrbaNe '12, pages 37–42, New York, NY, USA, 2012. ACM.

Rapid Assessment of Population Movements in Crises: The Potential and Limitations of Using Nighttime Satellite Imagery and Mobile Phone Data

Nita Bharti[a], Xin Lu[b,c,d,e], Linus Bengtsson[b,c], Erik Wetter[b,f], and Andrew Tatem[g]

[a]Department of Biology, Center for Infectious Disease Dynamics and Huck Institute of Life Sciences, Penn State University University, University Park, PA 16801, USA; [b]Flowminder Foundation, 17177 Stockholm, Sweden; [c]Department of Public Health Sciences, Karolinska Institutet, 17177 Stockholm, Sweden; [d]Department of Sociology, Stockholm University, 10691 Stockhom, Sweden; [e]Department of Information Systems and Management, National University of Defense Technology, 410073 Changsha, China; [f]Department of Management and Organization, Stockholm School of Economics, 11383 Stockholm, Sweden; [g]Geography and Environment, University of Southampton, SO17 1BJ, UK

Introduction

Societal instability and crisis events have been shown to prompt rapid, large-scale human movements. These movements are poorly understood and difficult to measure but can strongly impact health and access to resources for vulnerable, displaced populations in areas where they resettle. Data on these types of movements are rare but particularly important to inform emergency response planning to increase access to vital resources during humanitarian crises.

Previous research on this topic has largely been confined to surveys of small groups of people, such that the scale of displacement during crises is often unknown. Two recent approaches to measuring population densities and flows are nighttime satellite imagery and anonymized mobile phone call record data (CDR). Here we make a first attempt at applying these methods to a humanitarian crisis caused by internal strife. The results indicate that both methods, especially when used together can indicate population movements and thus be used as a planning tool by relief and development agencies. We conclude with some limitations and challenges as well as suggestions for further research and methodological development.

Background

In the aftermath of the presidential elections of 28 November 2010, armed conflict broke out in Côte d'Ivoire [1]. A total of almost one million persons are estimated to have been displaced at the end of March 2011, many relocated across international borders [2].

Large parts of Côte d'Ivoire were stabilized after the military victory of the UN-backed pro-Ouattara forces in April 2012, but the return to normality was a slow process. As security improved, UNHCR and IOM facilitated the return of refugees and internally displaced persons (IDPs) to the West. Large numbers of returning IDPs are believed to have come back without external assistance, although valid data on such return movements across large areas are extremely difficult to collect [3, 4].

The extent to which IDPs return to their original area of living after a conflict has ended provides vital information for the allocation of resources to returnees and to understand how the population perceives living conditions and security in the return area. Here, we study the movement patterns of a random sample of half a million Orange mobile phone subscribers during December 2011 to April 2012, to investigate to what extent mobile phone operator data can provide information on returning IDPs in the heavily affected areas of west Côte d'Ivoire. We contrast our analyses of the mobile phone data with data from interviews of UN staff in Abidjan, Côte d'Ivoire and as well as data and reports on population return movements from IOM, UNHCR and UNOCHA [5].

Crisis Movement 2011

The unrest cause by the onset of the second civil war in Côte d'Ivoire led hundreds of thousands of Ivorians to seek refuge in neighboring countries to escape violence. As the instability subsided in late 2011, people began to return to Côte d'Ivoire. The movement patterns of these individuals are extremely important to understand; as they settle in new locations or return to former towns, they determine the basal level of demands on health care and education facilities. As the country moves forward, understanding the movements and social networks of its people becomes increasingly important in ensuring

the proper supply and access to services and goods. The D4D Challenge (http://www.d4d.orange.com) presents a unique opportunity to analyze mobile phone usage data to understand movement patterns in Côte d'Ivoire following a crisis, during a time of rebuilding. We augment these detailed phone usage data with satellite imagery of anthropogenic light (following approaches outlined in Bharti et al [6]) to gain a broader understanding of the movement patterns in this area over a longer period of time and across national boundaries (details in Methods section).

Methods

Phone usage data

Orange has provided four data sets of phone usage spanning from December 2011 to April 2012. Below, we describe the two datasets that were used for this study.

Anonymized cell phone data has been shown to be a cost efficient proxy indicator for population displacement following the Haiti 2010 earthquake [7, 8]. Here the cell phone data were used to study displacement in the context of social instability, and thus the absence of localized single events to induce large population movements.

Aggregate communications between mobile phone towers are used to determine the availability of resources between regions and subprefectures. Similarly, the aggregation between cell towers is calibrated to stable nighttime brightness values from satellite imagery (see next section for information on the data). Understanding how phone usage is related to composite brightness values from satellite imagery provides a sustainable tool for measuring human presence that goes beyond the availability of the cell phone usage data. This is a useful tool for understanding resource availability and distribution as well as regional conflicts.

Mobility traces based on cell phone data are used to understand changes in urban populations. This data set includes phone usage by tower for 500,000 individual users for two weeks each, collectively spanning from Dec 2011 to April 2012. Displacement from and return to urban settlements is known to have occurred during this period of instability

and this may be reflected in the volume of phones using urban towers over time. We also look at the corresponding values of light emissions for these cities in nighttime satellite imagery. By comparing these two data sources, we attempt to overcome some of the biases inherent to each data set. This analysis looks at timing and magnitude of displacement and return, during the availability of the phone usage data and beyond this time period with the use of satellite imagery.

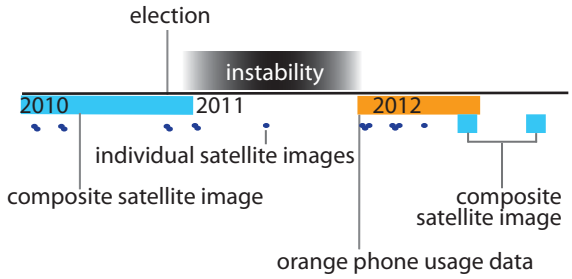
Satellite imagery

To understand changes in populations and movement in Côte d'Ivoire and neighboring areas, we measured changes in satellite-derived anthropogenic nighttime light emissions, a direct indicator of human presence [9] from the years preceding the civil war, during the civil war of 2011, and during the recovery period.

Operational Linescan System (OLS) instruments onboard Defense Meteorological Satellite Program (DMSP) satellites can detect areas of anthropogenically derived light sources (electric lighting and fires), which indicate the presence of human settlements. The spatial resolution of DMSP OLS (~1 km or 0.00833 decimal degrees, resampled from the 2.7-km native resolution at the SPIDR) imagery permits analyses of changes within cities, which are not often possible from reports of displaced persons. DMSP-OLS 2010 composite data are used to identify stable lights for defining urban areas and Visible Infrared Imaging Radiometer Suite (VIIRS) composite data [10] are used to measure stable lights for settlements in 2012. Observations in the 2012/4/18-26 and 2012/10/11-23 time periods were used. Cloud screening was done based on the detection of clouds in the VIIRS M15 thermal band [10].

For non-composite serial images, DMSP satellites provide georeferenced visible and thermal-infrared images to visualize nighttime lights and detect cloud cover. The processed, images, obtainable from the Space Physics Interactive Data Resource (SPIDR) [11], are captured twice a day; once during the daytime and once at night. To assess nighttime light brightness, individual DMSP-OLS images from the F18 satellites are used for this study, providing images from before, during, and after the period of instability.

Images are acquired from the National Oceanic and Atmospheric Administration (NOAA) National Geophysical Data Center and are screened to remove environmental elements that can contaminate brightness measurements. To avoid contamination from lunar illumination, images captured during bright moon phases are avoided. To avoid solar contamination and reduce the impact of variability in human behavior (extinguishing fires and lights while sleeping), we use images that were captured between 7 p.m. and 10 p.m. local time. Cloud contamination can also affect brightness measurements and we examine each TIR image and select only images that are free of cloud cover over all pixels of the area of interest (conservative numerical threshold value of 200). We extract brightness values for each pixel in the area of interest. For each image, the brightness value is measured for each pixel (as in Agnew *et al*., 2008 [12]).



Schematic 1. Timeline from 2010-2012 of events (above line) and proxy measures of human presence and movement from two data sources.

Here, we measure areas within Côte d'Ivoire for decreases in brightness due to displacement, increases in brightness due to repatriation, and areas across the national border in Liberia where hundreds of thousands of refugees fled during the instability.

We have previously demonstrated the use of satellite imagery of visible lights in public health by establishing a strong relationship between seasonal changes in urban populations and infectious disease transmission in three cities of Niger, West Africa [6]. Here, we adapt and apply this method to understand the displacement due to a civil war in Côte d'Ivoire and the return and redevelopment of the country following the instability. Although satellite imagery has been used to estimate sizes of refugee camps and other displaced populations [13], this is the first time that satellite images of light emissions have been used to understand rapid displacement during a humanitarian crises and the subsequent redevelopment of a nation.

Results

Coarse correlation between data sources

To calibrate the levels of light emissions and phone usage, we compared a five month period of data on aggregate communication between cell towers at the regional level and compare it to the regional levels of mean brightness values from composite satellite imagery from 2010 and to approximately three weeks of 2012 (see schematic 1). We found a strong positive correlation between the 2010 brightness values and the phone usage data (cor=0.93 at regional level, cor=0.56 for subprefectures), indicating that lights and phones are being used similarly in Cote d'Ivoire (Fig 1A,B). We also found a strong positive correlation between the 2012 brightness values and the phone usage data (cor=0.84 at regional level, cor=0.54 for subprefectures, not shown).

We then compared spatial cell tower density with stable areas of visible nighttime lights, indicating permanent settlements. We found that cell towers tended to be present where stable settlements were found across the landscape (Fig 1C). High spatial agreement between areas with consistent, detectable light emissions, and cell towers indicates overlapping resource availability. To measure this relationship, we plot distance to nearest tower against pixel brightness to reveal a negative correlation (Fig 1D, cor = -0.41), indicating that pixel brightness and distance to nearest cell tower have a negative relationship. Towers are placed in unlit locations to maintain network coverage along major roads, which may explain the variation in distance to towers for low-lit pixels.

Based on these results, we confirmed that these two data sources were likely measuring proxies of similar human presence and movement and we proceeded with higher resolution comparisons between the two data sets.

Urban areas by region

The election that led to the civil war divided the country geographically, coarsely separating the north from the south, though at high spatial resolution the south contains areas of support for both candidates. To measure the change in population in the cities of Côte d'Ivoire against the election results, we looked at brightness level of urban areas

across the country, beginning one year before the election and ending a few months after the war was officially declared over. The brightness levels in the south, primarily Gbagbo supporting cities, appeared more stable throughout this time period (Fig 2, black), while the northern cities, primarily supporting Outtara, show large fluctuations in brightness (Fig 2, red), in some cases experiencing decreases of nearly to 80%, as in the case of Tingrela.

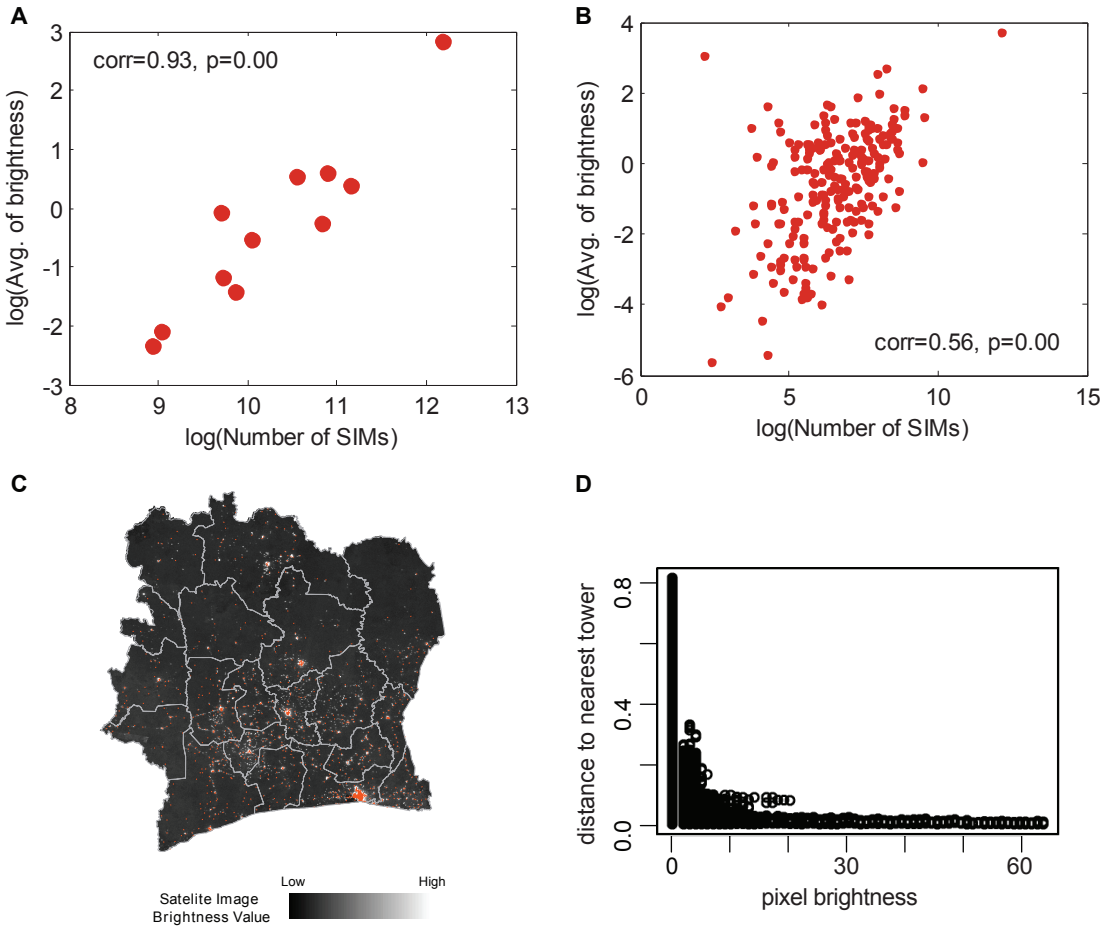


Figure 1. Correlation between number of cell phone users and the average night satellite image brightness value. A) For each of 11 regions in Côte d'Ivoire, the number of phones present from December 2011-April 2012 against the mean brightness value of each region from a 2010 composite image. B) Same as A for 255 subprefectures. C) Map of locations of cell towers (each orange point is a tower) overlaid on areas of stable brightness (white indicates bright areas, black indicates dark areas) from a 2012 composite image (VIIRS). Grey polygons outline 11 regions plotted in A. D) Tower distance and pixel brightness; the distance to the nearest cell tower decreases as pixel brightness increases.

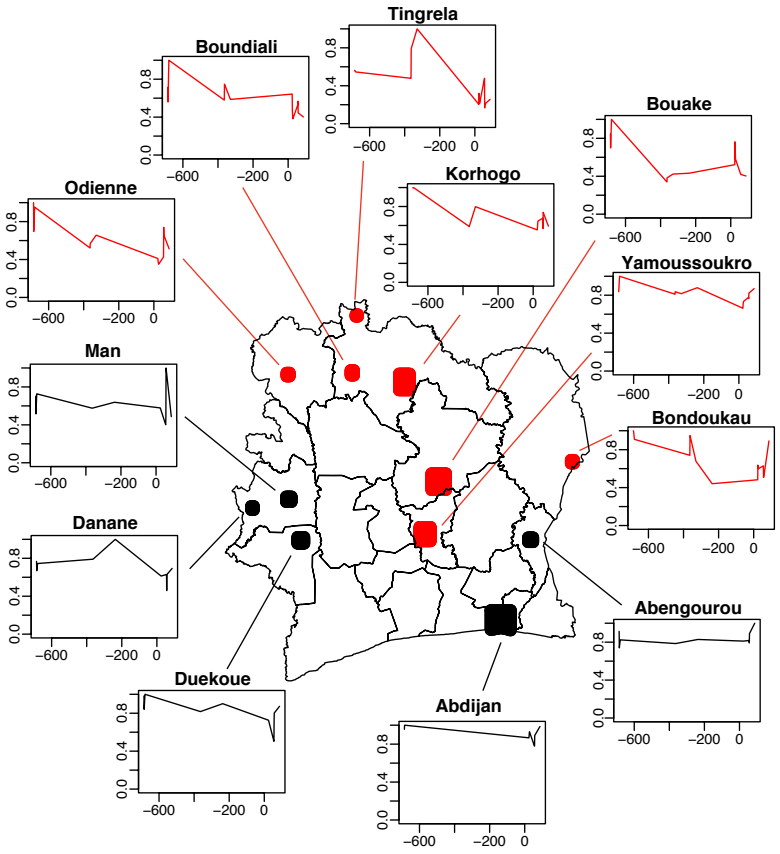


Figure 2. Map of Côte d'Ivoire showing changes in brightness levels for twelve cities. Colors indicate the election results (red primarily supported Outtara; black primarily supported Gbabgo). Y-axis on all plots is 0 to 1 and x-axis indicates date of image capture; Dec 1 2011 is day 0, indicating the start of the phone usage data. Instability occurred through most of 2011.

Refugee camps

Although the phone usage data are highly detailed in both spatial and temporal resolution, they do not provide information beyond national boundaries. In this particular case, we are interested in locating the hundreds of thousands of refugees who crossed Côte d'Ivoire's borders to neighboring countries during the time of instability. According to UNHCR reports, Liberia received the greatest number of refugees, recording over 175,000 Ivorian refugees at the height of the instability (Fig 3A). The majority of these refugees went to the Liberian counties of Grand Gedeh and Nimba. Using ten individual nighttime light satellite images, we measured the mean brightness levels of each of those two Liberian counties before, during, and after the civil war in Côte d'Ivoire (Fig 3B,C). We found that the increase in recorded refugees crossing the border into Liberia and the

increase in brightness for Grand Gedeh and Nimba occurs at the same time. Although the increase and decrease of brightness and recorded refugees occur at the same time, the magnitude of these changes following the period of instability do not match precisely.

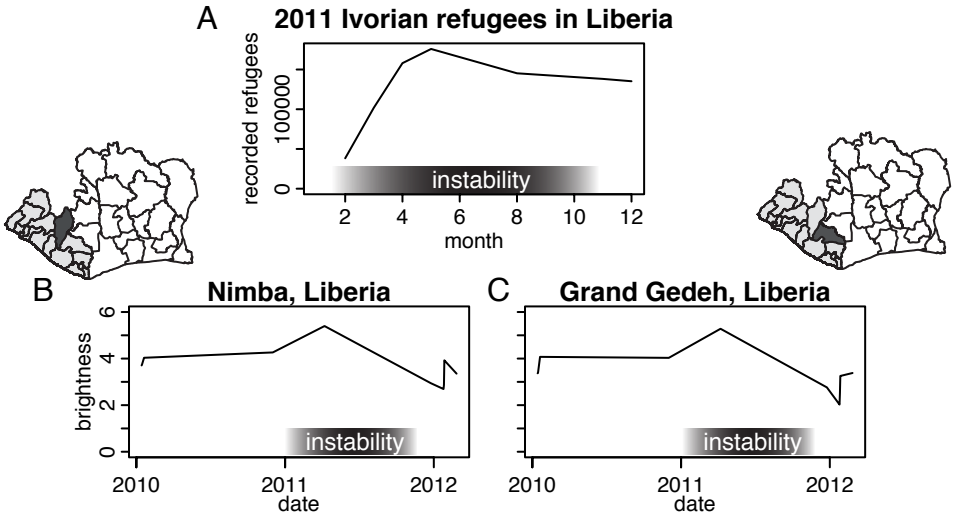


Figure 3. A) UNHCR recorded numbers of refugees by month of 2011 crossing the border from Côte d'Ivoire into Liberia. B) Brightness values from Nimba, Liberia and C) Grand Gedeh, Liberia from 2010-2012; darkened area on inset maps indicate location of each county; time of conflict noted along x-axis in all three plots.

Discussion

Coarse correlation between data sources

In measuring human presence and movement via a proxy, as we did here with both phone usage data and anthropogenic nighttime brightness values, it is important to understand the relationship between the usage and availability of the resources we are measuring. It was surprising that at larger spatial scales and therefore lower spatial resolution we found higher correlation between the phone usage and composite brightness data sets. It is possible that this level of spatial merging reduces noise and outlier measurements, as the regional data are broadly representative of areas with similar poverty indices. This observation requires further consideration to eliminate the possibility that these statistics are averaging across meaningful variation yet resulting in statistically significant correlation values.

The strong similarity in tower placement and stable settlements is not at all surprising and is deliberate such that areas with the largest demand for phone usage will have towers to supply sufficient service. These are likely to change in response to each other to remain similar over time. It does not appear that the conflict affected the locations of stable settlements seen in 2012 away from tower locations.

Urban areas by region

Some of the most staggering examples of violence come from the western parts of the country, particularly the March 29 2011 Duekoue massacre, which killed an estimated 800 to 1000 people and displaced tens of thousands [14] [15], and the battle of Abidjan a few days later, when heavy fighting occurred in the city and the UN intervened. Given these reports, it was surprising that we did not see a dramatic decrease in brightness in these regions corresponding to these large-scale events. However, for many of these urban conflicts, the internally displaced persons reportedly sought shelter at local camps and churches by the thousands. Additionally, many UN IDP camps were located at the outskirts of the big cities during this civil war. These aggregations of displaced person would emit anthropogenic light indistinguishable from a previously stable settlement. In a promising advance towards measuring these changes, the high resolution of VIIRS makes it possible to detect smaller spatial scale changes than was previously possible, particularly those on the edges of settlements, which have traditionally been difficult to distinguish from light ‘blooms.’

Refugee camps

We measured the changing brightness levels of border counties in Liberia to detect the presence of refugees crossing the border from Côte d'Ivoire. Although the timing of the increase and decrease of refugee presence and abundance in Liberia matches the change in brightness for Grand Gedeh and Nimba, the magnitude does not. As the instability draws to an end, refugees return to Côte d'Ivoire and brightness levels in the two counties with refugee camps decreases. However, while many refugees appear to have remained in Liberia until at least December of 2012, the brightness values of the counties with refugees decreases to below pre-instability levels.

Summary

The approach detailed here couples two powerful and technologically advanced data sources to address movement and displacement in conflict areas. Both offer near-realtime situational awareness, as well as historical time series which allow for baseline measurements and pattern identification. While very promising, each of these data sources has its own biases and limitations.

Temporal and spatial limitations

In this particular study, our data were limited by timeframe and availability. Temporally, the phone usage data were available following the conclusion of the civil war, permitting analyses of resettlement and return patterns, but phone records were not available during the displacement events themselves. The satellite image availability in this region was also limited, as cloud cover was frequently present along the coastal areas of Côte d'Ivoire. As an unfortunate result, temporally overlapping data between phone usage and satellite imagery were scarce, complicating our calibration attempts. One approach to overcome this problem is to use composite satellite imagery; these products were available for 2010 and 2012 (partial) but were lacking entirely for 2011 and were not comprehensively available for 2012. As additional satellites continue to capture this type of imagery, composite products may become more rapidly available but population changes in areas with consistent cloud cover will remain challenging to measure with nighttime satellite imagery.

Spatially, the phone usage data capture very high-resolution movements but here do not capture movement across international borders, as complementary roaming data would be needed. In this case, we were particularly interested in the large-scale movement of refugees into neighboring Liberia. We were able to use satellite imagery to assess population changes in these locations, though this analysis would have been improved with more frequent measures from images While these images are very useful for rapidly detecting large changes in population density or human presence, it is impossible to track individual movement patterns with these images. Therefore, very small villages and

nomadic populations cannot be measured in these images and very short distance movements cannot be detected.

Availability and accessibility

Cell phone usage continues to increase in under-resourced areas as people rely on mobile technology where infrastructure is lacking. With rising usership, phone usage data become increasingly informative and less biased by wealth. The high spatial and temporal resolution of phone records provides detailed information that cannot be matched by other data sources. According to our interviews, it is important to be aware that networks from time to time might not be operational in areas with violent conflicts, due to power outages or forced shutdowns by fighting parties which could be a source of bias in conflict regions. Additionally, due to privacy concerns and data ownership, the phone companies determine the availability and distribution of phone records, which are divided among competitor companies in many markets. With no monetary incentives to share phone records, the availability of phone data are currently dependent on the goodwill of phone companies, and thus will vary from country to country. Going forward, an area of great potential is to map cross-border movement using cell phone data, which will require more complex agreements and coordination with multiple operators.

High resolution, frequently captured nighttime satellite images from different sensors are publically available in near-real time but remain sensitive to environmental factors, particularly cloud and light contamination. Areas with consistent cloud cover remain difficult to visualize in nighttime satellite imagery even as image resolution increases, through sensors such as VIIRS. Light emissions are also biased by wealth; GDP and brightness are strongly correlated [16] [17], complicating efforts to calibrate brightness and population size.

In measuring proxies for human presence and movement, phone usage data and nighttime satellite imagery of visible lights are highly complementary. These data sources can augment each other and enhance our understanding of the biases inherent to each measurement. In this study, we were able to gather data beyond the spatial and temporal

restrictions imposed by each data set by sampling from the other. This approach is very promising for future applications. As we move forward, we aim to develop a formalized method for combining complementary data sources like these.

References

1. United Nations Operation in Côte d'Ivoire. *Post-election crisis 2013* 2013 Februrary 2013]; Available from: http://www.un.org/en/peacekeeping/missions/unoci/elections.shtml.
2. UNHCR. *Côte d'Ivoire on the Edge. A New Displacement Crisis in West Africa*. . 2013 February 2013]; Available from: http://www.unhcr.org/pages/4d831f586.html.
3. UNHCR, *Update No. 39. Cote d'Ivoire Situation*, 2011.
4. UNHCR, *Update No. 25. Côte d'Ivoire Situation*. 2011.
5. Bengtsson, L., *Interviews performed by Bengtsson in Abidjan*, February 6-8 2013.
6. Bharti, N., et al., *Explaining seasonal fluctuations of measles in Niger using nighttime lights imagery*. *Science*, 2011. **334**: p. 1424-1427.
7. Bengtsson, L., et al., *Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: a post-earthquake geospatial study in Haiti*. Plos Medicine, 2011. **8**(8): p. 9.
8. Lu, X., L. Bengtsson, and P. Holme, *Predictability of population displacement after the 2010 Haiti earthquake*. Proceedings of the National Academy of Science USA, 2012. **109**(29): p. 11576:11581.
9. Sutton, P., et al., *Census from heaven: an estimate of the global population using nighttime satellite imagery*. International Journal of Remote Sensing, 2001. **22**(16): p. 3061-3076.
10. NOAA National Geophysical Data Center. *VIIRS Nighttime lights 2012*. [cited 2013; Available from: http://www.ngdc.noaa.gov/dmsp/data/viirs_fire/viirs_html/viirs_ntl.html.
11. National Geophysical Data Center and NOAA Satellite and Information Service. *Space physics interactive data resource*. Available from: http://spidr.ngdc.noaa.gov/spidr/.
12. Agnew, J., T.W. Gillespie, and J. Gonzalez, *Baghdad nights: evaluating the US military 'surge' using nighttime light signatures*. Environment and Planning A, 2008. **40**: p. 2285-2295.
13. Checchi, F. and C. Grundy, *Satellite imagery for rapid estimation of displaced populations: a validation and feasability study*. 2012. **Final project report**.
14. United Nations Operation in Côte d'Ivoire, *Summary of UNOCI weekly press conference: POST-ELECTION VIOLENCE CLAIMS MORE THAN 1000 LIVES IN WESTERN COTE D'IVOIRE, ACCORDING TO UNOCI REPORT, 2011: Abidjan*, 2011.
15. BBC News Africa. *Ivory Coast: More than 100 bodies found, says UN*. 2011 February 2013]; Available from: http://www.bbc.co.uk/news/world-africa-13013082.

16. Ebener, S., et al., *From wealth to health: modelling the distribution of income per capita at the sub-national level using night-time light imagery*. International Journal of Health Geographics, 2005. **4**(1): p. 5.
17. Noor, A., et al., *Using remotely sensed night-time light as a proxy for poverty in Africa*. Population Health Metrics, 2008. **6**(5): p. 13.

Analysing and mapping population movements from anonymous cellphone activity data

15 February 2013

Hayden Glass, Sapere Research Group, hglass@srgexpert.com (primary contact)

Iain Kirkpatrick, kirkpatrick.iain@gmail.com

Aaron Schiff, covec, aaron.schiff@covec.co.nz

Abstract

We discuss how to turn anonymised cellphone activity data into maps of populations. We demonstrate two analyses that are possible from these data: stocks analysis, which provides a way to count populations by location over time; and flows analysis, which provides a way to analyse and visualise population flows between locations. We outline the promise of and challenges to development of a standardised software application to enable speedy, user-driven analysis and visualisation of population stocks and flows from anonymised cellphone activity data, with a primary focus on uses for emergency response. We consider pathways for further research, including analysis of movements and automated recognition of 'normal' versus 'unusual' movements.

Introduction

This paper is an entry in the Data for Development Challenge. We have focused on the locational datasets, particularly the SET2 data, which has location observations for a sample of anonymised users at cellsite level. We discuss how to use these data to analyse people's locations and visualise their movements over time. We also discuss the development of some generic analytical software tools to enable speedy, user-driven analysis of the data. Our primary focus has been on emergency response applications for this information.

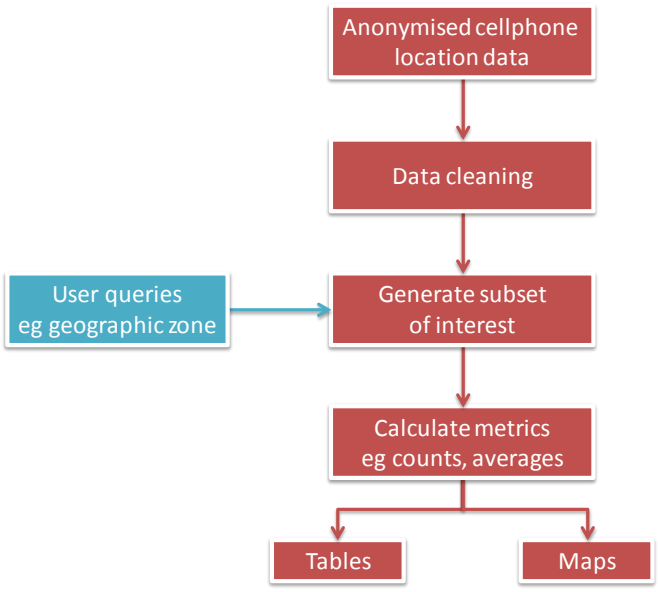
Two basic types of analysis are possible:

- **Location analysis** – representations of the locational observations of individuals and the corresponding geographic distribution of a population at a given point in time or during a given time period, and
- **Movement analysis** – representations of movements of individuals or groups of individuals satisfying some condition (e.g., presence in a given town on a specific morning) computed from the locational observations.

In this paper we focus on location analysis, and we suggest some directions for future research on movement analysis.

Figure 1 shows the high level process diagram for analysis. The first step is to clean the anonymised location data to remove invalid observations, e.g., where the cellsite is not recorded. A query is then applied to the data to restrict it to a subset of interest, e.g., all of the users who were observed in a given geographic zone on a particular day. From this, metrics of interest can be calculated, such as geographic population distribution at that time and on subsequent days. These metrics can be represented as numerical tables and maps, where the maps make use of the geospatial component of the data.

Figure 1 Process diagram



Location analysis

In this section we discuss analysis based on locational observations of cellphone users. The basic unit of data on which our analysis is based is an observation of an individual at a location at a point in time.[1] We use the high-resolution spatial dataset (SET2) where each record consists of three pieces of information: a location, a timestamp, and an arbitrary ID number assigned to each cellphone. A record is created each time a phone sends or receives a call or a text message. **Table 1** summarises the basic characteristics of the dataset.

Table 1: Characteristics of the Set2 data

Number of cellsite locations	1,238
Number of unique users	500,000
Number of location fixes	55,319,911
Average activity observations per cellsite	44,685
Average activity observations per unique user	111

In the dataset, a location observation is the approximate location of the cellsite that handled the user's activity. Out of necessity, we assume that users are located at these exact points at the time their activity is recorded. We also group cellsite locations into larger geographic zones based on Cote d'Ivoire administrative boundaries. This is shown in

Figure 2, with the 1,238 cellsites mapped across 236 sous-prefectures. Sous-prefectures are also grouped into departments, and departments in turn into still-larger regions.[2]

Because a record is created only when a call or text message is transmitted, we have more complete location data for users who use their phones more, and we do not know anything about the location of users in between the times they use their phones. The random frequency of locational observations becomes an issue for analysing movements calculated from these observations. Inferring users' locations during the time gaps requires modelling location over time in some way.

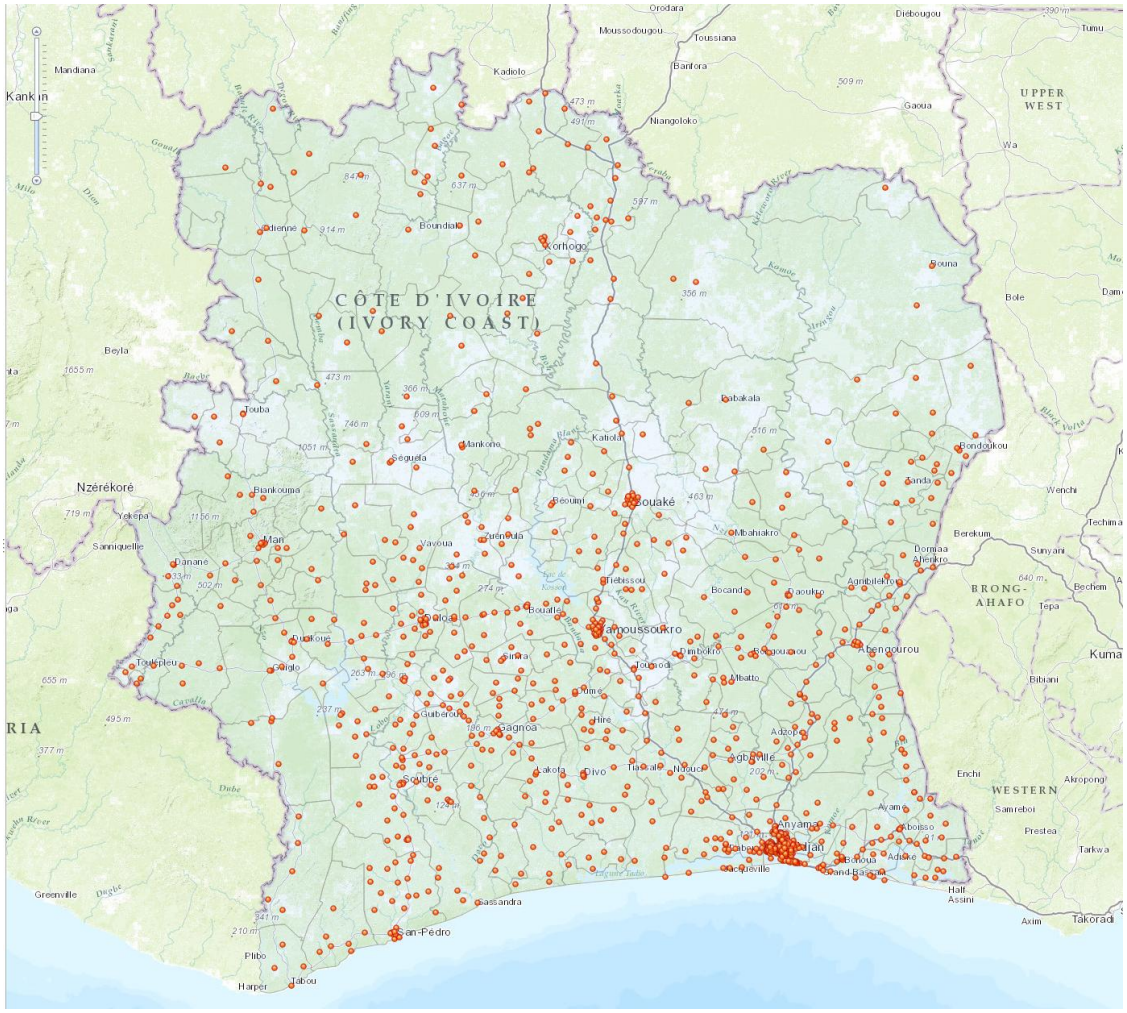
In addition, the distribution of cellsites is not uniform across the country. This means that we have better data on location in Abidjian, a sous-prefecture with 379 cellsites, than in Tengrela, a sous-prefecture with just one cellsite, and no data at all from sous-prefectures with no cellsites. We assume that the distribution of cellsites follows population, i.e., cellsites are constructed where there are users to be served, but lack of universal coverage remains a challenge.

In most cases, we group timestamps together into hour long spans. The SET2 data contains call records for 50,000 cellphones for each of five two-week periods. For privacy reasons, the set of 50,000 users was changed every two weeks, thus we cannot follow individual movements for longer than this. However, given that the samples are randomly distributed geographically, we can assume

[1] Note that cellphones may sometimes be shared among multiple users and that a single individual might have more than one cellphone. We assume the locational observations of a cellphone correspond to a unique user.
[2] We draw on the administrative boundaries data of the UN OCHA-maintained Common Operational Datasets http://cod.humanitarianresponse.info/search/node/cote%20d%27ivoire.

that the dataset is representative of the spatial distribution of users throughout the entire ten-week sample, even when the underlying user sample changes.

Figure 2: Boundaries of sous-prefectures and approximate locations of cellsites



For location analysis, we use the following terminology:

- A geographic **zone** is a region defined by the area served by one or more cellsites.
- A **timeframe** is a continuous span of time, which we usually define to be at least one hour.

For any given zone and timeframe, it is straightforward to query SET2 to determine the number of users who were observed in that zone during that timeframe. Most location analysis is derived from variations of this type of query.

We have developed two types of analysis using location data:

- **Stocks** – Counts of how many people are observed in geographic zones in a given timeframe, which can be used to produce geographic population distributions.

- **Flows** – Counts of people arriving and/or departing a zone during a given timeframe and analysis of the origins and destinations of these people.

If the selected timeframe is sufficiently short relative to the overall dataset (e.g., one hour or one day) we can visualise cross sections of stocks and flows over time. By observing changes in stocks and flows, it is possible to analyse simple dynamics of population movements.

Analysis of stocks

The stock of users is simply the number of unique users observed in a given geographic zone, within a given timeframe. This can be calculated from the SET2 data by counting the number of unique user IDs observed in the relevant cellsite(s) during the relevant timeframe. This will give a representation of the spatial distribution of population during a given timeframe. If the analysis is also repeated for multiple timeframes then changes in population distribution over time can be visualised through tables and animated maps.

Two user controls are required in the software application to facilitate this analysis:

- **Selection of the geographic resolution**: Choosing to analyse stocks across cellsites, sous-prefectures or administrative regions. In a mapping application, this can be chosen simply by setting the zoom level of the map.

- **Selection of a temporal resolution**: Choosing to analyse stocks calculated over a given amount of time, such as an hour, day or week. The analysis can be repeated across multiple timeframes of the same duration, to facilitate analysis of changes in stocks over time.

This allows two types of analysis to be performed – a static representation of population distribution at a given point in time, and an 'animation' of how this distribution changes over time. The results of this analysis can be represented visually on a map, and numerically in a table. For example, **Table 2** shows the distribution of the number of unique users observed in a selection of sous-prefectures between 9am and 12pm on 12 December 2011.

Further simple calculations can be performed over time to show changes in population distribution. For example, **Table 3** shows the average number of unique users observed in each sous-prefecture by hour between 9am and 5pm, across the entire dataset. For any given hour, the respective column in **Table 3** shows the distribution of unique users across sous-prefectures.

Trends across the rows in **Table 3** largely reflect changes in cellphone usage during the day, rather than changes in users' locations. To correct for this, the distribution of users in a given hour across sous-prefectures can be indexed based on the total number of unique users observed across all sous-prefectures in that hour. This is illustrated in **Table 4** and also graphically in **Figure 3**.

This type of analysis allows rapid and flexible counting of populations, as well as the visualisation and analysis of population changes over time.

Table 2: Number of unique users observed in each sous-prefecture between 9am and 12pm on 12 December 2011

Sous-Prefecture	Unique users
ABENGOUROU	504
ABIDJAN	18,429
ABOISSO	360
ADIAKE	87
ADZOPE	130
AFFERY	29
AGBOVILLE	367
AGNIBILEKROU	147
AGOU	48
AKOBOISSUE	13
AKOUPE	91
ALEPE	181
AMELEKIA	19
ANDO-KEKRENOU	3
ANGODA	24
ANIASSUE	52
ANOUMABA	16
…	…

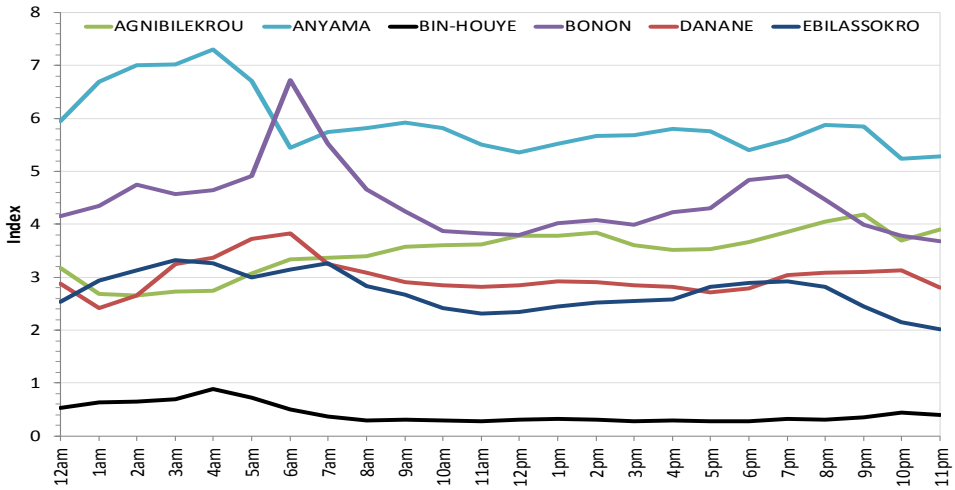
Table 3: Average number of unique users observed in each sous-prefecture, by hour

Sous Prefecture	9am	10am	11am	12pm	1pm	2pm	3pm	4pm	5pm
ABENGOUROU	142	136	136	141	127	121	117	124	124
ABIDJAN	6,129	6,068	5,990	5,794	5,301	5,138	5,115	5,435	5,452
ABOISSO	99	96	93	93	89	86	86	91	91
ADIAKE	26	25	24	23	23	23	22	22	23
ADZOPE	46	44	44	45	42	39	38	41	41
AFFERY	9	8	8	8	7	7	7	7	7
AGBOVILLE	101	95	92	92	86	82	81	87	86
AGNIBILEKROU	49	47	46	48	44	44	41	43	43
AGOU	17	15	15	15	14	13	14	14	15
AKOBOISSUE	4	4	4	4	3	4	4	4	4
AKOUPE	28	26	27	25	24	23	23	25	25
ALEPE	47	43	41	42	40	39	40	42	42
AMELEKIA	9	7	7	6	7	7	7	7	8
ANDO-KEKRENOU	3	2	2	2	2	2	2	2	2
ANGODA	7	6	6	6	6	5	6	6	6
ANIASSUE	13	12	12	11	11	11	11	11	12
ANOUMABA	5	5	5	4	4	4	4	4	5
…	…	…	…	…	…	…	…	…	…

Table 4: Index of the average number of unique users observed in each sous-prefecture, by hour.

Sous Prefecture	9am	10am	11am	12pm	1pm	2pm	3pm	4pm	5pm
ABENGOUROU	10	10	11	11	11	11	10	10	10
ABIDJAN	446	467	469	461	453	453	452	450	450
ABOISSO	7	7	7	7	8	8	8	8	8
ADIAKE	2	2	2	2	2	2	2	2	2
ADZOPE	3	3	3	4	4	3	3	3	3
AFFERY	1	1	1	1	1	1	1	1	1
AGBOVILLE	7	7	7	7	7	7	7	7	7
AGNIBILEKROU	4	4	4	4	4	4	4	4	4
AGOU	1	1	1	1	1	1	1	1	1
AKOBOISSUE	0	0	0	0	0	0	0	0	0
AKOUPE	2	2	2	2	2	2	2	2	2
ALEPE	3	3	3	3	3	3	4	3	3
AMELEKIA	1	1	1	1	1	1	1	1	1
ANDO-KEKRENOU	0	0	0	0	0	0	0	0	0
ANGODA	0	0	0	0	0	0	1	1	0
ANIASSUE	1	1	1	1	1	1	1	1	1
ANOUMABA	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...

Figure 3: Index of number of unique users observed in each sous-prefecture, by hour

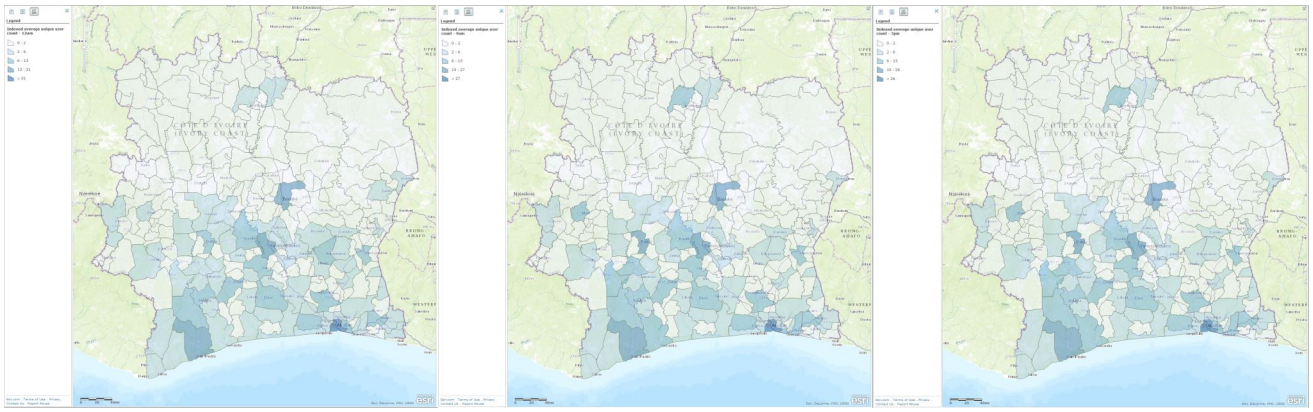


It is straightforward to represent these distributions on a map, using colours to represent different levels of observed population, as shown in **Figure 4** in a way that is visually similar to the Geofast service.[3] We have grouped cellsites into sous-prefectures for this analysis, but the same analysis can be made and visualised at the cellsite level.

[3] See http://sites.uclouvain.be/geofast/

Figure 4: Map representation of geographic distribution across sous-prefectures

12am **9am** **3pm**



Analysis of flows

Flows represent movements of people in and out of a geographic zone within a timeframe. For a given zone X and timeframe, a cellphone user is determined to have arrived in zone X from zone Y if they are observed in X during the specified timeframe and their most recent observation prior to that is in zone Y. Flows can be calculated from SET2 by sorting the observations by user ID and then by time, and comparing each observation for each user with the previous observation for that user.

User controls in the software application would allow selection of the location and timeframe of interest, and then the application would filter the data based on those settings and display the results in tables and animated maps.

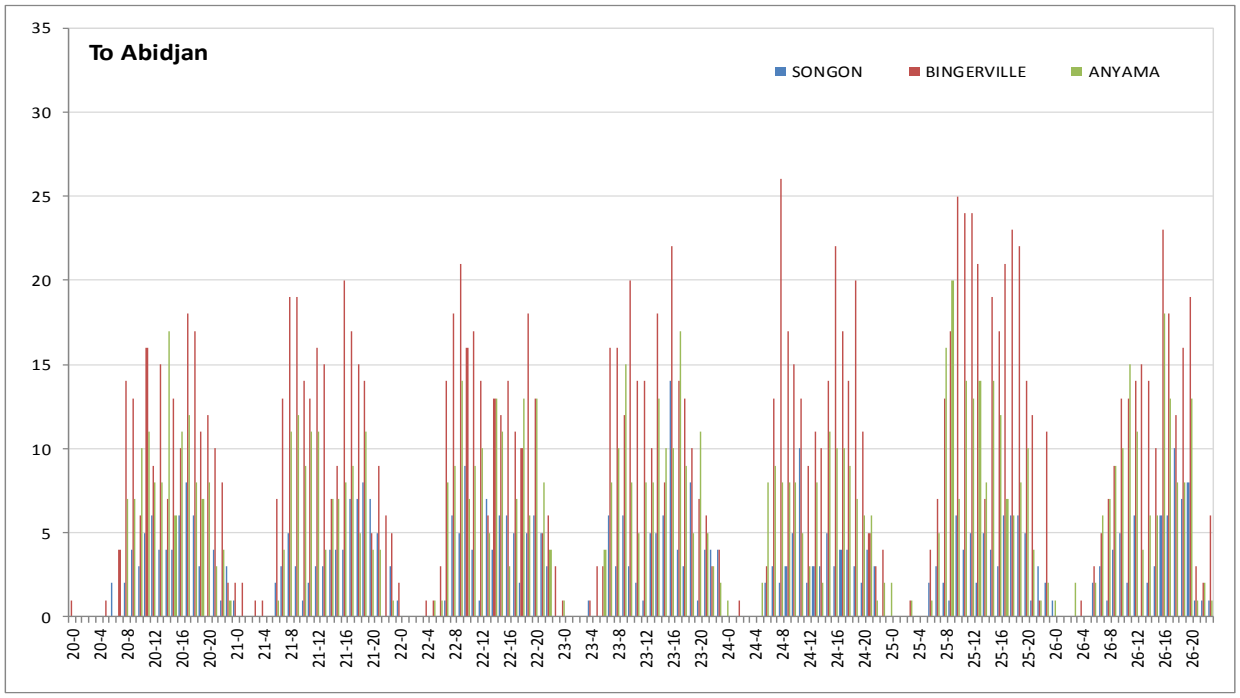
This type of analysis can reveal population movements that are difficult or impossible to analyse using more traditional survey methods. For example, it is straightforward to isolate a subset of people who were in a particular area at a particular time, and to quickly reveal where those users came from and where they went to. Bengtsson *et al* (2011) is an example of using this type of analysis to explore population movements after a cholera outbreak in Haiti, and to look at how quickly people returned to the capital following an earthquake.

As an example, **Table 5** shows the number of movements observed between Abidjan and the three surrounding sous-prefectures during the week beginning 20 February 2012. **Figure 5** shows the movements in to Abidjan from the three surrounding sous-prefectures during the same week on an hourly basis.

Table 5: Movements between Abidjan and surrounding sous-prefectures

To Abidjan From		
Songon	Bingerville	Anyama
489	1,563	973
From Abidjan To		
Songon	Bingerville	Anyama
545	1,571	960

Figure 5: Hourly movements to Abidjan from surrounding sous-prefectures



Implementation

We envisage developing a custom software application that can turn anonymised cellphone activity data into maps and analysis of populations.

It would have a software stack following the basic open-source 'LAMP' structure – an operating system and software on a server, with an SQL database storing the data which is accessed and manipulated for visualisation via a web interface. Access to the database would be asynchronous for a faster and smoother user experience. This kind of application could be built entirely using open-source technologies, reducing cost and restrictions, and allowing knowledgeable end-users to make their own modifications.

We can see at least three uses:

- **Emergency response** – Anonymised cellphone activity data can be used to reveal where people went after an emergency and help emergency responders guide the distribution of aid to the affected population. Analysis of this data can provide a faster and more flexible way to look at population movements compared with traditional survey methods. As noted above, Bengtsson *et al* (2011) is an example.

- **City and transport planning** – Possible applications include measuring the influence of changes to transport infrastructure on people's commuting patterns, and helping city authorities more easily gather feedback on how their changes are being experienced by city dwellers. Other applications include understanding movements within urban areas to help design infrastructure and public spaces. Ahas *et al* (2010) is an example of using location data to study commuter behaviour.

- **Tourism and events analysis** – Anonymised cellphone activity data can provide information on attendance at special events, on where attendees came from, and the routes that they took to and from the event. It can also be used to explore the routes that tourists take within a country, something that is difficult to study accurately using traditional means such as surveys. Ahas *et al* (2009) is an example of using cellphone activity data to track attendance at a modestly-sized agricultural fair in rural Estonia.

In many cases, anonymised cellphone activity data is more useful than traditional survey approaches to population mapping:

- Cellphone ownership is widespread and cellphone network coverage is near ubiquitous – The ITU (2012) reported that there were nearly six billion active mobile phone users out of a world population of seven billion at the end of 2011.

- The relevant data is collected as a matter of course by operators (although it requires some specific effort to anonymise the data and secure access to it), and

- Anonymised cellphone data provides larger samples than survey-based approaches, it can be more timely, and it makes possible the study of larger geographic areas using relatively simple tools (see Ahas *et al* (2008)). Cellphone data is also more flexible, since it is straightforward to manipulate the data to focus only on specific subsets of relevant users.

For example, it is straightforward to identify the data associated with only those users who were in a particular town on a particular morning.

That said, there are many issues to overcome to develop such an application and deploy it in practice (Ahas *et at* 2008). Amongst these issues are:

- **Getting access to the data itself** – We understand that mobile operators are generally cautious about making available this information, and are also looking to use it for their own commercial purposes.[4] Defining simple standards for this cellphone activity data could help encourage data availability. We note in particular the efforts of Flowminder.org, which we understand aims to systematise the collection and distribution of anonymised cellphone location data for use in emergency situations.

- **Ensuring anonymity** – There is an important distinction between mapping populations where data is anonymised and individuals cannot be identified, and tracking or predicting the movements of individuals, which raises a host of protection concerns in the case of emergency response applications. There are several well-known examples where datasets thought to be anonymised were shown to reveal information about identifiable individuals when combined with other public information.[5]

- **Scaling up to a population** – Counting or mapping cellphone users is not the same as counting populations. For example, although ownership is widespread, not all people have cellphones and some have more than one. Phone ownership may be patterned in various ways, including by age or by geographic location. To use cellphone activity data in each case, a way needs to be found to relate counts of cellphones to counts of population. Bengtsson *et al* (2011) find a high degree of alignment between their cellphone activity based population counts and the totals from official surveys. Ahas *et al* (2011) discuss the issues involved in using analysis of cellphone activity data in official statistics in Estonia. There are limits on the types of analysis that can be done with cellphone activity data because it does not include any additional demographic information.

- **Practical deployment challenges** – In our efforts thus far we have used standard spreadsheet and SQL software and the GIS packages ArcGIS and ArcGIS Online to manipulate, analyse and visualise the data. Some aspects are straightforward: once data is in the appropriate format, GIS software can recognise the locations of cellsites for each call record, geo-reference them automatically, and represent them on a map. It is trivial to add layers of additional geographic information to the same map, e.g., to show the region, department and sous-prefecture boundaries. GIS software can also interpret the temporal variance of the calls, and generate a 'time-slider' that controls which slice of time the visualisation shows.

[4] We note, for example, global mobile operator Telefonica's recent launch of Smart Steps, a product that 'will use fully anonymised and aggregated mobile network data to enable companies and public sector organisations to measure, compare, and understand what factors influence the number of people visiting a location at any time'. See http://blog.digital.telefonica.com/?press-release=telefonica-launches-telefonica-dynamic-insights-a-new-global-big-data-business-unit

[5] See for example http://www.wired.com/threatlevel/2009/12/netflix-privacy-lawsuit/ in relation to a competition run by Netflix, a United States video rental service.

Practical challenges emerge because of the size of the datasets involved and therefore the difficulties of storage and transmission, and for ensuring that the application is responsive to user input. Possible approaches include using server arrays such as Amazon's EC2 services, which scale services automatically to match website demand, or pre-processing 'slices' of the dataset to cut down on query times. Different challenges emerge in emergency response environments with limited internet access where we think this information could be most helpful.

Future research

There are many possibilities for further analysis of anonymised cellphone activity data.

Normal and unusual locations

As a further development of the location stocks analysis, it would be possible to compute the average location for each individual user over given timeframes, for example each day. This average location can be treated as a random variable with associated variance. 'Unusual' locations can then be highlighted where the user's location deviates at some point in time or during some timeframe from a statistical confidence region defined by the average location and its variance.

Such analysis could be particularly useful for early detection of events of interest, including emergencies that might potentially cause a large number of people to move to 'unusual' locations in a relatively short space of time. By establishing thresholds for such population movements, it may be possible to automatically detect and highlight situations of interest.

Movement analysis

The above discussion has focussed on analysis of the location data. Two or more locational observations for the same user can be converted into movements. A simple movement is defined by:

- An **origin** point and a **start time**; and
- A **destination** point and an **end time**.

These pieces of information can be identified from the location data, by searching for instances where activity for a given user is recorded in one location and subsequent activity is recorded in another location. From the origin and destination points and times, it is also possible to calculate the average speed and direction of the movement. With intermediate observations, or some inference about movement paths, it is possible to calculate the areas moved through between the origin and the destination.

Analysis of movements based on mobile network activity data present challenges due to the fact that locations are not observed at regular time intervals but rather only when user activity occurs. This means that the observed movement start time will often be prior to the actual physical movement, and likewise the observed end time will often be later. Thus movements calculated from network activity data will generally appear to be longer duration and slower speed than in reality.

The operation of mobile networks will also cause some distortions in estimated movements, as for technical reasons user activity may switch frequently between neighbouring cellsites, generating the appearance of short (and high speed) movements, where no actual movement has occurred.

With these caveats in mind, some additional analysis is possible once locational observations for each user are converted into movements. The analysis of movements will help to highlight *changes* in population distribution over time in ways that is less easily observable from the locational data.

Examples of such analysis include:

- Average speeds: Computation of the average speed within a given geographic zone, during a given timeframe. This could be useful for transport and urban planning. In addition, if observed speeds are unusually high or low, this could indicate an event of interest at that location.

- Direction: Computation of the average or most frequent direction of movement in or out of a given geographic zone during a given timeframe. This would also be useful for transport and urban planning, and for the analysis of crowd movements after organised events. The movement of an unusually large number of people in the same direction could indicate an event of interest at a location opposite to the movement.

- Sinks and sources: A 'sink' is a geographic zone that has a high number of people arriving during a given timeframe, while a 'source' is a geographic zone with a high number of people departing during a given timeframe. Identification and analysis of sinks and sources associated with regular and unusual movements could be useful for urban planning and for analysis of response to emergencies.

Other extensions

Further extensions to the analysis are possible when locational observations are combined with other data that can be obtained from mobile network activity. In particular, it would be possible to use call records to understand social networks and use this information to explain movements.

Furthermore, social networks could be used to *forecast* movements, on the assumption that at least some movement will be related to people's social interactions. In the case of emergencies, social networks could be used to predict where people will re-locate to. Bengtsson *et al* (2011) shows that movements during non-emergency times are good predictors of movements during emergencies.

References

Ahas R, Aasa A, Roose A, Mark U, Silm S (2008) Evaluating passive mobile positioning data for tourism surveys: an Estonian case study. Tourism Manage 29: 469–486.

Ahas R, Pechlaner H, Nilbe, K (2009) Developing Event Marketing Strategies with Mobile Telephone Position Data: Case Study with LindoraAgricultural Fair in Estonia, Proceedings of the European Cities Marketing Annual Conference & General Assembly, 80

Ahas R, Aasa A, Silm S, Tiru M (2010) Daily rhythms of suburban commuters' movements in the Tallinn metropolitan area: Case study with mobile positioning data, Transportation Research Part C 18 45–54

Ahas R, Tiru M, Saluveer E, Demunter C (2011) Mobile telephones and mobile positioning data as source for statistics: Estonian experiences

Bengtsson L, Lu X, Thorson A, Garfield R, von Schreeb J (2011) Improved Response to Disasters and Outbreaks by Tracking Population Movements with Mobile Phone Network Data: A Post-Earthquake Geospatial Study in Haiti.PLoS Med 8(8): e1001083. doi:10.1371/journal.pmed.1001083

Gething P, Tatern A (2011) Can mobile phone data improve emergency response to natural disasters? PLoS Medicine. 8.8 (Aug. 2011)

ITU (2012) Measuring the Information Society. Available online from http://www.itu.int/dms_pub/itu-d/opb/ind/D-IND-ICTOI-2012-SUM-PDF-E.pdf

MTS (2007) Mobility: A nation under siege: An insight into mobile communications during the 2006 Lebanon conflict

Tatern AJ, Qiu Y, Smith DL, Sabot O, Ali AS, et al. (2009) The use of mobile phone data for the estimation of the travel patterns and imported Plasmodium falciparum rates among Zanzibar residents. Malaria J8: 287.

Wesolowski A, Eagle N (2010) Parameterizing the dynamics of slums. AAAI Spring Symposium 2010 on Artificial Intelligence for Development

Zook M, Graham M, Shelton T, Gorman S (2010) Volunteered Geographic Information and Crowdsourcing Disaster Relief: A Case Study of the Haitian Earthquake, World Medical & Health Policy Vol. 2: Iss. 2, Article 2

Spotted: Connecting People, Locations and Real-World Events in a Cellular Network

Ramona Trestian
Performance Engineering Laboratory
School of Electronic Engineering
Dublin City University, Ireland
ramona@eeng.dcu.ie

Faisal Zaman
Performance Engineering Laboratory
School of Electronic Engineering
Dublin City University, Ireland
Faisal.Zaman@dcu.ie

Gabriel-Miro Muntean
Performance Engineering Laboratory
School of Electronic Engineering
Dublin City University, Ireland
munteang@eeng.dcu.ie

ABSTRACT

Being able to react fast to exceptional events such as riots protests or disaster preventions is of paramount importance, especially when trying to ensure peoples' safety and security, or even save lives. In this paper we study the use of fully anonymized and highly aggregate cellular network data, like Call Detail Records (CDRs) in order to connect people, locations and events. The goal of this study is to see if the CDR data can be used to detect exceptional spatio-temporal patterns of the collective human mobile data usage and correlate these 'anomalies' with real-world events (e.g., parades, public concerts, soccer match, traffic congestion, riots protests etc.). These observations could be further used to develop an intelligent system that detects exceptional events in real-time from CDRs data monitoring. Such system could be used in intelligent transportation management, urban planning, emergency situations, network resource allocation and performance optimization, etc.

Keywords

Cellular Networks, Human Mobility, Call Detail Records

1. INTRODUCTION

In the ever-evolving telecommunication industry, smart mobile computing devices have become increasingly affordable and powerful, leading to a significant growth in the number of advanced mobile users and their bandwidth demands. This, together with the improved next generation telecommunications infrastructure, motivates the continuing uptake of the mobility around the world. People can now connect to the Internet from anywhere at any time, while on the move (e.g. on foot, in the car, on the bus, stuck in traffic etc.) or stationary (e.g., at home/office/airport/coffee bars, etc.). The number of mobile users increases continuously as the penetration of both fixed and mobile broadband solutions becomes more affordable for the masses and more accessible around the globe. The connection to the Internet is possible and can be done via wireline or wireless solutions. Depending on the user location, wireless connectivity is enabled by different Radio Access Technologies (RATs) such as: Global System for Mobile Communications (GSM), Enhanced Data Rates for GSM Evolution (EDGE), Universal Mobile Telecommunications System (UMTS), High Speed Packet Access (HSPA), Long Term Evolution (LTE), Worldwide Interoperability for Microwave Access (WiMAX), Wireless Local Area Networks (WLAN), Wireless Personal Area Network (WPAN), etc. Use of all these RATs is rapidly spreading, covering various geographical locations in an overlapping manner.

Additionally, this increasing expansion of the telecommunication infrastructure could bring economic, social and technological benefits especially to the far reaching regions. For example, it can bring education to the remote regions; it can contribute to enabling innovations in healthcare (e.g., remote monitoring and diagnostics), smart grid solutions, social networking sites, economy, etc.

One of the key characteristics of these mobile networks and the mobile computing devices is that every time they are used a digital signature is recorded. Voluntarily or not, whenever people interact with the telecommunications networks or any type of social media platform, they leave behind digital traces. All these traces have become a powerful tool to analyze human behavior patterns. For example, the data collected by the cellular telecommunications systems referred to as Call Details Records (CDRs) is done in regular bases for billing and troubleshooting purposes. Moreover these CDRs contain the information details about every call carried within the cellular network, including information about the location, call duration, call time, and both parties involved in the conversation. Thus there is an increase interest on making use of the information provided by the CDRs in order to analyze the human mobility cheaply, frequently and especially at a very large scale. In general, understanding the human mobility patterns could have broad applicability on a wide range of areas, such as: network resource optimization, mobile computing, transportation systems, urban environment planning, events management, epidemiology, etc.

In this work we explore the use of anonymized Call Detail Records (CDRs) containing both voice-calls and SMS activities, from a cellular network in Ivory Coast in order to connect people, locations and events. The goal of this study is to identify the exceptional spatio-temporal patterns of the collective human activity from fully anonymized and highly aggregate cellular network data, like CDRs, and correlate these 'anomalies' with real-world events (e.g., parades, public concerts, soccer match, traffic congestion, etc.). These observations could be further used to develop an *intelligent system that detects exceptional events in real-time* from CDRs monitoring. The benefits of such systems could be threefold: (1) the *network operators* could benefit by detecting congested cells and optimize their network resources in advance of an exceptional event, e.g., make use of the wi-fi offloading solutions, enabling adaptive bandwidth allocation to their radio cells, etc.; (2) *the society* could benefit from intelligent transportation and urban planning and management; (3) *the individual* could benefit from traffic information and prediction, emergency management. For example, a real-time event detection system could be used in case of emergency situations, such as riots protests which could be more efficiently handled if detected and handled on time.

NetMob, May 1–2, 2013, MIT Media Lab, Cambridge, Boston.

Within this context, our research questions are: can the CDR data be used to detect exceptional spatio-temporal patterns of the collective human mobile data usage? Can we correlate these exceptional usage patterns to real-world events?

2. RELATED WORKS

Recently, there has been extensive academic research related to the use of user-generated traffic in mobile communications networks as a powerful tool to analyze human behavior. Several studies have shown that the use of massive mobile phone data sets collected from the cellular networks could have great potential in several areas, such as: definition of the universal law [1][2], urban planning [3][4], real-time traffic forecast [5], human localization and mobility patterns [6][7][8], context-aware applications [9][10].

The high predictability of human mobility patterns from cellular network traces was studied in [2] and [7]. In most of these studies the primary source of data used is CDRs. The authors in [7] look at the entropy of the locations and they show that 14 days is a sufficient time frame to analyze the human behavior. Whereas the authors in [2] show that there is a probability of 93% that the human location could be predicted regardless of how far the person travels within the preferred locations.

One of the first studies that provided evidence of geographic correlation between users' interests within a cellular network was conducted by Trestian et al. in [11]. The authors categorized the user interests into six groups, such as: mail, social networking, trading, music, news, and dating. The main focus of the study was to correlate these users' interests with their location, e.g., home or work. Their results showed that in general the users tend to spend a significant fraction of their time in their top three locations only. Another study that focuses on identifying important locations in humans' live from mobile data traces was conducted by Isaacmanc et al. in [12].

Gonzalez et al. in [1] conducted a study on mobility traces of 100,000 mobile customers and showed that there is a regular trend in human mobility, both temporal and spatial and with a high probability the users return to their few preferred locations. Characterizing the correlation between the customers' interests and their location or mobility pattern is crucial for location-based services. For example, Keralapura et al. in [13] used the IP flow traces from a cellular network operator, in order to study distinct behavior patterns of the mobile users when web browsing.

Looking at two major cities with an advanced economic level, such as Los Angeles and New York Cities, Isaacman et al. [3] studied the different mobility patterns of the users in terms of daily and maximal travel distance. Whereas Tso et al. in [14] presented an empirical study on the performance of High Speed Packet Access (HSPA) networks in Hong Kong. The authors run extensive field test under different movement scenarios.

Sagl et al. in [15] propose a visual analytics approach that investigates the spatio-temporal pattern of the collective human mobility and offer a comparison between different urban environments

All these studies have shown that understanding the humans' mobility patterns could be a crucial component in several areas, such as: network optimization opportunities for cellular network operators in handling the explosive growth in traffic observed from CDRs; transportation planning and management, modeling commuting flows, content delivery services and context-aware applications, etc. However not much focus was put on studying the correlation between the user activity in a cellular network from CDRs with real-world events, such as: parades, riots protests, football games, etc.

3. DATA COLLECTION METHODOLOGY AND CHARACTERISTICS

3.1 Data Collection and Preprocessing

In this paper we utilize the anonym CDR provided by the Orange Group within the Orange *Data for Development* (D4D) challenge. The CDRs are anonymized phone calls and SMS exchanges between five million Orange customers in Ivory Coast. The anonymized CDRs were collected from a random set of cellular phones over 150 days, between December 1, 2011 and April 28, 2012. The territorial expanse of the dataset on Ivory Coast is illustrated in Figure 1[1]. The Ivory Coast is located in West Africa having an area of 322,462 square kilometers and an estimated population of 20 million inhabitants. The political capital of Ivory Coast is Yamoussoukro whereas the biggest city is the port city Abidjan. The country telecommunications sector is dominated by mobile telephony with Orange being one of the leaders in the market, recording around five million customers (a quarter of the entire population). For the purpose of this study, there are four sets of data provided by Orange Group and described in the following sections.

Figure 1. Territorial expanse of the dataset – Ivory

3.2 Dataset 1: Antenna-to-Antenna Communication

The first dataset contains the aggregated number of calls as well as the calls durations within one hour, between any antennas pair. The dataset was stored in 10 files each corresponding to a 14 days interval. All the datasets are provided in Tabulation Separated Values (TSV) file format. For the Antenna-to-Antenna dataset each line stores information about the date, time, originating antenna, terminating antenna, number of voice calls, and the duration of the voice calls in minutes for a given hour.

3.3 Dataset 2: Individual Trajectories – High Spatial Resolution Data

The second dataset provides high resolution individual movement trajectories of 50,000 randomly sampled customers split into consecutive two-week periods. The data was stored in 10 TSV files, and in order to protect the customers' privacy new random

[1] https://maps.google.com/

identifiers for each customer are chosen in every two-week time period. Each line in the file contains information about the customer identification number, the connection date and time and the antenna identification number they are connected to.

3.4 Dataset 3: Individual Trajectories – Long Term Data

The third dataset contains the long term, low spatial resolution trajectories of the 50,000 randomly selected customers. The low spatial resolution is obtained by replacing the antennas identifiers with the sub-prefectures of the antenna the customer is connected to. Ivory Coast has a total number of 255 sub-prefectures administrative regions. Figure 2 illustrates the sub-prefectures and their identifiers along with the Orange antennas locations as provided in the datasets. There are a total number of 1238 antennas, and not all the sub-prefectures have cell phone towers.

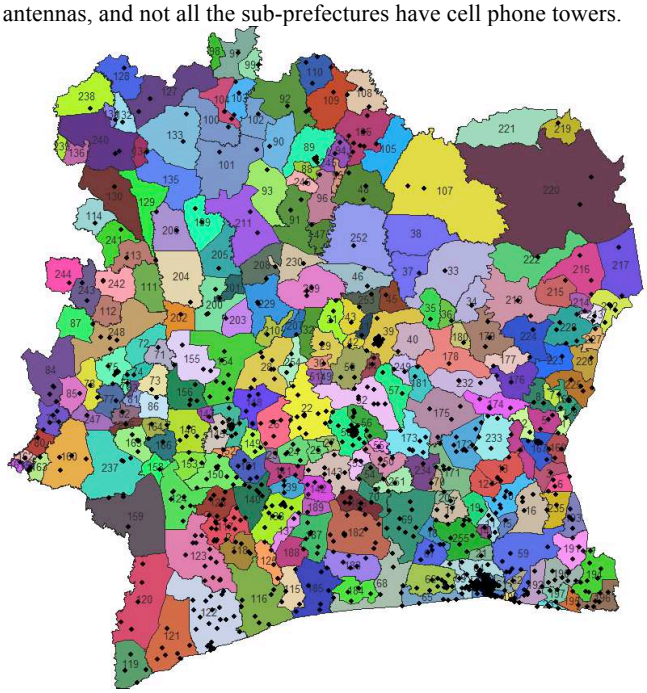


Figure 2. Ivory Coast sub-prefectures and Orange antennas location.

Each file in this dataset contains information on customer identification number, the connection date and time and the sub-prefectures identifier that contains the antenna the user is connected to.

3.5 Dataset 4: Communication Subgraphs

The fourth dataset contains information on the communication subgraphs of 5,000 randomly selected individuals (egos). The communication between the randomly selected egos and their second order neighbors was divided into two-weeks periods starting December 5, 2011 over the 150 days of the observation period. Each file indicates if there was communication between every customer's pair by providing the source customer id and the destination customer id.

3.6 Limitations of the Datasets

Even though the Call Detail Records represent a good source of location information they have several significant limitations:

- they are generated only when the mobile device is engaged in a voice call or exchanges text messages, thus no information about application usage type (voice/text/data) is available.
- the location granularity is at cell tower level or sub-prefectures, no information about the exact user location is provided.
- no information about the call duration is provided.

4. IDENTIFYING HOTSPOTS

The goal of this section is to analyze the datasets provided in order to correlate people, locations and events during the 150 days of recorded data.

The study aims on identifying hotspots with highly active antennas, analyze the human activity pattern and correlate the 'anomalies' in the patterns with certain real-world events. The starting date and the end date of each 2 week period file is listed in Table 1 for the data in Dataset 2 and Table 2 for the data in Dataset 3, along with the notations used in this paper for each of the periods.

Table 1. Dataset 2 Recorded Periods

File	Start Date dd/mm/yyyy	End Date dd/mm/yyyy	Notation
SAMPLE 0	05/12/2011	18/12/2011	Week1-2
SAMPLE 1	19/12/2011	01/01/2012	Week3-4
SAMPLE 2	02/01/2012	15/01/2012	Week5-6
SAMPLE 3	16/01/2012	29/01/2012	Week7-8
SAMPLE 4	30/01/2012	12/02/2012	Week9-10
SAMPLE 5	13/02/2012	26/02/2012	Week11-12
SAMPLE 6	27/02/2012	11/03/2012	Week13-14
SAMPLE 7	12/03/2012	25/03/2012	Week15-16
SAMPLE 8	26/03/2012	08/04/2012	Week17-18
SAMPLE 9	09/04/2012	22/04/2012	Week19-20

Table 2. Dataset 3 Recorded Periods

File	Start Date dd/mm/yyyy	End Date dd/mm/yyyy	Notation
SAMPLE A	01/12/2011	15/12/2011	A
SAMPLE B	16/12/2011	30/12/2011	B
SAMPLE C	31/12/2011	14/01/2012	C
SAMPLE D	15/01/2012	29/01/2012	D
SAMPLE E	30/01/2012	13/02/2012	E
SAMPLE F	14/02/2012	28/02/2012	F
SAMPLE G	29/02/2012	14/03/2012	G
SAMPLE H	15/03/2012	29/03/2012	H
SAMPLE I	30/03/2012	13/04/2012	I
SAMPLE J	14/04/2012	28/04/2012	J

4.1 Identifying HotSpot Antennas

We define as HotSpot Antennas or highly loaded antennas, the antennas that present the highest number of active users over the 150 days period, including the activity of the same user. On the other side we define as Popular Antennas the antennas that were visited by distinct users over each of the two-week period, thus excluding the repeated activity of the same user. Consequently,

the Popular Antennas are the antennas with the highest user diversity.

From Dataset 2 we compute the overall activity of each antenna over the full monitoring period. Figure 3 illustrates the antennas locations and activity over the 150 days period, with their size and color representation reported to the load intensity. Thus, heavily loaded antennas are represented by wider points and higher intensity color.

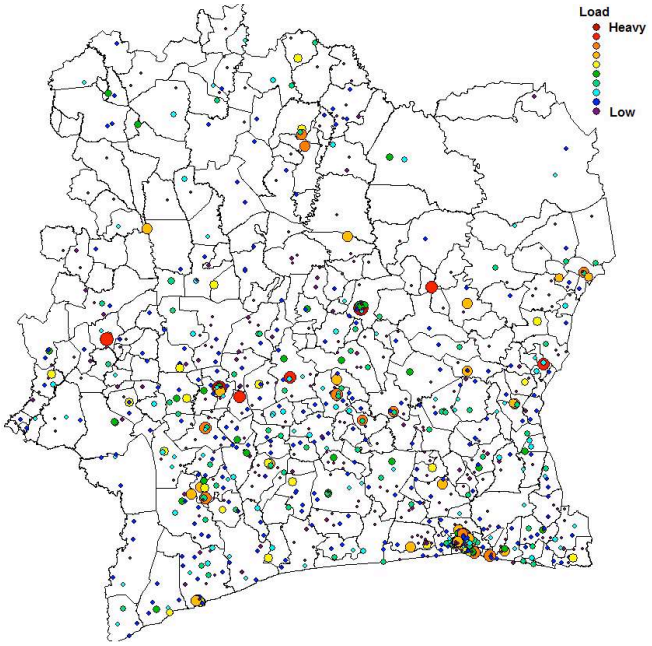


Figure 3. Orange antennas location and activity over 150 days.

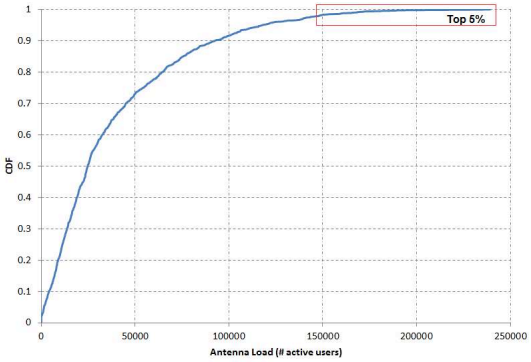


Figure 4. CDF of the number of active users at each antenna over 150 days.

The Cumulative Distribution Function (CDF) of the number of active users recorded at each antenna over the 150 days period is illustrated in Figure 4. The average load over the 150 days equals to 38333 active users and the standard deviation is about 38138 active users. This means that around 70% of the antennas have a load within one standard deviation to the mean, as seen in Figure 4. Looking only at the top 5% of the antennas, the number of active users is within 147000-235000 over the full 150 day period. The location of the top 5% of highly loaded antennas on the Ivory Coast is illustrated in Figure 5.

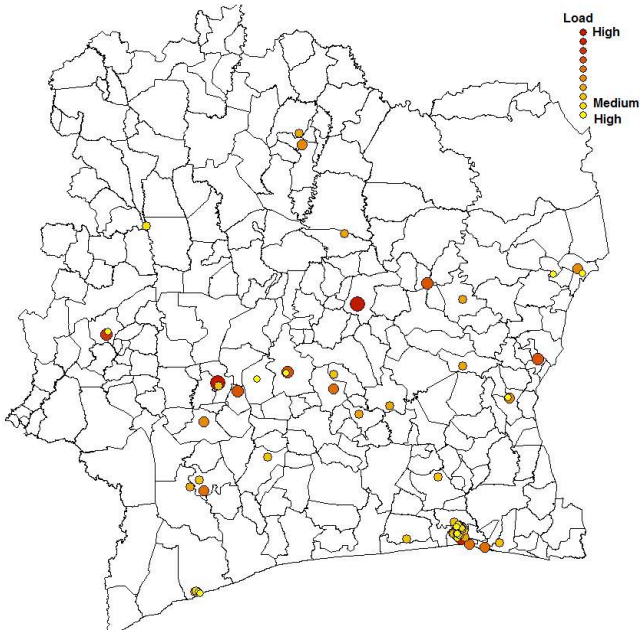


Figure 5. Top 5% of HotSpot Antennas over 150 days.

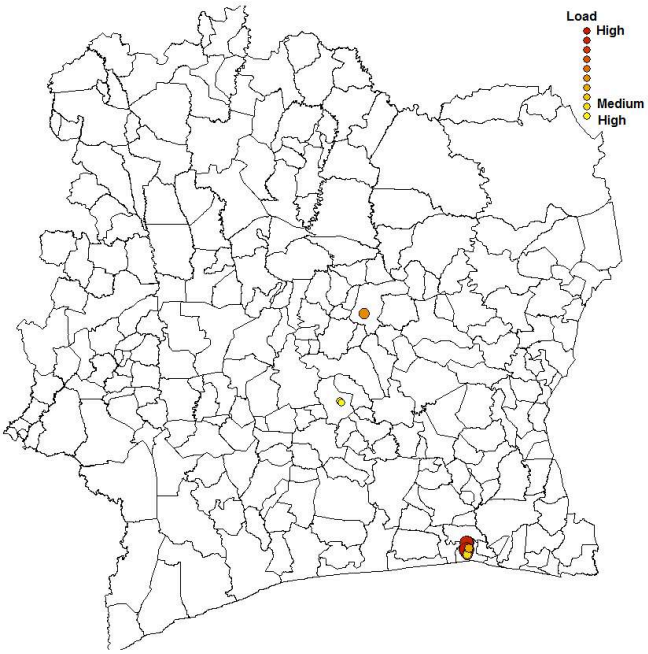


Figure 7. Top Most Popular Antennas over 150 days.

Figure 6 illustrates the average antennas activity for each of the two-week period. As it can be seen there is an increase in the activity at the beginning of February with a peak at the end of March beginning of April.

Figure 6. Average Antennas Activity over 150 days.

When analyzing the most Popular Antennas it was not possible to make the computation over the overall 150 days period as the dataset provides different random user identification number for each two-week period and we need the number of distinct users that were connected to each antenna over the time period. Thus we computed the top 5% of each two-week period and we took their intersection as the most Popular Antennas over the 150 days period. The top most Popular Antennas location is illustrated in Figure 7.

Figure 8 illustrates the average antennas user diversity for each of the two-week periods. It can be seen that antennas user diversity follows more or less the same distribution with the antennas activity. It can be seen that the peak in average antennas user diversity is at the end of March beginning of April.

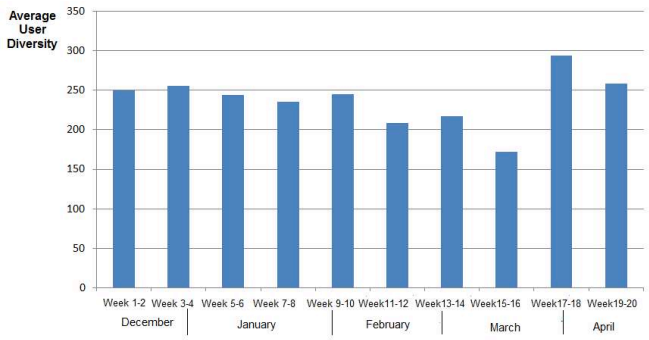


Figure 8. Average User Diversity over 150 days.

4.2 Identifying HotSpot Sub-Prefectures

From Dataset 3 we compute the overall activity of each sub-prefecture over the full monitoring period of 150 days. Figure 9 illustrates the sub-prefectures of the Ivory Coast and their activity over the 150 days period, with their color representation reported to the load intensity. Thus, heavily loaded sub-prefectures are represented by higher intensity color. When comparing the sub-prefectures activity map in Figure 9 with the indicative population map of Ivory Coast provided by Map Action[2] in Figure 10, it can be noticed that data usage activity is mostly registered in densely populated areas, as expected.

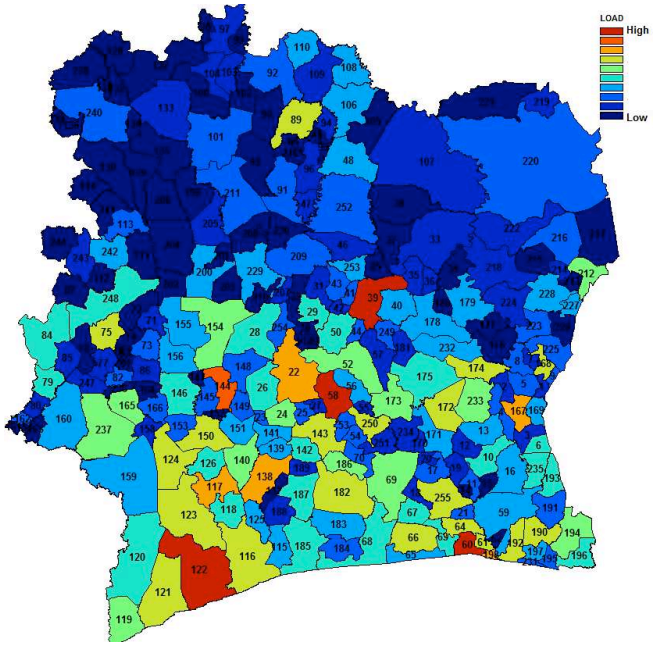


Figure 9. Sub-Prefectures Activity over 150 days.

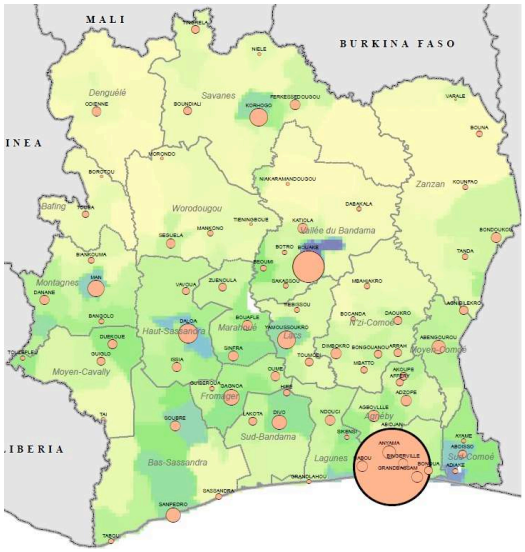


Figure 10. Ivory Coast Indicative Population Map.[2]

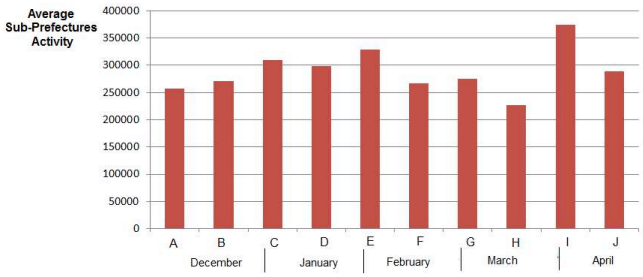


Figure 11. Average Sub-Prefectures Activity over 150 days.

[2] http://www.mapaction.org/component/mapcat/mapdetail/2383.html

The average Sub-Prefectures Activity over each two-week periods of recorded data is illustrated in Figure 11. The Figure shows a more clear distribution of the traffic usage, with several noticeable peak periods at the end of December, beginning of February and beginning of April.

5. CONNECTING PEOPLE AND LOCATIONS

Looking only at the study on Antennas Activity and User Diversity, by intersecting the Top identified HotSpot Antennas with the Top identified Popular Antennas we detected a set of highly loaded and high user density antennas over the full 150 days period. Looking at the location of the antennas, it was possible to identify their location within a certain sub-prefecture hence city. The results show that the antennas from the set are spread across three cities, such as Bouake, Yamoussoukro, and Abidjanas, as illustrated in Figure 12. These findings have significant impact and they can be correlated to the important cities of the country. Bouake is the second largest city in the Ivory Coast, Yamoussoukro is the official political and administrative capital, while the economic capital of the country and the city high the highest population density is Abidjanas.

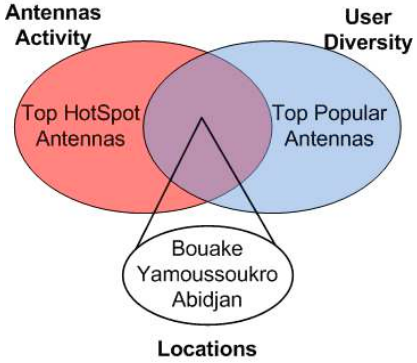


Figure 12. Connecting people and locations.

These observations led to the correlation between antennas activity and user density within a cellular network and their geographical location. Thus by analyzing the user activity and their mobility patterns within a cellular network only, it is possible to identify the major cities/locations within a country/city.

Understanding the people-location interaction could represent a potential for location-based services. For example time-independent interactions refer to overlapping trajectories between distinct people irrespective of the actual time of overlap. This information is very useful in social recommender systems which are based on location-based tagging services [16]. In these social recommender systems, users make use of the location-based services in order to obtain and share information (tags) about the points of interest in their surroundings.

In order to explore the idea on how many distinct people are likely to meet each other in a time-independent manner, we analyze the user diversity over each two-week period for the full 150 days. The results show that for a two-week period, a maximum number of 4.5% of the total 50,000 customers share the same location at a certain cell tower irrespective of the time and day.

From a lower resolution point of view, when looking at the study on the sub-prefectures activity as illustrated in Figure 9, the city with the highest activity is Abidjanas, followed by Yamoussoukro, Bouake and San Pedro.

6. CONNECTING PEOPLE, LOCATIONS AND REAL-WORLD EVENTS

From the spatio-temporal patterns of the collective customers' activity within the mobile network traffic datasets introduced previously, we analyze the correlation between people, locations and events. Specifically we are interested in studying the correlation between exceptional patterns detected in the mobile usage within a cellular network and real-world events such as public concerts, parades, soccer matches, riots protests, etc.

Understanding the exceptional data usage patterns could significantly improve the spatial and temporal awareness when taking decisions. An example would be in the case of event management, when organizing parades/carnivals/concerts, etc.

From the previous observations on the sub-prefectures activities we detected several peak periods which are highlighted below:

- *New Year's Eve*

The first observed substantial increase in the sub-prefectures activity as noticed in Figure 11, is at the beginning of January. This increase in users' activity is correlated with the New Year's Eve period. The detailed view of the overall Antennas activity during the New Year's Eve period is illustrated in Figure 13. It can be notice that there is a 40% increase in the users' activity on the first day of the New Year when compared with the previous day. Thus, the activity pattern of the customers perfectly correlates with the New Year's Eve.

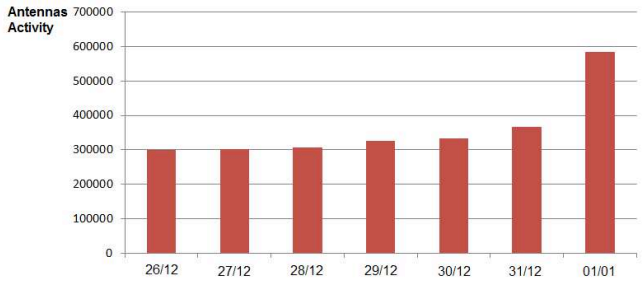


Figure 13. Antennas Activity over New Year's Eve.

- *Soccer Match*

The second substantial increase in the sub-prefectures activity form Figure 11 is at the end of January, beginning of February. This increase in user activity correlates with the duration of the Africa Cup of Nations 2012, also known as Orange Africa Cup of Nations[3] (Figure 14), which took place between 21st of January and 12th of February. In this competition Ivory Coast was defeated in the final, by Zambia, after a dramatic penalty shootout

[3] http://asia.eurosport.com

Figure 14. Orange Africa Cup of Nations.[3]

- *Carnivals/Parades*

The outstanding activity peak recorded over the 150 days period is at the beginning of April. Taking the antenna with the highest load and high user diversity we plot the average activity and user diversity over each two-week period covering the 150 days of recorded data in Figure 15 and Figure 16, respectively. It can be noticed that both sets follow more or less the same distribution with the highest user activity and diversity recorded at the end of March, beginning of April. Looking at the location of this antenna we see that is located within Bouake city.

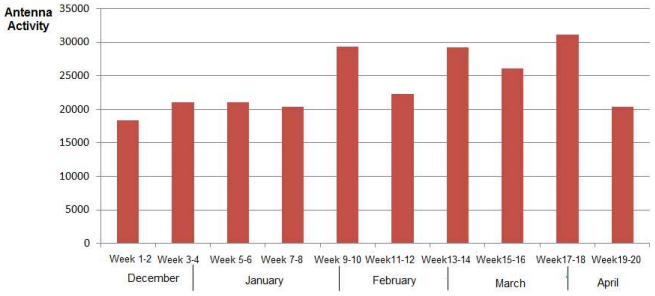


Figure 15. Top HotSpot Antenna Activity over 150 days.

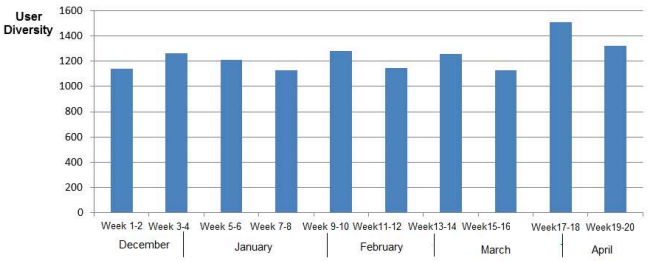


Figure 16. Top HotSpot Antenna User Diversity over 150 days.

Taking these observations and looking at the real-time events happening in that specific location during exactly that period, we come to know about the annual Bouake Carnival (Figure 17[4]). Thus these pattern exceptions in the antenna usage are perfectly correlated with the real-world event, such as Bouake Carnival. The Bouake Carnival is a week-long carnival happening each year around the ending of March through the first week of April and attracts thousands of visitors[4].

[4] http://face2faceafrica.com/

Figure 17. Bouake Carnival.[4]

Figure 18 illustrates the Bouake Antenna Activity and User Diversity per each day over Week17-18. Looking at the results, it can be seen that there is a significant increase in both activity and user diversity for Week 18 when compared with the previous Week17. This is because of the Bouake Carnival which is happening throughout the first week of April, when people come from all over the world to attend the festival, with a peak activity on Friday and Saturday as spotted in Figure 18.

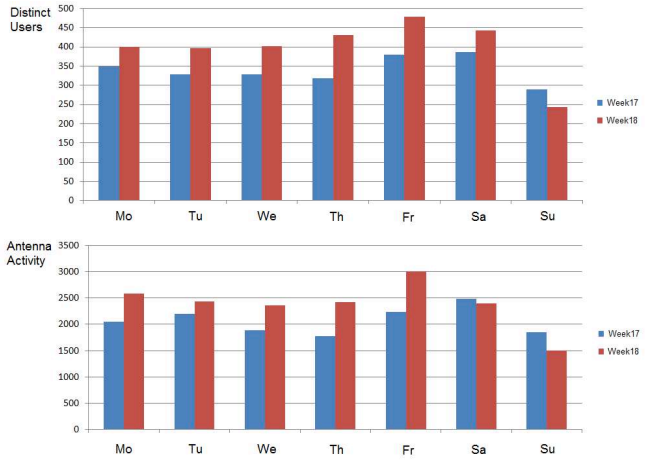


Figure 18. Bouake Antenna Activity and User Diversity for each day of Week17-18.

- *Weekends*

Figure 19 presents the average antennas activity for each week day over the 150 days monitored period. It can be noticed that users tend to have an increased cellular activity towards the end of the week, with the highest peak on Fridays, whereas the lowest activity is recorded on Tuesdays.

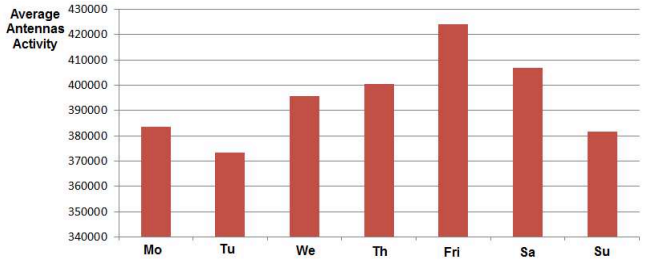


Figure 19. Average Antennas Activity on Week Days over 150 days period.

7. USERS MOBILITY AND ACTIVITY

In order to get a more general view of the user mobility, we computed the CDF of the number of distinct locations visited by any of the 50,000 customers over Week17-18 period. The Week17-18 period was selected as is the period with the highest recorded antennas activity and user diversity, thus the highest user mobility period. The CDF is illustrated in Figure 20. The average number of distinct locations is around 8 and the standard deviation is 9.81. This means that 70% of the users have less than 8 distinct locations, whereas around 9% of the customers have been seen in more than 20 distinct locations.

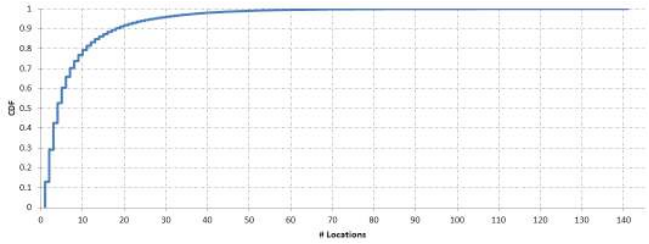


Figure 20. CDF of the number of distinct locations visited by each user over Week17-18.

Looking at the customers' activity during week days as illustrated in Figure 19, we notice that the users tend to be more active as the weekend is approaching and less active as the working days start.

In order to have a higher resolution on customers activity, Figure 21 illustrates the average antennas activity over the 150 days period for as recorded on Wednesdays, Fridays and Sundays. It can be seen that during the night hours there is not much traffic during work days with a small increase during weekends. We can notice that the users' activity during working days (Wednesdays and Fridays) follows the same distribution with a higher increase on Fridays. The users tend to communicate more before and after working hours with an almost constant activity during work. However in weekend, there is a reduced activity during morning hours when compared with the working days.

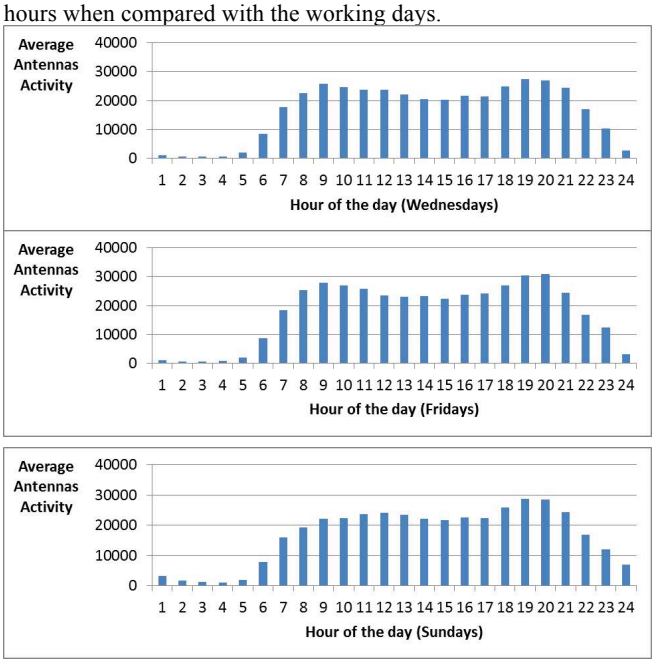


Figure 21. Average Antennas Activity on Wednesdays, Fridays, and Sundays over 150 days period.

8. CONCLUSIONS

In this work we explore the use of anonymized Call Detail Records (CDRs) containing both voice-calls and SMS activities, from a cellular network in Ivory Coast in order to connect people, locations and events.

The study presented in this paper, shows that CDR data can be used to detect exceptional spatio-temporal patterns of the collective human mobile data usage and that these 'anomalies' in the usage patterns can be correlated to real-world events (e.g., soccer match, parade/carnival, etc.). Understanding the exceptional data usage patterns could significantly improve the spatial and temporal awareness when taking decisions and this knowledge could be further used to develop an *intelligent system that detects exceptional events in real-time* from CDRs monitoring. For example, a real-time event detection system could be of crucial importance to ensure people's safety in case of emergency situations, such as riots protests which could be more efficiently handled if detected on time.

9. ACKNOWLEDGMENTS

We would like to thank the D4D Challenge committee and the Orange Group for Development and Initiative for the datasets provided in this study.

10. REFERENCES

[1] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns", *Nature*, 453:779-782, 2008.

[2] C. Song, Z. Qu, N. Blumm, and A.-L. Barabasi, "Limits of predictability in human mobility", *Science*, 327(5968):1018-1021, 2010.

[3] S. Isaacman, R. Becker, R. CÃa¸ceres, S. Kobourov, J. Rowl, and E. Varshavsky, "A Tale of Two Cities", in *Proc. Workshop on Mobile Computing Systems and Applications (HotMobile),* 2010.

[4] G. Di Lorenzo and F. Calabrese, "Identifying human spatio-temporal activity patterns from mobile-phone traces," in *14th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, 2011.

[5] F. Calabrese, C. Ratti, M. Colonna, P. Lovisolo, D. Parata, "A system for real-time monitoring of urban mobility: a case study in Rome", *IEEE Transactions on Intelligent Transportation Systems,* 12(1) 141–151, 2011.

[6] Y. De Mulder, G. Danezis, L. Batina, and B. Preneel "Identication via location-proling in gsm networks" in *Proc. of the 7th ACM workshop on Privacy in the electronic society*, pp 23-32, 2008.

[7] H. Zang and J. Bolot, "Mining call and mobility data to improve paging efficiency in cellular networks", in *Proc. of the 13th annual international conference on Mobile computing and networking, (ACM MobiCom)*, 2007.

[8] R. Becker, R. C´aceres, K. Hanson, J. M. Loh, S. Urbanek, A. Varshavsky, and C. Volinsky, "Classifying routes using cellular handoff patterns", *Proc. of Netmob,* 2011.

[9] Z. M. Mao, C. D. Cranor, F. Bouglis, M. Rabinovich, O. Spatscheck, and J. Wang "A Precise and Efficient Evaluation of the Proximity between Web Clients and their Local DNS Servers", in *Proc*. *USENIX Annual Technical Conference*, 2002.

[10] S. Mathur, T. Jin, N. Kasturirangan, J. Ch, W. Xue, and M. Gruteser, "ParkNet: Drive-by Sensing of Road-Side Parking Statistics" In *Proc*. *ACM MOBISYS*, 2010.

[11] I. Trestian, S. Ranjan, A. Kuzmanovic, and A. Nucci, "Measuring serendipity: connecting people, locations and

interests in a mobile 3G network", in *Proc. of ACM IMC*, pp. 267-279, 2009.

[12] S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, M. Martonosi, J. Rowland, and A. Varshavsky, "Identifying important places in people's lives from cellular network data," in *Proceedings of the 9th international conference on Pervasive computing*, Berlin, Heidelberg, 2011.

[13] R. Keralapura, A. Nucci, Z.-L. Zhang, and L. Gao, "Profiling users in a 3g network using hourglass co-clustering," *MOBICOM,* 2010.

[14] F. P. Tso, J. Teng, W. Jia, and D. Xuan "Mobility: A Double-Edged Sword for HSPA Networks", in *Proc*. *ACM MOBIHOC*, 2010

[15] G. Sagl, M. Loidl, and E. Beinat, "A Visual Analytics Approach for Extracting Spatio-Temporal Urban Mobility Information from Mobile Network Traffic," *ISPRS International Journal of Geo-Information*, vol. 1, no. 3, pp. 256–271, 2012.

[16] C. Biancalana, F. Gasparetti, A. Micarelli, G. Sansonetti. "Social Tagging for Personalized Location-Based Services" in *Social Recommender Systems (SRS), a Workshop of The 2011 ACM Conference on Computer Supported Cooperative Work*, 2011.

Towards an early warning system: the effect of weather on mobile phone usage
A case study in Abidjan

João Pedro Craveiro[1], Fernando M. V. Ramos[1], Eiman Kanjo[2], and Nour El Mawass[2]

[1] Universidade de Lisboa, Faculdade de Ciências, LaSIGE (Lisbon, Portugal)

jcraveiro@lasige.di.fc.ul.pt, fvramos@fc.ul.pt

[2] Information Systems Department, CCIS, King Saud University (Riyadh, Saudi Arabia)

ekanjo.c@ksu.edu.sa

Abstract: In mobile networks, traffic activity within particular network cells follows predictive patterns. Relevant changes in these patterns may indicate a local problem (an emergency or any other sporadic event). Detecting these changes could therefore be used as an early warning system. The hypothesis we aim to address is whether we are able to detect unexpected changes in weather quickly, by monitoring changes in cell patterns. The motivation is the fact that it is usually easier and cheaper to prevent damage provoked by weather conditions than to reverse the damage. This is particularly relevant in the socioeconomic context of developing countries such as Ivory Coast. In this paper, and as a first step towards the development of an early warning system, we jointly analyse mobile data and historic records of weather conditions. We employ exploratory factor analysis to reveal latent variables in the weather data, and spectral analysis to exploit the periodicity of mobile and weather data (both independently and jointly). From these results we derive a model which, in spite of its current limitations, hints that our hypothesis may be viable if conducted with higher quality weather data.

1 Introduction

In mobile networks, traffic activity within particular network cells follows predictive patterns [20]. Each cell has therefore its own "traffic signature". Relevant changes in the signature of a particular cell in a particular time period may indicate a local problem (an emergency or any other sporadic event). Detecting anomalies in a "cell signature" (changes in the normal pattern) could therefore be used as an early warning system. A relevant spike in the data, for example, may be used to detect an emergency situation.

The question we aim to address by analysing the D4D mobile dataset is to understand if we are able to detect unexpected changes in weather quickly, by monitoring changes in cell signatures. The motivation is the fact that it is usually easier and cheaper to prevent damage provoked by weather conditions than to reverse the damage. This is particularly relevant in the socioeconomic context of developing countries such as Ivory Coast. We have used the mobile dataset from Orange in conjunction with datasets containing historic records of weather conditions for the analysis. The main hypothesis we intend to explore is that *there is an influence of weather conditions on mobile phone usage which allows to predict, to some extent, the former from the latter*. Our motivation is the fact that the coverage of mobile network antennas is wider than the presence of weather stations, especially in Africa [17], and also that this may serve as a fallback mechanism in the event of failure in a weather station (either due to malfunctioning or communication problems).

Contributions In this paper, we jointly analyse mobile data and historic records of weather conditions towards an early warning system capable of detecting unexpected changes in weather conditions. Our contributions are:

- we have reproduced Sagl et al. [20]'s experience, both to validate the datasets we have available and to confirm the authors' results; additionally, we present what we believe to be a more correct interpretation of the exploratory factor analysis results presented in that paper;
- with the final aim of generating a predictive model of weather conditions we have generated and evaluated an autoregressive integrated moving average (ARIMA) model and its respective forecast.

Outline In Section 2, we briefly survey related work. In Section 3, we describe the case study, the datasets

used, and how we processed them. In Section 4, we perform exploratory factor and spectral analysis centred on our weather data, in a similar vein as Sagl et al. [20]'s work, although presenting what we believe to be a more accurate interpretation of the results. In Section 5, we derive a model to tentatively predict overall weather conditions from mobile data. In Section 6, we interpret and discuss our results and the quality thereof. Finally, Section 7 closes the paper with concluding remarks and future work directions.

2 Related Work

Sagl et al. [20] explored the influence of weather on mobile phone usage and (indirectly) human behaviour. The authors explore three types of area — urban, coastal, and mountain — in Northern Italy. In this paper, we reproduce and extend their analysis, providing a different view on some results, namely those from their exploratory factor analysis. We present more details in Section 4. Sagl et al. [21] extend this experiment by applying analysis methods from the time-, space-, and frequency domain.

Lu et al. [16] also considered the use of mobile phone tracking as an emergency response system, by analysing the movements of Haitians in the January 2010 earthquake. Despite sharing similar goals, we perform a different type of analysis, not least because the period covered by the mobile data we analyse does not include any such extreme natural event (at least not one we are aware of).

Phithakkitnukoon et al. [19] study the interaction between weather and mobile phone usage with an emphasis on the social relationships aspect, a different focus from our work.

Contemporary to the work described in this paper, Overeem et al. [17] use mobile network antennas for real-time rainfall monitoring. However, they rely on measurements of the signal's rain-induced attenuation between transmitter and receiver, and not on call data records.

3 Case Study

Abidjan is the economic capital of Côte d'Ivoire. Between 1933 and 1983 it was the political and administrative capital of the country, until President Félix Houphouët-Boigny transferred that status to Yamoussoukro. As such, it is a urban area which has experienced growth over the years and, despite no longer being the capital, it is still the greatest and most populated city in the country, and the centerpoint of many activities — government, economics, business, industry.

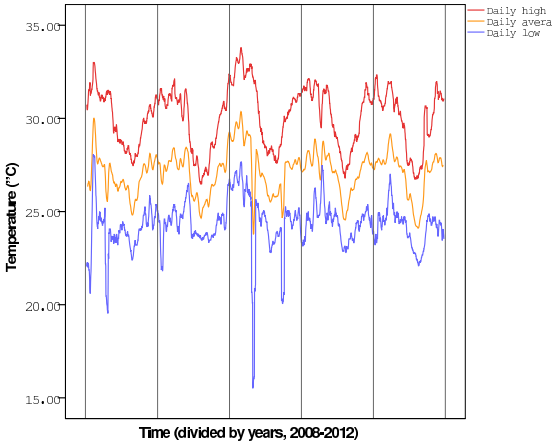


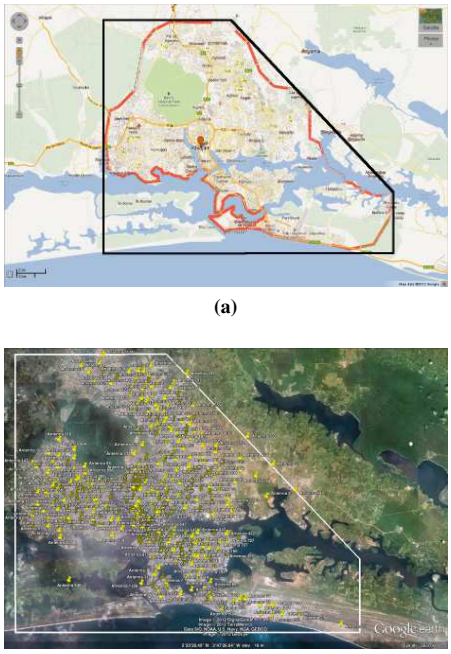
Figure 1: Daily maximum, mean and minimum temperatures in Abidjan throughout the years 2008–2012. Data source: [4]

The city's temperature does not vary considerably, as is typical near the Equator. Maximum temperatures never exceed 35 degrees Celsius, and minimum temperatures are seldom below 20 degrees Celsius; daily mean temperatures are consistently between 25 and 30 degrees Celsius (Figure 1). Precipitation is mostly concentrated in two distinct periods: May–June and October–November [1, 2].

3.1 Datasets used

3.1.1 Mobile network traffic data

The original data collection took place in Côte d'Ivoire over a five-month period, from December 2011 to April 2012. The original dataset contains 2.5 billion records, calls and text messages exchanged between 500 000 users. The customer identifier was anonymised by Orange Côte d'Ivoire. All subsequent data processing was completed by Orange Labs in Paris. We had access to four sets of processed data [7]. We now describe the one we used for this work: *Antenna-to-antenna (SET1)*. For this dataset, the number of calls as well as the duration of calls between any pair of antennas have been aggregated hourly. Calls spanning multiple time slots are considered to be in the time slot they started in. Antennas are uniquely identified by an antenna id and a geographic location. This data is available for the entire observation period. Communication between Orange customers and customers of other providers have been removed. The antenna-to-antenna traffic data is provided in ten TSV (tab-separated values) files, each corresponding to 14 days. Each line in one of these TSV files provides the number of calls as well as the traffic



(a)

(b)

Figure 2: Maps of Abidjan (a) the considered bounding polygon over the administrative border provided by Google Maps, and (b) geographical distribution of these antennas in Abidjan with the bounding polygon superimposed (Google Earth).

intensity (expressed in calls.seconds [1]) between a pair of antennas for a given hour. The relation between antenna identifiers and (for security reasons, approximate) geographic coordinates is established in a separate TSV file [7]; this file contained the location of 1231 antennas, labeled 1 to 1238 (with gaps).

We further processed this data in the following manner:

1. we established a convex polygon approximately delimiting the city of Abidjan; Figure 2a shows the considered polygon laid over the administrative border of Abidjan on Google Maps[2].

2. we processed the antenna positions' TSV file to collect the identifiers of antennas lying inside the considered convex polygon, using the exterior edges strategy to test if each antenna is inside the polygon [12]; this resulted in a set of 385 antennas (see Figure 2b for the geographical distribution of these antennas in Abidjan);

3. we computed the total amount and traffic intensity of calls *originating* in the antennas which were found to be inside Abidjan.

[1] A traffic intensity of, for instance, 60 calls.seconds corresponds to 1 call with a duration of 60 seconds, or 2 calls with an average duration of 30 seconds each, etc.

[2] `http://maps.google.com/`

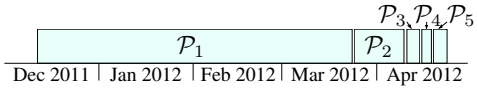


Figure 3: Time periods $\mathcal{P}_1$ to $\mathcal{P}_5$.

The result of our processing was a TSV file with the following kind of information for the whole period:

```
2012-04-22 21:00:00     72484   8450304
2012-04-22 22:00:00     18230   3849108
2012-04-22 23:00:00     29428   14512650
```

We will henceforth refer to these two mobile data variables as Number and Erlang. Other than punctual missing values, there were four significant gaps (between 15 and 20 hours each) in the available data. For this reason, we broke our data in 5 periods (graphically represented in Figure 3):

$\mathcal{P}_1$ 12 December 2011 0h00 (inclusive) to 23 March 2012 23h00 (exclusive);

$\mathcal{P}_2$ 24 March 2012 15h00 (inclusive) to 9 April 2012 23h00 (exclusive);

$\mathcal{P}_3$ 10 April 2012 18h00 (inclusive) to 14 April 2012 23h00 (exclusive);

$\mathcal{P}_4$ 15 April 2012 20h00 (inclusive) to 18 April 2012 23h00 (exclusive);

$\mathcal{P}_5$ 19 April 2012 15h00 (inclusive) to 23 April 2012 0h00 (exclusive).

For this case study, we use only data pertaining to period $\mathcal{P}_1$, whereupon we replaced missing values of Number and Erlang in 2 cases (0.08 %) through linear interpolation.

3.1.2 Weather data

Historical weather data pertaining to Abidjan for the considered period was downloaded from Wunderground [4] in CSV (comma-separated values) files. Each line of the files is structured as follows:

```
8:00 AM,26.0,25.0,94,1009,4.5,SSW,16.7,-,N/A,
     Thunderstorm,Light Thunderstorms,200,2012-04-09
      08:00:00
```

presenting, in order the following information: time of collection, air temperature (in degrees Celsius), dew point (in degrees Celsius), relative humidity, sea level pressure (in hectopascals), visibility (in kilometres), wind direction (expressed as a cardinal point), wind speed (in kilometres per hour), wind gust speed (in kilometres per hour), precipitation (in millimetres), events (e.g. "Rain", "Thunderstorm"), conditions (e.g. "Mostly Cloudy", etc.), wind direction (in degrees), and

Table 1: Rain and thunderstorm — correspondence between ordinal variable values and qualitative measurements

Value	Rain variable	Thunderstorm variable
0	No rain	No thunderstorm
1	Light Rain	Light Thunderstorm
2	Rain	Thunderstorm
3	Heavy Rain	Heavy Thunderstorm

date/time (UTC). Local time in Côte d'Ivoire corresponds to UTC all year round, as Côte d'Ivoire does not observe daylight savings time, so no special consideration was needed. The data was collected by a weather station located in the Port Bouet Airport (Félix Houphouët-Boigny International Airport).

Regarding precipitation, we were not able to rely on the quantitative measurements of rainfall, since the weather station which collects the data does not provide values for precipitation in millimetres in the observed period (for unknown reasons it began providing them in the daily reports on 9 April 2012 — i.e., at the end of period $\mathcal{P}_2$). For this reason, we only draw considerations regarding rainfall and the occurrence of thunderstorms using the qualitative information present in the data: the variables "events" and "conditions". We establish two ordinal variables, whose values correspond to qualitative indications, extracted from the textual events and conditions indications, as shown in Table 1. For the considered period ($\mathcal{P}_1$, we had to replace missing values in the air temperature, relative humidity, and sea level pressure variables in, respectively, 6 (0.24 %), 6 (0.24 %), and 8 (0.32 %) cases, through linear interpolation.

4 Exploratory Data Analysis

Sagl et al. [20] explore the influence of weather on mobile phone usage and (indirectly) human behaviour. The authors explore three types of area — urban, coastal, and mountain — in Northern Italy. They perform factor analysis to extract hidden components in the correlated input variables, and then perform spectral analysis to find correlations between these components and mobile phone usage data. We now reproduce their experience, with two main goals:

1. to assess the quality of our data, and
2. to compare results, since Abidjan is both an urban and coastal setting.

4.1 Assumptions for Factor Analysis

One of the major uses of factor analysis is to identify "latent variables in large data sets that are represented by highly correlated variables" [8]. We expect, as an hypothesis, our weather data to have this property. We restrict our analysis to five variables in the original data: air temperature (AT), relative humidity (RH), sea-level air pressure (AP), and the qualitative empirical measurements of rainfall (RE) and thunderstorm occurrence (TE) coded as the ordinal variables described in Table 1. We chose these variables mostly to have a setting as similar as possible to that used in [20].

Before advancing to an exploratory factor analysis, we should verify that our data meets the assumptions for such type of analysis. We resort to two statistical measures for this purpose: Bartlett [6]'s test of sphericity and the Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy (MSA) [14, 15].

The results for both tests are presented in Table 2. Bartlett's test of sphericity reveals a significant ($\mathrm{Sig.} < .05$) correlation between the considered variables. The KMO MSA falls slightly below the minimum suggested of .6 for a good factor analysis [18]. The same applies to the measures of sampling adequacy (MSA) in the anti-image correlation matrix (Table 3). As the difference is relatively small, we nevertheless proceed with exploratory factor analysis. We discuss this issue further in Section 6.

Table 2: Kaiser-Meyer-Olkin (KMO) and Bartlett's Test

KMO Measure of Sampling Adequacy		.518
Bartlett's Test of Sphericity	Approx. χ^2	2144.545
	df	10
	Sig	.000

Table 3: Anti-image Correlation Matrix

	AT	RH	AP	RE	TE
AT	.488[a]	.346	.040	.079	-.010
RH	.346	.496[a]	-.021	.022	-.020
AP	.040	-.021	.986[a]	.013	.021
RE	.079	.022	.013	.858[a]	-.301
TE	-.010	-.020	.032	-.301	.874[a]

a. Measures of Sampling Adequacy (MSA)

4.2 Factor Analysis

We extracted the factors using Principal Component Analysis, selecting the components using the Kaiser's criterion — eigenvalues of 1.0 and above. As we can see in Table 4, we extracted two components which explain 62.132% of the total variance.

The rotated component matrix in Table 5 summarises the final loading of the five considered variables for the two extracted components. The overall patterns of our

Table 4: Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
PC	**Total**	**% Var.**	**Cumul. %**	**Total**	**% Var.**	**Cumul. %**	**Total**	**% Var.**	**Cumul. %**
1	1.803	36.059	36.059	1.803	36.059	36.059	1.728	34.564	34.564
2	1.304	26.073	62.132	1.304	26.073	62.132	1.378	27.569	62.132
3	.959	19.188	81.320						
4	.646	12.926	94.245						
5	.288	5.755	100.000						

Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization.

Table 5: Rotated Component Matrix

	Principal Component (PC)	
	1	**2**
AT	-.903	-.136
RH	.903	.089
AP	.299	-.181
RE	.076	.806
TE	-.040	.818

Extraction Method: Principal Component Analysis
Rotation Method: Varimax with Kaiser Normalization

components are similar to those observed by Sagl et al. [20] for the urban and coastal areas in Northern Italy — which is an interesting effect, since Abidjan is an area which is both urban and coastal. We nevertheless disagree with the interpretation done there for the extracted components. Factor analysis is employed to reveal hidden variables (factors) in data with highly correlated input variables; for instance, as in the example given in [8], one can use factor analysis on an input consisting of grades on multiple test grades and find correlations between test grades (e.g., Reading and Spelling, or Arithmetic and Geometry) which reveal hidden variables (e.g., respectively, linguistic and quantitative capabilities). Then, any possible considerations given how "good" or "bad" a case is may be extracted, on a factor basis, from the score of said case for each factor. A factor itself does not have an intrinsic "good" or "bad" nature. Hence, it is our understanding that the Nice Weathervs. Bad Weatherinterpretation by Sagl et al. [20, 21] is misleading.

Our Principal Component 1 is heavily loaded by air temperature (negatively) and relative humidity (positively). These variables are strongly associated with how humans "perceive" the weather, and we will therefore term this component ThermalPerception. Cases with high scores on this factor will correspond to times at which the weather is perceived as "cold", whereas lower scores correspond to weather perceived as "warm". Moreover, the heavy loadings with opposite signs of air temperature and relative humidity are consistent with the effect relative humidity has on how humans beings perceive temperature.

Principal Component 2 is heavily positively loaded by our empiric rain and thunderstorm variables, and we will therefore term this component Precipitation. Cases with higher scores on this factor will correspond to more inclement weather, whereas lower scores correspond to milder weather, with near or total absence of either rain or thunderstorm.

Note that, depending on its value, a single factor (be it ThermalPerception or Precipitation) can on its own correspond to either "nice weather" or "bad weather", which we believe clearly demonstrates how the original interpretation by Sagl et al. [20, 21] is misleading.

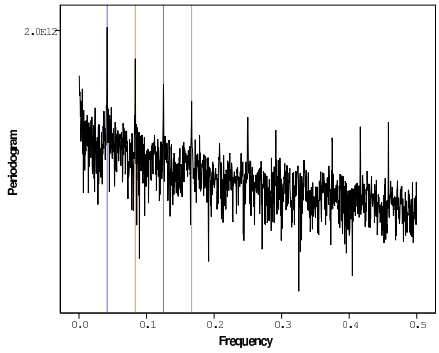
We estimated the components' scores using the Anderson–Rubin [5] approach, saving them as variables to be used in the remaining analysis.

4.3 Univariate Spectral Analysis

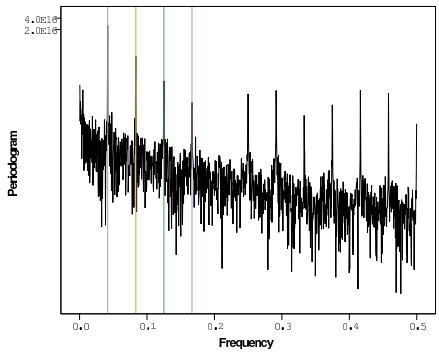
Univariate spectral analysis serves the purpose of revealing periodical components in time series. As Sagl et al. [20], we perform spectral analysis on the meteorological factor scores (resulting from factor analysis) and on mobile data.

The peaks in the periodogram for the mobile data variables (Figure 4) are consistent with those found by Sagl et al. [20]. Most significantly, the first and most pronounce peak at frequency $0.041666\ldots = 1/24$ highlights a circadian pattern in mobile traffic.

Regarding weather data, namely the scores for the extracted components, the periodograms in Figure 5 also provide some interesting information. The peaks in the periodogram for the ThermalPerception component also reveal a circadian pattern, which comes naturally from the obvious influence of daylight on temperature. On the other hand, the periodogram for the Precipitation component does not feature such isolated peaks, which

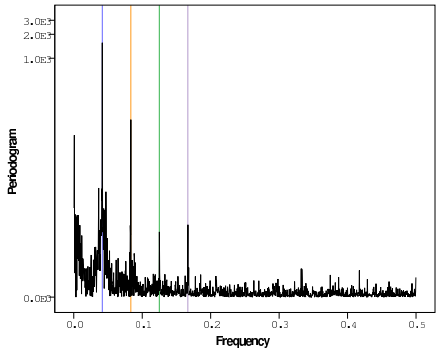


(a)

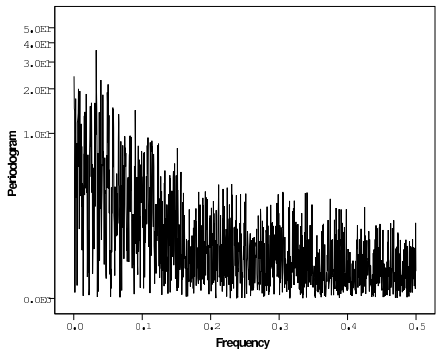


(b)

Figure 4: Periodogram of mobile data by frequency: (a) Number, and (b) Erlang. X-axis reference lines at frequencies corresponding to periods, from left to right: 24 hours (blue), 12 hours (orange), 8 hours (green), and 6 hours (purple).



(a)



(b)

Figure 5: Periodogram of weather data by frequency: (a) ThermalPerception component, and (b) Precipitation component. X-axis reference lines at frequencies corresponding to periods, from left to right: 24 hours (blue), 12 hours (orange), 8 hours (green), and 6 hours (purple).

is also natural given that rain and thunderstorm are not strongly influenced by the day–night cycle. Had we analysed data for a larger time interval, we would expect peaks at lower frequencies.[3]

4.4 Bivariate Spectral Analysis

Bivariate Spectral Analysis techniques, of which we focus on the squared spectral coherence (γ^2) between two variables, allow evaluating how the latter are correlated at specific frequency bands. We obtain the squared spectral coherence plots between the scores for each of the components (ThermalPerception and Precipitation) and the Number and Erlang variables.

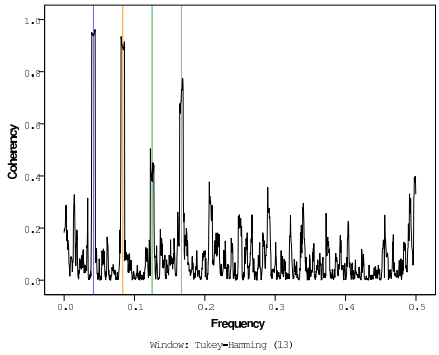
[3]In fact, we performed a simple analysis with daily Abidjan weather for 2008–2012 [4], and the periodograms for both the occurrence of rain and thunderstorms present their highest peak at 6-months periodicity, whereas mean daily values of air temperature and humidity feature a yearly periodicity.

4.4.1 ThermalPerception component

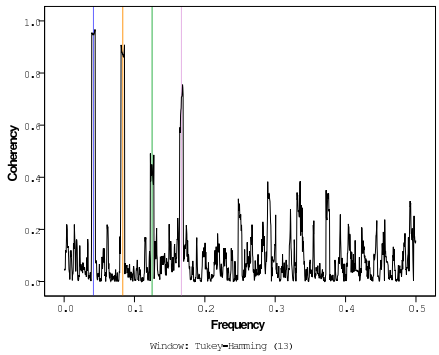
Figure 6 shows the squared spectral coherence plots the ThermalPerception component scores and Number (Figure 6a)/Erlang (Figure 6b). Just as in [20], the two most significant peaks are at frequencies corresponding to 24-hour and 12-hour periods; also similar is the fact that these two peaks are at similar levels (at approximately $\gamma^2 = .9$). However, our third peak is at the frequency corresponding to a 6-hour period, rather than an 8-hour one. We conjecture this to be due to differences in social aspects between Northern Italy and Côte d'Ivoire; for example, Abidjan and other cities in Côte d'Ivoire still have the main points of entry and exit closed between midnight and 5am, as a reminiscence of the curfews in civil war [3].

4.4.2 Precipitation component

Figure 7 shows the squared spectral coherence plots the Precipitation component scores and Number (Figure 7a)/Erlang (Figure 7b). Here, the most prominent

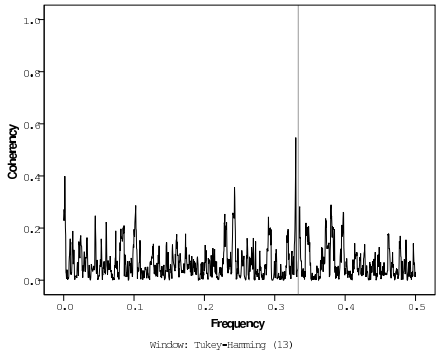


(a)

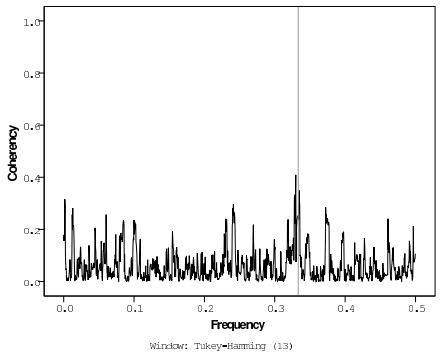


(b)

Figure 6: Squared coherence, by frequency, of the ThermalPerception component and: (a) Number; (b) Erlang. X-axis reference lines at frequencies corresponding to periods, from left to right: 24 hours (blue), 12 hours (orange), 8 hours (green), and 6 hours (purple).



(a)



(b)

Figure 7: Squared coherence, by frequency, of the Precipitation component and: (a) Number; (b) Erlang. X-axis reference line at the frequency corresponding to a 3-hour period (gray).

peak is found at the frequency corresponding to a 3-hour period, and even so at small values ($\gamma^2 \sim= .5$ for Number and $\gamma^2 \sim= .4$ for Erlang). We do not find the peaks at 24-hour and 12-hour periods which Sagl et al. [20] did. This is most likely due to the coarser granularity of our precipitation data (qualitative empirical measure vs. precipitation in millimetres).

4.5 Prognosis for data modeling

Mobile data and the ThermalPerception component of weather data present high spectral coherence at the 24-hour and 12-hour periods. This hints that there is some ground to try and explore these data towards a predictive model. Unfortunately, the same does not happen for the Precipitation component. We are unsure if this is related with the very coarse granularity of our weather data, and this is something we want to investigate further. Nevertheless, we will derive models for both components of weather data to confirm these indications.

5 Data Modelling

5.1 ARIMA and Seasonal ARIMA

A (non-seasonal) ARIMA model is defined as ARIMA(p, d, q), where p, d and q correspond to the orders of the **A**uto**R**egressive, **I**ntegrated and **M**oving **A**verage factors, respectively.

A seasonal ARIMA model is defined as

$$(p, d, q) \times (P, D, Q)_s \ ,$$

i. e., an ARIMA(p, d, q) model with an additional seasonal part, (P, D, Q). The structure of the seasonal part is the same as the non-seasonal (autoregressive, integrated, and moving average factors), with the difference that the factors operate over multiples of lag s (where s is the number of periods in a season). In our case, the "seasonal" aspect corresponds to the circadian cycle in our hourly mobile and weather data — $s = 24$. The autoregressive and moving average parts of the model, both seasonal and non-seasonal, concern to the consideration of past values in the time series in the predictions. We explain them more in the detail with the models we obtain.

Table 6: Model Description

PC	Model	Model Type
FAC1 (ThermalPerception)	Model 1	ARIMA$(2,0,2)$ $\times(1,1,1)_{24}$
FAC2 (Precipitation)	Model 2	ARIMA$(1,0,10)$ $\times(1,0,1)_{24}$

We opt for Seasonal ARIMA models because both mobile data and weather data show evidence of being both non-stationary (i.e., the mean, variance and/or autocorrelation structure of the time series change over time) and seasonal (periodic).

5.2 Model generation

To create the models, we invoke the Expert Modeler in SPSS, which deals with all the process of model creation including some preliminary steps (e.g., differencing to achieve stationarity). Another aspect which the Expert Modeler includes is automatic detection of outliers in time series [13]. We do not use this feature to create our models, since we are interested in capturing possible effects of observations which diverge significantly from the rest of the time series. We use the data for the first 8 weeks in period $\mathcal{P}_1$ as the estimation period (training set), with the remaining of period $\mathcal{P}_1$ being the forecast period (testing set).

The resulting Seasonal ARIMA models are described in Table 6. Regarding Model 1, generated for the ThermalPerception component, the non-seasonal autoregressive factor ($p = 2$) specifies that, to predict the value for a given hour, the value of the series two hours in the past is used. A null non-seasonal integrated factor ($d = 0$) indicates that the time series did not exhibit a trend to be removed by differencing. The non-seasonal moving average factor $q = 2$ specify that deviations from the mean value of the series from each of the last two hours are considered when predicting a value of the series. Regarding seasonal factors:

- $P = 1$ specifies that, to predict the value for a given hour, the value of the series one day in the past at the same time (i.e., 24 hours in the past, as expected) is used;
- $D = 1$ indicates that a linear trend in the successive values for the same hour of each day was removed by differencing;
- $Q = 1$ indicates that the model considers the deviations from the mean value of the series from the previous day at the same time.

As for Model 2, generated for the Precipitation component, its non-seasonal and seasonal factors specify that:

Table 7: Model Statistics (excerpt)

	No. of	Stat.	Ljung–Box Test		
Model	**Predictors**	R^2	**Q(18)**	**DF**	**Sig.**
1	0	.563	6.047	12	.917
2	1	.144	25.732	11	.007

- the value for the previous hour is used in the prediction ($p = 1$);
- the time series did not exhibit a trend to be removed by differencing ($d = 0$);
- deviations from the mean value of the series from each of the last ten hours are considered ($q = 10$);
- the value of the series 24 hours in the past (previous day at the same time) is used in the prediction ($P = 1$);
- the time series did not exhibit a trend in the successive values for the same hour of each day ($D = 0$), and;
- the model considers the deviations from the mean value of the series from the previous day at the same time ($Q = 1$).

These are interesting results, in the sense that they correspond quite closely to the way people empirically try to predict how the weather is going to be. For instance, $p = 2$ in Model 1 shows parallel with the assumed principle of locality that near future temperature predictions can be made from the temperature observed a couple of hours before. On the other hand, $p = 1$ in Model 2 is consistent with the notion that we do not have the same degree of confidence w.r.t. to precipitation — we cannot so easily infer that it will (resp. will not) rain just because if was (resp. was not) raining a couple of hours ago. The difference between both models' moving average factors ($q = 2$ for Model 1, and $q = 10$ for Model 2) can be interpreted under the same light.

The statistics for our models are provided in Table 7. According to these results, Model 1 does not use past values from any of the mobile data time series to predict values of the ThermalPerception time series — it only relies on past values of the ThermalPerception time series itself (number of predictors = 0). On the other hand, we can see that Model 2 uses one predictor, either the Number or the Erlang time series. Only by looking into the saved model (in the XML file generated by SPSS) do we find that this predictor is Erlang, although its weight on the prediction is very low:

```
<PredictorEffect variableID="Erlang_1">
  <Transformation delay="6"/>
    <Numerator><NonSeasonalFactor><ZeroLagTerm>
```

```
      <EstimatedParameter>
        1.48809994077751e-08
        5.8349316286678e-09
      </EstimatedParameter>
    </ZeroLagTerm></NonSeasonalFactor></Numerator>
</PredictorEffect>
```

This shows that, although to a small extent, mobile traffic intensity (Erlang) has some influence in the predictive modeling of the Precipitation component time series.

Stationary R^2 is a measure of model goodness-of-fit which compares the stationary part of the model to a simple mean model. Both values are positive, which means that both our models are better than this simple mean model. Model 2 presents a very low absolute value, meaning that it does not provide a significant added value compared to the series mean. One possible partial explanation for this is the coarser granularity of our precipitation data — a possibility we aim to explore in the future. To understand if the mobile traffic data was adding value to the predictive model, we decided to make a simple experiment by ordering the Expert Modeler to generate a model for Precipitation without providing independent variables to be used as predictors (i.e., without using the traffic data for weather prediction. As a result we get a model of the same kind (Seasonal ARIMA$(1, 0, 10) \times (1, 0, 1)_{24}$) but with a lower stationary R^2 (.138), which confirms the small but anyway confirmed influence of Erlang as adding some value as a precipitation predictor. Model 1 has a middling value for stationary R^2; this moderately good result was somewhat expected, due to the influence of the daily cyclicity of temperature.

The Ljung–Box test has the null hypothesis (H_0) that the model does not exhibit lack of fit. The result is significant at the $\mathrm{Sig.} < .05$ level for Model 2, which means we are able to reject H_0 — i. e., Model 2 exhibits lack of fit. On the other hand, we are not able to reject H_0 at the $\mathrm{Sig.} < .05$ level for Model 1, which means that our model to predict the ThermalPerception component has fit.

5.3 Forecasting

Figure 8 shows the Model Fit (blue on the left) and Model Forecast (bolder blue on the right) plotted over the observed time series (red) for the ThermalPerception and Precipitation scores. Let us focus on the forecasts.

Model 1 forecasts correctly the daily patterns of high–low ThermalPerception; this is the main reason why the model is considered fit. However, it is not able to forecast the peaks that happen in the testing set.

The forecast of Precipitation by Model 2 is too conservative, which explains the good scores in the goodness-of-fit measures (such as the mean absolute percentage error) — it gets predictions right most of the time because it does not rain most of the time. However, it never predicts the peaks, which are in fact what matters — i. e., rain and thunderstorms.

6 Discussion

Some of the assumptions for factor analysis (Section 4) were not met, although by a small margin — namely, the Measures of Sampling Adequacy. This is arguably due to the low quality of our weather data, and more specifically to the coarse granularity of the precipitation variable. The results of the exploratory factor analysis and spectral analysis we have performed are consistent with those by Sagl et al. [20]; we also identified two components, which we termed ThermalPerception and Precipitation due to the variable loadings on each one. We performed spectral analysis on our mobile data (number of calls and traffic intensity) and on these components, and concluded:

- mobile variables and ThermalPerception have high peaks of coherence at the frequencies corresponding to 24-hour and 12-hour periods — this highlights the circadian nature of both phenomena;
- mobile variables and Precipitation have overall low coherence, with a small but mostly isolated peak at the frequency corresponding to a 3-hour period (we do not have an explanation for this peak).

We replicated Sagl et al.'s work Sagl et al. [20] almost entirely, with a small difference: the coherence between mobile variables and Precipitation. We cinjecture this difference to arise from two factors: the lower quality of our precipitation data, and differences in social aspects between Northern Italy and Côte d'Ivoire.

We used the SPSS Expert Modeler to generate Seasonal ARIMA models to try and predict future values for ThermalPerception and Precipitation based on past values thereof, and also tentatively using past values of the mobile data time series. The resulting models' parameters are consistent with the way we commonly antecipate the weather conditions for the near future. Due to the limited time frame of our data, the generated models look into relatively recent past values; especially for the Precipitation component, it is expected that models based on wider estimation periods (ideally, at least six months) provide better forecasts. The model to predict ThermalPerception does not take mobile data into account, and does a limited job at forecasting future values; we argue this to be due to the strength of the cyclic character of temperature and related variables, which most likely overpowers any possibly existent relation between temperature and mobile usage variations. The model to predict Precipitation is not able to predict rain or thunderstorms; however, it does take

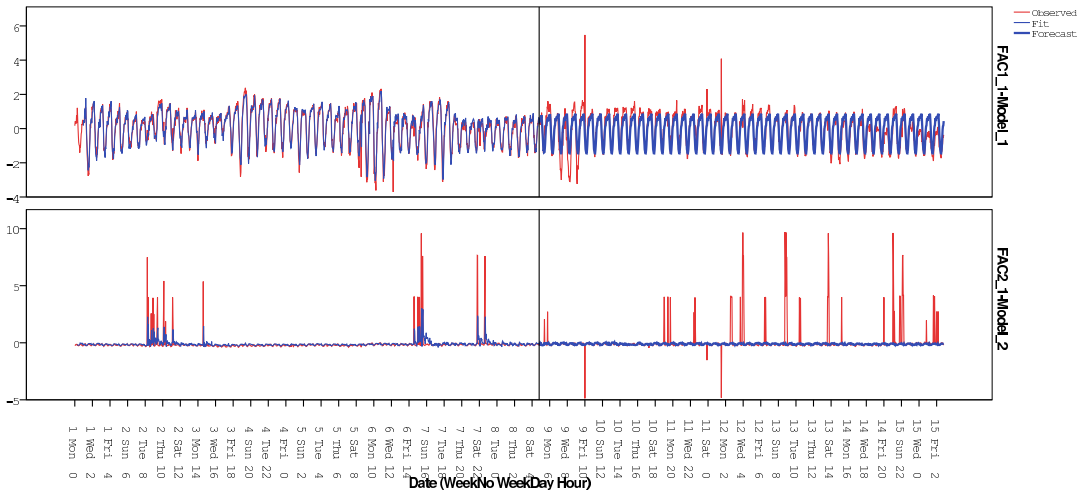


Figure 8: *Top panel:* Model 1 Fit (blue on the left) and Forecast (bolder blue on the right) plotted over the observed time series (red) for the ThermalPerception scores. *Bottom panel:* Model 2 Fit (blue on the left) and Forecast (bolder blue on the right) plotted over the observed time series (red) for the Precipitation scores.

past mobile data into account in its prediction, which yields a very small (but factual) improvement.

7 Conclusion and future work

In this paper we analysed mobile data and historic records of weather conditions. We employed exploratory factor analysis to reveal hidden variables in weather data, and spectral analysis to exploit the periodicity of mobile and weather data (both independently and jointly). From these results we derived models which, albeit with limited efficiency, hint that our hypothesis of predicting weather conditions from mobile phone usage is viable. We hope to further explore the hypothesis of predicting precipitation events with the help of mobile data, for which we aim to improve the considered data in two fronts: *(i)* finer-grained precipitation data, and *(ii)* finer-grained mobile usage data (e. g., more than just traffic intensity in the whole city, consider mobile usage divided into different areas as a proxy for overall user mobility). As for the first front, it depends on meteorological data we at the moment do not have (and would appreciate any help into obtaining it). As for the second one, we think it may be feasible with further analysis of the data we already possess, and hence we hope to be able to study the D4D mobile datasets for an extended period. With this additional data we hope to develop models to ultimately predict, with reasonable accuracy, unexpected weather events.

Still in the mobile–weather scope, future work includes using accessing mobile data from other locations to reproduce these experiments and to help us devise and adjust a predictive model. We are already in the works towards accessing the mobile dataset used by Phithakkitnukoon et al. [19] (Lisbon, Portugal), from which we have a small and limited sample, and also data from a telecom operator in Saudi Arabia.

Finally, we also intend to couple the data analysis of mobile datasets (namely, user mobility patterns) wth compositional scheduling analysis [22] with support to parallelism [10, 9] to aid public transport network planning. Again, this would be particularly interesting for the case of Abidjan and similar socioeconomic-wise cities.

Acknowledgments The authors would like to thank Tiago Guerreiro for some technical and bibliographical help with statistics. This work is partially supported by Fundação para a Ciência e a Tecnologia (FCT, the Portuguese national science foundation), through funding to LaSIGE research unit (strategic project PEst-OE/EEI/UI0408/2011). J. P. Craveiro's activities are further funded by FCT Doctoral Grant SFRH/BD/60193/2009.

References

[1] Historical weather for 2011 in Abidjan, Côte d'Ivoire - WeatherSpark, 2011. URL `http://weatherspark.com/history/28547/2011/Abidjan-Lagunes-Cote-d-Ivoire`. Verified on February 18, 2013.

[2] Historical weather for 2012 in Abidjan, Côte

d'Ivoire - WeatherSpark, 2012. URL http://weatherspark.com/history/28547/2012/Abidjan-Lagunes-Cote-d-Ivoire. Verified on February 18, 2013.

[3] Abidjan travel guide - Wikitravel, 2012. URL http://wikitravel.org/en/Abidjan#Stay_safe. Verified on February 18, 2013.

[4] Weather Forecast & Reports - Long Range & Local | Wunderground | Weather Underground. URL http://www.wunderground.com.

[5] T. W. Anderson and H. Rubin. Statistical inference in factor analysis. In J. Neyman, editor, *Econometrics, Industrial Research, and Psychometry*, volume 5 of *Proceedings of the Third Berkeley Symposium on Mathematical. Mathematical Statistics and Probability*. Cambridge University Press, London, England, 1956.

[6] M. S. Bartlett. A note on the multiplying factors for various chi-square approximations. *Journal of the Royal Statistical Society*, 16(Series B):296–298, 1954.

[7] V. D. Blondel, M. Esch, C. Chan, F. Clérot, P. Deville, E. Huens, F. Morlot, Z. Smoreda, and C. Ziemlicki. Data for Development: the D4D challenge on mobile phone data. *CoRR*, abs/1210.0137, 2012. URL http://arxiv.org/abs/1210.0137.

[8] S. Boslaugh and P. Watters. *Statistics in a Nutshell: A Desktop Quick Reference*. IN A NUTSHELL. O'Reilly Media, 2009. ISBN 9780596510497.

[9] J. P. Craveiro and J. Rufino. Towards compositional hierarchical scheduling frameworks on uniform multiprocessors. Technical Report TR-2012-08, University of Lisbon, DI–FCUL, Dec. 2012. URL http://hdl.handle.net/10455/6891.

[10] A. Easwaran, I. Shin, and I. Lee. Optimal virtual cluster-based multiprocessor scheduling. *Real-Time Systems*, 43:25–59, 2009. ISSN 0922-6443. 10.1007/s11241-009-9073-x.

[11] A. Field and G. Hole. *How to Design and Report Experiments*. SAGE Publications Ltd, London, UK, 2003. ISBN 978-0-7619-7383-6.

[12] E. Haines. Point in polygon strategies. In P. S. Heckbert, editor, *Graphics Gems IV*, Graphics Gems, chapter I.4, pages 24–46. Academic Press, Boston, 1994.

[13] *IBM SPSS Forecasting 20*. IBM Corporation, 2011.

[14] H. Kaiser. A second generation Little Jiffy. *Psychometrika*, 15:173–190, 1970.

[15] H. Kaiser. An index of factorial simplicity. *Psychometrika*, 39:31–36, 1974.

[16] X. Lu, L. Bengtsson, and P. Holme. Mobility and predictability of population movements after the Haiti 2010 earthquake. In *NetMob 2011*, May 2011.

[17] A. Overeem, H. Leijnse, and R. Uijlenhoet. Country-wide rainfall maps from cellular communication networks. *Proceedings of the National Academy of Sciences*, 2013. doi: 10.1073/pnas.1217961110.

[18] J. Pallant. *SPSS Survival Manual: A step by step guide to data analysis using SPSS*. Open University Press. McGraw-Hill Education, 2010. ISBN 9780335242399.

[19] S. Phithakkitnukoon, T. W. Leong, Z. Smoreda, and P. Olivier. Weather effects on mobile social interactions: A case study of mobile phone users in Lisbon, Portugal. *PLoS ONE*, 7(10):e45745, Oct. 2012. doi: 10.1371/journal.pone.0045745.

[20] G. Sagl, E. Beinat, B. Resch, and T. Blaschke. Integrated geo-sensing: A case study on the relationships between weather and mobile phone usage in Northern Italy. In *ICSDM 2011*, pages 208–213, June/July 2011. doi: 10.1109/ICSDM.2011.5969033.

[21] G. Sagl, T. Blaschke, E. Beinat, and B. Resch. Ubiquitous geo-sensing for context-aware analysis: Exploring relationships between environmental and human dynamics. *Sensors*, 12(7):9800–9822, 2012. ISSN 1424-8220. doi: 10.3390/s120709800.

[22] I. Shin and I. Lee. Compositional real-time schedulability analysis. In I. Lee, J. Y.-T. Leung, and S. H. Son, editors, *Handbook of Real-Time and Embedded Systems*, Computer and Information Science. Chapman & Hall / CRC, 2007.

Does conflict affect human mobility and cellphone usage? Evidence from Côte d'Ivoire

Sera Linardi*[1], Shankar Kalyanaraman†[2], and Daniel Berger‡[3]

[1]Graduate School of Public and International Affairs, University of Pittsburgh
[2]Department of Computer Science, New York University
[3]Department of Government, University of Essex

Abstract

Ever since the disputed elections of 2010, in which then President Gbagbo lost to current President Ouattara, Côte d'Ivoire has seen incidents of sporadic and tumultuous violence. In most cases, these incidents were incited by forces loyal to the two political rivals, and along largely ethnic lines. In this paper, we make use of the large dataset of cellphone usage records made available via the D4D project to investigate how and whether violence and ethnic conflicts affect cellphone usage and mobility. Following violent events, we observe a slow upward trend in the number of callers, peaking several days after the event. This is inconsistent with people simply using the phone to inform their contacts that the event has occurred and that they are unharmed. We also explore the interaction of violence and mobility. We show that violence occurs in areas which are already experiencing a) an upward trend in mobility *from*, and b) a downward trend in mobility *to*. Only the first of these appears to be exacerbated by the actual violent events.

Understanding and investigating mobility and other consequences of conflict forms a core component of research in social science, and the almost universal-access to cellphones across the world makes this line of work more important now than ever before. In this paper, we focus on understanding the effect of violence in Côte d'Ivoire on cellphone usage and whether there are any observable changes in human mobility as a consequence.

We code the violence data using reports from the UN mission to the Côte d'Ivoire. "ONUCI hebdo" (UNOCI weekly) is a weekly summary of events in the Côte d'Ivoire, with a focus on the actions, concerns, and effects of the UN post-conflict peacekeeping mission.[1] We code all cases of violence (including, but not limited to fatal violence) which is given a description and includes a specific location. This excludes general events (there are reports of tensions along the Liberian border) and violence which is not reported by the United Nations. Since UNOCI has a mandate to focus on the post-crisis situation, this effectively filters for political violence. We explicitly do include subprefectures where voting had to be canceled due to intimidation or violence. While we might be concerned that the UN would avoid the most violent areas, thus missing the very areas we are most interested in, evidence [RDG11] shows that UN peacekeepers focus on areas where the conflict has been historically most explosive. The cellphone usage dataset we used was provided by Orange Telecom as part of the D4D Challenge.[2] We focus on the subprefecture-level long-term records comprising call detail records for 500,000 cellphone users over 143 days between December 1, 2011 to April 28, 2012. A record $r(u, t, s)$ has the following fields: randomized user ID $u \in \mathbb{N}$, call date t,[3] caller's subprefecture ID $s \in \{1, \ldots, 255\}$. Each subprefecture s is associated with a latitude and longitude. We have three dependent variables to measure call activity and mobility in subprefecture s and time t. Intuitively, we think of mobility at the

*linardi@pitt.edu
†shankar@cs.caltech.edu
‡dberger@essex.ac.uk

[1]These can be found at `http://www.onuci.org/spip.php?article5548`
[2]`http://www.d4d.orange.com/home`
[3]Call time was provided but is not used in this analysis.

subprefecture level as measuring population movements over a period of time. Letting $U(s,t)$ denote the set of users who are recorded as making calls from subprefecture s at time t, our dependent variables are defined as follows:

Definition 1. *For time period t, the number of callers in subprefecture s is denoted $c(s,t)$ and is given by $c(s,t) = |U(s,t)|$.*

Definition 2. *For a time period (t_1,t_2), the inward mobility of a subprefecture s is denoted $m^{in}(s,t_1,t_2)$ and is given by $m^{in}(s,t_1,t_2) = |U(s,t_2) - U(s,t_1)|/|U(s,t_1)|$.*

Definition 3. *For a time period (t_1,t_2), the outward mobility of a subprefecture s is denoted $m^{out}(s,t_1,t_2)$ and is given by $m^{out}(s,t_1,t_2) = |U(s,t_1) - U(s,t_2)|/|U(s,t_1)|$.*

Our primary independent variable are violent incidents. Let T denote the length of a time-window, $(t-T, t+T)$, and R denote the radius of the circle centered at point p within which we wish to understand the impact of violent incidents. For an incident occurring at (p_0,t_0), let $B(p_0,R) = \{p : |lat(p) - lat(p_0)| \leq R, |lon(p) - lon(p_0)| \leq R\}$ denote the 'R-neighborhood' of p. We set up a dummy variable $h_{s,t}$ for all subprefecture-time pairs in the R-neighborhood of an incident's position p_0 as follows:

$$h_{s,t} = \begin{cases} 1 & s \in B(p_0,R), t \in (t_0 - T, t_0 + T) \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Our model regresses the number of callers c, inward mobility m^{in} and outward mobility m^{out} against $h_{s,t}$. We account for subprefectural level and temporal variation in call volume and mobility[4] by including time and subprefecture level fixed effects (s_i and t_j, respectively). Our regression equations are as follows:

$$c_{s,t} = \alpha_s + \beta_t + \delta^c h_{s,t} + \epsilon_{s,t} \tag{2}$$
$$m^{in}_{s,t} = \alpha_s + \beta_t + \delta^{in} h_{s,t} + \epsilon_{s,t} \tag{3}$$
$$m^{out}_{s,t} = \alpha_s + \beta_t + \delta^{out} h_{s,t} + \epsilon_{s,t} \tag{4}$$

We also wish to estimate a priori and a posteriori effects of violence on cellphone usage and mobility. For instance how does the immediate runup to a violent incident at p_0 on date t_0 affect cellphone usage and mobility afterwards? We extend the models in (2)-(4) by considering three dummy variables $h^{R,-1}_{s,t}, h^{R,0}_{s,t}, h^{R,+1}_{s,t}$ to denote indicators for time periods $(t_0 - T, t_0), (t_0, t_0 + T/2), (t_0 + T/2, t_0 + T)$. This allows us to differentiate the effect of violence on call patterns within radius R of the incident before the violence occurence ($h^{-1}_{s,t}$), right after the occurence ($h^{R,0}_{s,t}$) , and a few days afterwards ($h^{R,+1}_{s,t}$). Let $\delta_{-1}, \delta_0, \delta_{+1}$ indicate the corresponding coefficients. This gives us:

$$c_{s,t} = \alpha_s + \beta_t + \delta^c_{-1} h^{R,-1}_{s,t} + \delta^c_0 h^{R,0}_{s,t} + \delta^c_{+1} h^{R,+1}_{s,t} + \epsilon_{s,t} \tag{5}$$
$$m^{in}_{s,t} = \alpha_s + \beta_t + \delta^{in}_{-1} h^{R,-1}_{s,t} + \delta^{in}_0 h^{R,0}_{s,t} + \delta^{in}_{+1} h^{R,+1}_{s,t} + \epsilon_{s,t} \tag{6}$$
$$m^{out}_{s,t} = \alpha_s + \beta_t + \delta^{out}_{-1} h^{R,-1}_{s,t} + \delta^{out}_0 h^{R,0}_{s,t} + \delta^{out}_{+1} h_1 + \epsilon_{s,t} \tag{7}$$

Figure 1 overlays the coded violent incidents (in orange) on the log of call volumes (in grey) for all of Côte d'Ivoire. Table 1 shows the effect of violent events on call volumes. It is immediately clear from the table that there is no generalized increase in call volumes around a violent event. However, the division of the event effect into immediate pre-event, immediate post-event and intermediate post-event tells a more interesting story. There is no change in call volumes in the days leading up to violence. This is consistent with violent events neither being carefully orchestrated nor being generally seen coming by the local population. After the violent events, there is a small bump in call volumes, which is consistent with people calling friends and family to inform them of the events and reassure them that the caller is unharmed. The surprising observation is that the medium term effect is even greater than the short term effect. Phones are used more several days on from an episode of violence than right

[4]For instance, call volume and mobility tend to be very high in high population density subprefectures near Abidjan, the capital of Côte d'Ivoire. National holidays such as New Year's Day also has spikes in call volume.

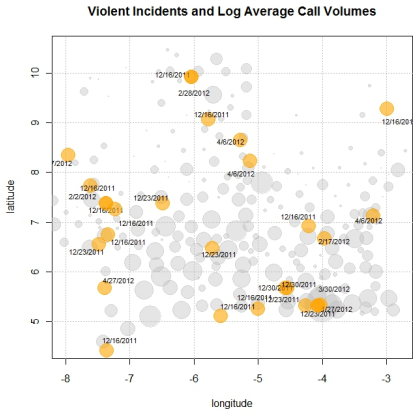


Figure 1: Violent incidents (orange) and call volumes (grey) on Côte d'Ivoire map

	δ	δ_{-1}	δ_0	δ_{+1}
Num. callers	1.828	-3.27	3.37^{*}	7.8^{***}
(T=7,R=0.5)	(1.17)	(2.19)	(1.96)	(2.28)
Num. callers	0.8	-1.8	2.61^{**}	4.865^{***}
(T=14,R=1)	(0.72)	(1.12)	(1.07)	(1.07)

Note: Sample size is 2% over 500,000 callers. Dummy variables for dates and subprefectures were part of the model. Standard errors in parentheses. ***,**,* indicate statistical significance at 1%,5%,10% respectively.

Table 1: Violence effect on number of callers

	δ	δ_{-1}	δ_0	δ_1
Inward mobility	-0.072^{***}	-0.086^{**}	-0.073^{**}	-0.064^{**}
(T=7,R=1)	(0.016)	(0.029)	(0.025)	(0.029)
Inward mobility	-0.093^{***}	-0.062^{***}	-0.085^{***}	-0.095^{***}
(T=14,R=1)	(0.01)	(0.02)	(0.02)	(0.02)
Outward mobility	0.04^{***}	0.046^{***}	0.035^{*}	0.033
(T=7,R=0.5)	(0.007)	(0.013)	(0.011)	(0.013)
Outward mobility	0.019^{***}	0.018^{***}	0.015^{*}	-0.03
(T=14,R=1)	(0.004)	(0.006)	(0.006)	(0.006)

Note: Sample size is 2% over 500,000 callers. Dummy variables for dates and subprefectures were part of the model. Standard errors in parentheses. ***,**,* indicate statistical significance at 1%,5%,10% respectively.

Table 2: Violence effect on mobility

afterwards. Table 2 shows analogous estimates of coefficients for estimating violence effects on mobility. It can be seen that there is a significant negative effect of a violent incident on inward mobility (number of callers coming to a subprefecture), and a positive effect on outward mobility (number of callers leaving a subprefecture). However, the estimates of the time evolution of this migration are different from that of the call volume. First, even *before* the violent events, outmigration increases and inmigration falls. While it appears that inmigration falls even more after violence than before it, there is no evidence that outmigration increases. Therefore, it looks unlikely that the event is directly causing an increase in internal displacement.[5] In conclusion, we find that even when controlling for day and subprefecture fixed effects violent incidents are still correlated with increases in call volume and migration. However, the patterns in the correlations are very different. Phone call volumes fail to anticipate the violence, and slowly ramp up, peaking several days after a violent event. Significant net outmigration, however, commences significantly before violent events, suggesting that the tensions which lead to violence are already having effects on the population's behavior.

References

[RDG11] Andrea Ruggeri, Han Dorussen, and Ismene Gizelis. Sub-national dynamics of UN peacekeeping. Working Paper, 2011.

[5]This is *not*, however, implying that the tensions and threat of violent events do not cause internal displacement.

D4D Challenge – Report

Symbolic clustering of users and antennae

Monika Cerinšek · Jernej Bodlaj · Vladimir Batagelj

February 15, 2013

Abstract Large D4D Challenge datasets about mobile connectivity in Ivory Coast can be analyzed in many different ways. We present an attempt to produce a typology of users/antennae using symbolic clustering. This approach allows clustering of symbolic objects – description of units of data set in which the properties (variables) can take structured values (in our case discrete distributions). We used symbolic clustering for antennae from dataset Set 2 and for users from Set 3. Both sets (antennae and users) were clustered according to their mobile activity through days in a week and through hours in a day. We distinguished groups of antennae (and users) based on a similarity of their activity through week and day.

Keywords Mobile · Symbolic clustering · Large data

1 Introduction

Orange, France Telecom, published a D4D challenge providing 4 datasets derived from data on the Orange phone calls in Ivory Coast in the time period

M. Cerinšek
Hruška d.o.o., Kajuhova 90, 1000 Ljubljana
Tel.: +386-15423614
Fax: +386-59022240
E-mail: monika@hruska.si

J. Bodlaj
Hruška d.o.o., Kajuhova 90, 1000 Ljubljana
Tel.: +386-15423614
Fax: +386-59022240
E-mail: jernej.bodlaj@hruska.si

V. Batagelj
University of Ljubljana, FMF, Department of Mathematics, Jadranska 19, 1000 Ljubljana, Slovenia
E-mail: vladimir.batagelj@fmf.uni-lj.si
URL: http://pajek.imfm.si

from December 2011 to April 2012. This base data consists of 2.5 billion phone calls and messages that were exchanged between 5 million users. So real time data became available for us to analyze it and to solve a problem from real life.

We decided to try to produce typologies of antennae and users based on their activity patterns. The pattern of phone usage during a day and during a week were selected as the properties that distinguish different groups of antennae and users. There are groups of users that make most calls around noon, some users that use phone mostly on the weekend, etc. This can help the Orange with setting new or renewing existing subscription plans. The classification of antennae might also help with that. According to a placement of similar antennae the Orange might determine where similar users are located. And so they can put there more commercials about the subscription plans that are more appropriate for the targeted group of users.

In our analysis we used the second and the third set from the D4D challenge datasets and a transformation of data to networks as presented in [Bodlaj, J. et al. (2013)].

The Set 2 consists of 10 similar datasets. They all store information about calls – for each call is given an identification of user that made a call, timestamp of a call and an identification of antenna that "send" a call. The difference between the 10 datasets is that they all cover different 14 days time periods. The identifcations of users change from dataset to dataset, but the identifications of antennae are the same in all datasets. Another dataset is contained in Set 2 – coordinates of antennae. We used them to place the antennae on the map of Ivory Coast. Because identifications of users changes through datasets, we cannot track them. So we focus on antennae. We used the Set 2 for the clustering of antennae according to their usage through a day and through a week.

The Set 3 also stores the information about calls. For each call we have the identification of user, the timestamp of a call and the subregion, from where the calls was made. The information about all calls is stored in a single dataset, so we can use this dataset for the clustering of users according to their phone activity through a day and through a week.

The rest of this report is sectioned in three parts. In the following section we introduce a clustering of symbolic objects. In next section we present the results of clustering antennae from Set 2 and in the last section we present the result of clustering of users from Set 3.

2 Clustering of symbolic objects

Given a set of units $\mathcal{U}$ the clustering is a process of organizing units into groups – clusters of similar units. For analyzing the D4D data we shall use an approach to clustering of (very) large data sets of structured units based on representation of units by *symbolic objects* (SOs) [Billard, L., Diday, E.

(2006)]. The SOs can describe either single units or groups of initial units condensed into SOs in a pre-processing step or during the clustering process.

An SO X (representing user(s) or antenna(e)) is described by a list $X = [\mathbf{x}_i]$ of descriptions of variables V_i. In our case, each variable is described with frequency distribution (*bar chart*) of its values [Korenjak-Černe, S. et al. (2002), Korenjak-Černe, S. et al. (2008), Kejžar, N. et al. (2011)]

$$\mathbf{f}_{xi} = [f_{xi1}, f_{xi2}, \ldots, f_{xik_i}].$$

With $\mathbf{x}_i = [p_{xi1}, p_{xi2}, \ldots, p_{xik_i}]$ we denote the corresponding probability distribution

$$\sum_{j=1}^{k_i} p_{xij} = 1, \quad i = 1, \ldots, m$$

We approach the clustering problem as an optimization problem over the set of *feasible* clusterings Φ_k – partitions of units into k clusters. The *criterion function* has the following form

$$P(\mathbf{C}) = \sum_{C \in \mathbf{C}} p(C). \tag{1}$$

The *total error* $P(\mathbf{C})$ of the clustering $\mathbf{C}$ is a sum of *cluster errors* $p(C)$. In this paper we shall assume a model in which the error of a cluster is a sum of differences of its units from the cluster's *representative* T

$$p(C, T) = \sum_{X \in C} d(X, T). \tag{2}$$

Note that in general the representative needs not to be from the same "space" (set) as units. The best representative is called a *leader*

$$T_C = \operatorname*{argmin}_T p(C, T). \tag{3}$$

Then we define

$$p(C) = p(C, T_C) = \min_T \sum_{X \in C} d(X, T). \tag{4}$$

The SO X is described by a list $X = [\mathbf{x}_i]$. Assume that also representatives are described in the same way $T = [\mathbf{t}_i]$, $\mathbf{t}_i = [t_{i1}, t_{i2}, \ldots, t_{ik_i}]$.

We introduce a dissimilarity measure between SOs with

$$d(X, T) = \sum_i \alpha_i d(\mathbf{x}_i, \mathbf{t}_i), \quad \alpha_i \geq 0, \quad \sum_i \alpha_i = 1, \tag{5}$$

where

$$d(\mathbf{x}_i, \mathbf{t}_i) = \sum_{j=1}^{k_i} w_{xij} \delta(p_{xij}, t_{ij}), \quad w_{xij} \geq 0. \tag{6}$$

This is a kind of a generalization of the squared Euclidean distance. In our analyses we shall use the dissimilarity

$$\delta(p_x, t) = (p_x - t)^2$$

The weight w_{xij} can be for the same unit X different for each variable V_i (needed in descriptions of ego-centric networks, population pyramids, etc.).

For clustering of SOs two classical clustering methods were adapted [Batagelj,-V. et al. (2013)]:

- *leaders method* (a generalization of k-means method [Hartigan, J. A. (1975)], dynamic clouds [Diday, E. (1979)]).
- Ward's *hierarchical clustering method* [Ward, J. H. (1963)].

Both adapted methods are based on the *same* criterion function – they are solving the *same* clustering problem. With the leaders method the size of the sets of units is reduced to a manageable number of leaders. The obtained leaders can be further clustered with the compatible agglomerative hierarchical clustering method to reveal relations among them and using the dendrogram also to decide upon the right number of clusters.

2.1 Leaders method

Leaders method is a generalization of a popular nonhierarchical clustering k-means method. The idea is to get "optimal" clustering into a pre-specified number of clusters with the following iterative procedure:

determine an initial clustering
repeat
  determine leaders T_C of the clusters C in the current clustering $\mathbf{C}$;
  assign each unit to the nearest new leader – producing a
    new clustering
until the leaders stabilize.

Given a cluster C, the corresponding leader T_C is the solution of the problem

$$T_C = \underset{T}{\operatorname{argmin}} \sum_{X \in C} d(X, T) = \big[\underset{\mathbf{t}_i}{\operatorname{argmin}} \sum_{X \in C} d(\mathbf{x}_i, \mathbf{t}_i)\big]_{i=1}^{m}$$

Therefore $T_C = [\mathbf{t}_i^*]$ and $\mathbf{t}_i^* = \operatorname{argmin}_{\mathbf{t}_i} \sum_{X \in C} d(\mathbf{x}_i, \mathbf{t}_i)$. To simplify the notation we omit the index i.

$$\mathbf{t}^* = \underset{\mathbf{t}}{\operatorname{argmin}} \sum_{X \in C} d(\mathbf{x}, \mathbf{t}) = \big[\underset{t_j \in \mathbb{R}}{\operatorname{argmin}} \sum_{X \in C} w_{xj} \delta(p_{xj}, t_j)\big]_{j=1}^{k}$$

Again we omit the index j

$$t^* = \underset{t \in \mathbb{R}}{\operatorname{argmin}} \sum_{X \in C} w_x \delta(p_x, t)$$

This is a standard optimization problem with one real variable. The solution has to satisfy the condition

$$\sum_{X \in C} w_x \frac{\partial \delta(p_x, t)}{\partial t} = 0 \tag{7}$$

For $\delta(p_x, t) = (p_x - t)^2$ we get from (7)

$$t^* = \frac{\sum_{X \in C} w_x p_x}{\sum_{X \in C} w_x} = \frac{P}{A}$$

where

$$A_{ij} = \sum_{X \in C} w_{xij} \quad \text{and} \quad P_{ij} = \sum_{X \in C} w_{xij} p_{xij}$$

Let $w_{xij} = w_{xi}$ then for each $i = 1, \ldots, m$:

$$\sum_{j=1}^{k_i} t^*_{ij} = 1$$

The leaders' components are *distributions*.

Let further $w_{xij} = n_{xi}$ then for each $i = 1, \ldots, m$:

$$t^*_{Cij} = \frac{\sum_{X \in C} n_{xi} p_{xij}}{\sum_{X \in C} n_{xi}} = p_{Cij} \tag{8}$$

The leader of a cluster is a list of distributions of variables' values on the cluster.

Given leaders $\mathbf{T}$ the corresponding optimal clustering $\mathbf{C}^*$ is determined from

$$P(\mathbf{C}^*) = \sum_{X \in \mathcal{U}} \min_{T \in \mathbf{T}} d(X, T) = \sum_{X \in \mathcal{U}} d(X, T_{c^*(X)}) \tag{9}$$

where

$$c^*(X) = \operatorname*{argmin}_k d(X, T_k)$$

We assign each unit X to the closest leader $T_k \in \mathbf{T}$.

2.2 Hierarchical agglomerative clustering

The hierarchical agglomerative clustering procedure is based on a step-by-step merging of the two closest clusters.

each unit forms a cluster: $\mathbf{C}_n = \{\{X\}: X \in \mathcal{U}\}$;
they are at level 0: $h(\{X\}) = 0,\ X \in \mathcal{U}$;
for $k = n - 1$ **to** 1 **do**
  determine the closest pair of clusters
    $(u, v) = \text{argmin}_{i,j:i\neq j}\{D(C_i, C_j): C_i, C_j \in \mathbf{C}_{k+1}\}$;
  join the closest pair of clusters $C_{(uv)} = C_u \cup C_v$
    $\mathbf{C}_k = (\mathbf{C}_{k+1} \setminus \{C_u, C_v\}) \cup \{C_{(uv)}\}$;
    $h(C_{(uv)}) = D(C_u, C_v)$
    determine the dissimilarities $D(C_{(uv)}, C_s), C_s \in \mathbf{C}_k$
endfor

$\mathbf{C}_k$ is a partition of the finite set of units $\mathcal{U}$ into k clusters. $h(C_{(uv)})$ is the level of the cluster $C_{(uv)} = C_u \cup C_v$.

Therefore the computation of dissimilarities between new (merged) cluster and the rest has to be specified. To obtain the compatibility with the adapted leaders method, we define the dissimilarity between clusters C_u and C_v, $C_u \cap C_v = \emptyset$, as [Batagelj, V.(1988)]

$$D(C_u, C_v) = p(C_u \cup C_v) - p(C_u) - p(C_v).$$

$\mathbf{u}_i$ and $\mathbf{v}_i$ are components of the leaders of clusters C_u and C_v.

For $\delta(p_x, t) = (p_x - t)^2$ we get

$$D(C_u, C_v) = \sum_i \alpha_i \sum_j \frac{A_{uij} \cdot A_{vij}}{A_{uij} + A_{vij}} (u_{ij} - v_{ij})^2 \tag{10}$$

a *generalized Ward's relation*.

Instead of the squared Euclidean distance other dissimilarity measures $\delta(x, t)$ can be used (see [Kejžar, N. et al. (2011)]). Relations similar to Ward's can be derived for them.

The proposed approach is implemented in the R-package ***Clamix*** [Batagelj, V. et al. (2010)].

3 Antennae

In Set 2 are given the data about calls made in 10 14-days periods. One possible analysis is the analysis of an activity of antennae. Given exact timestamp of each call enables us to see the activity of each antenna at selected hour in a day or at selected day in the week. Antennae with no activity are not included in this set.

We transform data from the Set 2 into two-mode networks and one of them is also a network of antennae as one set of vertices and days in a week as another set of vertices. Another two-mode network that we produced is a network of antennae as one set of vertices and hours in a day as another set of vertices.

We analyzed the activity of antennae through a day and through a week in the report about a visualization [Bodlaj, J. et al. (2013)]. In this analysis we do not distinguish between different days in a week and different hours in a day. We take a sum of days in a week in which an antennae was active and similar for hours. But maybe some antennae are similar to each other according to their activity through week – maybe some antennae are much more active in first few days in a week than in the weekend. Maybe there are some antennae that are active mostly in the morning or in the evening. To distinguish different groups of antennae that are similar to each other according to their activity, we clustered them using symbolic object clustering.

A SO describing an antenna consists of two symbolic variables: daily activity and weekly activity. Both are described by the corresponding frequency distributions of calls. First we merged using the leaders method the original 1215 antennae into 20 clusters represented by the corresponding leaders L1 – L20. (also SOs). The obtained leaders were further clustered using agglomerative hierarchical method. The clustering process is presented with the dendrogram in Fig. 1. Fig. 2 presents visual representations of the 20 leaders ordered as in the dendrogram.

From the dendrogram we can see the "natural" numbers of clusters. The obvious numbers are 2 and 4. The second cluster in the clustering into 4 clusters is much larger than the others. We decided to split it by increasing the number of clusters to 5.

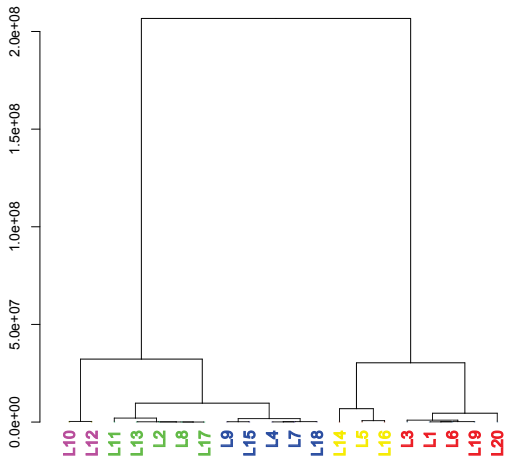


Fig. 1 Dendrogram of 20 leaders of antennae clusters.

We cut the dendrogram of leaders on Fig. 1 in five clusters. Their representatives are shown in Fig. 3. For each representative its name, color, number of calls and number of antennae are shown on the right side. The first diagram

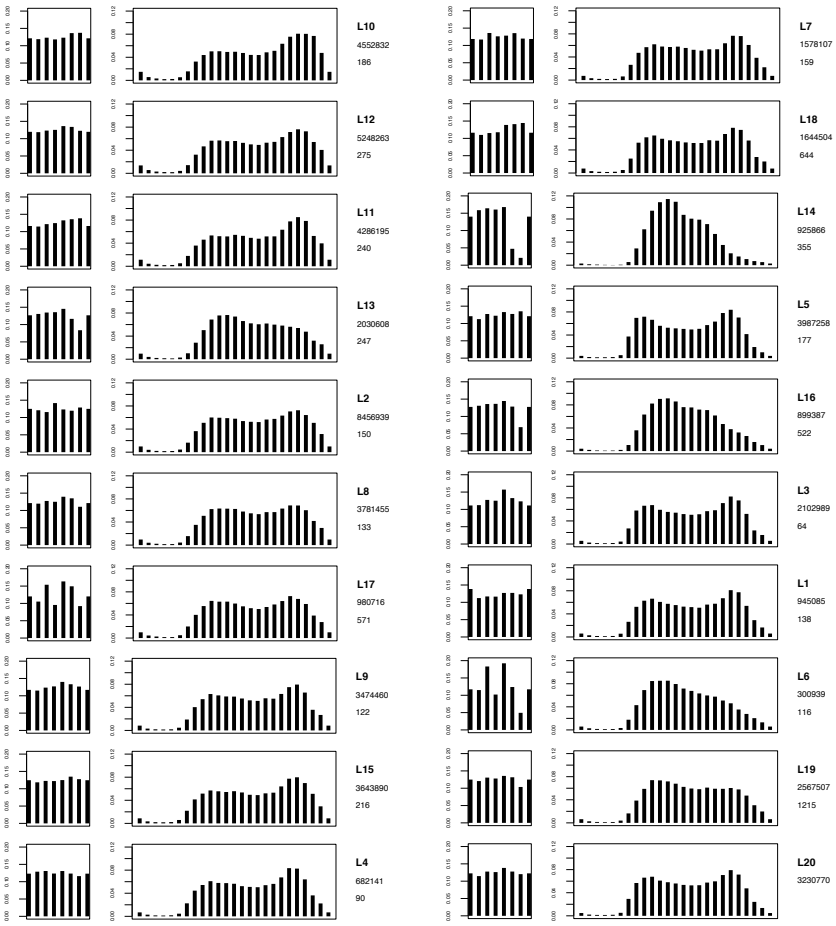


Fig. 2 All 20 leaders of antennae.

presents the weekly distribution (Sunday – Saturday). The second diagram presents the daily activity (0 – 23).

The cluster C5 is less common than others, because it contains 29.2% of all antennae and only 16.5% of all calls were made from those antennae. More then one third (35.3%) of all calls were made from antennae in group with the representative R2 with only 20.3% of all antennae.

All five representatives have quite similar distribution of calls during a day – they all have more calls around noon and in the evening, a bit less calls during a day and almost no calls during a night. But one can still notice differences among them. The representatives of clusters C1, C2 and C3 have the highest peak in the evening around 19.00 o'clock, but differ in the activity till midnight. The representatives of clusters C4 and C5 have stronger morning

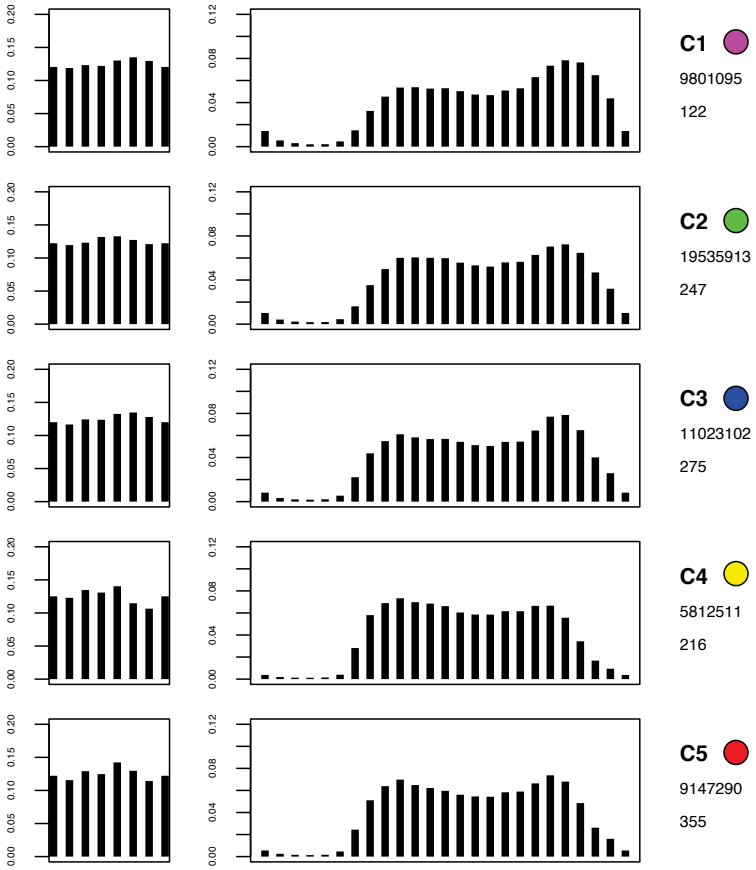


Fig. 3 Representatives of five groups of leaders.

activity and weaker after evening activity. The evening peak is moved to 18.00 o'clock.

Let us take a look at the weekly distribution of calls for the representatives of clusters of antennae. Weekly distributions for the first three representatives are almost the same for all days in a week. The week starts with Sunday, so we can see that most calls are made in the middle of a week (Thursday, Friday). The fourth and fifth representatives have more diverse weekly distributions, but still relatively close to the uniform distribution.

We draw antennae on the map of Ivory Coast to see if clusters from symbolic object clustering form groups also on the map. Fig. 4 presents antennae that are colored according to the color of their representative (Fig. 3), we used the same method for drawing as in [Bodlaj, J. et al. (2013)]. Few magenta,

yellow and green groups can be found on the map. Red and blue antennae are spread all over the country. The surrounding of Abidjan is almost white because of large amount of antennae in that area.

Fig. 4 Groups of antennae according to symbolic object clustering.

4 Users

From the data in Set 3 we constructed 500000 SOs describing users by their daily and weekly activity patterns derived from 865515875 calls. As in the case of antennae we used symbolic cllustering to determine the typology of users activity. We first reduced using the leaders method the 500000 SOs to 50 clusters C1 – C50 represented by leaders L1 – L50. The leaders were further

clustered using the agglomerative hierarchical clustering. The clustering process is presented by a dendrogram in Fig. 5. For further reference also some internal nodes in dendrogram are numbered.

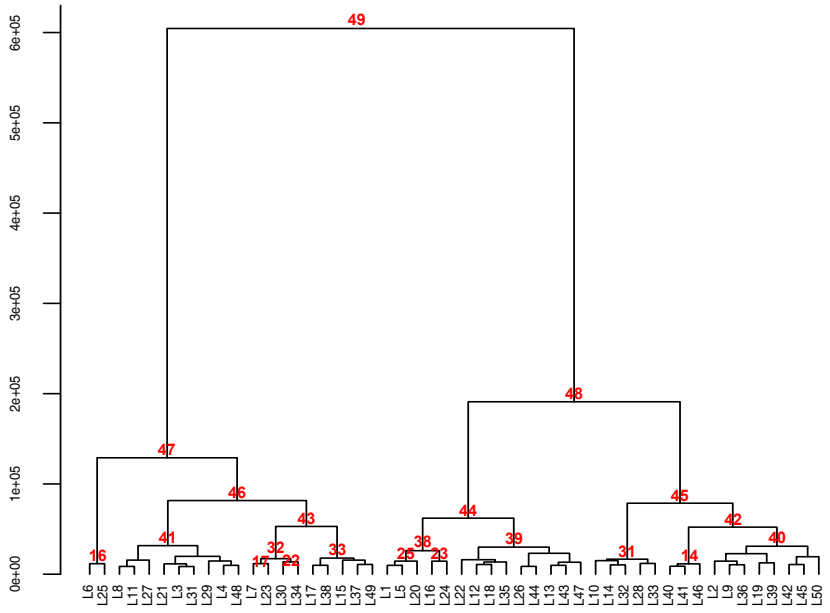


Fig. 5 The dendrogram of 50 leaders. Some interval merging points are also marked.

The top node 49 corresponding to all SOs in one cluster is presented in Fig. 6. In the weekly pattern we see that there is slight increase of activity towards the end of the week (Thursday, Friday, Saturday). The daily pattern shows almost no activity the first six hours after the midnight; an increase of activity towards morning, peak around 9 o'clock; slight decrease in the afternoon; and the second peak around 19 – 20 o'clock.

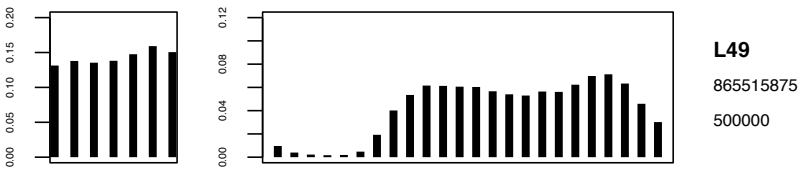


Fig. 6 Representative of all in one cluster C49 in dendrogram in Fig. 5.

There are different "natural" numbers of clusters: 2 (C47, C48), Fig. 7; 4 (C16, C46, C44, C45), Fig. 8; and 9 (C16, C41, C32, C33, C38, C31, C14, C40), Fig. 9 and Fig. 10.

The clustering in 2 clusters splits the users to C47 – late evening (17.3%) and C48 – morning users. The late evening users have a slight increase of use over the weekend.

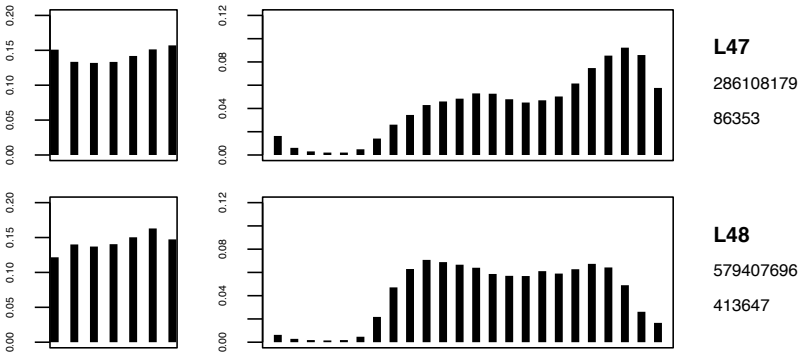


Fig. 7 Representatives of leaders of main two clusters in dendrogram in Fig. 5.

In the clustering in 4 clusters both clusters split into 2 subclusters. The late evening users contain a cluster C16 of night users (only 1391) that have very intense use in the period 21 – 1 o'clock.

The main subcluster C45 of morning users shows intensive morning use and a decrease over the weekend.

In the clustering in 9 clusters the lare evening cluster splits into four subclusters (C16, C41, C32, C33), Fig. 9. All of them have the highest peak in evening , but in different positions. In clusters C32 and C33 the activity is increased over weekends.

The morning cluster splits in five subclusters (C38, C39, C31, C14, C40). The first three have also strong evening activity. The last two are the most active in the morning and decrease their activity aafter afternoon.

5 Conclusions

Data about phone calls and messages in Ivory Coast inside the Orange mobile network give opportunity for a large variety of analyses. In the report symbolic clustering is presented as a method for an application of producing a typology of antennae or users.

The symbolic clusterings of antennae from Set 2 and users from Set 3 were presented. We focused on classification of users because this might help the Orange to identify the typical groups of users that are using their mobile network and to define appropriate subscription plans.

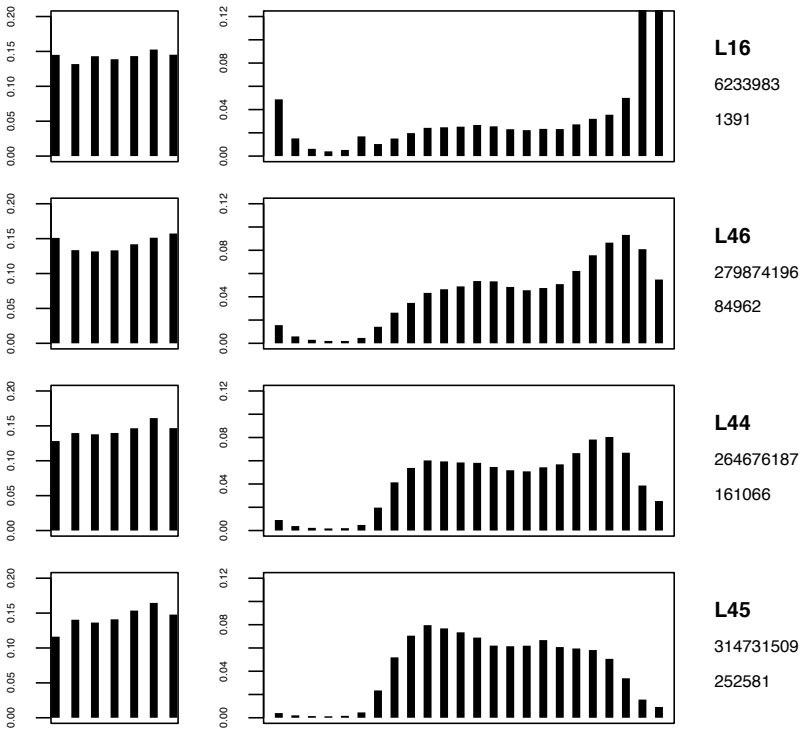


Fig. 8 Representatives of leaders of main four clusters in dendrogram in Fig. 5.

In future research the SOs could be extended alos with symbolic variable describing the use of antennae – for example individually the first 5 the most used antennae (in decreasing order for a given SO) followed by data for selected segments of antenae.

More interesting typologies could be obtained if we would have access to additional data, for example age and gendre of users.

Acknowledgements The first and the second author were financed in part by the European Union, European Social Fund.

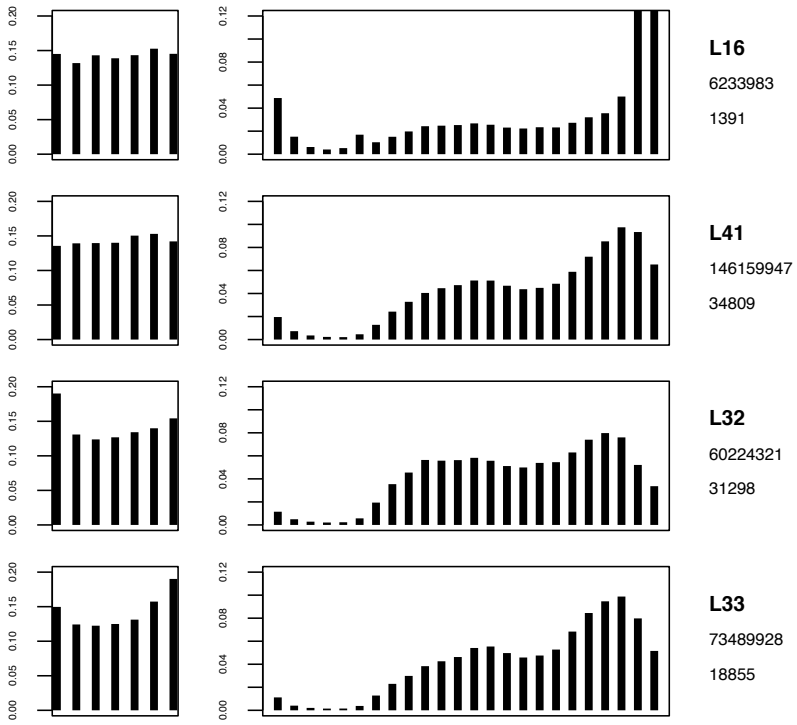


Fig. 9 Representatives for groups in the first half of dendrogram in Fig. 5.

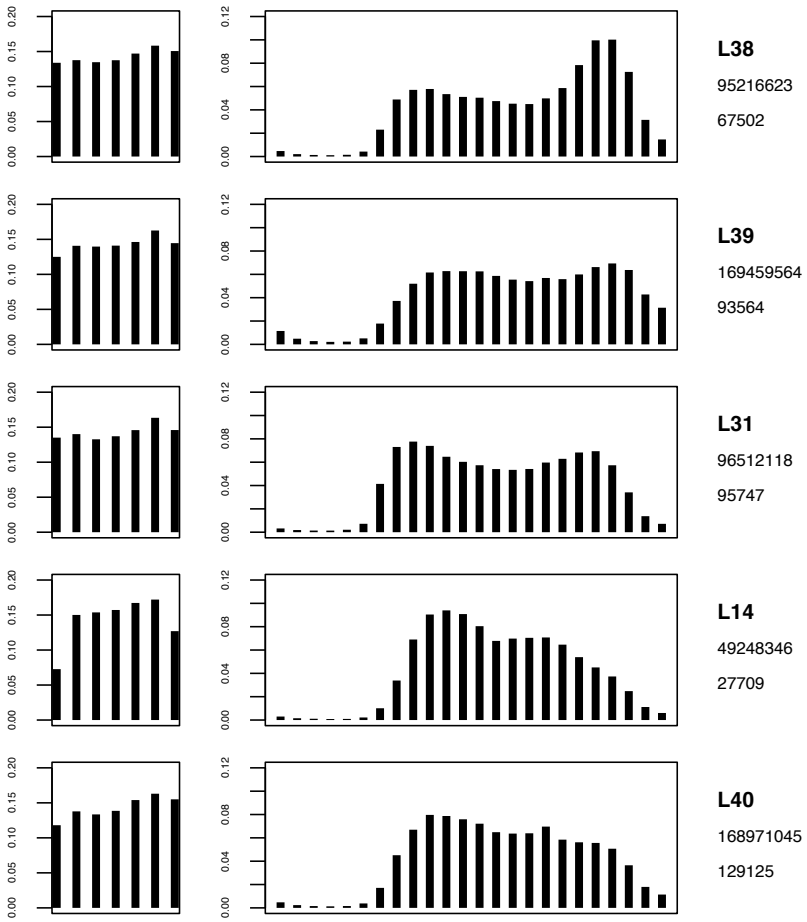


Fig. 10 Representatives for groups in the second half of dendrogram in Fig. 5.

References

[Anderberg, M. R. (1973)] Anderberg, M. R. (1973). *Cluster Analysis for Applications*. New York: Academic Press.

[Batagelj, V.(1988)] Batagelj, V. (1988). *Generalized Ward and Related Clustering Problems*. Classification and Related Methods of Data Analysis. H.H. Bock (editor). North-Holland, Amsterdam, 1988. p. 67-74.

[Batagelj, V. et al. (2010)] Batagelj, V., Kejžar, N., Korenjak-Černe, S. (2010). Clamix – tools for clustering of modal valued symbolic data. `https://r-forge.r-project.org/projects/clamix/` .

[Batagelj, V. et al. (2013)] Batagelj, V., Kejžar, N., Korenjak-Černe, S. (2013). Clustering of Modal Valued Symbolic Data. In preparation.

[Billard, L., Diday, E. (2006)] Billard, L., Diday, E. (2006). *Symbolic data analysis. Conceptual statistics and data mining*. New York: Wiley.

[Blondel, V. D. et al. (2012)] Blondel, V. D., Esch, M., Chan, C., Cleroty, F., Deville, P., Huens, E., Morloty, F., Smoreday, Z., Ziemlickiy, C. (2012). 'Data for developement: The D4D Challenge on mobile phone data. Retrieved from: *arXiv:1210.0137v1 [cs.CY]*, 29 Sep 2012.

[Bodlaj, J. et al. (2013)] Bodlaj, J., Cerinšek, M., Batagelj, V. (2013). D4D Challenge – Report: Visualization of traffic. Manuscript.

[Diday, E. (1979)] Diday, E. (1979). *Optimisation en classification automatique*, Tome 1.,2.. INRIA, Rocquencourt (in French).

[Hartigan, J. A. (1975)] Hartigan, J. A. (1975). *Clustering algorithms*, Wiley-Interscience: New York.

[Kaufman, L., Rousseeuw, P. (1990)] Kaufman, L., Rousseeuw, P. (1990). *Finding groups in data: an introduction to cluster analysis*. New York: John Wiley & Sons.

[Kejžar, N. et al. (2011)] Kejžar, N., Korenjak-Černe, S., Batagelj, V. (2011). Clustering of discrete distributions: A case of patent citations. *Journal of Classification, 28(2)*, 156–183.

[Korenjak-Černe, S. et al. (2011)] Korenjak-Černe, S., Batagelj, V., Japelj Pavešič, B. (2011). Clustering large data sets described with discrete distributions and its application on TIMSS data set. *Journal Statistical Analysis and Data Mining, 4(2)*, 199–215.

[Korenjak-Černe, S. et al. (2008)] Korenjak-Černe, S., Kejžar, N., Batagelj, V. (2008). Clustering of population pyramids. *Informatica, 32(2)*, 157–167.

[Korenjak-Černe, S. et al. (2002)] Korenjak-Černe, S., Batagelj, V. (2002). Symbolic data analysis approach to clustering large datasets. In: Jajuga, K., Sokołowski, A., Bock, H-H. (eds.). 8th Conference of the International Federation of Classification Societies, July 16-19, 2002, Cracow, Poland. Classification, clustering and data analysis. Berlin: Springer, p. 319-327.

[Ward, J. H. (1963)] Ward, J. H. (1963). Hierarchical grouping to optimize an objective function, *Journal of the American Statistical Association, 58*, 236–244.

Discovering common structures in mobile call data: An efficient way to clustering ego graphs

Syed Agha Muhammad and Kristof Van Laerhoven

Embedded Sensing Systems
Technische Universität Darmstadt
Germany
{muhammad,kristof}@ess.tu-darmstadt.de

Abstract. This paper conducts a study on a mobile call data set from 5000 individuals in order to examine what type of prototypical calling behaviors tend to occur for individual users. By representing the call data with methods from graph analysis, several features are suggested to characterize the shape and type of neighborhood graph around each mobile phone user. By cluster analysis of these features for all mobile users, we show that the data set contains seven distinct types of so-called neighborhood graphs or ego graphs. This categorization allows concise analysis of users' call data as they change over time and might be used as a sociological tool to detect rhythms and outliers in call behavior.

Keywords: Social computing, Graph analysis, Ego graph, Feature space, Fast independent component analysis (FastICA), k-means clustering, Clustering validation

1 Introduction

Social network analysis has been proposed as a key instrument in modern sociological studies. The elementary units in these networks are the so-called ego nodes representing mobile users, and structural ties depicting a relationship between them. Traditionally, social interaction data is gathered using labour-intensive methodologies that often are time-consuming and constrained by a limited amount of study participants observed. The use of data from mobile devices in the study of social data collection is gaining grounds, as especially mobile phones can relatively easily be equipped to collect call and location data. The low threshold of capturing such data provides a huge scientific opportunity to study the structure and dynamics of larger social networks at different levels; From the small-scale individuals perspective to the large-scale collective behavior of groups, with an unprecedented degree of reach and accuracy.

This work proposes a set of features taken from graph analysis to characterize and categorize the basic element in a social network: A user and his immediate network of close contacts. For this we consider the neighborhood graph or ego graph, which is a sub-graph that is formed around a single user node and comprises the nodes that was in close connection to during the observed period. For mobile call data, we thus categorize users according to their contact network.

Until several years ago, the empirical work on ego graphs have been confined to a limited amount of users. The advent of several technologies to observe and manage social data has since then led to studies conducted on much larger scale. The works on ego graphs have mostly focused on the prediction of personality traits of users[1–3]. Stoica et al. [4] used different structural measures on mobile data to analyze the neighborhood of the ego graphs. Mostly, the ego graphs are collected from social networking websites since their data are available in large chunks and their reference structure can be gathered in a straightforward manner. Akoglu et al. [5] have proposed a technique to find anomalies in ego graphs, and similarly, [13, 14] have focused on determining different social ties between participants from the social networking data.

In this paper, we use the calling pattern data from the Orange D4D mobile dataset taken from 5000 individuals over a period of five months. The clustering of the ego graphs is performed based on the neighborhood structure for each user. A key challenge for achieving this comes from the fact that the data provided does not contain qualitative information about said users apart from anonymized source and destination of calls made. The contributions of this paper are three-fold: (1) To cluster ego graphs, we selected 15 measures from graph theory as features that describe the ego graph. (2) After feature transformation, we apply independent component analysis to allow visual inspection of the feature space, and (3) We applied clustering analysis in the feature space to find emerging categories within the data. In real world scenarios, the social graph formation always follows skew distribution, and their structures can change abruptly.

The remaining of the paper is structured as follows: The input feature space measures are discussed in section II. The results of applying different measures on graphs are discussed in section III. In section IV, we perform visual analysis of the call data by applying independent component analysis on the transformed feature data. In section V, we discuss clustering results. We discuss the characteristics and properties of emerging clusters, and show how clusters for participants can morph over time. We discuss the conclusions of this paper in section VI.

2 Feature Extraction from Call-log data

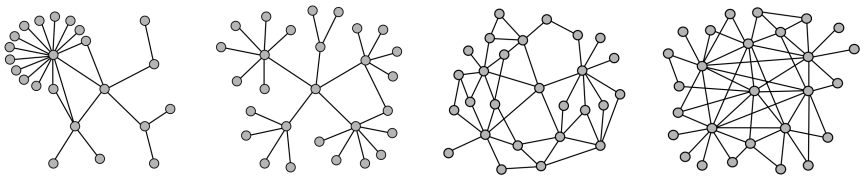
In this section, we give a brief introduction of the Orange dataset, we show how the mobile call data is represented with graphs, and introduce the proposed feature space as a basis for the clustering of the ego graphs.

2.1 Orange Mobile Data Set

For analysis, we use dataset 4 of the orange D4D challenge mobile dataset [6]. It contains call data (in call source and destination) for approximately 5000 mobile phone owners over a period of five months. The data is structured as source and destination numbers of calls made by the participants over singular time spans of two weeks. It does not provide any other type of information for the egos.

2.2 Representation of Mobile data with Graph Theory

A graph G is a pair of sets (V,E), where V denotes the finite set of non-empty nodes and E denotes the set of edges between the nodes. Mostly, graphs have some real value associated with the edge called the weight; however, the graphs provided in this dataset have no weights associated to them. The degree of a node is defined as the number of edges incident to the vertex. Figure 1 shows some of the possible combinations from the orange dataset. Figure 1(a) represents a graph with one of the node having many neighbors. Figure 1(b) represents a star shaped graph, where one of the node controls the communication between the remaining nodes. Similarly, Figure 1(c) and 1(d) represent the graph with many connections between the nodes in different branches.



(a) Strong neighbor (b) Star graph (c) Linked neighbors (d) Dense graph

Fig. 1: Some illustrations of ego graphs, depicting the ego node in the middle, with connections to first- and second-degree neighbors. This work focuses on the automatic categorization of such ego graphs according to their graph structure.

2.3 Feature Space

In the remainder of this section, we discuss several concepts from graph theory, such as centrality measures, small-world model, transitivity, clique, k-core and sub-graph similarity features, that can be applied to describe ego graphs.

Centrality Measures There are different centrality measures available in literature, but most famous amongst them are degree, betweenness and closeness centrality [7].

Degree centrality is the number of edges adjacent to a node, with which it is in direct contact. Any node, whose position permits him to be directly in contact with many other nodes is perceived to be as a major channel of communication. Closeness centrality is based upon the degree to which a node is close to other nodes in the graph. Normally, a higher closeness suggests the capability of a node to send information quickly across its neighbors. A node is considered to be central with respect to time and cost efficiency, if it has minimum distance to all the other nodes in the network.

The two measures mentioned above are directly based on how close the ego node is to the other nodes, while betweenness centrality is based on the geodesic distance between the specific node and the remaining nodes. Betweenness centrality is based on the frequency with which a node is placed between two nodes on the shortest path connecting them. One of the important roles of betweenness centrality is that it acts as boundary spanner between different nodes, that can not communicate with each other directly.

For the call log data, centrality measures provide an intuitive prospective about the characteristics of the graph as a whole and any node in particular. A person having high degree might have a more active call behavior for the observed period. Similarly, higher closeness and betweenness describe the importance of a node for diffusing the information in the graph, and the importance of node ties within the network respectively.

Efficiency measures Efficiency measures introduced by Latora and Marchiori [8] can be used to find out how efficiently information is exchanged over the network, and to characterize the closeness of the ego to the small-world model.

A small world network is defined as a graph in which most of the nodes are not direct neighbors to each other, but where most nodes can be reached by a small number of steps. Small world networks are highly clustered, like regular lattices, and have small characteristic paths like random graphs. The global efficiency of the network E(G) is defines as:

$$E(G) = \frac{1}{N(N-1)} \sum_{i \neq j \epsilon G} \frac{1}{d_{ij}} \tag{1}$$

d_{ij} denotes the shortest path length between i and j, and 1/N(N-1)is the normalizing factor, with a value between 0 and 1, where 1 represents high and 0 represents low efficiency.

For each node i in the graph, the local efficiency is defined as

$$E_{loc} = \frac{1}{N} \sum_{i \neq j \epsilon G} \frac{E(G_i)}{(G_i^{ideal})} \tag{2}$$

where for each node i $E(G_i^{ideal})$ is the efficiency of the ideal case, when G_i has all possible $k_i(k_i\text{-}1)/2$ edges, where k_i represent the edges incident with i.

The local efficiency describes how fault tolerant the system is, which means in case when node i is removed from the graph, how efficient the communication between the first neighbors of i remains. The concept of fault tolerance is described from the prospective of the immediate neighbors of the ego, and not from the overall graph prospective. Higher values of global and local efficiency suggest a model which is nearer to small world model.

Transitivity Transitivity measures the probability that the neighborhood of an individual node is connected. We used both global transitivity and local transitivity in particular:

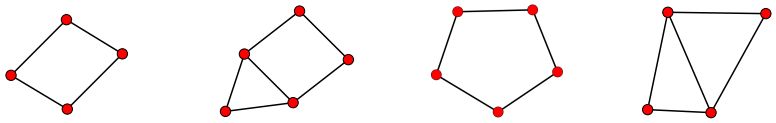
Both use the concept of triple, which is a set of three nodes, that can be closely connected to each other (close loop), or two out of three nodes are connected (open triple). The global transitivity of a given graph G is then defined as the ratio between the number of closed triples in G and the total number of triples. It gives clustering value at the level of the entire graph. Normally, for global transitivity the number of triples are counted in the ego graph. For each ego, presence of triples gives an indication of the clustering in a network, and is often referred to as clustering coefficient. The clustering coefficient of a node is the fraction of pairs of its neighbors that have edges between one another. The local transitivity of a node measures how concentrated its neighbors are to forming a clique and the graph to a small-world network.

Cliques A clique [9] is defined as a group of nodes that are tightly connected to each other. In graph theory, a clique is defined as a maximal complete sub-graph of graph. Every node in the sub-graph has a direct connection to every other node in the graph. A clique is composed of many overlapping and close triples. Since we are interested in knowing the community structures around an ego, it is always better to scan through some ranges of k and then monitor how communities change. Finding large number of cliques in such small graphs is to some extent a strict parameter to cluster an ego; however, they reflect cohesiveness present in the graph, and are also a reasonable approximation for the presence of close complete sub-graphs on the ego.

k-core A k-core of G is a maximal connected sub-graph of G with all nodes having a degree of at least k. It is one of the connected components of the sub-graph of the graph by deleting all the nodes having a degree less than k [10].

Sub-graph Isomorphism In the previous two sections, we discussed clique and k-core to detect presence of dense structures and shapes from the ego graphs. However, real graphs have many complicated structures that are sometimes slightly difficult to recognize, especially the ones showing ties between the second order neighborhood of the ego. We defined four such pattern, as shown in Figure 2, and apply VF2 [11] to detect those patterns from the graphs.

As an illustration of the feature space, we applied all aforementioned features on four different ego graphs of the type shown in Figure 1. The resulting values are represented in Table 1 below, showing the results for different features (rows) extracted from the four graphs (columns). Highlighted is the situation when the ego node does not have the highest degree or between-centrality in the graph. We added another feature to detect highly populated nodes from the data, especially the ones with very high degree and dense structures. Such a case could be interesting to analyze, and different interpretations might be possible based on the social context. Certain features show distinct variations for different graph structures, especially clique, k-core, sub-graph matching results. The star-shaped graph has the highest betweenness amongst all, as it is the center



(a) Rectangle (b) Penta and tri (c) Penta (d) Double triangle

Fig. 2: The chosen shapes to be matched within the graph, depicting typical communication patterns that are present in different first- and second-order neighborhoods of the ego graph.

for communication in the network. For the denser graphs, results indicated very strong structure.

		Strong neighbour	Star	Linked neighbour	Dense
Centrality	Degree	**0.208**	**0.2**	**0.161**	**0.18**
	Betweenness	**0.403**	0.79	0.467	0.361
	Closeness	0.558	0.55	0.534	0.515
Efficiency	Global	0.0007	0.0005	0.0003	0.0001
	Local	0.534	0.533	0.517	0.683
Transitivity	Global	0.008	0.0	0.095	0.097
	Clustering	0.1	0.0	0.1	0.162
Clique	number of cliques	1	0	8	20
k-core	size of core	7	0	19	21
Sub-graph	Rectangle	1	0	1	1
	Penta and Tri	1	0	1	1
	Penta shape	0	1	1	0
	Double triangle	0	0	3	8
Ego neighbours	size of the neighbours	5	5	5	9
Populated node	boolean	1	0	0	0

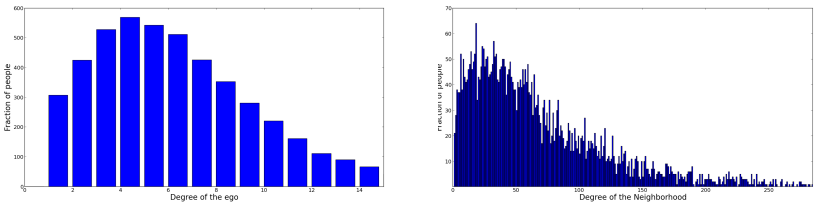
Table 1: Features illustrated for ego graphs of the four types from Figure 1.

3 Analysis of the Features

In this section, we apply the features to the call-log data to study the structural properties of the ego graphs and different trends in them. We discuss especially the degree distribution of the ego node and its neighborhood, centrality measure, clique, and k-core results.

Degree distribution is a fundamental measure used for the study of networks. Figure 3 shows the degree distribution for the ego and its neighborhood. In this Figure, x-axis and y-axis represent the degree and fraction of mobile users sharing a specific degree in the network respectively. We only report the most relevant trends from the data. Figure 3(a) shows that almost 80% of the participants have

less than 10 immediate neighbors. The median for each ego node is 6. Figure 3(b) suggests that neighbors typically have a higher degree than the ego node. The median value for the first order neighborhood are slightly higher than the ego node's median value. The average median value is 8 for each neighbor. These two measures are a first key determinant to describe the graphs, as they describe the graph in terms of their structural ties.



(a) Ego degree distribution (b) Neighbours degree distribution

Fig. 3: Degree distribution for the ego node and its neighborhood

Figure 4 shows the results for betweenness and closeness centrality. Betweenness centrality varies significantly for ego networks: Around 60% of the mobile users share betweenness centrality between 0.50 to 0.80. Higher betweenness represents those graphs where the ego node has no second order neighborhood, and therefore tends to lack dense, cohesive structure. Similarly, lower betweenness represents graphs where one of the neighbors has much more nodes than the rest of the graph, in that case the first neighbor with a highly populated structure has the highest betweenness. Closeness centrality results are more stable: For around 90% of the users, its values are more or less within a defined range. The stability comes from the fact that ego has a maximum hop distance of 2 to access any node within the graph.

Figure 5 shows the results for the number of cliques and k-core structures present in the data. In this Figure, x-axis and y-axis represent the value for k and number of cliques, k-cores respectively. We vary the values of k for cliques and k-cores. It suggests on average 3 thousand participants have a clique of at least a set of 3 members in the graph. As we increase the value for k, the number of cliques and k-cores starts to decrease. Figure 5 shows that there exist very few large dense structures in the data.

4 Fast Independent Component Analysis (FastICA)

In this section, we present visualization results by using FastICA [12] to reduce the dimensionality of the feature space. The technique is based on a fixed point iteration scheme increasing non-Gaussianity as a measure of statistical independence. To reiterate, the proposed feature space consists of following 15 features

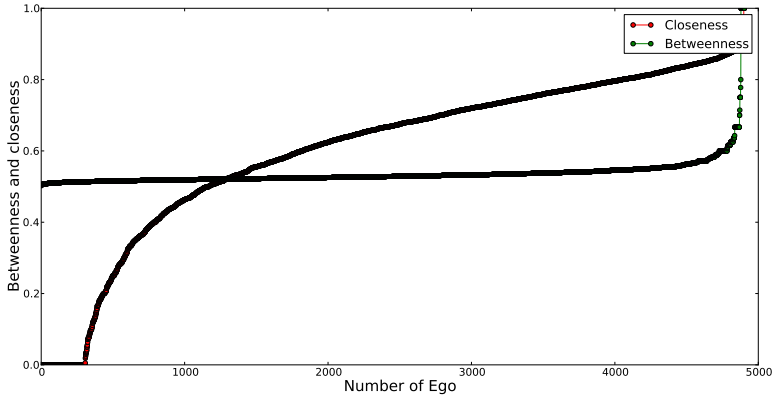


Fig. 4: Betweenness and Closeness centrality for all ego nodes.

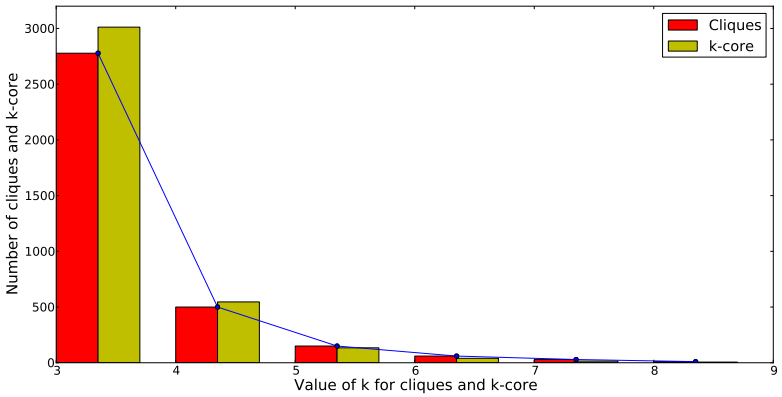


Fig. 5: Average cliques and k-core in the data.

from all ego graphs: a) 3 centrality measures; b) 2 small world measures; c) 2 transitivity measures; d) clique; e) k-core; g) sub-graphs matching with 4 shapes h) first order neighbors i) populated nodes. We normalized the features prior to applying the algorithm. Figure 6 shows the results for FastICA for different time periods over a period of 5 months, when combining all data and afterwards transforming it with FastICA for the given periods. The visual inspection suggests that overall three general patterns have emerged from the data. Each plot represents the data for two weeks. For the first two weeks, the graph has roughly twelve clusters emerging. For the next 12 weeks, the results shows approximately six consistent clusters. For the last two weeks, the clusters have shifted slightly.

In the next section, we will discuss our use of the k-means clustering algorithm for this data to automatically detect these emerging clusters.

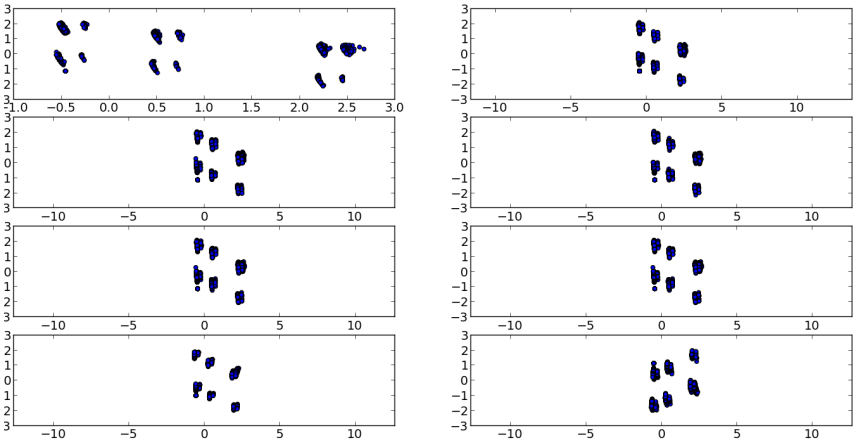


Fig. 6: FastICA results over different time periods with reduced dimensionality.

5 Clustering Results

In this section, we discuss clustering results from the k-means algorithm applied to the features discussed previously. We discuss the properties of different emerging clusters and also the clustering trends from the call-log data. We concatenated the input feature space of the five months data, and performed k-means clustering to study their characteristics at different temporal spaces. We discuss the properties of different emerging clusters and also the clustering trends from the call-log data.

In the visual inspection, we observed mostly six clusters for the different 2-week time spans. To optimally cluster the data and avoid any clustering error, we performed an elbow test on k-means data for different value of k as shown in Figure 7. The figure's x-axis and y-axis represent the different values of k and percentage of variance between different k values, respectively. The elbow method provides a reasonable approximation to select the value of k. It measures the marginal gain between consecutive clusters, where a high value of marginal gain between two clusters means a proper division of the clusters, which at some point can be expected to drop: The number of clusters are selected at that point. It suggests that marginal gain is increasing till k is 7, and then it

decreases. The value of k was thus set to 7 and we applied k-means to the 15-dimensional feature space. The data was clustered in 7 groups, each cluster with certain characteristics. We perform next a thorough study of each of the emerging clusters. The detected 7 prototypical clusters have the following properties and characteristics:

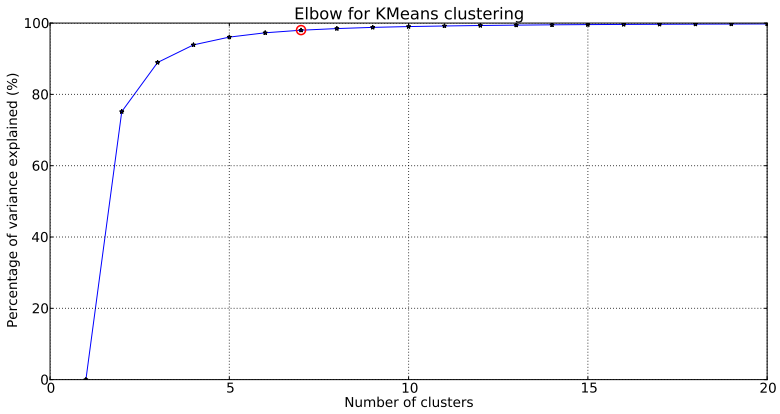


Fig. 7: Cluster validation for k-means.

Cluster 1 The ego has very few immediate neighbors, and there is no connection between any of the ego neighbors. The provided graphs have no second order neighborhood apart from few cases. Normally, such clusters are identified by a high value of centrality measures and the graphs have no clique and k-core in them. The sub-graphs matching algorithm finds no matching shapes.

Cluster 2 The ego graph has a reasonable size immediate neighborhood, but a very strong second order neighborhood with very high degree. In many cases, the graphs have highly populated nodes. The provided graphs are rich in terms of their structure, but have a very few densely connected sub-graphs. For some cases, the sub-graphs matching algorithm finds matching shapes.

Cluster 3 The ego has on average between 7 and 10 immediate neighbors. The network has some dense, and completed sub-graph connections, but mostly in the second order neighborhood. The graphs have many sub-graph matches. The graphs have many over populated and dense structure nodes .

Cluster 4 The ego has very high first, and second order ties. Overall, the graphs are densely populated with many strong cohesive networks.

Cluster 5 The provided graphs have overall small size network, and overall the dense structures are not significant in the graph. The graphs have very few first and second order neighborhood. The graphs have no highly populated nodes.

Cluster 6 The ego has a very high number of first order neighborhood, and reasonable second order networks. The graphs have not too many cliques, or k-cores. Mostly the graphs have high sub-graph matching results, especially for rectangle and penta shape. It shows that different nodes of the second order neighborhood are connected to each other.

Cluster 7 The ego has very few first order network, and one of the alters has an extremely high populated structure. In such a case, the specific alter is the most powerful node within the network. The specific alter has a very high value for centrality and small-world measure.

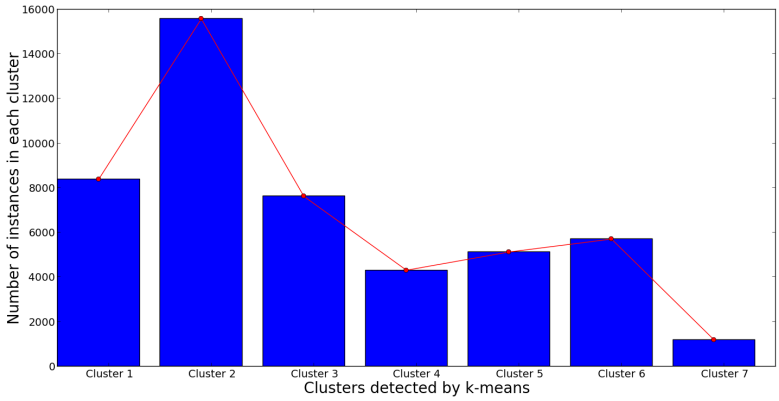


Fig. 8: Number of instances classified in each cluster.

Figure 8 shows the number of instances clustered in each of the seven clusters. Most of the calling patterns are clustered into the first three clusters, which have some very interesting properties. The highest number of instances are associated with cluster 2, where the neighborhood of the ego has many nodes, but the graphs does not have many dense structures. There are very high instances of cluster 1, which indicates that a bigger chunk of the call-log data has no well-defined structure. Similarly, cluster 3 indicates the graphs with dense structures.

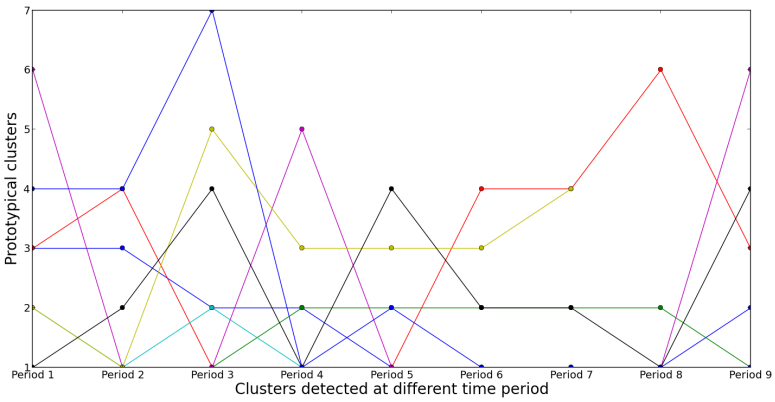


Fig. 9: Morphing of participants clustering shapes at different time period (at random only 10 participants selected).

Minimum instances are found for cluster 7, where one of the immediate neighbors of the ego is more powerful than the ego.

We observed variations in the clustering characterization of the egos at different temporal spaces, the same ego has different calling patterns and assigned to a different cluster. To get an overview of the clustering pattern at different temporal spaces, we selected a sample of 10 participants with their clustering patterns as shown in Figure 9. We selected a small sample to avoid complexity, but the exhaustive study of the clustering results suggested that calling pattern for the remaining participants have also more or less fluctuated during different temporal spaces. Figure 9 shows that for some participants the calling patters are consistent, but for some participants the pattern has been consisting changing. For some participants, the graphs have fluctuated between 4,5 time or even more during the 5 months period. The variations mainly be due to the expected skewness in the real world scenarios. The emerging clustering structures and their characteristic can be further used in sociological studies to assign meaningful interpretations to them.

6 Conclusion and contributions

In this paper, we clustered the ego graphs for 5000 participants based on their calling patterns. The paper aimed to provide a data-driven instrument to be used by sociologists for graph interpretations. We formed an input space of a large number of graph theory measures, especially those features addressing the properties of the ego networks. After extracting features from the graphs, we applied independent component analysis to visually inspect the possible clustering structures within the data. A visual inspection by means of dimensionality reduction with the FastICA algorithm suggested 6 clusters for most of the time periods. Similarly, we tested the high dimensional input space with k-means clustering to

detect the prototypical clusters from the data. This clustering analysis produced 7 most prototypical emerging clusters, each cluster is characterized by certain characteristics and properties; such as denser graph, structure based on different shapes, richness of the ego and its neighbors. Clustering structures for participants were morphing with temporal spaces, resulting in abrupt fluctuations in the graph structures. In the future, we would like to apply the same feature space and clustering technique on graphs with some defined reference structure, as it will provide a facility to interpreting them from social prospectives.

7 Acknowledgements

The work has been fully supported and funded by Graduate College, 'Topology of Technology' (GRK 1343).

References

1. Wehrli, S.: Personality on Social Network Sites: An Application of the Five Factor Model. ETH Zurich Sociology Working Papers 7 (2008)
2. Chittaranjan, G., Blom, J., and Gatica-Perez, D.: Whos who with Big-Five: Analysing and Classifying Personality Traits with Smartphones. ISWC (2011)
3. J. Staiano, B. Lepri, N. Aharony, F. Pianesi, N. Sebe, and A. Pentland, Friends dont Lie: Inferring Personality Traits from Social Network Structure, Proc. of ACM UbiComp, 2012.
4. A. Stoica and C. Prieur, Structure of neighborhoods in a large social network, SocialCom, 2009.
5. L. Akoglu, M. McGlohon, and C. Faloutsos. oddball: Spotting anomalies in weighted graphs. In PAKDD (2), pages 410421, 2010.
6. Blondel V, Esch M, Chan C, Clerot F, Deville P, et al.. (2012) Data for Development: the D4D Challenge on Mobile Phone Data. arXiv : 1210.0137.
7. Freeman, L.: Centrality in social networks: Conceptual clarification. Social Networks, 1(3), 215-239 (1979)
8. Latora, V. and Marchiori, M.: Economic Small-World Behavior in Weighted Networks. Eur. Phys. Journ. B Condensed Matter 32(2) (2003)
9. Palla, G., I. Derenyi, I. Farkas, and T. Vicsek. 2005. Uncovering the overlapping community structure of complex networks in nature and society. Nature 435: 814 818.
10. V. Batagelj and M. Zaversnik, An O(m) Algorithm for Cores Decomposition of Networks, cs.DS/0310049 , 2003.
11. L.P. Cordella, P. Foggia, C. Sansone, M. Vento. A (sub)graph isomorphism algorithm for matching large graphs, IEEE Transactions on Pattern Analysis and Machine Intelligence, 26 (10) (2004), pp. 13671372
12. A. Hyvrinen. Fast and robust fixed-point algorithms for independent component analysis, IEEE Transactions on Neural Networks, 10 (3) (1999), pp. 626634
13. Tang J, Lou T, Kleinberg J. Inferring social ties across heterogeneous networks. In: WSDM12
14. W. Tang, H. Zhuang, and J. Tang. Learning to infer social relationships in large networks. In ECML/PKDD11, 2011.

DAMMoPD

Data Analysis and Mining of Mobile Phone Dataset

TEDECO[1]

Carlos Badenes Olmedo, Susana Muñoz Hernandez

cbadenes@gmail.com, susana@fi.upm.es

Tel.+34 91 336 7455, Fax +34 91 336 3669

Facultad de Informática, Universidad Politécnica de Madrid

Campus de Montegancedo, 28660 Boadilla del Monte (Madrid)

1 "Technology for the Development and the Cooperation" Cooperation group of the Technical University of Madrid. http://tedeco.fi.upm.es

1 Proposal

Create a learning machine, based on the content of four databases, which is able to generate notifications about any identified incident and even be able to predict future events by behavior patterns.

The knowledge base is enriched by the information collected along with information calculated, such as:

- Areas with higher and lower use of *sms* and voice calls
- Patterns of daily travel per person in a given area in a particular time
- Average distance in every voice calls and *sms*
- Identification of cell towers with high and low voice and SMS traffic

The knowledge gained could identify whether a cell tower is saturated, for example, when there are consecutive calls with the same origin and destination which could mean that the line went dead. Detect the quality of communication routes between home and work calculating the average time to travel from home to work and so decide to install public transport for more used zones. Generate an alarm when a cell tower has more traffic than threshold of normal load mainly in unoccupied areas. Analyze the number and duration of voice calls on towers which has more SMS traffic because could be there coverage problems.

2 Scope

We analyze the four mobile phone datasets provided by Orange, along another additional information to discover some valuable knowledge that contribute to the socio-economic development of the Ivory Coast population.

This analysis has focused on the proposal of a development primarily **descriptive**, but a **predictive** model that should be built to complete a sustainable solution based on these previous results.

So we start discovering different communication patterns and then we try to associate these patterns to geographic locations with enough accuracy to be relevant. With these results we analyze the specific situation on these places to get a useful knowledge that explains why these patterns occur and find solutions to improve communications.

3 Data Preparation

3.1 DataSet

The datasets are based on anonymized call detail record (cdr) of phone calls and SMS exchanges between five million of Orange's customers in Ivory Coast between December 1, 2011 and April 28, 2012.

1. **Antenna-to-antenna traffic on an hourly basis**

date_hour	originating_ant	terminating_ant	nb_voice_calls	duration_voice_calls

Table 1: SET1: Antenna-to-antenna

2. **Individual Trajectories for 50,000 customers for two week time windows with antenna location information**

user_id	connection_datetime	antenna_id

Table 2: SET2:Individual Trajectories: High Spatial Resolution Data

antenna_id	longitude	latitude

Table 3: SET2: Individual Trajectories: Antenna location

3. **Individual Trajectories for 50,000 customers over the entire observation period with sub-prefecture location information**

user_id	connection_datetime	subpref_id

Table 4: SET3: Individual Trajectories: Long Term Data

subpref_id	longitude	latitude

Table 5: SET3: Individual Trajectories: Subpref location

4. **Communication graphs for 5,000 customers**

source_user_id	destination_source_id

Table 6: SET4: Communication Subgraphs

3.2 Data Cleaning and Transforming

To make more effective this analysis we need to review all collected data to complete sometimes with more detailed information and to remove less significant fields other times. Association rules can't work with all types of information, specific format and values are required to discover facts, principles or relationship.

Those "*Dimensions"*, or more important sets of information, have been extended, called *aggregation levels*, to get more in-depth information. With all of this we build a *multidimensional model* to organize data around *facts*, whose *attributes*, or measures, could be seen more or less detailed according these *dimensions*.

3.2.1 Time Dimension

Time-dependent data is especially important to get patterns of behavior or practices. This type of information is key, and so we transform these fields: '*date_hour'* of '*Antenna-to-antenna*' dataset, and '*connection_datetime*' of '*Individual Trajectories*' datasets to new attributes with more relevant values.

Figure 1 illustrates the new types of time fields. These types of data are required by association rules so we will transform the *timestamp* values to new range of values: *week days*, *workday* or *public holiday* and a *time interval* into a day.

A *Time-Interval* is defined by:

NAME	INTERVAL
Morning	4:00 a.m – 11:59:59 a.m
Afternoon	12:00 p.m. - 7:59:59 p.m.
Evening	8:00 p.m. - 3:59:59 a.m.

Table 7: Time Interval

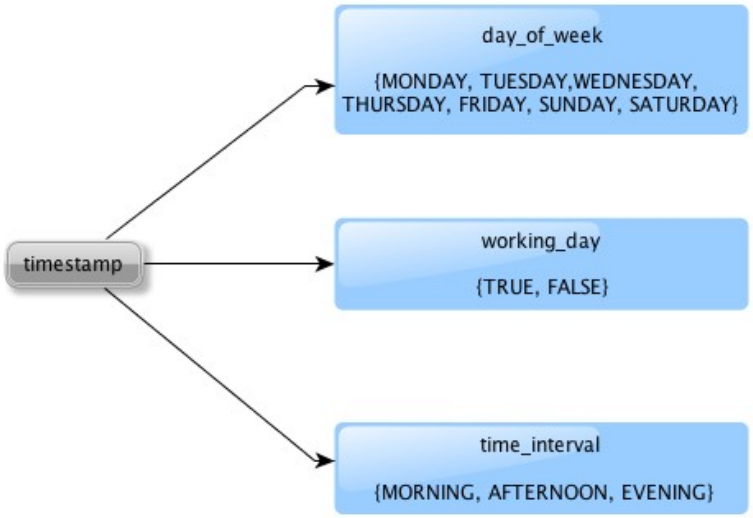


Figure 1: Date transformation

The next table illustrates the public holidays of Ivory Coast for 2012:

Day	Day of Week	Description
January 01	Sunday	New Year's Day
February 05	Sunday	The Day after the Prophet's Birthday
February 13	Monday	Public Holiday (Post African Cup of Nations Recovery)
April 09	Monday	Easter Monday
May 01	Tuesday	Labor Day
May 17	Thursday	Ascension Day
May 28	Monday	Whit Monday
August 07	Tuesday	Independence Day (National Day)
August 15	Wednesday	Assumption Day
August 16	Thursday	The Day after Lailatou-Kadr (Quran Revaiation)
August 19	Sunday	Korite / Aid-el-Fitr (End of Ramadan)
October 26	Friday	Tabaski / Ai-El-Kebir (Feast of Sacrifice)
November 01	Thursday	All Saints' Day
November 15	Thursday	National Peace Day
November 15	Thursday	Islamic New Year
December 25	Tuesday	Christmas Day

Table 8: Public Holidays

3.2.2 Duration Dimension

The mean of duration for each voice call could be as useful as number of voice calls, so we create a new field named '*mean_duration_voice_calls*' calculated from '*nb_voice_calls*' and '*duration_voice_calls*' existing fields.

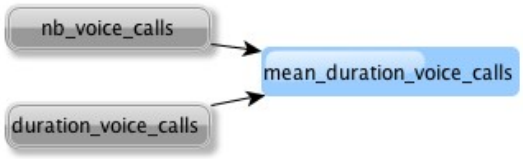


*Figure 2: attribute '**mean_duration_voice_calls**'*

3.2.3 Location Dimension

Google Geocoding API v3 (*https://developers.google.com/maps/documentation/geocoding/*) is designed for geocoding static addresses for placement of application content on a map.

We use this service to locate antennas into three administrative levels: area, region and country. In this way, we can study the communication between users by area or region or country and analyze the results grouped in these three levels.

A Java client is developed to use the remote RESTful web service:

http://maps.google.com/maps/api/geocode/json?latlng=<*lat*>,<*long*>&*sensor=false*

As described in D4D Mobile Phone Data document, we use geographic coordinates of antenna positions for locate them on the map. At this point, is important to say that, because the accuracy level is good but not the best, some antennas have been located in neighborhood countries as Ghana or Mali:

- **GHANA**: Juabeso, Bia y Jomoro
- **MALI**: Kadiolo

This is not a problem because these regions will be processed as any other region from Ivory Coast.

We persist these locations in the file: '*antenna_location.arff*' to be used and related in each analysis that required it.

3.3 Exploring and Selection

We want to have mining views, tables with important fields only, optimized to every particular study. These views are in ARFF format (Attribute-Relation File Format) understandable by WEKA (http://weka.wikispaces.com) , an open-source data mining tool.

This has made it necessary to build a Java application which, using the WEKA API libraries, to read received data from Orange in TSV format (*Tab Separated Values*) and map, reduce and calculate new fields to write them in ARFF files. The source code is available in Github:

https://github.com/cabadol/dammopd

The following datasets have been grouped by a key field: users, antennas and locations.

3.3.1 User by Antenna Driven Data

From dataset '*Individual Trajectories: High Spatial Resolution Data*' the fields related with users and antennas have been selected and filtered together with calculated fields to build this view stored in the file: '**user_by_antenna_driven_data.arff'.**

The next table shows the fields of this view. More details about these fields can be found in APPENDIX A:

FIELDS	
user_id	num_calls_made_tuesday
num_calls_made	num_calls_made_wednesday
mean_calls_made	num_calls_made_thursday
num_calls_made_holiday	num_calls_made_friday
num_calls_made_working	num_calls_made_saturday
day_of_calls_made	num_calls_made_sunday
num_calls_made_morning	weekday_of_calls_made
num_calls_made_afternoon	num_antennas
num_calls_made_evening	most_used_antenna
time_of_calls_made	antenna_relationship
num_calls_made_monday	

Table 9: User by antenna driven data fields

3.3.2 User by Sub-prefectures Driven Data

From dataset '*Individual Trajectories: Long Term Data*' the fields related with users and sub-prefectures have been selected and filtered together with calculated fields to build this view stored in the file: '**user_by_subpref_driven_data.arff'.**

The next table shows the fields of this view. More details about these fields can be found in APPENDIX A:

FIELDS	
user_id	num_calls_made_tuesday
num_calls_made	num_calls_made_wednesday
mean_calls_made	num_calls_made_thrusday
num_calls_made_holiday	num_calls_made_friday
num_calls_made_working	num_calls_made_saturday
day_of_calls_made	num_calls_made_sunday
num_calls_made_morning	weekday_of_calls_made
num_calls_made_afternoon	num_subprefs
num_calls_made_evening	multi_subprefs
time_of_calls_made	most_used_subpref
num_calls_made_monday	subpref_relationship

Table 10: User by sub-prefectures Driven Data fields

3.3.3 Antenna Driven Data

From dataset '*Atenna-to-Antenna Data*' the fields have been grouped and have been joined with new calculated fields to build this view stored in the file: '**antenna_driven_data.arff'.**

The next table shows the fields of this view. More details about these fields can be found in APPENDIX A:

FIELDS			
ant_id	day_of_calls_recv	num_calls_made_monday	num_calls_recv_friday
num_calls_made	dur_calls_recv_holidays	num_calls_made_tuesday	num_calls_recv_saturday
num_calls_recv	dur_calls_recv_working	num_calls_made_wednesday	num_calls_recv_sunday
mean_calls_made	day_longer_calls_recv	num_calls_made_thrusday	weekday_of_calls_recv
mean_calls_recv	num_calls_made_morning	num_calls_made_friday	dur_calls_recv_monday
num_calls	num_calls_made_afternoon	num_calls_made_saturday	dur_calls_recv_tuesday
calls_type	num_calls_made_evening	num_calls_made_sunday	dur_calls_recv_wednesday
dur_calls_made	time_of_calls_made	weekday_of_calls_made	dur_calls_recv_thrusday
dur_calls_recv	dur_calls_made_morning	dur_calls_made_monday	dur_calls_recv_friday
dur_calls_type	dur_calls_made_afternoon	dur_calls_made_tuesday	dur_calls_recv_saturday
mean_dur_calls_made	dur_calls_made_evening	dur_calls_made_wednesday	dur_calls_recv_sunday
mean_dur_calls_recv	time_longer_calls_made	dur_calls_made_thrusday	weekday_longer_calls_recv
num_calls_made_holiday	num_calls_recv_morning	dur_calls_made_friday	num_antennas_called
num_calls_made_working	num_calls_recv_afternoon	dur_calls_made_saturday	num_antennas_recv
day_of_calls_made	num_calls_recv_evening	dur_calls_made_sunday	multi_caller
dur_calls_made_holiday	time_of_calls_recv	weekday_longer_calls_made	multi_receiver
dur_calls_made_working	dur_calls_recv_morning	num_calls_recv_monday	most_called_antenna
day_longer_calls_made	dur_calls_recv_afternoon	num_calls_recv_tuesday	most_recv_antenna
num_calls_recv_holiday	dur_calls_recv_evening	num_calls_recv_wednesday	
num_calls_recv_working	time_longer_calls_recv	num_calls_recv_thrusday	

Table 11: Antenna Driven Data Fields

3.3.4 Area Driven Data

Based in the last view: '*Antenna Driven Data*', we're going to build a new view centered in the administrative area where every antenna is located. For this, we'll use the dataset with locations of antennas obtained from Google Geolocation API. Fields will be stored in the file: '**area_driven_data.arff'.**

The next table shows the fields of this view. More details about these fields can be found in APPENDIX A:

FIELDS
area
num_antennas
num_calls
num_calls_by_antenna
num_calls_made
num_calls_recv
mean_dur_calls_made
mean_dur_calls_recv

Table 12: Area Driven Data

4 Evaluation and Interpretation

4.1 Antenna workload

Analyzing the data of: '*area_driven_data.arff*' we can see that there are regions with less antennas that needed to complete successfully all traffic of communications and avoid situations where any communication could not be established because the antenna is saturated with others previously established communications.

The *mean* of calls handled by one antenna for this 5 months is: 599.099,0 calls. But this value is not completely adequate, because the *variance*: 87.624.878.794,0 and the *standard deviation*: 296.015,0 are too high. In our view, there is technical merit in using the *median:* 694.113,0 for a 95% *confidence interval* [623.233-764.993], since it is less sensitive to extreme values.

According to this confidence interval there are regions where should build more antennas:

Area	CURRENT Num. Antennas	PROPOSAL Num. Antennas	
		lower	upper
Bongouanou	**19**	19	23
Duekoue	**11**	11	13
Sakassou	**4**	4	5
Toumodi	**15**	15	19
Divo	**15**	16	19
Bounfie	**18**	19	23
Beoumi	**5**	5	7
Biankouma	**4**	4	5
Danane	**11**	12	15
Daloa	**29**	32	39
Daoukro	**12**	13	16
Bangolo	**2**	2	3
Bocanda	**3**	4	4
Korhogo	**16**	19	23
Abidjan	**396**	471	579
Yamoussoukro	**36**	43	53
Vavoua	**11**	14	17
Man	**14**	18	22
Grand-Bassam	**8**	10	13
Guiglo	**10**	13	16
Oume	**15**	20	24
Sassandra	**14**	18	23
Agnibilekrou	**8**	11	13
Abengourou	**23**	34	42
Sinfra	**7**	11	13
Tabou	**9**	14	17
San Pedro	**30**	50	61
Bouake	**26**	44	54
Soubre	**41**	73	89

Table 13: Recommendation of antennas per area

4.2 Calls Made and Received Pattern

Again, analyzing the data*: 'area_driven_data.arff'* discover a curious pattern between calls made and calls received. We calculate the rate of calls made over received for each region. Values under 1% are dropped.

Three types of areas according to the type of calls mainly established are defined:

- *Speaker*: Areas where have been recorded more calls made than received.
- *Recipient*: Areas where have been recorded more calls received than made.
- *Neutral*: The same number of calls made and received.

These are the speaker areas:

Area	Calls Made over Received %
Juabeso	14.26
Tengrela	12.90
Bia	12.51
Grand Lahou	4.45
Tabou	3.94
Sassandra	3.21
Oume	2.09
Jomoro	1.97
Kadiolo	1.80
Guiglo	1.79
Duekoue	1.27
Alepe	1.15

Table 14: percentage of calls made over received

You could think about wealth of a region to made more calls than receive, but if you take a look the number of antennas of areas like Juabeso, Tengrela or Bia you'll see that these regions only have one antenna and not too many communication traffic. In addition, when these areas are located on the map you can see that they are all in border zones. This is the major reason why this happens.

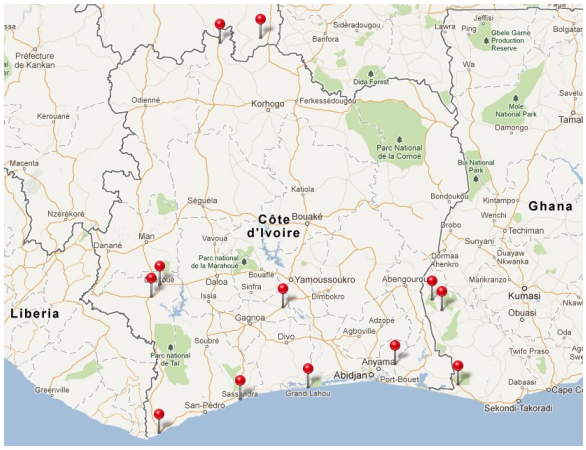


Figure 3: Speaker areas

In this areas could be interesting define a special Calling Plan which reduce the call set-up charge to facilitate the communication in these places.

These are the recipient areas:

Area	Calls Recived over Made %
Odienne	4.49
Tiassale	4.18
Lakota	4.03
Beounm	3.51
Daloa	2.89
Divo	2.53
Seguela	2.49
Niakaramandougou	2.18
Issia	2.11
Biankouma	2.10
Danane	2.09
M'Bahiakro	2.03
Man	1.73
Bongouanou	1.70
Mankono	1.70
Zuenoula	1.68
Bouake	1.60
Tanda	1.39
Grand Bassam	1.33
Tiebissou Department	1.21
Adzope	1.15
Dimbokro	1.14

Table 15: percentage of calls received over made

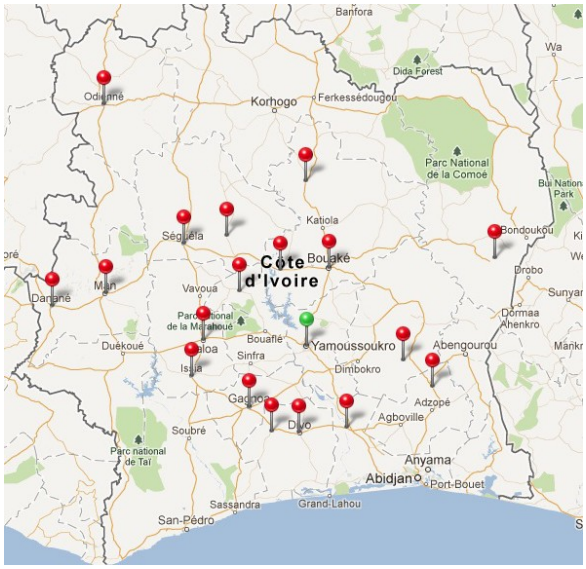


Figure 4: Recipient areas

The values are not as high as before. These locations are poor areas near the capital Yamoussoukro. It is very likely that people migrate in order to work or study to cities with more wealth leaving friends and family in their native cities. A special Calling Plan which define a group of number with low-cost for calls between them, could be an interesting option to encourage the calls from these areas.

4.3 Call Duration Pattern

In base of '*area_driven_data.arff*' *we have checked what mean of call duration are recorded for each administration area, making a distinction between calls made and received.*

Mean of call made duration is distributed from 96 seconds in Katiola to 174 seconds in Dumbokro. Its mean value is 125 seconds, standard deviation is 15,5 seconds, so its a uniform distribution.

Mean of call received duration is distributed from 93 seconds in Tengrela to 178 seconds in Dumbokro. Its mean value is 130 seconds, standard deviation is 17,3 seconds, so its a uniform distribution too.

By combining the values of calls made and calls received, you can see that both types of values grow together. The next figure shows the mean of calls made duration (X) and the mean of calls received duration (Y):

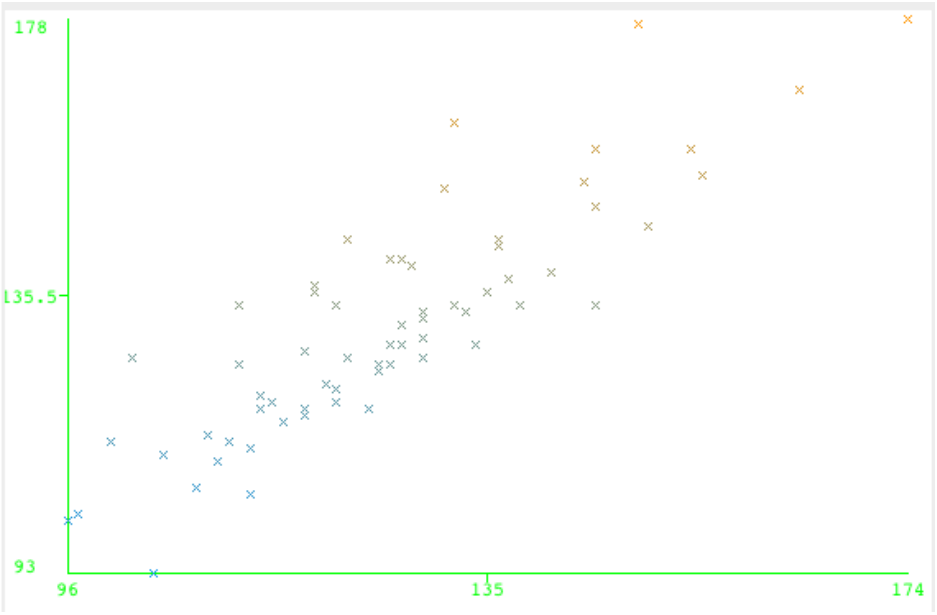


Figure 5: Mean duration of calls made and received

We think that this behavior is caused by all samples are in the same mobile operator. A sample with values in different mobile operators could show a less lineal graph, under the influence of various tariff plans.

4.4 Calls and Antennas Pattern

We have also analyzed the view of '*user_by_antenna_driven_data.arff*' to check what antennas a user has connected to and how many calls has established.

One user has completed 24.992 calls, far above the mean of the others users, which most likely mean that it's a public phone managed by a user. This pattern affects the analysis and so we have decided to remove it.

As can be seen from the graph in figure 6, it exists dependence between the number of connected antennas (Y)

and the number of calls made (X).

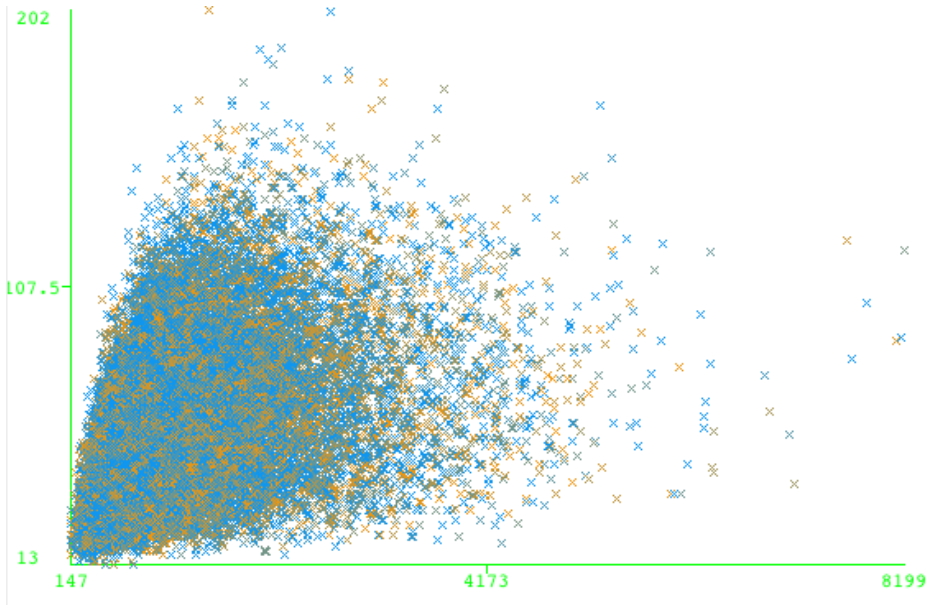


Figure 6: number of calls made and number of connected antennas

When the number of connected antennas grows then grows the number of calls made until a specific threshold. If the user has connected to 123 antennas then he could made more calls without connects to any other antenna. If the user has made 5.390 calls then he could connect to others antennas without complete any more calls.

According to this information, 2.922 users connected to 47 different antennas, 3 users connected to 13 antennas only, and 1 user connected to 202 different antennas. The mean of different antennas connected by a user is 53. This value is the *median*, because the standard deviation is too high, 22.

Could be interesting to define a tariff model appropriate for each type of user in base of the number of different connected antennas. For example, users who connect to more than 53 or 55 different antennas could reduce the cost of every calls because they usually made many more calls that the others users. Even could be defined a tariff plan by every mobility degrees, previously identified.

4.5 Time Pattern

In this section we relate calls and days, from the data:'*user_by_antenna_driven_data.arff*'.

The distribution of calls for each week day is:

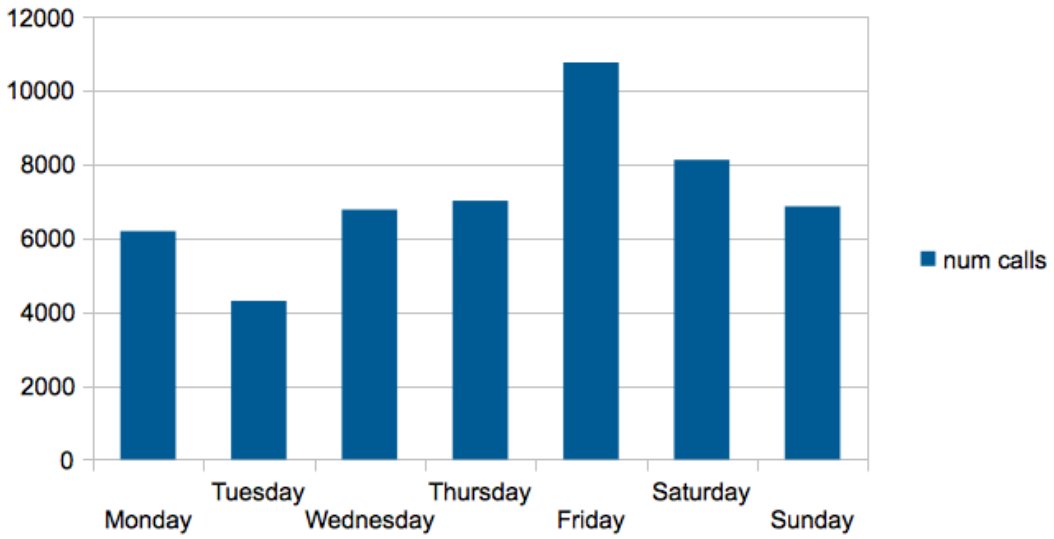


Figure 7: week day of calls

During the weekend are made more calls that others days, mainly in Friday are made more calls (21,52%) and in Tuesday (8,59%) is the lowest.

The distribution of calls in time interval is:

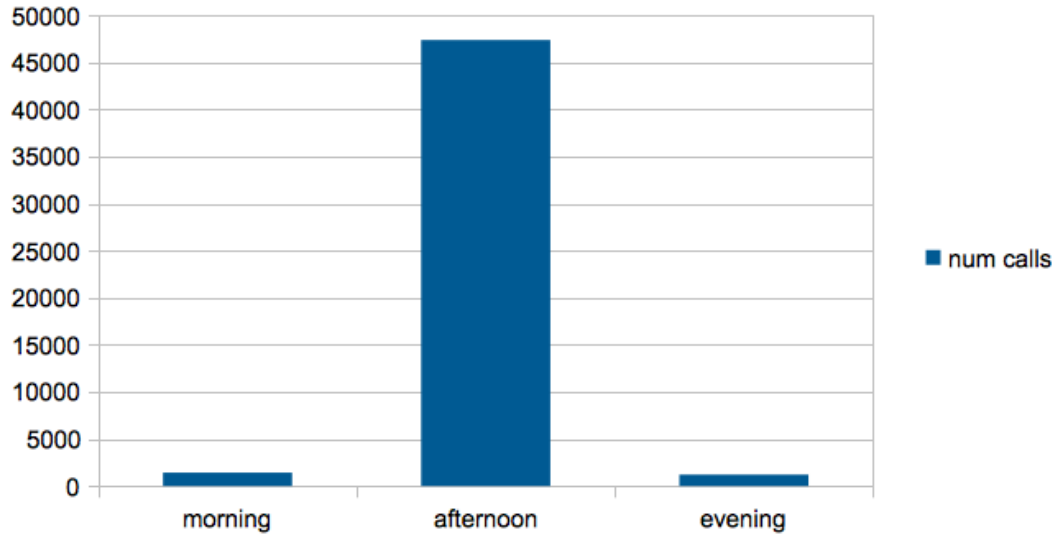


Figure 8: time intervals of calls made

During afternoon are mainly made calls, as expected. But we're going to analyze the calls in the morning and in

the evening.

In the morning, the calls are made mainly in Friday too. But Wednesday and Monday have more calls than Saturday. Tuesday is the lowest too:

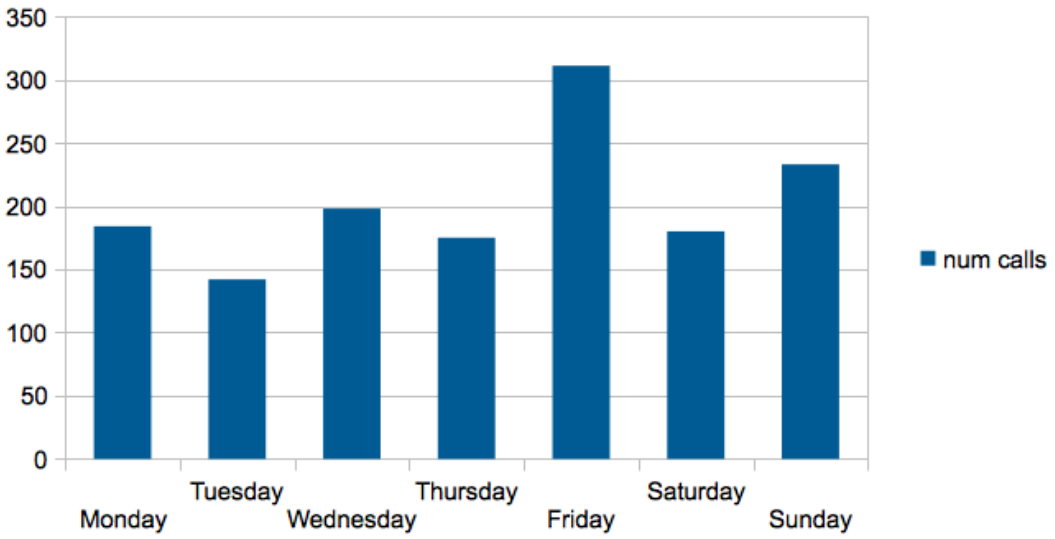


Figure 9: week day distribution for 'morning calls'

We group antennas with *morning profile* by administrative area, and calculate the rate of antennas having this profile of the total number of antennas in this area.

Área	Antennas with 'morning' profile	Number of Antennas	Rate %
Katiola	1	1	100
Tabou	8	9	88,89
Sinfra	5	7	71,43
Guiglo	7	10	70
Bocanda	2	3	66,67
Dabakala	2	3	66,67
San Pedro	20	30	66,67
Sassandra	9	14	64,29
Vavoua	7	11	63,64
Soubre	26	41	63,41
Bouake	16	26	61,54
Abengourou	14	23	60,87
Oume	9	15	60
Touba	4	7	57,14
Mankono	6	11	54,55
Gagnoa	15	28	53,57
Divo	8	15	53,33
Yamoussoukro	19	36	52,78
Bongouanou	10	19	52,63
Abidjan	198	396	50
Aboisso	22	44	50
Adiake	3	6	50
Agnibilekrou	4	8	50
Bangolo	1	2	50
Biankouma	2	4	50
Bouna	2	4	50
Bounfie	9	18	50
Daoukro	6	12	50
Grand Lahou	3	6	50
Man	7	14	50
Sakassou	2	4	50
Zuenoula	5	11	45,45
Daloa	13	29	44,83
Bondoukou	8	18	44,44
Tiassale	4	9	44,44
Issia	7	16	43,75
Korhogo	7	16	43,75
Adzope	9	21	42,86
Niakaramandougou	3	7	42,86
Tanda	5	13	38,46
Agboville	7	19	36,84
Danane	4	11	36,36
Dimbokro	2	6	33,33
Jaqueville	1	3	33,33

Department			
Toumodi	5	15	33,33
Boundiali	3	10	30
Odienne	3	10	30
M Bahiakro	2	7	28,57
Alepe	2	8	25
Seguela	3	12	25
Tiebissou Department	1	4	25
Beoumi	1	5	20
Toulepleu	1	5	20
Duekoue	2	11	18,18
Ferkessedougou	3	17	17,65
Dabou	3	18	16,67
Lakota	1	11	9,09

Table 16: Rate of antennas with 'morning calls' pattern

Now, we locate on the map the areas with more than 50% of antennas in *morning profile*:

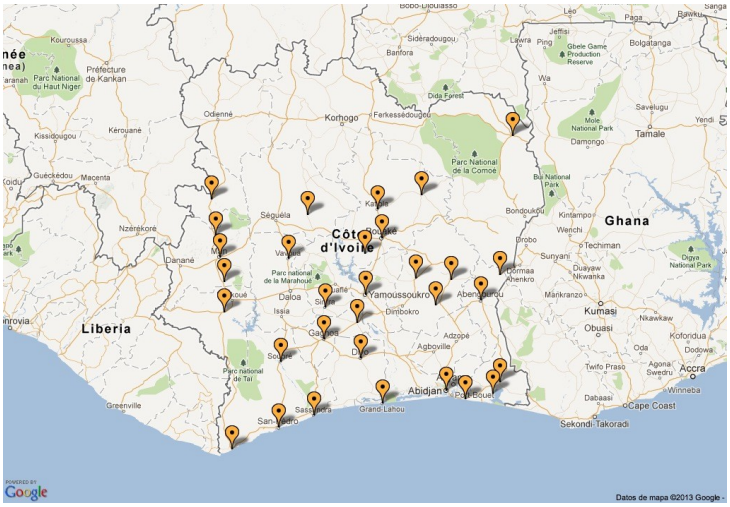


Figure 10: Areas with 'morning calls' pattern

On the previous map can be seen locations where made calls mainly between 4:00 a.m. and 11:59:59 a.m

Could be defined an adapted tariff plan with reduced cost for calls made in that time interval and then facilitate the communication in these areas.

Again, we're going to analyze users with 'evening calls' pattern. In this case, the week day with more calls made is the Sunday (218), and Wednesday (146) is the lowest.

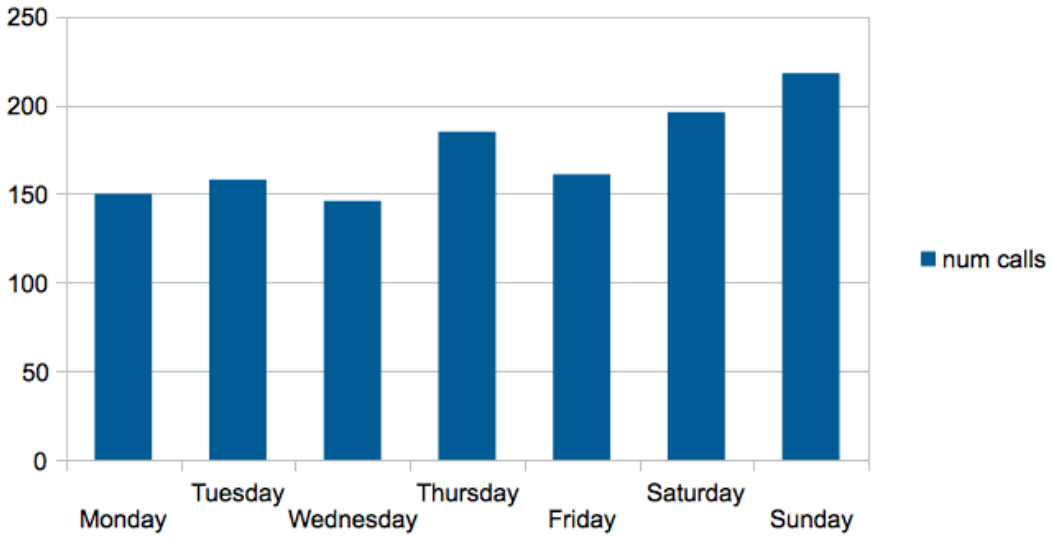


Figure 11: week day distribution for 'evening calls'

The next table shows the rate of antennas with 'evening calls' pattern for each administrative area:

Área	Antennas with 'evening' profile	Number of Antennas	Rate %
Jomoro	1	1	100
Grand Bassam	5	8	62,5
Beoumi	3	5	60
Bangolo	1	2	50
Abidjan	188	396	47,47
Bouake	12	26	46,15
Zuenoula	5	11	45,45
Yamoussoukro	16	36	44,44
Dabou	8	18	44,44
Tabou	4	9	44,44
Abengourou	9	23	39,13
Daloa	11	29	37,93
Korhogo	6	16	37,50
San Pedro	11	30	36,67
Gagnoa	10	28	35,71
Divo	5	15	33,33
Dimbokro	2	6	33,33
Bocanda	1	3	33,33
Aboisso	13	44	29,55
Man	4	14	28,57
Niakaramandougou	2	7	28,57
Sinfra	2	7	28,57
Bounfie	5	18	27,78
Duekoue	3	11	27,27
Soubre	11	41	26,83
Toumodi	4	15	26,67
Bongouanou	5	19	26,32
Daoukro	3	12	25
Bouna	1	4	25
Tiebissou Deparment	1	4	25
Agboville	4	19	21,05
Guiglo	2	10	20
Mankono	2	11	18,18
Danane	2	11	18,18
Ferkessedougou	3	17	17,65
Bondoukou	3	18	16,67
Grand Lahou	1	6	16,67
Tanda	2	13	15,38
Sassandra	2	14	14,29
Touba	1	7	14,29
Issia	2	16	12,50
Alepe	1	8	12,50
Agnibilekrou	1	8	12,50

Tiassale	1	9	11,11
Lakota	1	11	9,09
Seguela	1	12	8,33
Oume	1	15	6,67

Table 17: Rate of antennas with 'evening calls' pattern

Now, we locate on the map the areas with more than 50% of antennas in evening *profile*:

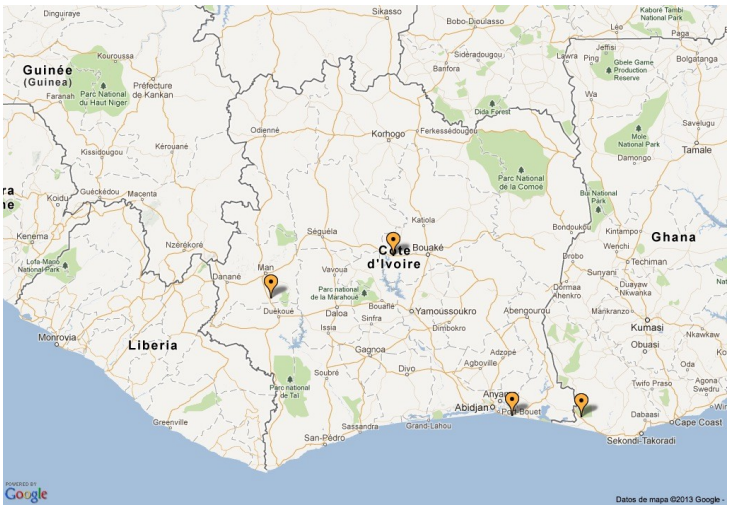


Figure 12: Areas with 'evening calls' pattern

On the previous map can be seen locations where made calls mainly between 8:00 p.m. and 3:59:59 a.m Could be defined an adapted tariff plan with reduced cost for calls made in that time interval and then facilitate the communication in these areas.

5 Conclusion

We should take advantage of these data to help and facilitate the quality of life of people who need it most and improve the service of the telecommunication companies to the population. Challenges like this, humanize technologies and allows to near the benefits got from them to the interest of people and enterprises.

We have manipulated the available information to obtain useful data to detect interesting patterns. We have developed the necessary software to obtain this useful data and we have interpret and evaluate the results obtaining some interesting conclusions and recommendations explicitly stated in the previous section. Apart from this, we can go further and say that we feel we could get more knowledge, more useful information that would produce changes and improvements in communications of Ivory Coast and so on in the daily lives of its citizens. With more available data and possible working with local partners (that could teach us some characteristics of the local population, infrastructure, etc.) we consider that we could be able to detect more patterns and achieve more interesting recommendations and conclusions. In particular, we believe that a good idea would be repeating or continuing this work in collaboration with any university of Ivory Coast and/or professionals from the telecommunication companies of this country, so we could get more effective results and we both could learn each other about technical and social related issues.

Now that we have designed a model for manipulating, studying and interpreting data, a proposed next step in this research could be building a system able to learn of events that could be found in the record of data, identify the corresponding pattern and be able to detect similar situations before they occur. So a system of preventing alarms could be built from our approach to get a complete system to fix communication errors.

6 APPENDIX A: Data Fields

Detailed information of fields of mining views:

FIELD	DESCRIPTION	TYPE
ant_id	Antenna identification	NUMERIC
antenna_relationship	Identify the set of antennas which user has connected	NUMERIC
Area	Admin area where antenna is located	{Abidjan..Yamoussoukro}
calls_type	Most common type of call	{ made, received }
day_longer_calls_made	Type of day with longer duration of calls made	{ holiday, working}
day_longer_calls_recv	Type of day with longer duration of calls received	{ holiday, working}
day_of_calls_made	Most common type of day for calls made	{ holiday, working}
day_of_calls_recv	Most common type of day for calls received	{ holiday, working}
dur_calls_made	Duration of calls received	NUMERIC
dur_calls_made_afternoon	Duration of calls made in the afternoon	NUMERIC
dur_calls_made_evening	Duration of calls made in the evening	NUMERIC
dur_calls_made_friday	Duration of calls made on Friday	NUMERIC
dur_calls_made_holiday	Duration of calls made on holidays	NUMERIC
dur_calls_made_monday	Duration of calls made on Monday	NUMERIC
dur_calls_made_morning	Duration of calls made in the morning	NUMERIC
dur_calls_made_saturday	Duration of calls made on Saturday	NUMERIC
dur_calls_made_sunday	Duration of calls made on Sunday	NUMERIC
dur_calls_made_thrusday	Duration of calls made on Thrusday	NUMERIC
dur_calls_made_tuesday	Duration of calls made on Tuesday	NUMERIC
dur_calls_made_wednesday	Duration of calls made on Wednesday	NUMERIC
dur_calls_made_working	Duration of calls made on Working Day	NUMERIC
dur_calls_recv	Duration of calls received	NUMERIC
dur_calls_recv_afternoon	Duration of calls received in the afternoon	NUMERIC
dur_calls_recv_day	Type of day with longer duration of calls received	NUMERIC
dur_calls_recv_evening	Duration of calls received in the evening	NUMERIC
dur_calls_recv_friday	Duration of calls received on Friday	NUMERIC
dur_calls_recv_holidays	Duration of calls received on holidays	NUMERIC
dur_calls_recv_monday	Duration of calls received on Monday	NUMERIC
dur_calls_recv_morning	Duration of calls received in the morning	NUMERIC
dur_calls_recv_saturday	Duration of calls received on Saturday	NUMERIC
dur_calls_recv_sunday	Duration of calls received on Sunday	NUMERIC
dur_calls_recv_thrusday	Duration of calls received on Thrusday	NUMERIC
dur_calls_recv_tuesday	Duration of calls received on Tuesday	NUMERIC
dur_calls_recv_wednesday	Duration of calls received on Wednesday	NUMERIC
dur_calls_recv_working	Duration of calls received on Working days	NUMERIC
dur_calls_type	Type of calls with longer duration	{ made, received}

FIELD	DESCRIPTION	TYPE
mean_calls_made	Average calls made per day	NUMERIC
mean_calls_recv	Average calls received per day	NUMERIC
mean_dur_calls_made	Average duration of calls made per day	NUMERIC
mean_dur_calls_recv	Average duration of calls received per day	NUMERIC
most_called_antenna	Anntena wich has been made more calls	NUMERIC
most_recv_antenna	Anntena wich has been received more calls	NUMERIC
most_used_antenna	Most common antenna used for user calls	NUMERIC
most_used_subpref	Most common subpref used for user calls	NUMERIC
multi_caller	Call made to more than one different antena	{ true, false}
multi_receiver	Call received from more than one different antena	{ true, false}
multi_subprefs	More than one subpref which user has connected	{ true , false }
num_antennas	Number of antennas which user has connected	NUMERIC
num_antennas_called	Number of antennas to have call made	NUMERIC
num_antennas_recv	Number of antennas to have call received	NUMERIC
num_calls	Total number of calls made and received	NUMERIC
num_calls_by_antenna	Number os calls for each antenna	NUMERIC
num_calls_made	Total Number of calls made	NUMERIC
num_calls_made_afternoon	Number of calls made in the afternoon	NUMERIC
num_calls_made_evening	Number of calls made in the evening	NUMERIC
num_calls_made_friday	Number of calls made on friday	NUMERIC
num_calls_made_holiday	Number of calls made on holidays	NUMERIC
num_calls_made_monday	Number of calls made on monday	NUMERIC
num_calls_made_morning	Number of calls made in the morning	NUMERIC
num_calls_made_saturday	Number of calls made on saturday	NUMERIC
num_calls_made_sunday	Number of calls made on sunday	NUMERIC
num_calls_made_thrusday	Number of calls made on thrusday	NUMERIC
num_calls_made_tuesday	Number of calls made on tuesday	NUMERIC
num_calls_made_wednesday	Number of calls made on wednesday	NUMERIC
num_calls_made_working	Number of calls made on working days	NUMERIC
num_calls_recv	Total Number of calls received	NUMERIC
num_calls_recv_afternoon	Number of calls received in the afternoon	NUMERIC
num_calls_recv_evening	Number of calls received in the evening	NUMERIC
num_calls_recv_friday	Number of calls received on Friday	NUMERIC
num_calls_recv_holiday	Number of calls received on Holiday	NUMERIC
num_calls_recv_monday	Number of calls received on Monday	NUMERIC
num_calls_recv_morning	Number of calls received in the morning	NUMERIC
num_calls_recv_saturday	Number of calls received on Saturday	NUMERIC
num_calls_recv_sunday	Number of calls received on Sunday	NUMERIC
num_calls_recv_thrusday	Number of calls received on Thrusday	NUMERIC
num_calls_recv_tuesday	Number of calls received on Tuesday	NUMERIC
num_calls_recv_wednesday	Number of calls received on Wednesday	NUMERIC

FIELD	DESCRIPTION	TYPE
num_calls_recv_working	Number of calls received on Working day	NUMERIC
num_subprefs	Number of subprefs which user has connected	NUMERIC
subpref_relationship	Identify the set of subprefectures which user has connected	NUMERIC
time_longer_calls_made	Most common time interval for longer duration of calls made	{ morning, afternoon, evening}
time_longer_calls_recv	Most common time interval for longer duration of calls received	{ morning, afternoon, evening}
time_of_calls_made	Most common time interval for calls made	{ morning, afternoon ,evening}
time_of_calls_recv	Most common time interval for calls received	{ morning, afternoon ,evening}
user_id	User identification	NUMERIC
weekday_longer_calls_made	Day of week with longer duration of calls made	{ monday...sunday}
weekday_longer_calls_recv	Most common day of week for longer duration of calls received	{ monday...sunday}
weekday_of_calls_made	Most common day of week for calls made	{ monday .. sunday}
weekday_of_calls_recv	Most common day of week for calls received	{ monday .. sunday}

First steps for a Synthetic Population of Ivory Coast

Andrea Apolloni[1], Anton Camacho[1], Ken Eames[1], John W. Edmunds[1], Sebastian Funk[1]

[1] *London School of Hygiene and Tropical Medicine, London, UK*

(February 15, 2013)

Keywords: D4D challenge, Synthetic population, Mobil phone data

Abstract

We used cell phone data to build up a synthetic population for major areas in Ivory Coast. In this framework individuals are endowed with demographic characteristics according to census data and a mobility pattern drawn from mobile phone data. In this article we show some of the analysis conducted, the procedure we applied and some of the results obtained.

The work conducted till now represents the first steps of a larger project of building realistic contact networks in developing countries that can be used for studying the diffusion of infections.

1. Introduction

The advent of new technologies as well as the increase in computational capabilities has increased the number of micro simulation models aiming to reproduce human behaviors.These models, called *Synthetic Populations*, have been applied in different fields, among them: transportation [6] [4], health [10] [15], water and power demand [18], land use [14]. Given some input information, these models produce a range of possible scenarios and for this reason are suitable for informing policy makers.

The philosophy of Synthetic Population is quite straightforward. Given whatever population, with a specific distribution of demographic characteristics, the actors are endowed with individual traits and activities to accomplish satisfying certain constraints [13] [16]. As a consequence, individual activities patterns drive the evolution of the contact network, in a way not predictable by theoretical models [10].

Although fascinating, this type of models require a huge amount of data in order to provide realistic scenarios. In some cases micro-data can be collected through census bureau or ad-hoc surveys. However, the collection of individual micro-data is not desirable or not possible in some countries [16].

In the recent years, researchers have found in cell phone data a gold mine of possibilities for uncovering mechanisms ruling human mobility [8] [17] [12] [9]. In order for a mobile phone to place outgoing calls and to receive incoming calls, it must periodically report its presence to nearby cell towers, thus registering its position in the geographical cell covered by one of the towers ("on-field" data). In this way, anonymized cell phone data, provides some information of the users' whereabouts during a normative day. Studying these datasets, researchers have found that human trajectories show a high degree of temporal and spatial regularity, in contrast with Levy flights and random walk models [8] [12] [9]*. Moreover cell phone data can be used to study mobility in low density rural areas collecting data otherwise not collectable [17].

In this article we present part of the work done in building a synthetic population using census and cell phone data. The work is still ongoing and aims to recreate a realistic contact and mobility network, in the spirit of [10], that can be used for epidemiological and prevention purposes. Although the data we are using to inform the model, are temporally and spatially detailed, the lack of users' demographic information has compelled to create ad-hoc methods for associating activity pattern to individuals. In order to develop a realistic synthetic population, information about purposes of movement and facilities in the antenna area should be included: although individuals are using the same antenna they are not in necessarily in close enough proximity to pass on an infection between each other.

2. Description of the Dataset

In order to re-create individual mobility pattern in Ivory Coast we used anonymized cell phone data made available by France Telecome/Orange Côte d'Ivoire in the context of the *D 4 D challenge* [2]. The dataset contains information

*this type of regularities have been previously found in the context of Synthetic Populations [10], where individual movements are dictated by activities routine

about 50×10^3 customers cell phone use for a time period of 21 weeks, from December 2011 till April 2012. Data have been given in 4 different datasets. In order to build up the synthetic population we exploited the dataset number 2 that contains Call Detail Records (CDR) for a period of 2 weeks [7]: in this period every time a customer is making/receiving a call ora SMS, the time and the position of the used antenna is recorded. The data represents almost 70×10^6 individual trajectories. In the following we use the expression "making a call" for indicating any activity of the user that has been collected. The France Telecome/Orange Côte d'Ivoire network consists of 1328 antenna and the company has provided the geographical location of the antenna.

3. Preprocessing

We performed a Voronoi tessellation, which partitions the space into cells based on the distance between each point and the closest antenna, in order to geo-locate customers during the day. In this way individual making calls using the same antenna can be located in the same tassel. As we can notice in fig.(1), on the left, the largest part of the antenna (95%) covers a small area (less than 10 Km^2). We included data from CIESN [1] about the population located around the antenna. CIESIN data are estimates using satellite of the population at specific locations. Earth is divided in a grid where each unit has a resolution of 15 min $\times$ 15 min of arc, and for each cell the number of individuals is estimated. Cells are then associated to each antenna and population in the antenna is evaluated. In case a cell belongs to different antennas, its population is equally partitioned among them. As can be easily seen in fig.(1), on the right, the average usage of antenna is less than 40×10^3 inhabitants, with just two antenna with more than 3 millions inhabitants served. These information give an estimate of the average usage of specific cell phone areas and will be exploited to assess the household distribution in the antenna area.

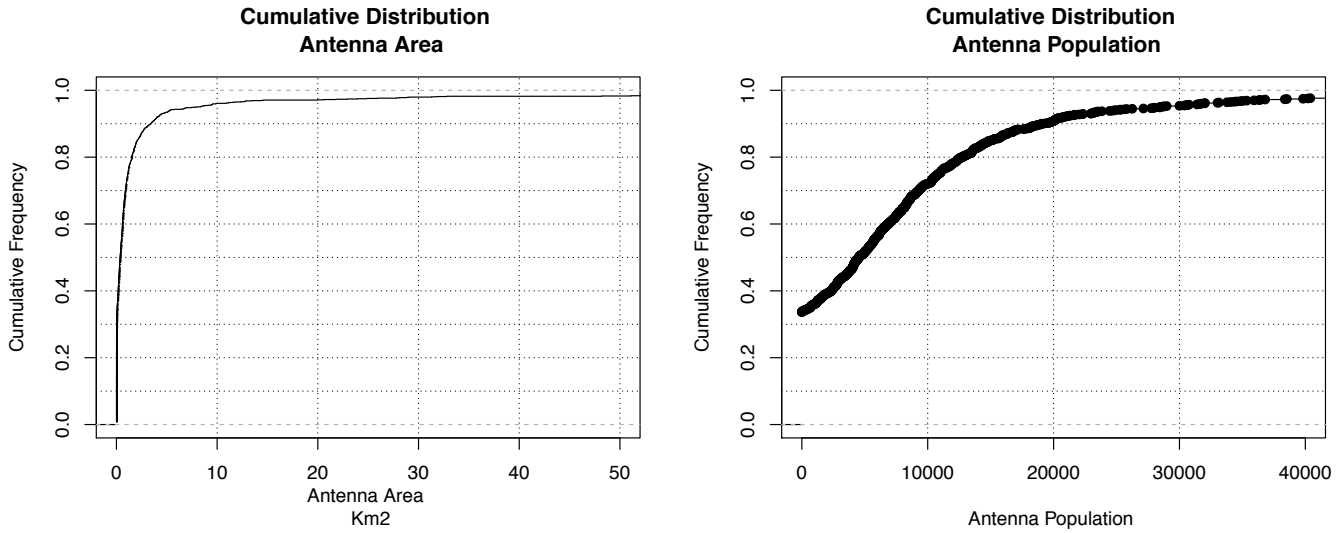


Figure 1: On the left the cumulative distribution of the Voronoi tessellation area. On the right the cumulative distribution of the average antenna usage in terms of population estimates in the antenna area

Before proceeding to the analysis of the dataset and the construction of the synthetic population, we apply some smoothing to the data. In the first step we aim to evaluate the actual position of the antenna where the call has occurred. Following Blondel [9], we assume that the user is making calls at time $t(1),\ldots,t(n)$ and the coordinate of the antenna serving the calls are respectively $x(1),\ldots,x(n)$. The smoothed positions $y(1),\ldots,y(n)$ are then evaluated as a weighted average:

$$y(i) = \sum_{j\in B^\delta(i)} w(j)x(j) \tag{1}$$

where $B^\delta(i)$ denotes the indices of antennas where a call has been occurred in a maximum time interval of $\delta = 60$ minutes from the current time. Further distant positions are then weighted using w:

$$w(j) = 1 - \frac{|t(i) - t(j)|}{\delta}. \tag{2}$$

Since the dataset records the time and the position a call occur, not the *on field* position at each time unit, we consider that the customer moves between two different calls on a straight line. This is the only possible approximation

we can apply since we don't have any information about the purpose of the movement and the transportation mean used. This approximation allows to estimate which antenna area the customer is crossing and how much time he /she is spending in the antenna area, fig(2).

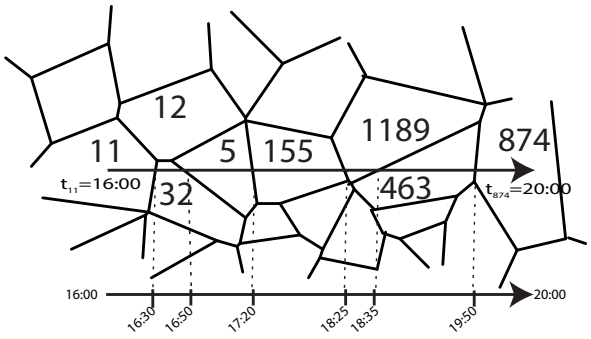


Figure 2: Individuals are supposed to move straight between two calls. The time they cross border between two antenna area is used to evaluate the permanence time in the specific area

4. Some results on Antenna Occupancy

After the data smoothing, we have evaluated the hourly occupancy of each antenna. In order to check the existence of possible patterns we have considered for each day of the week the number of individuals using a specific antenna at a certain hour of the day. Figure (3) shows the hourly occupancy of the antenna where the solid line corresponds to the average value over all the weeks of the datasets. Compared to previous works on this topic [8] [12] [9], our results are more noisy but in any case show the existence of a pattern that is repeating almost identically every day of the week: few individuals are in the antenna during the early morning, then around noon there is a peak of occupancy that is gradually fading.

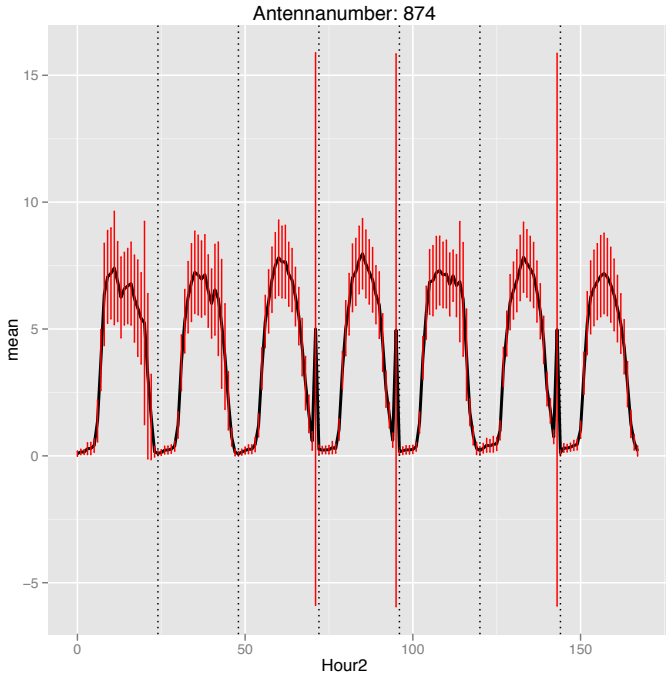


Figure 3: Fraction of customers in a randomly selected antenna during a week time

In order to capture similarities among antenna usage and individual activities pattern we perform k-mean cluster analysis [9] for different values of $k = (3, \ldots, 10)$. To each antenna we have associated a vector with 168 entries representing the occupancy at each hour of the week. The results shown here is for $k = 3$ although not appreciable differences from the other cases. As we can see in fig (4), on the left, there isn't any geographical cluster among antennas with the exception of the North-West area, a majority rural area.

Moreover, if we analyze the occupancy of the antenna of the different clusters , fig (4) on the right, we notice that the only difference among clusters is the number of occupancy, but there isn't any shift in peak time or duration that can be connected to a different use of that antenna area as seen in Blondel [9]. The absence of geographical clustering as well as of a specific occupancy pattern represent a major obstacle in inferring some information on individual activity pattern: a shift in the occupancy time between antenna of different clusters would have corresponded to a specific use of that seta antenna, thus distinguishing among antennas serving residential, business or shopping areas.

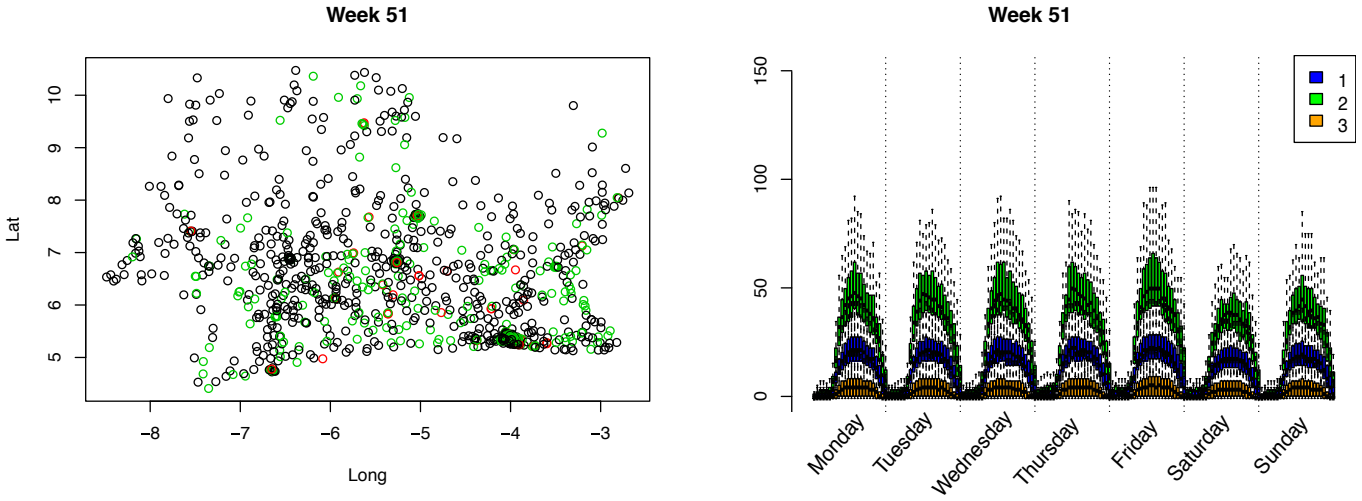


Figure 4: On the left Geographical distribution of antenna clusters. On the right the box plot of the hourly occupancy. The different colors correspond to different clusters

In the absence of this information we refers to individual data to extract some information about human mobility. We have considered for each pattern the number of different antennas an individual has spent some times in. As clearly shown in fig(5) most of individuals visit few antennas: 45% just one antenna; 50% between 2 and 5 and few more than 5. The distribution is pretty stable except on Saturday, when the fraction of people staying in one location is higher and less people are traveling. This is in accordance with previous studies showing that individuals tend to vista few places during a normative day.

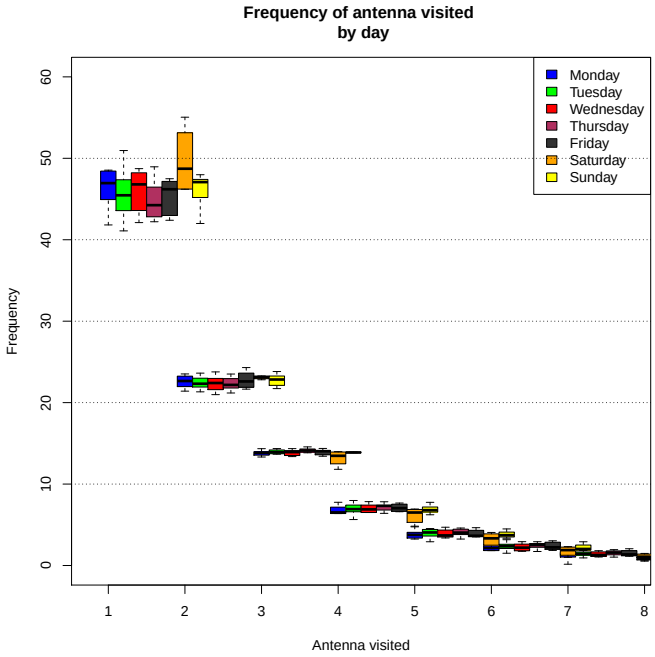


Figure 5: Percentage of individuals whose patterns involve a specific number of different antennas. Different colors correspond to different day of the week.

Since we haven't information about users home location, we make the assumptions location for houses, the antenna area here the first call is made by a user. This is quite a reasonable assumptions if we consider the large fraction of individuals not leaving the area and thus staying at home. Under this assumption the returning probability, evaluated

as the fraction of path with more than one antenna that start and end at the same house location in the same day, is almost 66% and increases to 73% if we consider patterns of users whose last call of the day is made at a different location than house, but the first of call of the day after is made at the house location.

5. Synthetic Population Construction

Using the definition for houses as above we then build up the synthetic population for a specific set of antennas n the metropolitan area of the capital city. The population belonging to this subset is about 85000 individuals.

The algorithm, consists of two parts: a first part where a population of synthetic households is built up; a second part where to each individual is associated a mobility pattern, meant as a list of antenna visited and time spent in each one. The algorithm is repeated for each antenna we are interested.

5.1. synthetic household construction

The first point consists in recreating a population of households whose members have demographic characteristics with respect to the data provided by our sources. For this purpose we use the estimate from U.N. [5] for the age and gender distribution of Ivory Coast population $P(a, g)$, and survey data from D.H.S. [3] regarding family size distribution $P(s)$. Unfortunately neither the census nor the survey contain information about household type distribution (i.e. Single parent, Couple of parents, Composite) and the household head age distribution[†]. However in the construction of the population we used other information provided such as: the gender of the household head $P(g)$, the marital status by age and gender $P(m)$ (Married, Single, Widow) and the age difference between partners $P(d)$. The procedure we followed is similar to the one used in [11] where household is build progressively by picking members accordion to the relation with the household head. A simplified version of the algorithm is reproduced:

1. Generate a list of individuals I whose size is equivalent to the population of the area.
2. Associate demographic traits, such as gender and age, to each individual with probability extracted from census data $P(a, g)$.
3. Create a list of possible head of the household H, corresponding to all the individuals older than 14 years.
4. Associate to each household head his own marital status according to $P(m)$.
5. Pick the size of the household according to $P(s)$.
6. Pick the gender of the household head according to $P(g)$.
7. If $s = 1$ the household head is either single or widow. The marital status is then picked according to $P(m)$ and the record of the household head is eliminate from H and I.
8. If $s > 1$ the marital status of the household head is chosen according to $P(m)$.
   - If $s = 2$ and the marital status is couple the partner of the head is chosen with probability given by $P(d)$ among individual of opposite sex. The record of the partner is then eliminated from H and I.
   - if $s > 2$ in the case of the couple of parents , the partner is chosen according to $P(d)$ and records eliminated, whilst other members are chosen based on $P(a, g)$ and their records eliminated from the lists they belong to.

The process is repeated as long as the lists are empty or a pre-fixed number of iteration is reached. If at any step, there is no individual in I satisfying the property, the members are put back in the list and a new attempt to build the household is launched. The algorithm aims to create the population of the household with the same size distribution, associating to children at least one adult in a family and respecting the overall distribution of relation between adults. However due to the lack of knowledge about type distribution we can not verify the composition of the household.

We have compared the synthetic population with the original data from the Census [5] and household survey[3]. As we can easily notice in the fig (6), there is a very good agreement between the original data and the ones produced by the previous algorithm.

[†]Data about household composition have been requested

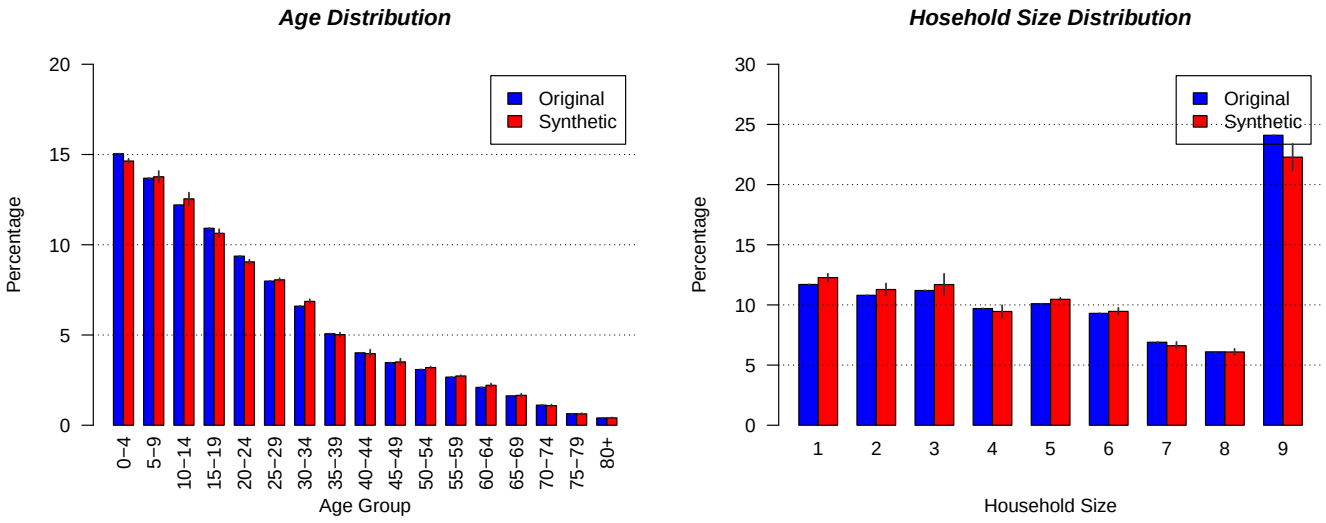


Figure 6: Comparison between the age (on the left) and the family(on the right) distribution for the data extracted from Census [5] and DHS survey [3] (blue) and the synthetic ones (red). Synthetic estimates are evaluated as the average over all the antennas of subset under examination

5.2. SYNTHETIC MOBILITY PATTERN CONSTRUCTION

Before the association to each individual of an activity pattern , we have estimated some quantities that describe the ensemble of paths starting at a specific house location:

- The length distribution $P(l)$ that is the fraction of paths starting from the house location that cross l antennas before coming back to home location. In this case we consider all the antennas, also if an individual is crossing the same antenna multiple times.
- For each length l we then evaluate the probability of a specific path p $P(l|p)$.
- At the home location we evaluate the fraction of individuals leaving the area at a certain hour τ, $P(\tau)$.
- For each of the antenna involved in at least one of the path we evaluate the rate of leaving after a certain time t' once arrived at time τ': $P(\tau'|t')$.

Having evaluated these quantities the paths are associated according to the following algorithm:

1. For each household we check if there are children less than 14 years old.
2. In the positive case to each child is associated a path of length 1, that is the children is not going out of the household location. Moreover one of the adults of the household, older than 14 years, is randomly selected and a path of length 1 is imposed (custody adult).
3. If there aren't children, to each member of the household is associated a path of length l according to $P(l)$.
4. For $l > 1$ a path p is associated to each individual according to $P(l|p)$.
5. For children, custody adults and adults with path of length $l = 1$ the time spent in the household location is equal to 24 hours.
6. For the other adults we associate a departing hour according to $P(\tau)$.
7. The arrival time in the m-th location is evaluated as $\tau'_m = t'_{m-1} + 1$ minute.
8. The departing time from the m-th antenna is then evaluated according to $P(\tau'_m|t')$.
9. In the case of l-th antenna the individual is supposed to stay there till the end of the day.

The algorithm aims to reproduce the occupancy pattern in each of the antenna where paths starting from a specific home location are passing through. As a first step we check if the algorithm produced sets of patterns comparable to the original one. In figure (7) we show the relative difference (in percentage) between the generated paths and the original ones in terms of their lengths (numbers of antenna crossed during the day). For a specific value of the length l, a positive value indicates that the algorithm has generated fewer paths with respect to the original ones. A negative value corresponds to the opposite situation. As we notice for paths of small length, where the statistic is higher, the fraction of generated paths is comparable to the original ones.

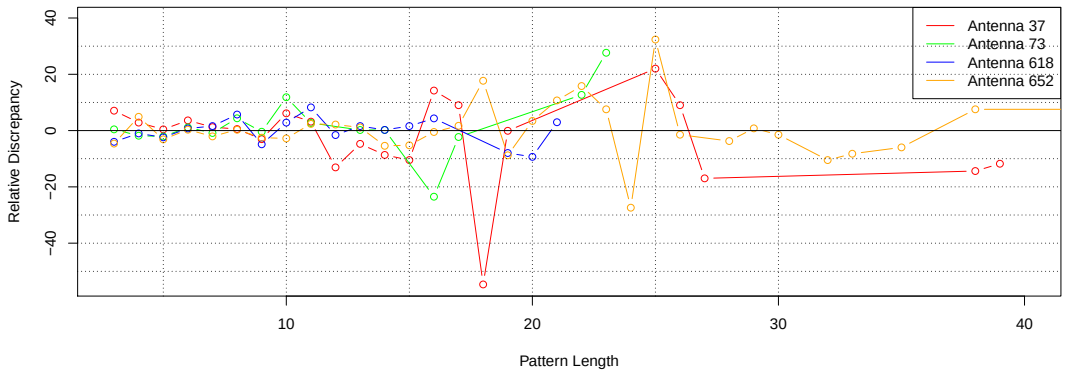


Figure 7: Relative discrepancy between generated and original path length distribution. We consider only path outside the household place. Each colors correspond to a different home location. A positive value indicates that the algorithm has generated less paths of a specific length in comparison to the original data, viceversa for the negative

Figure (8) shows the case of two antennas that don't correspond to home location. We evaluated the number of individuals present in each antenna at a certain hour, both in the synthetic population and the original case. In order to make comparison, due to the differences in population sizes, we have normalized the hourly occupancy to the total number of individuals passing through the antenna in a day. For the original case we have considered an average over all the day of the sample. Although there are some discrepancies in most of the cases the original and the synthetic distribution are comparable. A possible cause for the discrepancies can be traced back in the way origin data have been averaged. We are working in implementing the algorithm to improve the resemblance between the original data and the Synthetic Population.

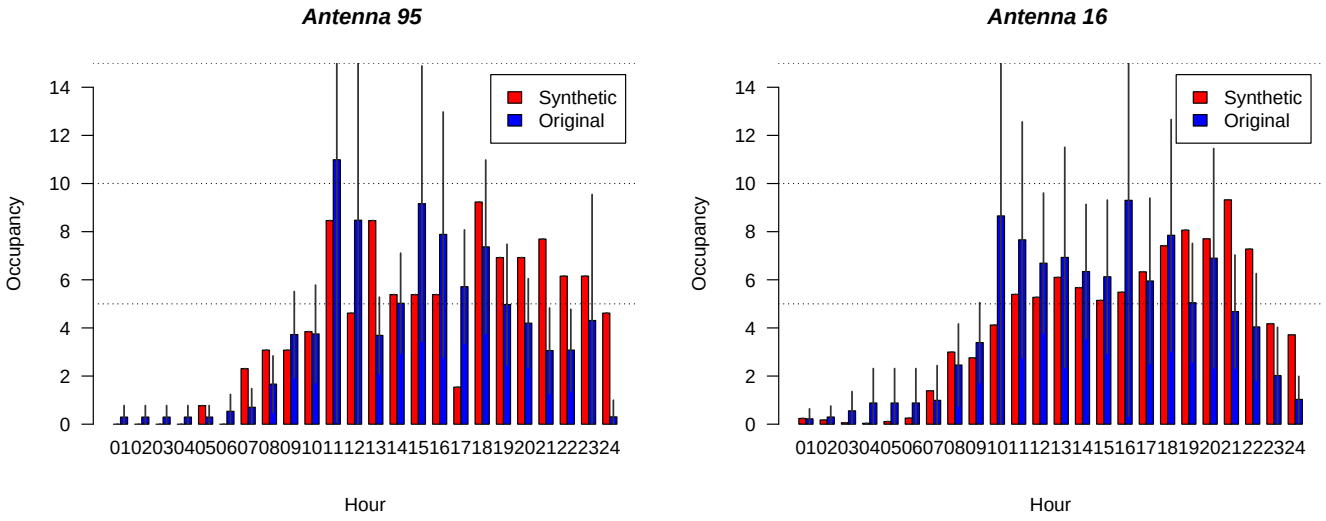


Figure 8: Comparison of the hourly occupancy of two antenna in a normative day. In blu we have indicated the average occupancy as extracted fro the smoothed data. In red we have considered the same quantity as from the synthetic population. In order to make comparison, due to the different population sizes we have normalized to the total number of individuals passing through the antenna in a day.

6. Conclusions and Future works

In this article we have shown some of the method and results obtained in developing a synthetic population for a specific area of Ivory Coast. The method strongly relies on the use of the data provided of Orange as a proxy for human mobility. However due to the lack of some critical information about household characteristic, user demographic data and land use the model developed is quite naive compared to more elaborated ones.

Several assumptions have been done to accompany the lack of users' demographical data. First we have considered adults any individual older than 14 years. This is related to the fact that there is a considerable fraction of individuals married at that age and possible household heads.

We have imposed that children spend all their times at home under the surveillance of an adult, thus reducing the number of adults moving. However this is not true since in most of the African countries, school age children in rural areas spend long period of time in schools far from home.

We aim to inform the model with information about age-use of cell phone as well as, school attendance statistics to improve the description children movement. The use of regression techniques applied to demographic status of the user could shed light over mechanisms that are fundamental for understanding human mobility and provide a more realistic representation.

At this moment we are able to reproduce the co-presence of individuals in each antenna area. However, in order to build the contact network, information about premises and facilities in the area should be included: only individuals sharing the same office/classroom/house can establish contacts. Further work will entail calibrating the model to epidemiological data to assess the spread of infections and how to control them.

7. Acknowledgement

Studies and Researches were performed using mobile communication data made available by France Telecom and Orange Côte dIvoire within the D4D Challenge. This work has been funded by Epiwork (EUFP7, Grant Agreement number 231807)

References

[1] Center for international earth science information network (ciesin), columbia university http://sedac.ciesin.columbia.edu//.

[2] D4d challenge: Data for development http://www.d4d.orange.com/.

[3] Measure dhs, demography and health surveys http://www.measuredhs.com/.

[4] Transims: An open source transportation modeling and simulation toolbox http://code.google.com/p/transims.

[5] United nations statistics division http://unstats.un.org/unsd/demographic/sources/census/.

[6] R J Beckman and K A Baggerly. Creating Synthetic baseline populations. *Transportation Research Part A: Policy and Practice*, 30(6):415–429, February 1996.

[7] Vincent D Blondel, Markus Esch, Connie Chan, Fabrice Clerot, Pierre Deville, Etienne Huens, Frédéric Morlot, Zbigniew Smoreda, and Cezary Ziemlicki. Data for Development: the D4D Challenge on Mobile Phone Data. *arXiv.org*, cs.CY, September 2012.

[8] J Candia, M C González, P Wang, T Schoenharl, G Madey, and A-L Barabási. Uncovering individual and collective human dynamics from mobile phone records. *arXiv.org*, physics.soc-ph, October 2007.

[9] Balázs Cs Csáji, Arnaud Browet, V A Traag, Jean-Charles Delvenne, Etienne Huens, Paul Van Dooren, Zbigniew Smoreda, and Vincent D Blondel. Exploring the Mobility of Mobile Phone Users. *arXiv.org*, physics.soc-ph, November 2012.

[10] Stephen Eubank, Hasan Guclu, V S Anil Kumar, Madhav V Marathe, Aravind Srinivasan, Zoltán Toroczkai, and Nan Wang. Modelling disease outbreaks in realistic urban social networks. *Nature*, 429(6988):180–184, May 2004.

[11] Floriana Gargiulo, Sonia Ternes, Sylvie Huet, and Guillaume Deffuant. An iterative approach for generating statistically realistic populations of households. *arXiv.org*, cs.MA, December 2009.

[12] Marta C González, César A Hidalgo, and Albert-László Barabási. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, June 2008.

[13] Alison Heppenstall Dianna Smith Kirk Harland and Mark Birkin. Creating Realistic Synthetic Populations at Varying Spatial Scales: A Comparative Critique of Population Synthesis Techniques. *2010:61:3*, pages 1–24, January 2012.

[14] Maxime Lenormand, Sylvie Huet, and Floriana Gargiulo. Generating French virtual commuting network at municipality level. *arXiv.org*, math.ST, September 2011.

[15] S Merler and M Ajelli. The role of population heterogeneity and human mobility in the spread of pandemic influenza. *Proceedings of the Royal Society B: Biological Sciences*, 277(1681):557–565, January 2010.

[16] Rolf Moeckel, Michael Wegener, and Klaus Spiekermann. Creating a synthetic population.

[17] Andrew J Tatem, Youliang Qiu, David L Smith, Oliver Sabot, Abdullah S Ali, and Bruno Moonen. The use of mobile phone data for the estimation of the travel patterns and imported Plasmodium falciparum rates among Zanzibar residents. *Malaria journal*, 8:287–287, 2009.

[18] P Williamson, M Birkin, and P H Rees. The estimation of population microdata by using data from small area statistics and samples of anonymised records. *Environment and Planning A*, 30(5):785–816, 1998.

Place Identification and Prediction in the D4D Data Set using Machine Learning

Negar Ghourchian and Doina Precup
McGill University
Montreal, Canada
{negar, dprecup}@cs.mcgill.ca

Abstract

We present an approach for place identification from cell phone call records using machine learning methods. We propose to use non-parametric Bayesian methods, specifically Hierarchical Dirichlet Processes, to cluster calls and antennas at different times of the day. The method adapts the number of clusters automatically to the data. Visualizing the resulting clusters over time gives insight on the activity patterns around the Ivory Coast, which could inform development planners.

1 Introduction

Cell phone data analysis is attracting a significant amount of attention, due to the information that phone usage patterns gives about the social behavior of the users. In this paper, we report on our work conducted as an application for "The Data for Development (D4D) challenge on mobile phone data" [Blondel *et al.*, 2012]. The goal of the project was to use Orange D4D data to find interesting locations, which are visited according to distinctive patterns, and (to the extent possible) model individual movement. These tasks are important for urban and community planning, as well as for assessing the state of the community and allocating resources (such as transportation, security, community services), as well as in safety circumstance (e.g. for issuing traffic warnings).

For the work presented in this paper, we used the first data set of D4D, which consist of the number of calls, as well as the duration of calls between any pairs of Orange cell towers in the Ivory Coast. Our goal was to discover correlations in the call traffic based on both geographic location of the towers and the time of the calls. We used a non-parametric Bayesian framework [Orbanz and Teh, 2010] to compute the probability distribution of call traffic between towers, dependent on the time, and to identify "important" locations using the entropy of the distribution. Intuitively, the locations of interest should depend greatly on time. For instance, users are more likely to be at home during the night and at work during the day; therefore, the corresponding detected locations are probably residential areas (suburb) and workplaces, respectively. In Section 2, we described the preprocessing of the data, during which we extracted temporal features such as weekday and weekend from the initial data. Next, in section 3 we describe the proposed clustering approach to identify the important places based on the call traffics of each antenna tower for each time slot. Section 4 contains empirical results, and in Section 5 we conclude and discuss ongoing work.

2 Data Preprocessing

The original data set contains 3600 hours of antenna-to-antenna cell phone call traffic records for 150 days in Ivory Coast. Antenna towers are uniquely identified by an antenna ID and a geographic location. As described in the D4D contest, for technical reasons, some antenna identifiers are not always available. The corresponding communications which were assigned to the code -1 have been removed from our work due to their misleading effects.

We assume that traffic is generated by different patterns during the day. The initial observations have been arranged hour by hour, so hourly important locations changes can be directly monitored; clustering is performed on hourly data. We observed that almost every day the number of calls decreases dramatically from midnight until 5:00 in the morning so we call this duration "night" time and label the rest of the day, e.g. from 6:00 to 23:00 as "day" time. We also grouped and indexed the data by the day of the week. This allows us to observe (potentially) weekly changes of important locations; also, later we will be able to evaluate fairly the performance of each day's clustering on a test set from the same day of the week. This is necessary, as weekday and weekend patterns are also different.

3 Hierarchical clustering for mining call patterns

In this section we present the probabilistic clustering algorithm that we used for finding the location of interests based on traffic call over each antenna. The approach we present is similar to that of [Frank *et al.*, 2012], but they focused on mapping Wifi positions of individual phones, rather than antennae. As described before, each instance contains information about the originating and terminating antenna IDs, as well as the total number of calls between these towers. We aim to compute the probability distribution of calls over each antenna to be able to highlight important clusters or locations for each time step. However, the number of clusters is not

known a priori and tends to grow as more data points are seen. Further, each antenna is been paired with various other antennas at each time step, which means that the clusters we compute have to share components.

Due to these requirements, we adopted Hierarchical Dirichlet Processes (HDP) [Teh *et al.*, 2006] which is a non-parametric Bayesian model for clustering multiple groups of data. HDPs allow dependencies across clusters and model each group of data with a mixture model. The number of components is itself an unbounded random variable, which is allowed to increase with the number of data points. This algorithm was originally motivated by the problem of document clustering, where each document is a collection of words (independent draws from underlying distribution), and the HDP aims to find topics (defined as probability distributions over words).

This problem formulation fits very naturally with the location clustering that we want to addres. We will consider each observation from a pair of antennae to correspond to a *document*. The *words* are the originating and terminating antenna IDs and number of calls corresponds to the word counts. Because we were looking for the most popular locations to which users call during a specific hour, we assigned the number of calls to the destination antennas. The HDP assigns to each observation a distribution over antenna locations (i.e. topics) which emphasizes the locations of most popular call receivers. However, the HDP model does not provide a unique cluster for each location. Instead, it computes clusters consisting of several antennae (i.e., spread over locations) for each hour of the day. For a better visualization of the average behavior of the clusters over a day, we empirically set a cut-off threshold to truncate very low-probability occurrences for calls. For the experiments, we used the HDP implementation provided by David Blei. HDPs require setting several parameters, which we picked by initial exploratory experiments using a small subset of the data.

4 Empirical Results

One of the challenging issues facing big data analysis is how to efficient sub-sample the existing data. Using the entire data set yields significant computational costs, but subsampling also leads to worse solutions, by ignoring some of the data. We examined two different methods of sampling which lead to different clustering results. In a first attempt, we uniformly sampled the all observations at each time step. This approach results in a fair distribution of both high and low number of received calls. However, our goal was to look for places of interest, which should be identified by high number of received calls, and this sampling method reduces the contrast between such places and the rest of the data. Thus, for second approach, we collected all the antennae with high numbers of calls (above a certain threshold proportional to the total number of calls recorded in the data for each time slot) and then drew uniformly at random from the rest of the observations. This approach worked significantly better, so we present results based on it.

Figure 1 presents the average log-likelihood over the training set and over an independent test set (from a different week), as a function of the number of iterations of the algorithm. As seen, the algorithm converges quickly and successfully to solutions which have good log-likelihood. Figures 2 and 3 present a visualization of the "important" clusters obtained at four different times during the day (9am, 12pm, 6pm and 10pm). We decided what to plot based on a threshold. As expected, the call patterns are very different at different times. Daytimes show significantly more calls, with more clusters and some very focused clusters. At night, there are significantly fewer clusters, and they are more spread out. It is worth noting that some areas are only visited at certain times during the day.

5 Discussion and Future Work

The pilot experiment that we ran showed the utility of Hierarchical Dirichlet methods for clustering cell phone data. The resulting clusters can be used to study qualitatively the movement of people around the country during the day. This information could be useful for authorities planning new development such as roads or neighborhoods. The only down side to the algorithm is that the number of clusters tends to proliferate as the number of calls increases. The algorithm is, of course, designed this way. However, it would be useful to explore ways of regularizing the number of clusters, beyond thresholding.

From a quantitative point of view, we aim to fit a time series model to the resulting clusters, to have a more precise characterization of population movement. Initially we aimed to use the method in [Bachman and Precup, 2012] to model jointly the spatio-temporal aspect of the data, but the current implementation of their code did not allow handling the required amount of data. We are currently working to extend the code for the big data setting of the D4D challenge.

References

[Bachman and Precup, 2012] Philip Bachman and Doina Precup. Improved estimation in time-varying models. In *ICML*, 2012.

[Blondel *et al.*, 2012] Vincent D. Blondel, Markus Esch, Connie Chan, Fabrice Clérot, Pierre Deville, Etienne Huens, Frédéric Morlot, Zbigniew Smoreda, and Cezary Ziemlicki. Data for development: the d4d challenge on mobile phone data. *CoRR*, abs/1210.0137, 2012.

[Frank *et al.*, 2012] Jordan Frank, Doina Precup, and Shie Mannor. Generating storylines from sensor data. In *Nokia MDC Challenge Worksho*, 2012.

[Orbanz and Teh, 2010] P. Orbanz and Y. W. Teh. Bayesian nonparametric models. In *Encyclopedia of Machine Learning*. Springer, 2010.

[Teh *et al.*, 2006] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

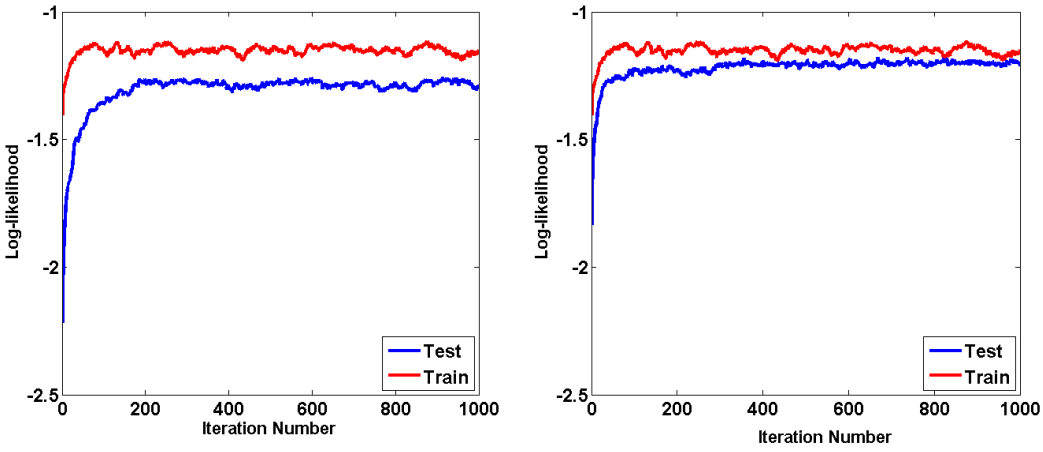


Figure 1: Log-likelihood over training and testing data, as the algorithm progresses. Learning is quick and converges to good solutions

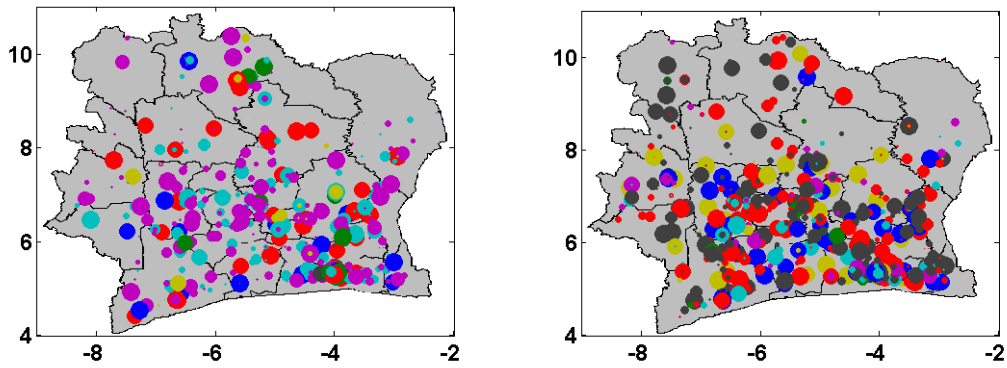


Figure 2: Visualization of antenna clusters (learned based on call numbers) for 9am (left) and 12pm (right)

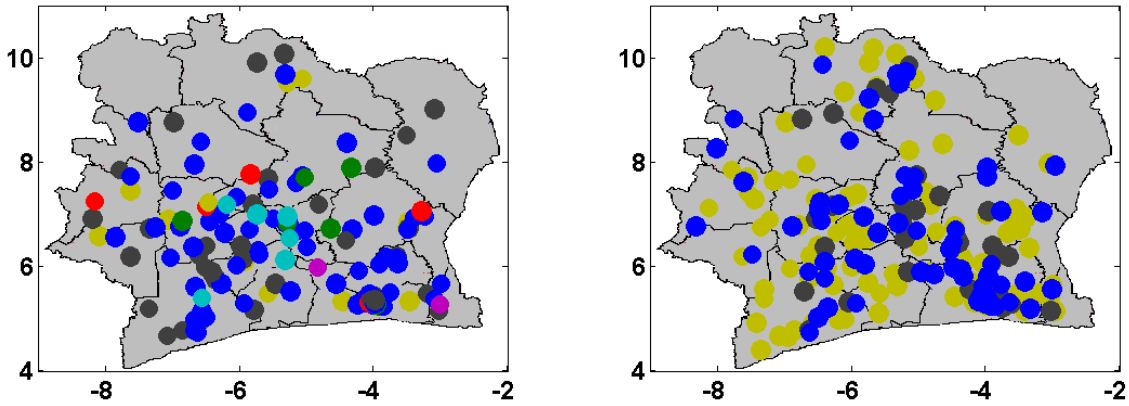


Figure 3: Visualization of antenna clusters (learned based on call numbers) for 6pm (left) and 10pm (right)

Patterns of Cell Towers in Mobile Cellular Network

Jiping Xiong[1], Gyan Ranjan[2], Lei Chen[3], Zhi-li Zhang[2]

1. Dept. of Communication Engineering, Zhejiang Normal University xjping@zjnu.cn

2. Dept. of Computer Science, University of Minnesota

3. School of Electronics and Information, TongJi University

Abstract: Currently, most of researches on the characteristics of mobile cellular network are relied on the analysis of individual mobile users, litter effects have been put on the cell towers' behaviors. Since the communications between cell towers are the aggregation effects of users, it is interesting to find out some particular characteristics. In this paper, based on the data sets obtained from D4D project, we find out that some distributions follow power law and heavy tail rules, such as the average durations of aggregated calls and the average number of calls per cell tower. We also unveil some statistics characteristics including that most of the calls are between short distance cell towers.

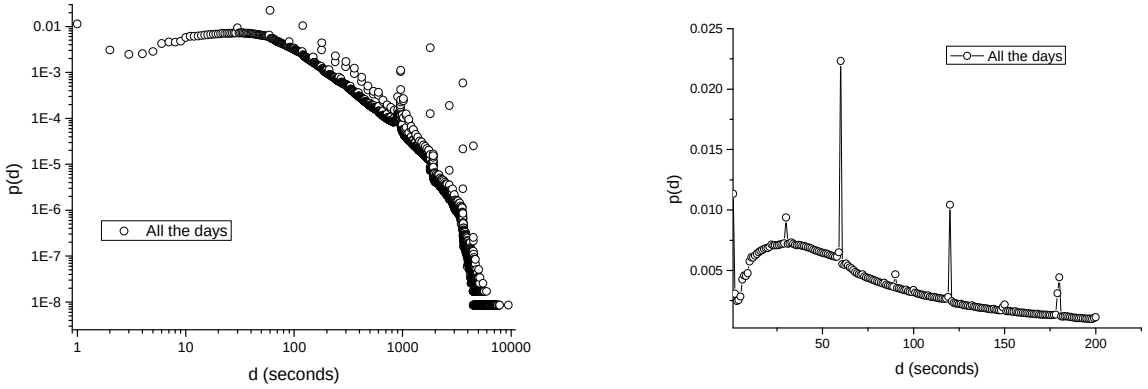
Key Words: Mobile Cellular Network, Power-law distribution, D4D Project

In this paper, we use the datasets from D4D project [1] to analysis the patterns in mobile cellular network. Mainly, we will focus on the dynamic behaviors of cell towers, such as the distribution of average call duration, number of calls between cell towers, relationship between distance and communication volumes, and other interesting statistic characteristics.

SET1 of D4D datasets is the aggregated calls and SMS information among cell towers across 140 days, and the number of the whole records is more than 171 millions. Each record in SET1 contains the number of calls or SMS and the duration in hour by hour basis, as well as the source and destination cell towers.

We find that there are more than 66 percent of records (the number is more than 116 millions) only have one number of call. It means we can assume each such record as a perfect individual Call Detail Records (CDR) of a mobile user.

Fig. 1(a) gives out the percent distribution of different call durations among those individual CDRs. The coordinate is logarithm scale, and x-axis is the duration of call, y-axis is the percent of related duration. We find a surprising but reasonable in real life phenomenon, that is the calls which duration is the multiple of 30 seconds dominate in the dataset. For example, the percent of the calls with 60 seconds duration is 2.2% which is the top percent. The spurs above the heavy line in Fig.1 (a) are the kind of these calls. This finding maybe should be considered in future reasonable model for mobile cellular network. In real life, people use to terminate the call before the end of minutes, and that results to the phenomenon. From Fig.1 (b) (linear coordinate), we can see the spurs more clearly, and also we can find that those spurs with multiple of 60 seconds have even more percent than those with multiple 30 seconds duration.



(a) all kinds of durations from 1s to 7848s

(b) part of durations from 1s to 200s

Fig. 1. Distribution of call duration among individual CDRs

If we get rid of the spurs, the distribution does exhibits power-law distributions in multiple ranges, for example the range from 100 second to 1000 second, which confirms previous works [2-3].

Fig.2 is the distribution of average call duration for all the CDRs among cell towers in SET1 including the individual CDRs. Fig.2 and Fig.1(a) are almost the same shape, that means the pattern of aggregation communications between cell towers can reflect exact individual behaviors of mobile user. To the best of our knowledge, we are the first one to claim this exact pattern matches between cell towers and mobile user.

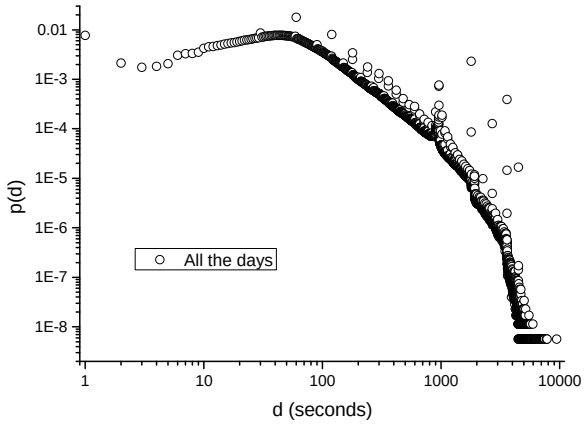


Fig. 2. Distribution of call duration among all CDRs

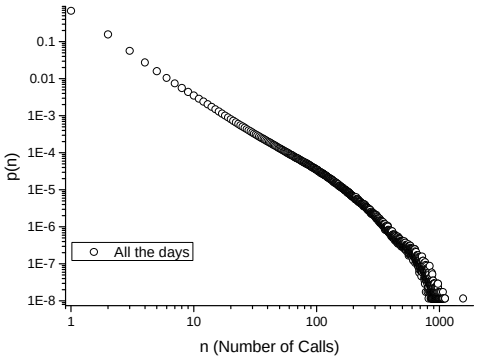


Fig. 3. Distribution of number of calls among cell towers

Fig.3 displays the distribution of number of calls among all cell towers across all days, and again it can be perfectly modeled by a power-law distribution with heavy tail.

From fig.4, as we can expected, the percent of calls decrease as the communication distance increase, and more than 65% communications happen within the range of 40 kilometers.

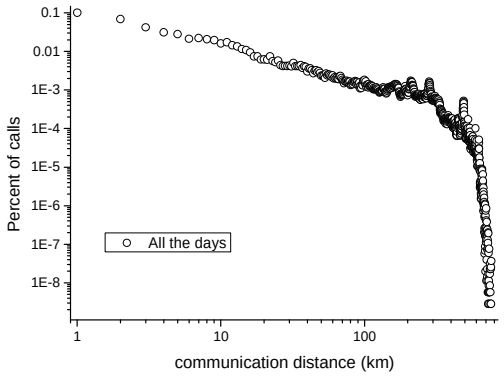


Fig. 4. Relationship between percent of number of calls and communication distance

We aggregate the communications day by day and plot the total number of calls each day from Dec 5, 2011 to April 22, 2012. Fig.5 shows the temporal communication volumes changes of the total 140 days. At the first day (Dec 5, 2011) the number of calls is very low, but it went high at following days and keep at a low volume from day 11 (Dec 15,2011) to day 47 (Jan 20, 2012). Day 124 (April 6, 2012) has the highest volume among the 140 days. We also find a surprising result, that almost more than 10% cell towers have zero communication volumes from day 1 to day 114 (Mar 27, 2012), Fig.6 gives an communication volumes changes example of one of these cell towers.

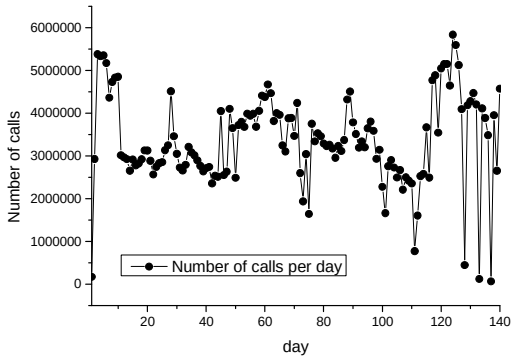


Fig. 5. Number of calls from Dec 5,2011 to April 22, 2012

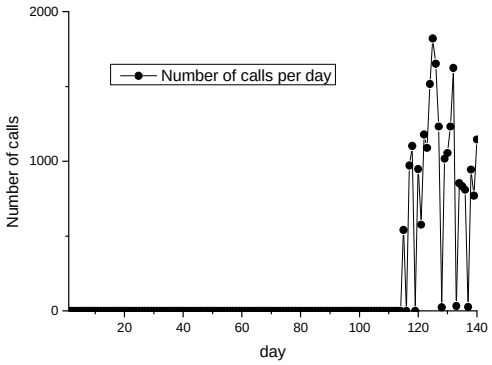


Fig. 6. Number of calls from Dec 5,2011 to April 22, 2012 of cell tower 1222

If there are communications between two cell towers, then the two cell towers are neighbors. We are interested in the dynamic changes of neighbors for each cell tower in a day by day basis. We also define a cell tower is a constant or stable neighbor of another cell towers if and only if this cell tower is the neighbor in all the 140 days (we need to remove those days with zero communications). We find out that the average number of stable neighbors is 5, and most of the stable neighbors are close to related cell tower.

We generate an interesting distribution of users according to the number of different cell towers that he connected during the two weeks using the individual trajectories for 50,000 customers for two week time in the SET2. Fig.7 is the result, and we can find that most users only connect to less than 3 cell towers during the whole two week trajectories. It means that the activity region of most users is fairly small. Also, this kind of distribution can be well modeled by power-law.

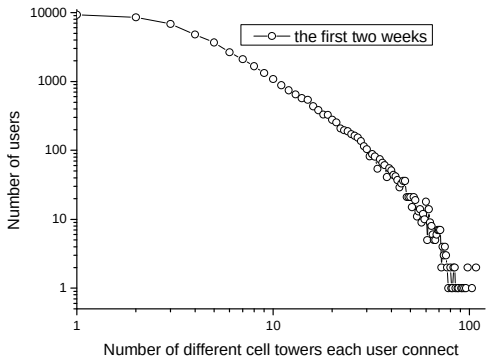


Fig. 7. Distribution of users according to the number of different cell towers

In this paper, we extensively exploit the distribution including number of calls, call durations and communication distance among cell towers. It is interesting that almost all the distributions can be modeled by power-law rule which is a very common distribution in complexly networks. We can also expect that the inter-call duration of the 50,000 users can also following this promising rule.

References

[1] D4D Challenge project, http://www.d4d.orange.com/home
[2] Siyuan Liu, Lei Li, Christos Faloutsos, Lionel M. Ni: Mobile Phone Graph Evolution: Findings, Model and Interpretation. ICDM Workshops 2011: 323-330
[3] Pedro O. S. Vaz de Melo, Leman Akoglu, Christos Faloutsos, Antonio Alfredo Ferreira Loureiro: Surprising Patterns for the Call Duration Distribution of Mobile Phone Users. ECML/PKDD (3) 2010: 354-369

Mobile Phone Data Analysis of Cote d'Ivorie

Didem Gündoğdu
Boğaziçi University, Computer Science, Istanbul Turkey
didem.gundogdu@boun.edu.tr

Introduction

Mobile phones became an extension of our daily lives. 1990's huge, clumsy, heavy phones evolved into smart, tiny, fashionable phones, with high computational abilities. Most of them are not just phone but a very efficient daily organizers, enhanced with many sensor capabilities carried on at all times These communication tools make life easier for scientists, who aim to collect data in order to understand human behaviors.

Today's telecommunication companies possess valuable data extracted from mobile phone usage, but there are severe privacy issues that must be resolved ere these data become available to the scrutiny of scientists. For this reason, controlled and ethically approved release of mobile data is extremely valuable. There are some recent campaigns to gather mobile data, which produced the Reality Mining Dataset [3] and the Nokia Mobile Challenge Dataset [4 which are the widely-known. These campaigns have collected data for from about 100-200.

Data for Development (D4D) is a very valuable challenge, for those aspects; being real data, and having thousands of participants. Data are not collected from a set up, the number of subscribers, are as much as 500.000 for the whole data sets.

The primary research interest of this report is to predict the age distributions of the subscribers. This is a useful information that can be correlated with many different social aspects of the life in the city. However D4D datasets do not have a ground truth annotation for the age of the subscribers. We have analyzed the data for understanding the demographic distribution, in order to discover patterns that may be used to improve the life conditions of the Cote D'Ivorie community.

Motivation and Background

Numerous researches and applications have been done in the last decade, for understanding human behavior, since new phones embody multi sensors for evaluating each step we take. However there are not so many studies related to demographic prediction from mobile data. Mo et al. "Your Phone Understands You", or Jia-Ching et al. "Demographic Prediction Based on User's Mobile Behaviors" from Nokia MDC[1], are all based on supervised techniques, which requires training phase. In those techniques content of the training data set is important; generalization may not work well while testing in other social groups. Aarthi S., et al [9] studied on Reality Mining data, with similar methods, but only for predicting the gender and income. One of the most recent study for predicting demographic data are from UC Berkeley, Blumenstock and Gillick with Nathan Eagle[5], with Rwanda telecommunication data. They observed that socio-economic status is correlated with the call duration, and they found that gender is quite an important factor for understanding mobility pattern.

If you have demographic data and location information just for a given time, you can predict many things. Some examples could be making dynamic advertisement just targeting your age group, traffic planning, urban area planning. If young generations are hanging out at some location at the evening, building a cinema, theatre or concert hall in that region may help to improve social life quality in that region.

[1] Nokia MDC: Nokia Mobile Data Challenge

The possible challenges would be socio economic situation of the Cote d'Ivoire. In 2011 Currently around %60 of the population is living below the poverty threshold. The mobile data may not exactly depict the Cote d'Ivorie's current snap shot.

Facts about Cote d'Ivorie

There are around 20 million people living in Cote d'Ivorie, 5 million of that population is residing in Abidjan, former capital of the country. Yamoussoukro is the capital [17] around 800.000 inhabitants living. %56 of the population is capable of reading and writing. The distribution of the population can be viewed in fig. 2.

Economy is mainly dependant on agriculture, as one of the major exporter of cocoa and palm oils. Cote d'Ivorie is one of the 20 poorest countries in the world according to IFAD[2].
Half of the population is living in the rural areas, and nearly %60 of the population is living below poverty threshold. Especially north eastern and western part of the country is suffering from extreme poverty. [18] [19] [20]

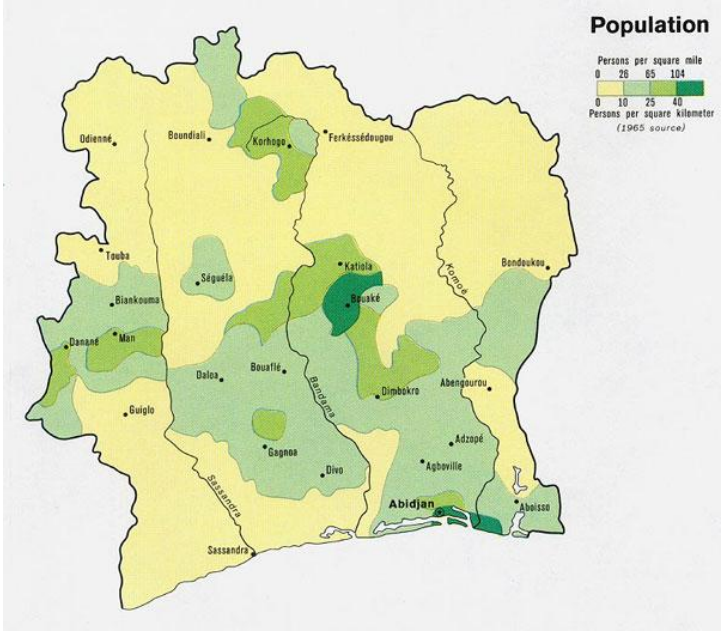


Fig. 2 Population Distribution

Mobile telecommunication licenses, have first delivered in 1998, and since then it is evolving with penetration rate of %5 according to ICT reports. [21]

Orange Telecom's position in Cote d'Ivoire

There are five main players in Telecomminication sector in Cote'd Ivorie. And Orange Telecom is holding slightly more than 6 million subscribers as a leader in telecomunication sector.However, MTN is the other strong operator in the country.

[2] The International Fund for Agricultural Development (IFAD), a specialized agency of the United Nations

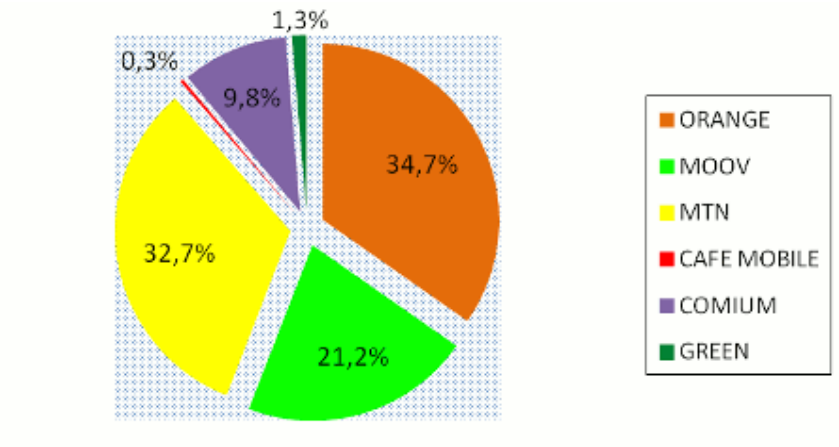


Fig.1 Market Share of Mobile Network Operators [15]

Related Work

We plot the Mobility versus Talkativiy diagram for 10 subsets, and found that characteristic of the users behaviors are similar as in fig. 3 .Each subset presents a power of law characteristic, very similar of Gonzales et.al. Findings, [6] in "Understanding Human Mobility Patterns".

In order to understand behavior pattern, we applied topic model,[1] like Farrahi et al. [2] [8] [10] who used topic model in order to understand human behavior patterns. They divided 24 hours in half hour periods, and labeled each period with respect to GPS tags as, Home (H), Work (W), and Other (O). 24 hour is divided into 8 time slot, and each word is constructed in that way.

Farrahi et.al. Used Reality Mining dataset [2], it includes not only call log, but also Bluetooth, SMS usage, and survey to take additional information of the subject.

Methodology

Set2 and Set3 are similar in data type; Set2 is a small subset of Set3. However Set3 consists of 5 months data, with around 2GB each. In Set2 each file is around 150 MB approximately 5million lines to process. Set2 has 10 individual sub datasets; each sub dataset contains 50.000 users. Our work based on Set2's first sub dataset, POS_SAMPLE_0 data file.

Data is preprocessed according to weekday, from Monday to Sunday- 7 and time slot, "Night, morning, day, evening". Aggregated according to time slots for analyze.

Time slots:
Slot 1 : 21:00 – 06:00 – Night
Slot 2 : 06:00 – 10:00 – Morning
Slot 3 : 10:00 – 17:00 – Daytime
Slot 4 : 17:00 – 21:00 – Evening

We first visualized the data, for understanding talk frequency, mobility, and location. The distribution of the talkativity and mobility is depicted, on the Cote d'Ivorie. The semantic of the call regularity, may give us some clues about socio economic situation of the given region. One of the challenges was, we do not have any information of the subscriber, when s/he has very low call frequency.

The first visualization was on Mobility versus Talkativity. Mobility is with respect to number of different Antenna ID's during the sampling period. Talkativity stands for the number of outgoing call. Mobility Versus Talkativity is plotted for subset of Dataset2, as shown in fig 2, they all have similar characteristic, which is tend to have less talk and less mobility, which is meaningful if we consider the daily expenditure of an average person is 1 Eur. [19]

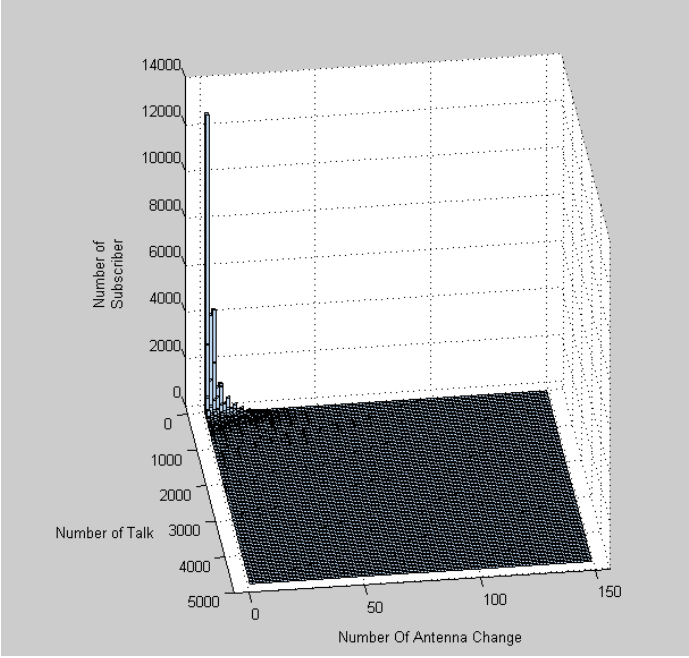


Fig 3 Talkativity Versus Mobility Distribution

In order to understand the subscriber's mobility, a pairwise longest distance found for each subscriber for the given period. The Antenna locations were given as latitude, and longitude information. The longest distance was taken as diameter, and the median of the two antennas become the circle of subscriber's mobility.

The visualization has been plotted with Tableu Software[16], with low mobility and with high mobility respectively. Tableu Software is very flexible and from the link of each plot, it can be visualized over Web, for different zoom and view options.

Around the city Abidjin, people tend to move less, however the capital of the Cote'd Ivorie Yamoussoukro, has people with higher mobility.

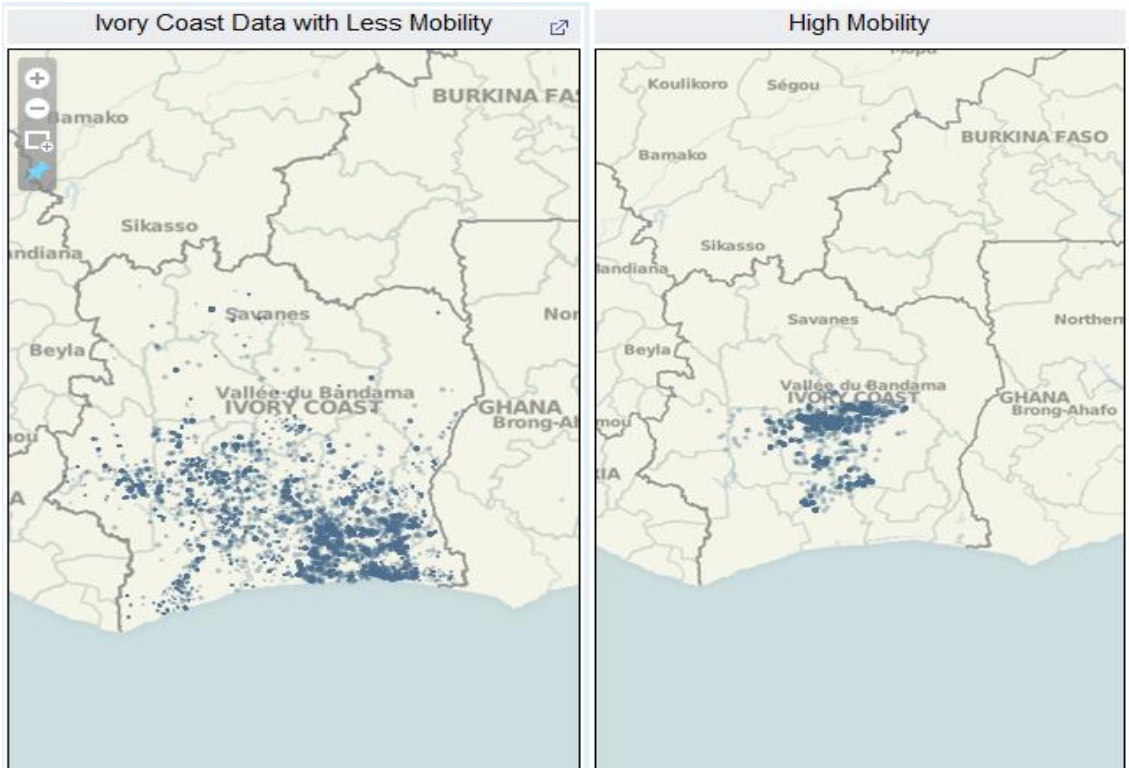


Fig. 4 – Mobility Diameters

http://public.tableausoftware.com/views/ivory_best2/Dashboard1?:embed=y

Possible Working and Home Locations

In order to understand the subscribers, significant locations, we assumed that if the user talks from the same antenna during the day, it is annotated as working place, if the antenna ID does not change during the night talks; it is predicted to be the home location. (At least %70 same antenna ID, in that time slot)

Here, we are taking Antenna Id's locations; this may mislead us, as neighbors will be shown as living in the same location.

In that computation, for 50.000 subscribers, around 14.000 subscribers have same daily and night location. It can give us, the information that they are not, working or house-wifes. The plot can be seen in fig 7 it is focused on Abidjan for giving an idea about poverty zones.

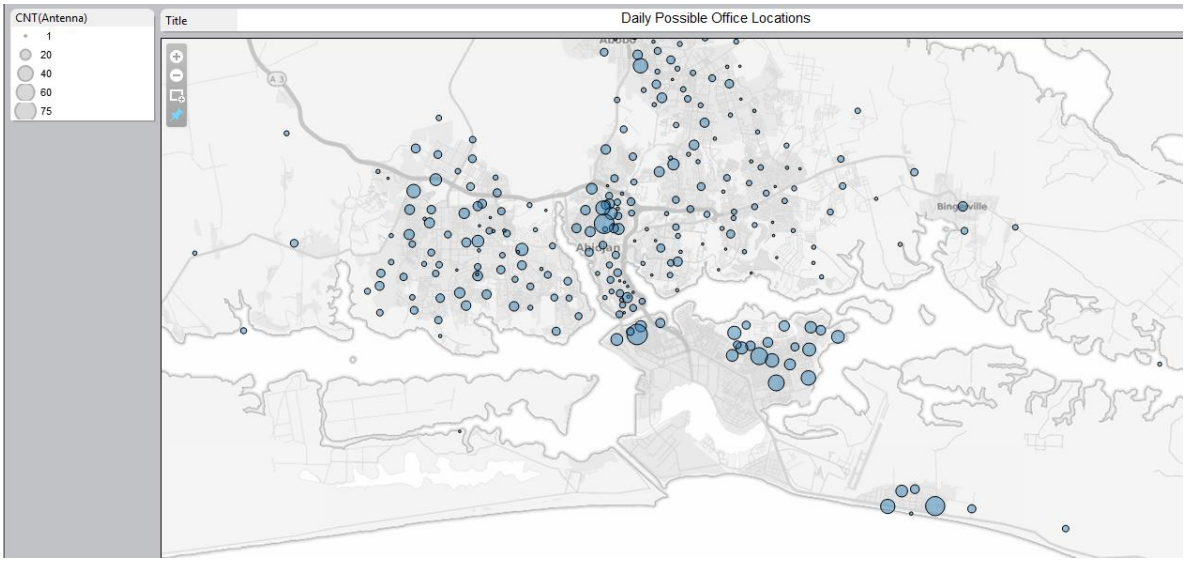


Fig 5. Daily Work Locations

http://public.tableausoftware.com/views/filteredWork/Sheet1?:embed=y

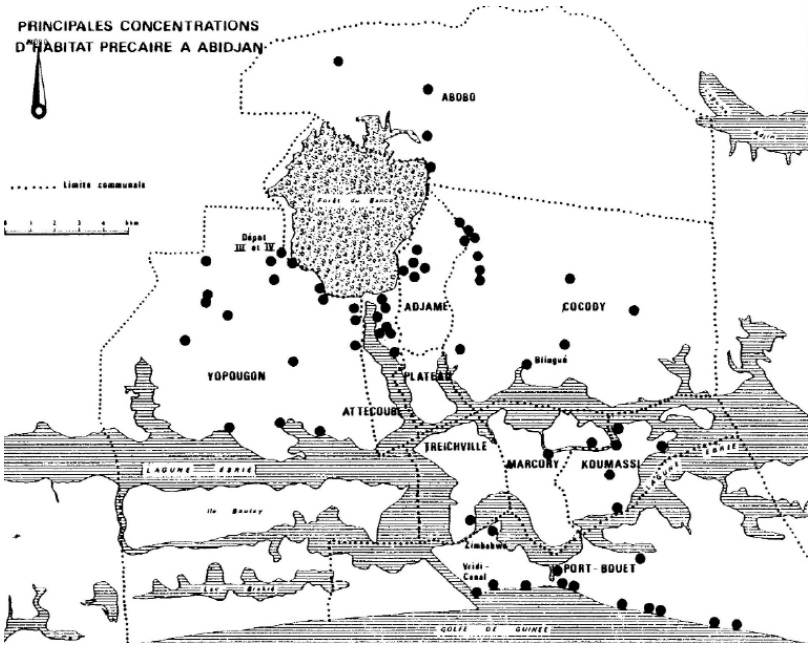


Fig 6. Abidjan rural residential area (BNETD) [22]

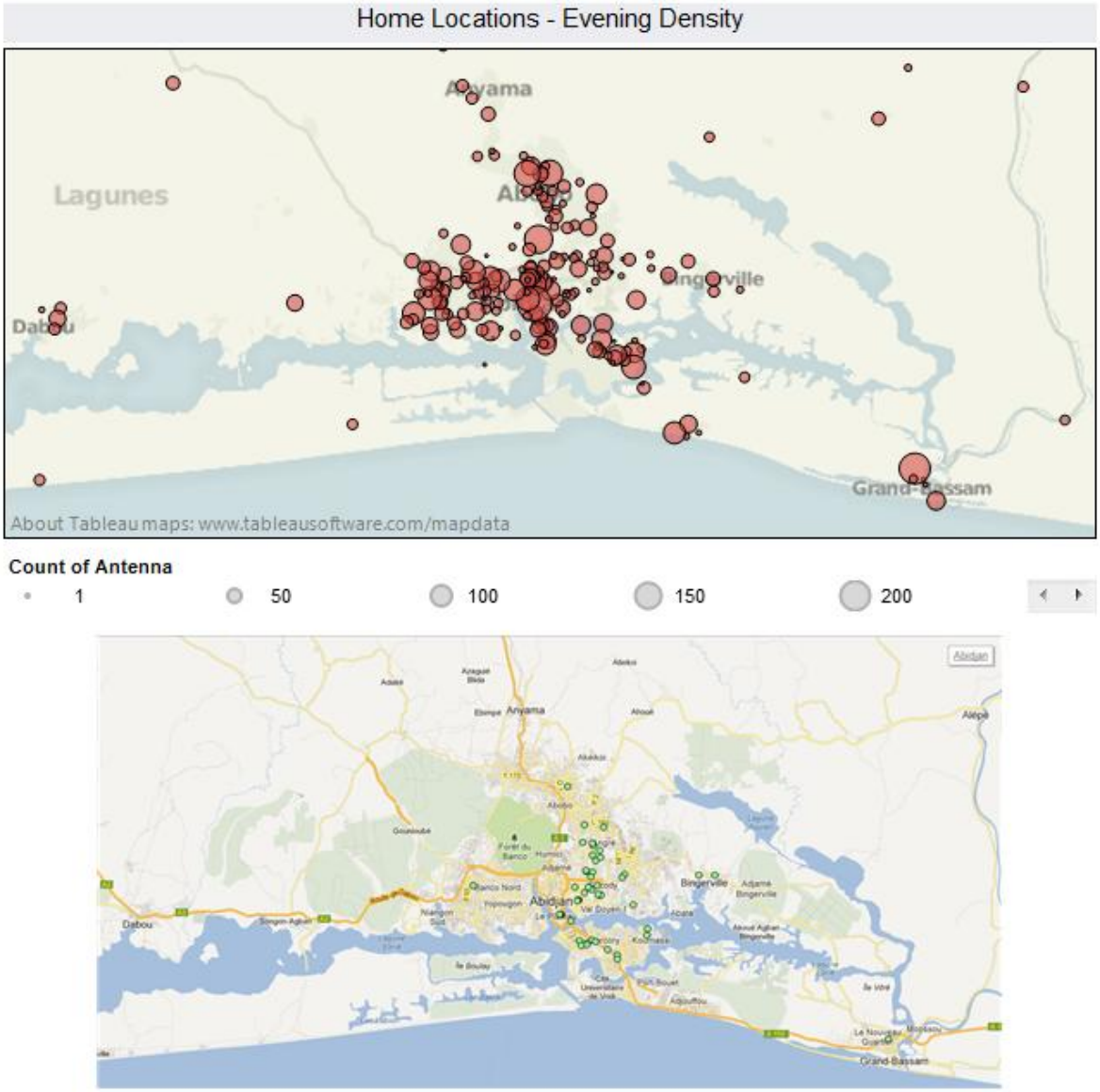


Fig 7 .Schools are plotted from Google Maps. & Home Locations

http://public.tableausoftware.com/views/HomeLocations/Dashboard1?:embed=y

Yopougon and Northern part of Abidjan have slums. You may notice from the fig. 7 above, dense home locations without schools. Even in non-worker's plot in fig.8. you can see the similar clusters of regions with low income.

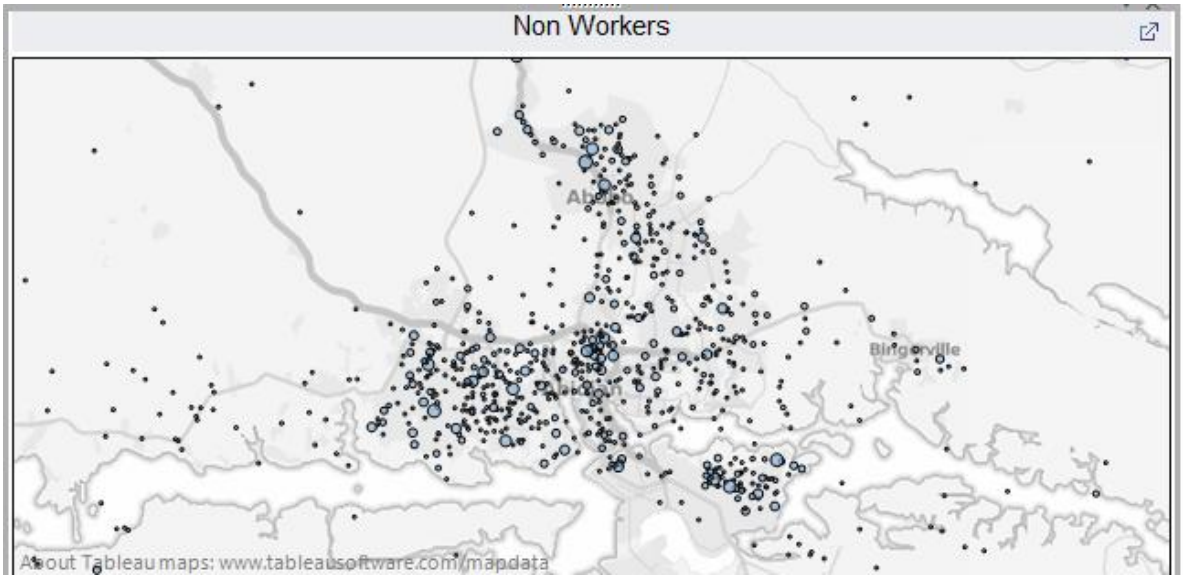
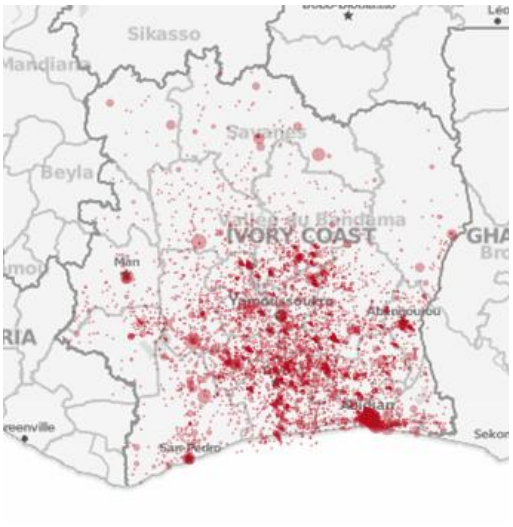


Fig 7. Non working population

http://public.tableausoftware.com/views/nonWorkers/Dashboard1?:embed=y

Poverty Plots

From OPHI(Oxford Poverty and Human Development Initiative) fig 8b, the poverty distribution of the Cote d'Ivorie is seen. Here we compare talkativity plot fig 8a with that distribution. The talkativity statistics is proving that diagonal regions of the country are capable of having a mobile phone and being able to have some calls. Northeast region is obviously less dense in population and suffering from poverty.



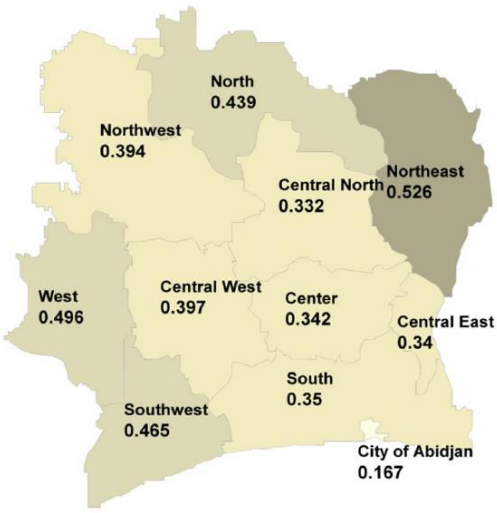


Fig 8a. Talking Density versus Poverty

http://public.tableausoftware.com/views/TalkLocation/Sheet1?:embed=y

Fig 8b Poverty Rates at the Sub-national Level, Oxford Poverty and Human Development Initiative (OPHI)

Call Behavior Understanding

In order to understand the call pattern of the subscriber's, LDA (Latent Dirichlet Allocation) [1] is used.

LDA is mainly applied into documents, for understanding the hidden semantic topic patterns. It is useful to classify vast amounts of documents automatically. Each document may be consisting of several topics, with different weights and distributions. The usage of LDA is not restricted only to documents, exploring the hidden semantic patterns of human behavior, human action recognition with bag of words image representations are some other application areas.

Farrahi's[2] work, a document is a day of a user, and author is the user. Author Topic Model, is a special extension of LDA. [1]

Implementation

In order to implement LDA, we should define Documents, Topics and Vocabulary in our domain.

Documents → User Mobile Data
Topics → Different Call Behaviors
Vocabulary → Daily Call Numbers as

Matlab Topic Toolbox [14] is used for implementation. Data is preprocessed as vector of words, which is a subset of vocabulary given below. The number of topics or call patterns should be given. We set as number of topics as 10, in our experiment.

Call Information

In first experiment, we set the number of calls during the day as follows,
0, 0-3, 4-7, 8 – 10, 20 <
However in the first experiment, it is observed that Cote'd Ivory mobile phone users are not very talkative, and this model does not discriminate the users from each other. So new range is set as,
0, 1-2, 3-5, 6-10, 11 <, and instead of dividing the day into 8 time slots, we divided the day into four slots, which represent, night, morning, daytime, evening.

Number of daily calls	*Number of Calls*	*Corresponding Word*
Nr. of daily calls in the slot1:	0/ 1-2/ 3-5 / 6 - 10 / 11 < :	A0/A1/A2/A3/A4
Nr. of daily calls in the slot2:	0/ 1-2/ 3-5 / 6 - 10 / 11 < :	B0/B1/B2/B3/B4
Nr. of daily calls in the slot3:	0/ 1-2/ 3-5 / 6 – 10 /s11 < :	C0/C1/C2/C3/C4
Nr. of daily calls in the slot4:	0/ 1-2/ 3-5 / 6 - 10 / 11 < :	D0/D1/D2/D3/D4

After running out the LDA with 10 topics, and with hyper parameters: Alpha = 0, 01 and Beta = 5 for 300 iterations, we got the probability distribution of call behaviors as shown below, fig 6.

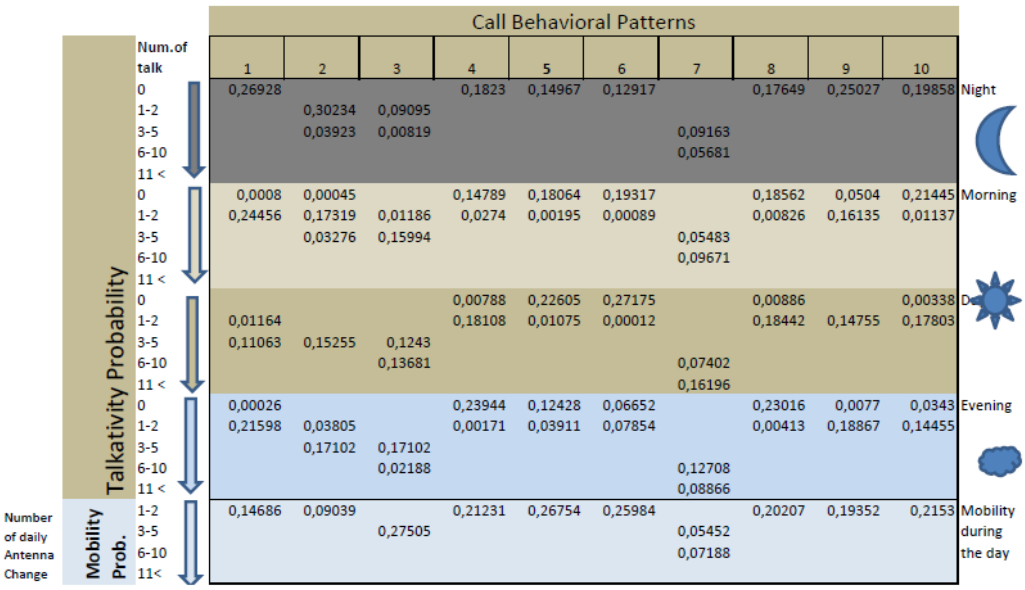


fig. 6 The distribution of topics

In call behaviors, we see that mainly people tend to speak and move less, except behavior pattern 7, which is more likely, talk during the day with the number greater then 10. In general the rest of the call behavior is, having 1-2 calls in each time slot.

And we could divide the day into three instead of 4, as in each slot the call characteristic remains same, instead of morning and day, we could make, night, day, and evening. Behaviors, 4-5-6-8-9-10 are quite similar. In 7 is quite different from the rest of the topics since it represents talk more, move more pattern. 2 and 3 are similar but 3 tend to talk more at evening and more mobile.

Conclusion

While we are working on data, there are some assumptions that we have to keep in mind, especially for special condition of Cote d'Ivorie, as one of the poorest countries in the world.

1. Orange is not the only telecom operator
2. Not everyone has one mobile phone
3. Even sometimes, one person can have more than one subscription to mobile operators
4. %50 of the population is living in rural areas, [17] and the mobile operator's coverage is not satisfactory in rural areas.[15]
5. Antenna ID, for location based computation may not be adequate.
6. Population's socio-economic conditions may lead preventing common technology usage

We analyzed data in talkativity, mobility and call pattern aspects. The findings are in parallel with the countries, socio economic situation. The northern part of the country is struggling with poverty, and even in main city Abidjin have some districts with very poor conditions. The new residential areas have inadequate coverage from government, as in fig 7, the school locations are far from the poor restricts.

Government may focus on home locations, or non-working people regions, to bring more education, health

services and recreational places to make life conditions better.

Although Northern part of the map has very little activity, it should be kept in mind, that the poverty hid the populations profile over there.

The socio economic situation of the region can be seen from talk frequency map. Northwest region with few talk regularity, may leads our analysis as one of the poorest district. However from the population distribution it can be due to low population density or other mobile phone operators, rather than Orange, may be leading that region.

If we analyze home location density map, and compare with fig. 5, which is possibly from an old source, the reference does not mention the date, we can see the how city is enlarging. It is enlarging through north and west. This information may help municipalities, for foreseeing new infrastructure planning.

As you can see in fig 3, general clustering algorithms are not working quite well, in such power law characteristics, so LDA model is used. Even in LDA, data tend to gather in less talk, less movement schema.

Future Work

As a future work, we can extend the data analyze, for the rest of the datasets. W e can not analyze the whole data, because of the computational and time restrictions. The LDA implementation may be enriched with additional words, such as weekend call usage, weekday call usage, day specific calls, etc.

If the subscriber's personal data could be given, more sophisticated analysis can be done. Without having a ground truth, researches may be misleading.

Reference

[1] David M. Blei, Andrew Y. Ng, Michael I. Jordan, Pervasive Computing, Newcastle, UK, June 2012. "Latent Dirichlet Allocation", Journal of Machine Learning Research 3 (2003) 993-1022, University of California

[2] Katayoun Farrahi and Daniel Gatica-Perez, "Probabilistic Mining of Socio-Geographic Routines From Mobile Phone Data, IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING, VOL. 4, NO. 4, AUGUST 2010"

[3] N. Eagle and A. Pentland. "Eigenbehaviors: Identifying Structure in Routine," Behavioral Ecology and Sociobiology 63:7, 1057-1066, 2009.

[4] Kiukkonen, N., Blom, J., Dousse, O., Gatica-Perez, D., and Laurila, J. 2010. "Towards rich mobile phone datasets: Lausanne data collection campaign." In Proc. ACM Int. Conf. onPervasive Services (ICPS). Berlin.

[5] N. Eagle, A. Pentland, and D. Lazer. "Inferring Social Network Structure, using Mobile Phone Data," PNAS, 106(36) 15274-8, 2009.

[6] M.C. Gonzalez, A. Cesar and A.L. Barabasi. "Understanding Individual Human Mobility Patterns" Nature 453(7196):779-782, 2008

[7] Tino Kreutzer, Internet and Online Media Usage on Mobile Phones among Low-Income Urban Youth in Cape Town, Centre for Film and Media Studies University of Cape Town

[8] K. Farrahi, D. Gatica-Perez, "Learning and Predicting Multimodal Daily Life Patterns from Cell Phones", In ICMI-MLMI'09, November 2–4, 2009, Cambridge, MA, USA

[9] Aarthi.S & Bharanidharan.S, Saravanan.M & Anand.V, "Predicting Customer Demographics In A Mobile Social Network", In 2011 IEEE, DOI 10.1109 / ASONAM.2011.13

[10] K. Farrahi, D. Gatica-Perez, "Probabilistic Mining of Socio-Geographic Routines from Mobile Phone Data", Ecole Polytechnique Fédérale de Lausanne (EPFL), 2010

[11] J.E. Blumenstock, D. Gillick, N. Eagle, "Who's Calling? Demographics of Mobile Phone Use in Rwanda", 2010, Association for the Advancement of Artificial Intelligence (www.aaai.org)

[12] G. Chittaranjan, J. Blom , D. Gatica-Perez, "Mining large-scale smartphone data for personality studies", International Symposium onWearable Computers, San Francisco, California, June 2011

[13] D. Ashbrook, T. Starner "Learning Significant Locations and Predicting User Movement with GPS", 6th International Symposium on Wearable Computers (ISWC.02)

[14] Matlab Topic Modeling Toolbox 1.4, by Mark Steyvers, University of California Irvine, Tom Griffiths, University of California Berkeley, http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm

[15] http://www.atci.ci/index.php/Service-mobile/abonnes-service-mobile.html

[16] http://www.tableausoftware.com/

[17] https://www.cia.gov/library/publications/the-world-factbook/geos/iv.html

[18] http://www.irinnews.org/Report/81804/COTE-D-IVOIRE-Poverty-getting-worse-study

[19] http://www.ruralpovertyportal.org/country/home/tags/cote_divoire

[20] http://hdr.undp.org/external/mpi/Cote-d-Ivoire-OPHI-CountryBrief-2011.pdf

[21] Arsène KOUADIO Konan , Côte d'Ivoire ICT Sector Performance Review 2009/2010

[22] The case of Abidjan, by Kouame Appessika, Bureau National d'Etudes Techniques et de Development(BNETD)

Properties of Dynamic Networks

Rajul Anand and Chandan K. Reddy
Department of Computer Science, Wayne State University
Corresponding author : rajulanand@wayne.edu

I. Introduction

The mobile communication graphs can reveal important information regarding the various dynamics happenings inside the network. The static network graph analysis of such data is highly popular and still considered as the pre-requisite for analyzing the network. However, knowing the dynamics of the network (what changes between different time points) can be useful for identifying such as network congestion, resource allocation for maintenance etc. We reviewed the communication graphs (SET 4) which contain sub-graphs for 5000 randomly selected individuals with neighborhood information. In order to analyze the changes between different time periods, maintaining the highly compressed and summarized form of the graph; this is useful for comparing the differences between snapshots.

A summarization of the network can be done in many ways. We believe in multi-level summarization [1] of the network which allows the user to (i) view the high-level summary of the graph (important nodes,connection flow between them) and (ii) delve into details of the summary to understand the phenomenon.

A. Clustering as summarization

Clustering is a natural way to summarize the various groupings inside the graph and yet the vast literature of clustering algorithms are not directly applicable for this mobile call graph. It is because the most of the clustering algorithms focus on minimizing the edge cut of the induced graph and present the results as clustering of the graph. Also, the clustering algorithm provide only very detailed level summary which focus only upon the grouping relation between nodes. It doesn't provide the grouping relation between collection of nodes. Finally, they ignore the characteristics of graphs like whether it's a communication graph or social graph of friends or a graph of citations. Our current communication graph needs to be analyzed by method which understands the flow dynamics and then provide the clustering summary. We have two candidates clustering algorithms for the same: (i) Markov Clustering(MCL) [2] and (ii) Map-Equation [3].

The MCL algorithm uses the concept of a stochastic graph flow and identifies the natural clusters present in the graph data. Since, the clustering is driven from the idea of flow inside the graph, it's suitable for a communication graph. However, MCL has an inherent tendency to find hubs in the form of attractor nodes. The two reason we didn't consider MCL for summarization is as follows : (i) provide clusters centered around hubs (ii) not readily scalable for very large graphs. The first problem is of more concern since the graph provided to us is already an ego graph with roughly 5000 connected components. Our initial runs of MCL with few components confirmed that MCL will most likely to provide us the connected components and will not compress the graph any further. The *hubs usually in this case where the user id's* who were chosen to create the neighborhood graphs; although not all the times user id's were selected as hubs in our limited results on smaller graph samples.

II. Clustering Algorithm

The map equation [3] is a flow and information theoretic measure which can be used to obtain the optimal network partitions. Random walk is used in map equation to describe the information flow in both directed and undirected networks. Using the ergodic visit frequencies of the nodes in the network, the two level Huffman encoding is applied where unique codes are given to modules. Within each module, the code words are reused so as to save the number of bits needed for representation.

However, **the basic principal of giving shortest length code to the most frequently visited node is followed**.

Let n be the number of nodes divided into m modules for a module partition M. The lower bound on the code length is defined to be L(M). Let Q_i represent the probability to exit module i and p_j represents the frequency of node j (For undirected networks, as in this case, the node visit frequency is simply the relative weight of the edges connected to the node.). Then map equation can be defined as :

$$L(M) = \sum_{i=1}^{m} Q_i \, log \sum_{i=1}^{m} Q_i - 2\sum_{i=1}^{m} Q_i \, log \, Q_i - \sum_{j=1}^{n} p_j \, log \, p_j + \sum_{i=1}^{m} (Q_i + \sum_{j \in i} p_j) \, log(Q_i + \sum_{j \in i} p_j) \quad (1)$$

The advantage of the above method is that it provide us meaningful summary based on the flow in the graph and to provide us multiple level summarization. Note that, there are also other methods [4] based upon modularity maximization which also provided the hierarchy of clusters in the network. We didn't include those methods for our analysis but might as well be applicable here.

III. Dataset Description

The network was reduced to second order neighborhood divided into period of two weeks. Roughly 5000 connected components in the network and their second order neighborhoods. The communication between second order neighbors was not included in the network. The user identifiers between 1 and 10,000 represent chosen users and id's 20,000 and onwards represents neighbors. The public phone usage pattern was already excluded from the graph. Only the edges between the nodes is provided and no other information was given. The ten snapshots of the graph with each snapshot containing two weeks of the graph information with anonymized identifiers is provided. Anonymized identifiers meaning the same user is provided different id (although unique) in each snapshot.

TABLE I
Properties of Graph snapshots (0-9)

Sl	*Node*	*Edges*	*Density*	*Avg*	*Conn*	*CC*	*Tr*
0	312790	326371	6.67e-6	2.08	4713	0.22	16545
1	248787	259111	8.37e-6	2.08	4358	0.22	13365
2	236800	244276	8.71e-6	2.06	4717	0.19	10993
3	240647	247617	8.55e-6	2.05	4883	0.18	10738
4	332805	347606	6.27e-6	2.08	4904	0.21	17337
5	245781	253300	8.38e-6	2.06	4855	0.18	10890
6	343891	358981	6.07e-6	2.08	4903	0.22	17864
7	207932	212507	9.83e-6	2.04	4779	0.16	8249
8	212507	355688	6.12e-6	2.08	4886	0.22	17187
9	198844	202582	1.02e-5	2.03	4723	0.15	7294

IV. Results

We did minimal processing of the data which involved separating the ten snapshots from a single file and convert it into PageK network format [1]. We ran our method ten times to obtain the best summarization of the method due to stochastic nature of the algorithm. However, we found that a *single run of the algorithm was enough to find the best code-length* and the performance improvement in code length with multiple runs was in order of 10^{-3} with no significant difference in major modules.

A. Original Graph Properties

Important properties of each snapshot are highlighted in Table I. The notations *Avg* is average degree, *Conn* is number of connected components, *CC* represents clustering coefficient of the network which depends upon T, the number of triangles in the network. We can easily conclude from the table that average degree remains nearly constant at around 2.08. The number of connected components are around 5000 because the graph was created from these many individual graph seeds. The number of connected components for each graph snapshot were equal to *number of modules found by our second level of detailed summarization*. The final graph is relatively dense than the other graph snapshots.

$GRAPH\_4, GRAPH\_6$ and $GRAPH\_8$ are more similar to each other having nearly the same clustering coefficient and number of triangles contained in the graph. Similarly, $GRAPH\_2, GRAPH\_3$ and $GRAPH\_5$ have comparable clustering coefficient and number of triangles in the graph.

We plotted the degree distribution of all the time periods together to notice whether the user

[1] http://netwiki.amath.unc.edu/DataFormats/PajekNetAndPajFormats

activity changes significantly during any of these time periods. Fig. 2, 4 compare the degree distribution in the original graph. We notice large tail in end for half of the snapshots signifying the scale-free networks behavior with **few large number of nodes communication with lots of other nodes**. We specifically found tails in *GRAPH*_3, *GRAPH*_4, *GRAPH*_6, *GRAPH*_7 and *GRAPH*_9 and specifically, *GRAPH*_9 having the longest tail among them. The other remaining snapshots(*GRAPH*_0,GRAPH_1, *GRAPH*_2,GRAPH_5 and *GRAPH*_8) have similar, very bumpy and small tail ending on the x-axis between region 6 and 7 confirming more similarity among these snapshots.

B. Clustered Graph Summary

We found consistently for all the graphs that a detailed summarization was found in two different ways: The first summarization involved all the nodes as their own group although the weightage of links was modified due to relative re-weighting based upon the number of neighbors. Also, the nodes who were not directly connected were also found to be connected if they share at least one neighbor. The weightage between such nodes depend upon the number of common neighbors they had. We believe this *summarization is more of re-weighting of network* which can be further utilized by any clustering algorithms using edge-cut minimization techniques.

The *second summarization consistently produced the number of connected components* in the network. We believe that any major network clustering algorithm will be able to provide this summarization although the re-weighting was done by the method to highlight the strength of nodes based upon the neighborhood graph.

Our **final level of summarization compressed the network within the connected components** which is ordinarily not possible with typical clustering algorithms as they focus over clean edge cuts. Our final summarization leads to more compressed network with lots of small modules centered around number of triangles. This is evident from increase in clustering coefficient for all the graph snapshots in Fig. 1. The original graph snapshots had lower clustering coefficient (around 0.22) whereas the new summary contains more small, compact clusters with ties broken with non-cluster members and new edges being formed due to cluster membership.

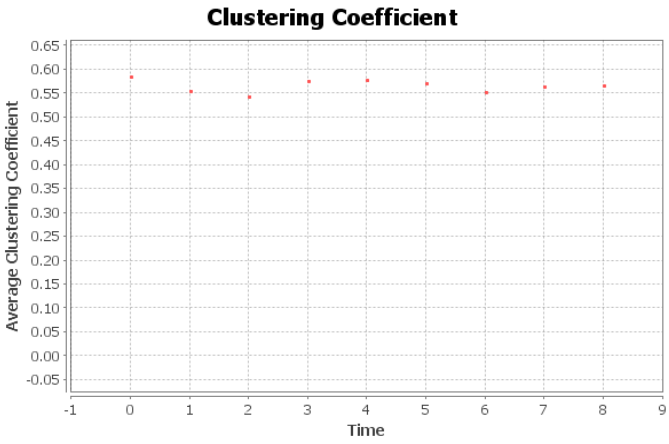


Fig. 1. Clustering Coefficient for Clustered Graph snapshots

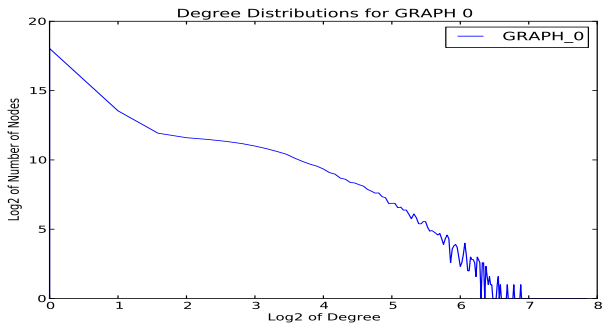
C. Compact Summary

We obtained the high level clusters and to visualize the high level summary of the graph snapshots, we retained only the most significant node of the cluster as representative so that providing the summary becomes easy. We believe some of the important node might not be available in summary because of this pruning process. However, we have made available the links for the final clustering results [2] and the pruned graph used for final high level summary [3]. Fig 6, 7 provide one summary for each graph snapshot; although we do have more detailed summary for some of the nodes in few graph snapshots, which can be found in Appendix.

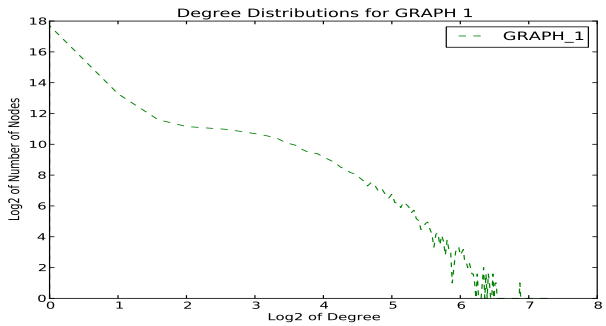
A large number of components with anonymization of node id's make it very difficult to observe the similarity or difference between different graph snapshots. However, we can clearly observe in **Fig 7 (d) and (e) having a common structure with two clique of size three having one node common between them**. The common node being important(influential) in both the graph snapshots also make it more likely. We also observed that most of the graph contained three node components with weak links with other node's. We can observe in Fig. 6 (a) that only the first graph snapshot has the only biggest, connected graph in the summary.

We also plotted the degree distribution of the resultant graph summary(Fig 3, 5) and noticed that the **summarized clustered graphs have retained their degree distribution** with respect to the original graph. So, we can conclude that the summa-

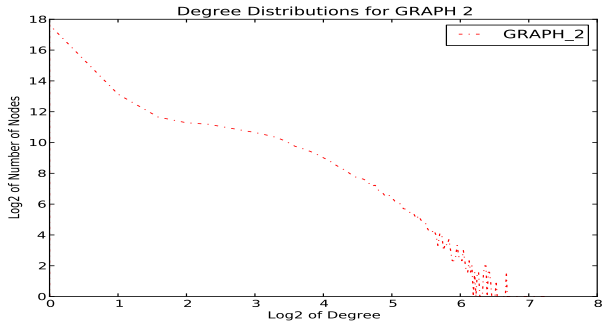
[2]https://www.dropbox.com/s/xh83xc43ll58i6b/level2.zip
[3]https://www.dropbox.com/s/27tb0aqbie8i606/simplified.zip



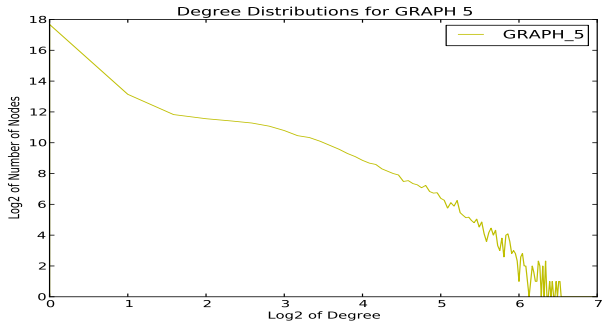
(a)



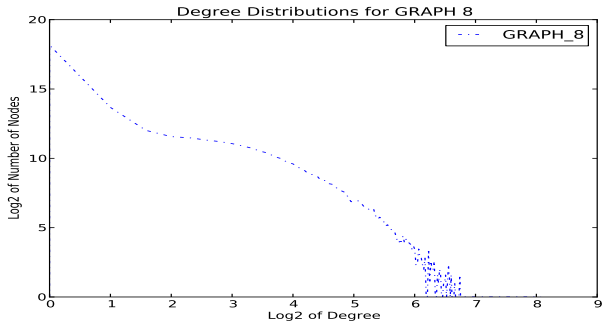
(b)



(c)

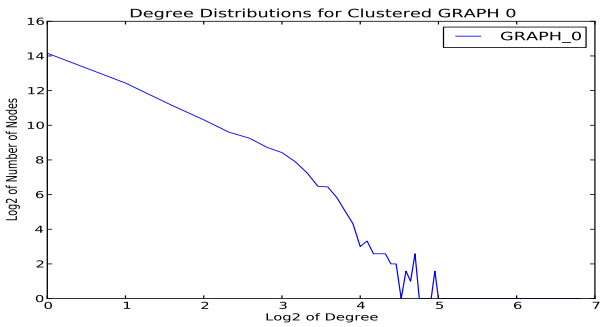


(d)

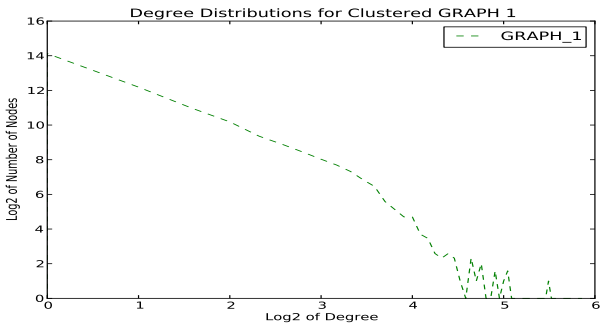


(e)

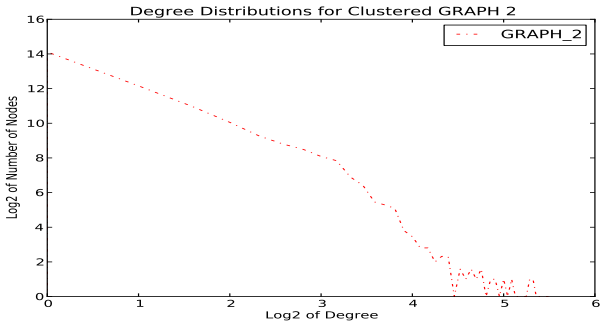
Fig. 2. Degree Distribution of original graph for (a) GRAPH_0 (b) GRAPH_1 (c) GRAPH_2 (d) GRAPH_5 and (e) GRAPH_8



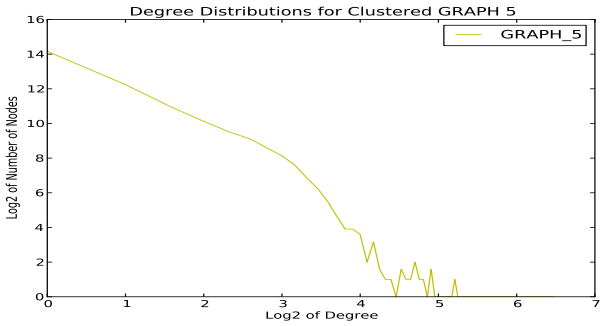
(a)



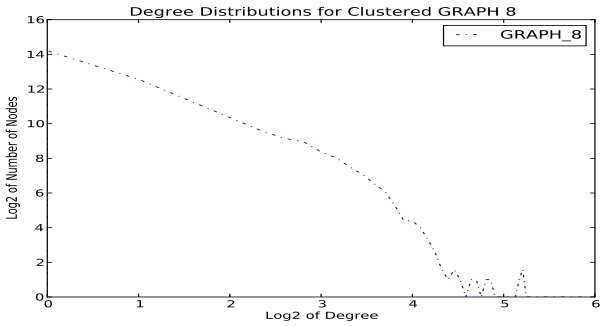
(b)



(c)

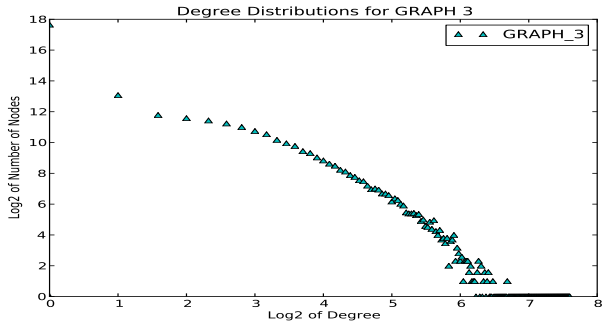


(d)

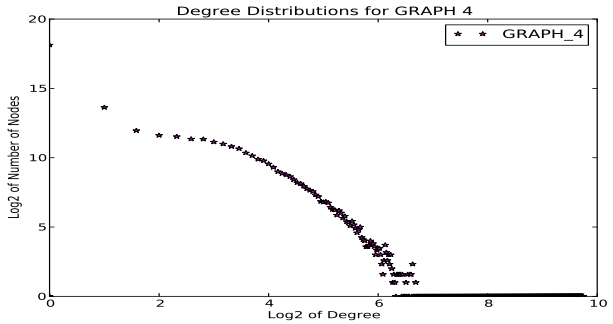


(e)

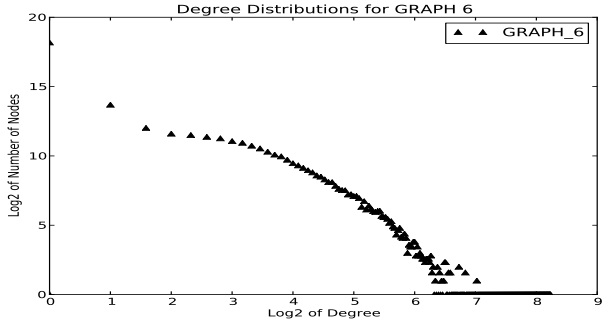
Fig. 3. **Degree Distribution of clustered graph** for (a) GRAPH_0 (b) GRAPH_1 (c) GRAPH_2 (d) GRAPH_5 and (e) GRAPH_8



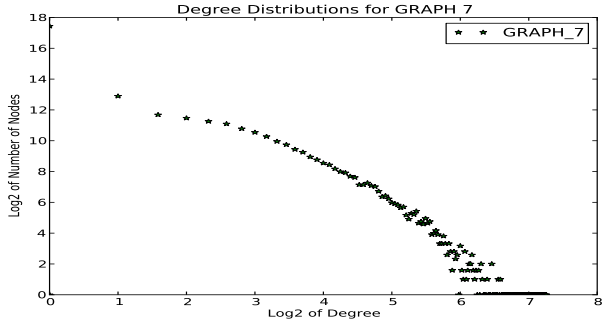
(a)



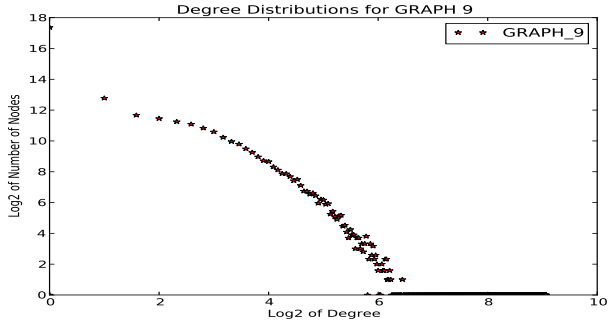
(b)



(c)

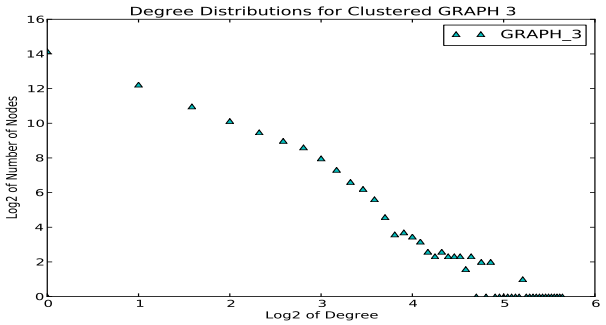


(d)

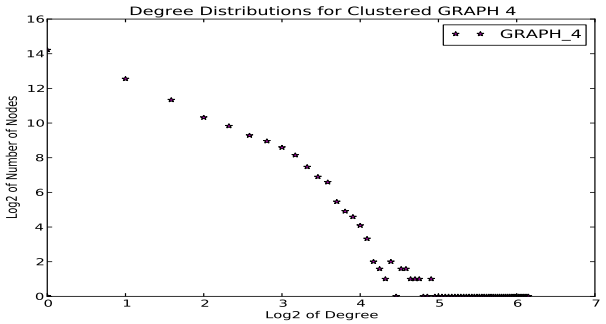


(e)

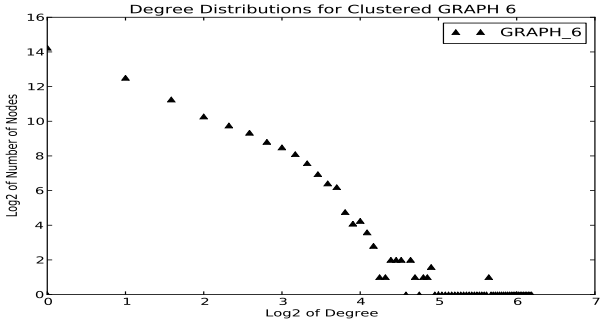
Fig. 4. Degree Distribution of original graph for (a) GRAPH_3 (b) GRAPH_4 (c) GRAPH_6 (d) GRAPH_7 and (e) GRAPH_9



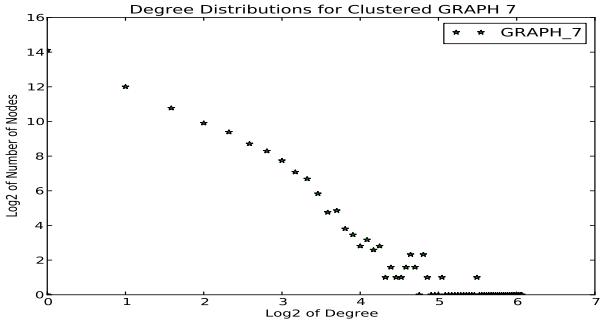
(a)



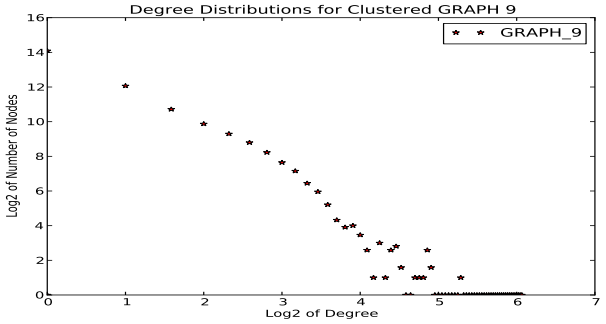
(b)



(c)

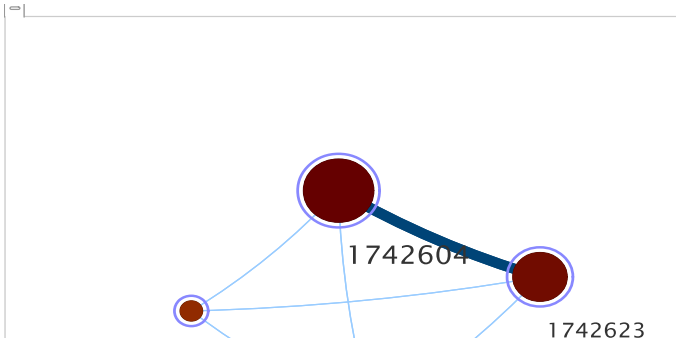


(d)

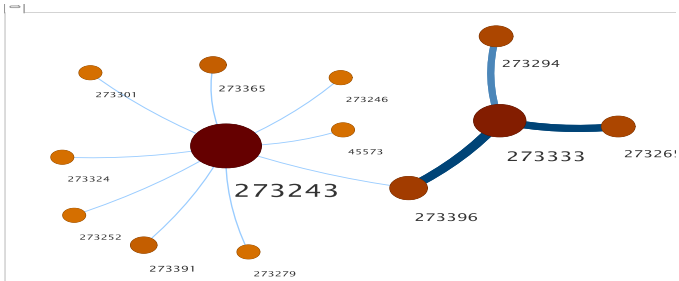


(e)

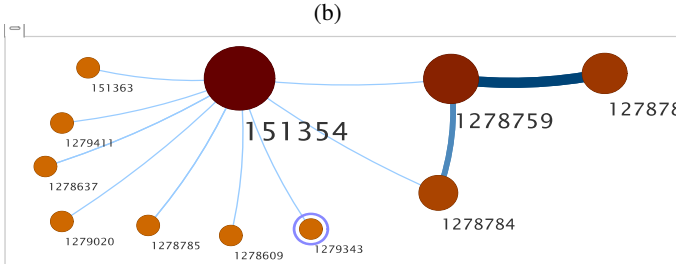
Fig. 5. **Degree Distribution of clustered graph** for (a) GRAPH_3 (b) GRAPH_4 (c) GRAPH_6 (d) GRAPH_7 and (e) GRAPH_9



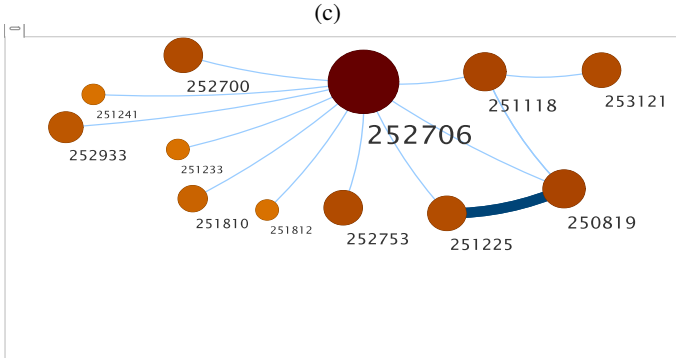
(a)



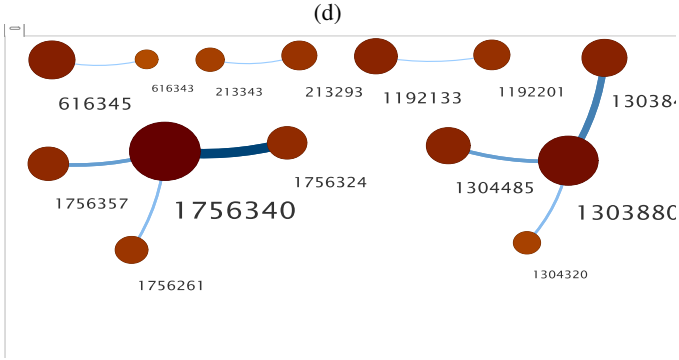
(b)



(c)

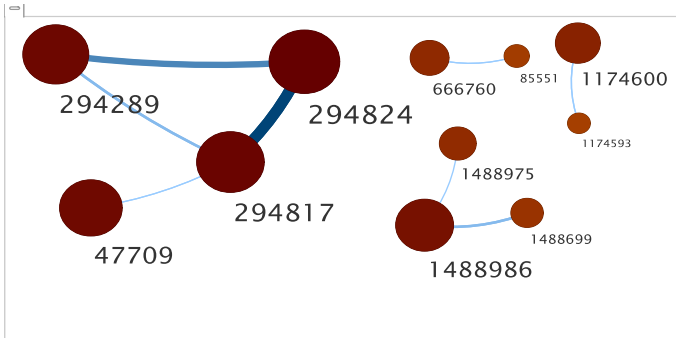


(d)

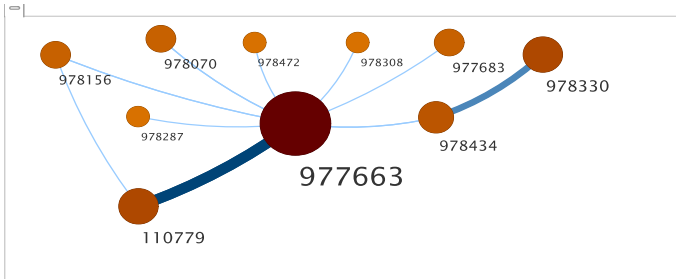


(e)

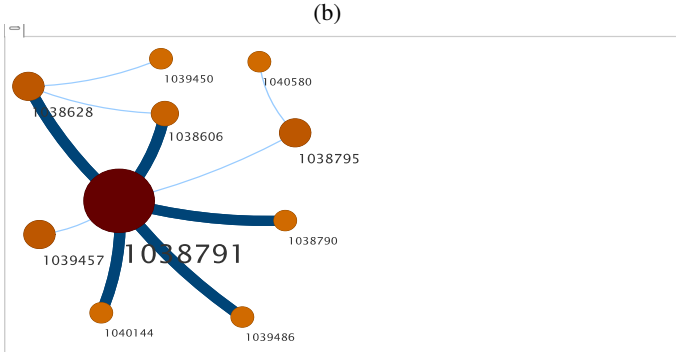
Fig. 6. High Level Cluster Summary for (a) GRAPH_0 (b) GRAPH_1 (c) GRAPH_3 (d) GRAPH_4 and (e) GRAPH_5



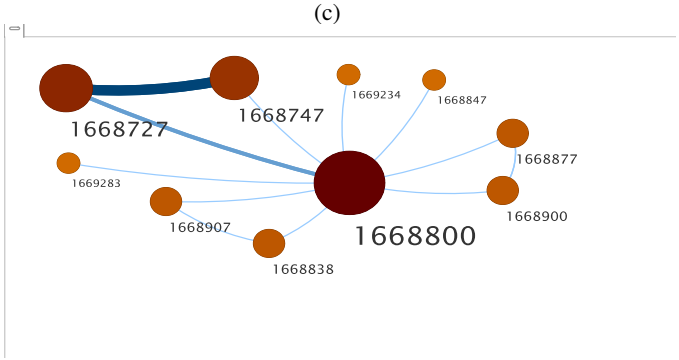
(a)



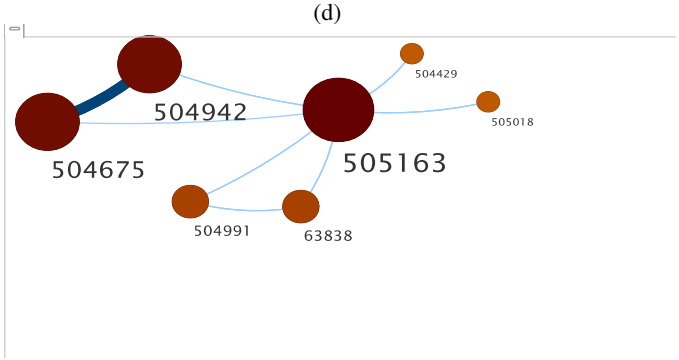
(b)



(c)



(d)



(e)

Fig. 7. High Level Cluster Summary for (a) GRAPH_5 (b) GRAPH_6 (c) GRAPH_7 (d) GRAPH_8 and (e) GRAPH_9

rized graph is good enough to study the properties of the network and only the non-important nodes were removed from the network. Improve clustering coefficient means that these summarized graph has better cohesiveness and yet retain the same degree distribution.

V. Conclusion

We analyzed the communication graph of the mobile network datasets and presented a high level, meaningful summary of the communication graph spanned over different snapshots in time. We used a flow-based information-theoretic method to obtain compact clusters in the graph. The clustering was based upon the idea of compressing the network with least amount of code-length necessary. Our method also provided the multi-level structure of the graph thus allowing both macro and micro-microscopic view of the graph. We noticed that despite clustering, the graph retained it's degree distribution and formed cohesive compact clusters centered around clique's of size three. We also provided our results for interested researchers to further study the property of this network.

References

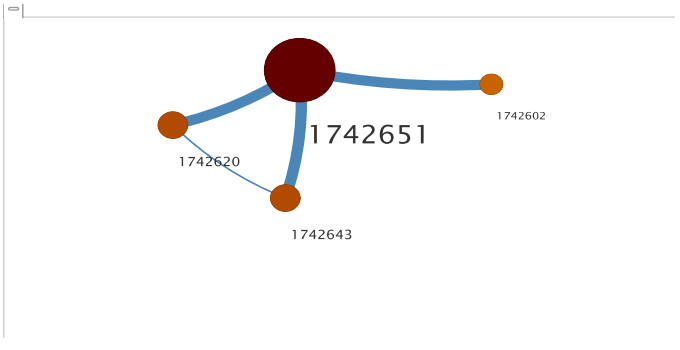
[1] G. Karypis, R. Aggarwal, V. Kumar, and S. Shekhar, "Multilevel hypergraph partitioning: applications in vlsi domain," *IEEE Trans. VLSI Syst.*, vol. 7, no. 1, pp. 69–79, 1999.

[2] S. van, "Graph clustering by flow simulation," in *PhD thesis, University of Utrecht, May 2000*, 2000. [Online]. Available: http://www.library.uu.nl/digiarchief/dip/diss/1895620/inhoud.htm

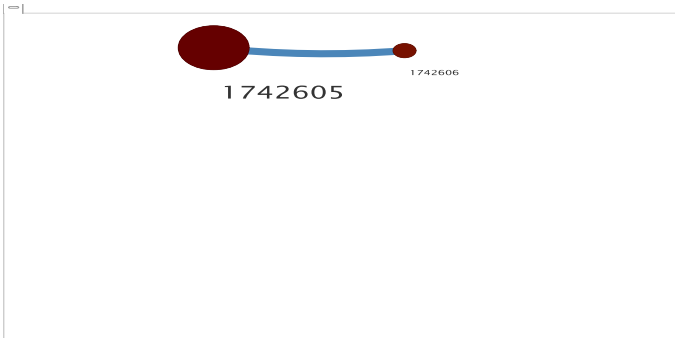
[3] M. Rosvall, D. Axelsson, and C. T. Bergstrom, "The map equation," *The European Physical Journal - Special Topics*, vol. 178, pp. 13–23, 2009.

[4] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, pp. P10 008+, Jul. 2008. [Online]. Available: http://dx.doi.org/10.1088/1742-5468/2008/10/p10008

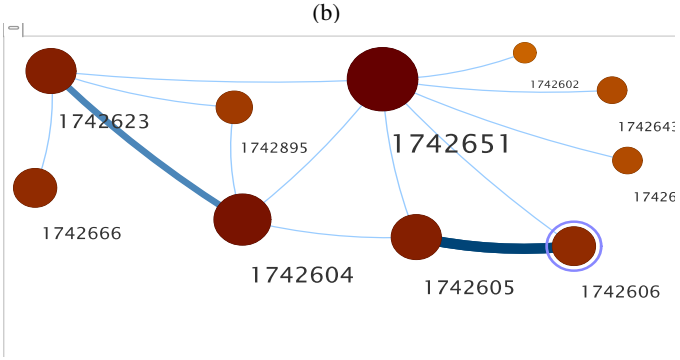
Appendix



(a)

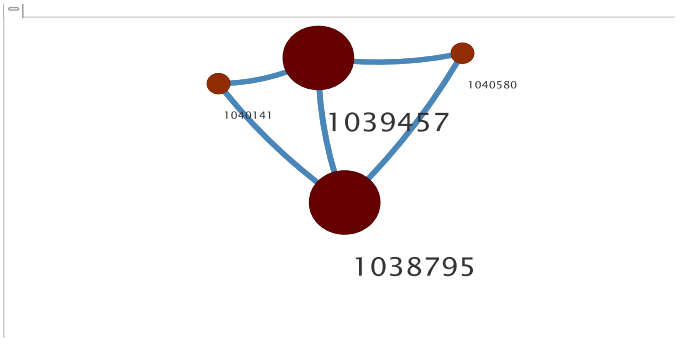


(b)



(c)

Fig. 8. Close-up summary of Nodes from Fig. 6 (a) **(GRAPH_0)**



(a)

Fig. 9. Close-up summary of Nodes from Fig. 7 (c) **(GRAPH_7)**

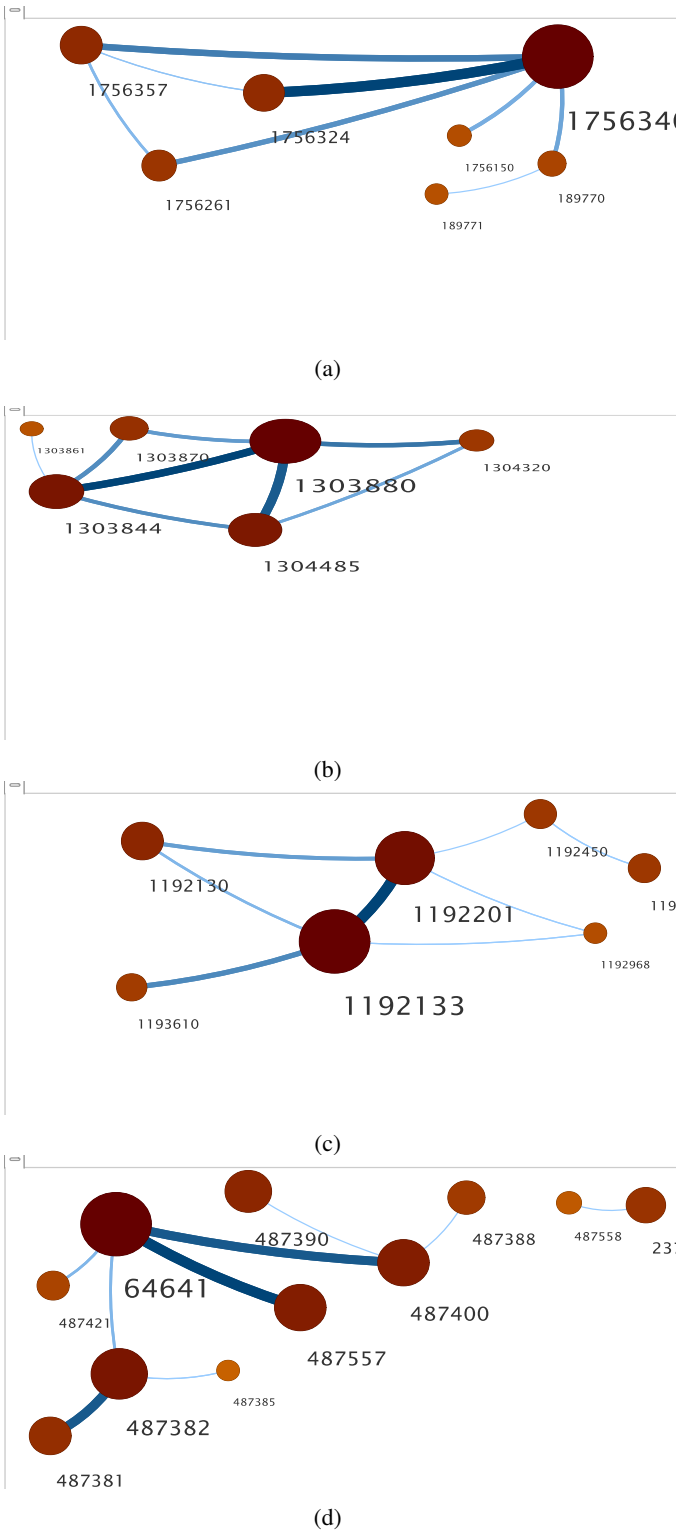


Fig. 10. Close-up summary of Nodes from Fig. 6 (e) **(GRAPH_2)**

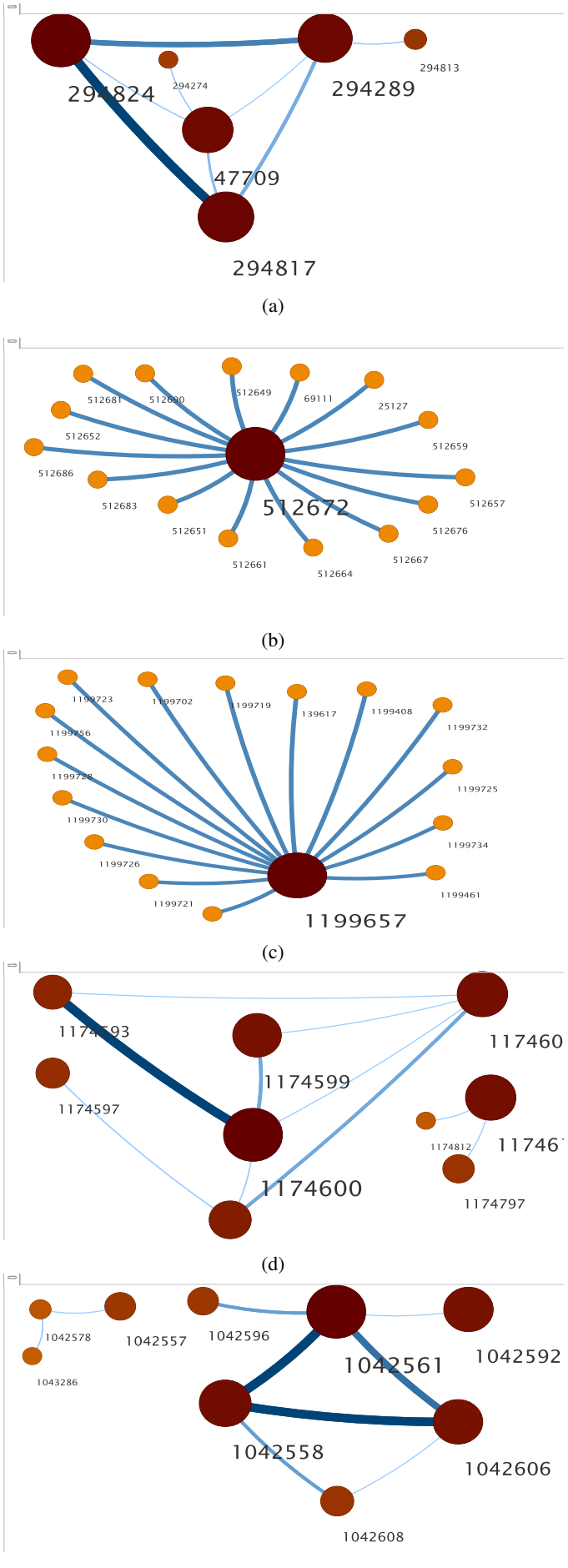


Fig. 11. Close-up summary of Nodes from Fig. 7 (a) **(GRAPH_5)**

Constrained link prediction on the D4D dataset

Bo Zong, Petko Bogdanov, Ambuj K. Singh
Department of Computer Science,
University of California at Santa Barbara
Santa Barbara, CA, USA
{bzong, petko, ambuj}@cs.ucsb.edu

ABSTRACT

Spatial data offers new opportunities and challenges for researchers to explore the factors that affect link formation in various networks. Graph models inspired by spatial data have been proposed to improve the support for applications, such as link prediction, community detection, and so on. In this paper, we focus on link prediction problems. In particular, we formulate a constrained link prediction problem: given a weighted graph G, a node v of G, a newly incoming node u, and a threshold h, the goal is to estimate the likelihood that node u and node v form an edge such that the weight of the edge is no less than h. We applied existing techniques to solve the link prediction problem: (1) a preferential attachment based method; (2) a preferential attachment and nodes' distance based method; and (3) a latent parameter model. Moreover, we conducted an experimental study to (1) investigate how these techniques perform on the D4D dataset by varying the threshold, and (2) analyze the factors that influence their performance.

1. INTRODUCTION

The increasing availability of spatial data brings new opportunities as well as challenges for researchers to uncover the factors that govern link formation in various networks. Efforts have been made to incorporate spatial information into graph models and facilitate a variety of applications, such as link prediction [7, 10] and community detection [3, 5, 7]. In this paper, we focus on the link prediction problem in spatial networks. In particular, we formulate and study a variant of link prediction tasks, referred to as *constrained link prediction*, on the D4D dataset.

A spatial graph is featured by node locations in addition to its link structure. A constrained graph G_h is an unweighted graph obtained from an edge weighted spatial graph G by only keeping those edges of weight exceeding a predefined threshold h. The constrained link prediction problem aims to estimate the likelihood that a newly incoming node u and a node v in G_h will establish an edge of weight exceeding h.

Several important applications are related to this problem. For example, in a data cleaning task, we are given the source node and the weight of an edge, and we are asked to recover the missing destination node. Another example is anomaly detection: if we have a model that accurately predicts links, a decrease in prediction performance might result from anomalous events.

There have been several techniques that model and predict link formation in spatial networks. First, one can apply a method based on preferential attachment [2] to approach the constrained link prediction problem, where the spatial information is simply ignored. More recent methods extend preferential attachment by considering the spatial distances among nodes [3, 5, 10]. These techniques have demonstrated modeling and prediction power on several spatial networks (*e.g.,* Internet [10], mobile networks [5], etc.). Third, Larusso et al. [7] proposed a latent radius model that incorporates preferential attachment, spatial information, and latent factors that affect link formation. This model provides new opportunities to improve modeling and prediction power for spatial networks. Moreover, it offers a new way to explore unknown factors that influence link formation. In this paper, we apply the above three types of techniques to the constrained link prediction problem, and compare their performance on the D4D dataset.

When we study and model how the adoption of technologies influences the economic growth in the developing world, we need to take a variety of factors (*e.g.,* culture and social economy) into consideration. The effect of these factors may be varied by how technologies are used, as well as the rate of their adoption [4]. In the specific context of mobile phone usage and the *D4D* challenge, such factors might be (1) the high price of long distance calls, (2) the skewed population distribution due to fast urbanization, (3) non-homogeneous infrastructure, and many others. As it is intractable to incorporate all such factors into a single model, one realistic solution is to apply latent parameter models to summarize these factors. In this project, we applied the state-of-the-art latent radius model [7], and investigate how much latent factors influence link formation in spatial networks among antennas.

The constrained link prediction problem, involving spatial information and latent factors, is the first step towards exploring unknown factors and their effect on the usage of communication networks. Its application may allow a better design/optimization on context-specific networks. In addition, the interpretation of the learned latent factors is an obligatory step towards understanding and optimizing a large scale

communication system in a developing country.

The rest of this paper is organized as follows. We formally define the constrained link prediction problem in Section 2. Data preprocessing is presented in Section 3. We provide an experimental comparison among existing techniques on the constrained link prediction problem in the context of the D4D dataset in Section 4. Section 5 concludes this paper, and discusses future work.

2. PROBLEM DEFINITION

In this section, we introduce the notation and define the constrained link prediction problem.

Spatial communication graph. A spatial communication graph G is represented by a tuple (V, E, s, f), where (1) V is a set of nodes, (2) $E \subset V \times V$ is a set of edges, (3) $s : V \to \mathbb{R} \times \mathbb{R}$ is a function that assigns a geographical coordinate to each node, and (4) $f : E \to \mathbb{R}$ is a function that assigns a numerical value to each edge as its weight. In particular, we focus on undirected communication graphs in this work.

Constrained graph. Given a spatial communication graph $G = (V, E, s, f)$ and a numerical value h referred to as a *threshold*, a constrained graph G_h is represented by a tuple (V, E_h, s), where $E_h = \{(u, v) \mid (u, v) \in E \wedge f(u, v) \geq h\}$. In other words, G_h is an *unweighted* subgraph of G containing those edges whose weights are no less than the threshold.

Now we are ready to define the *constrained link prediction* problem.

Definition 1 (Constrained link prediction). *Given a spatial communication graph $G = (V, E, s, f)$, a threshold h, a node $u \in V$ and a newly incoming node $v \notin V$, we aim to estimate the likelihood that v and u form a link on the constrained graph G_h.*

We consider the constrained link prediction problem for the following reasons. (1) Despite the difficulties of estimating the exact values of edge weights [1], one might have more interest in the scale of edge weights (*e.g.,* which are the top-k most likely links that a node u will build with communication duration longer than $10,000$ seconds?). Constrained link prediction provides a formulation of this kind of queries. (2) There have been techniques that model link formation on unweighted spatial networks [2, 5, 7, 10], and those techniques provides potential solutions to constrained link prediction. In this work, we consider unweighted constrained graph, and explore how well the existing techniques work on this problem.

3. DATA PREPARATION AND ANALYSIS

In this section, we describe how we process the data, construct spatial communication graphs, and obtain constrained graphs. Moreover, we analyze characteristics of the data by graph metrics.

3.1 Spatial communication graphs

We start with the methodology of constructing spatial communication graphs, and then we analyze the graphs' characteristics.

3.1.1 Building graphs

The spatial communication graphs are obtained from two datasets, SET1 and ANTPOS, in the D4D data: (1) SET1 contains 10 trunks of communication records among 1231 antennas, and each trunk of data includes records for two weeks; and (2) ANTPOS includes the antennas' geographical locations. Given a trunk of communication records spanning two weeks and ANTPOS, a spatial communication graph is constructed in four steps.

First, inconsistent records are removed. There are less than 3% of records in SET1 with missing destination antennas. In this work, we adopt a simple strategy to clean the data: remove the records with incomplete information.

Second, self-communication records are ignored. There are less than 2% of records in SET1 that are communication from an antenna to itself. In this work, we ignore those self-communication records.

Third, records for the same pair of antennas are merged. In each record, three entries are considered: (1) source antenna, (2) destination antenna, and (3) communication duration. We consider that the communication is undirected by ignoring the roles of source and destination, and merge communication records for the same pair of antennas by summing up their communication duration.

Fourth, a spatial communication graph is constructed based on the processed set of records: (1) nodes are antennas; (2) there is an edge between two nodes, if there is a record specifying the communication between the two nodes; (3) the edge weight of an edge is the total communication duration in the record; and (4) the geographical locations of nodes are extracted from the dataset ANTPOS.

3.1.2 Graph analysis

From the 10 trunks of records, we obtain 10 spatial communication graphs. Table 1 presents average and maximum node degree of the 10 spatial communication graphs. Moreover, node degree distribution and edge weight distribution of a spatial communication graph are demonstrated in Figure 1(a) and Figure 1(b), respectively. Since the 10 spatial communication graphs demonstrate similar distributions for node degree and edge weight, in this paper, we only show the distributions of the graph extracted from the first trunk of records.

The characteristics of the extracted spatial communication graphs are summarized as follows. (1) The spatial communication graphs are dense. As shown in Table 1 and Figure 1(a), the density of the graph is raised by a large portion of high degree nodes, and these nodes make the graph almost fully connected. (2) From Figure 1(b), we can see the weight (total call duration) distribution demonstrates a power law behavior [1, 8, 9].

3.2 Constrained graphs

In the following, we introduce how we obtain constrained graphs, and present the characteristics of these graphs.

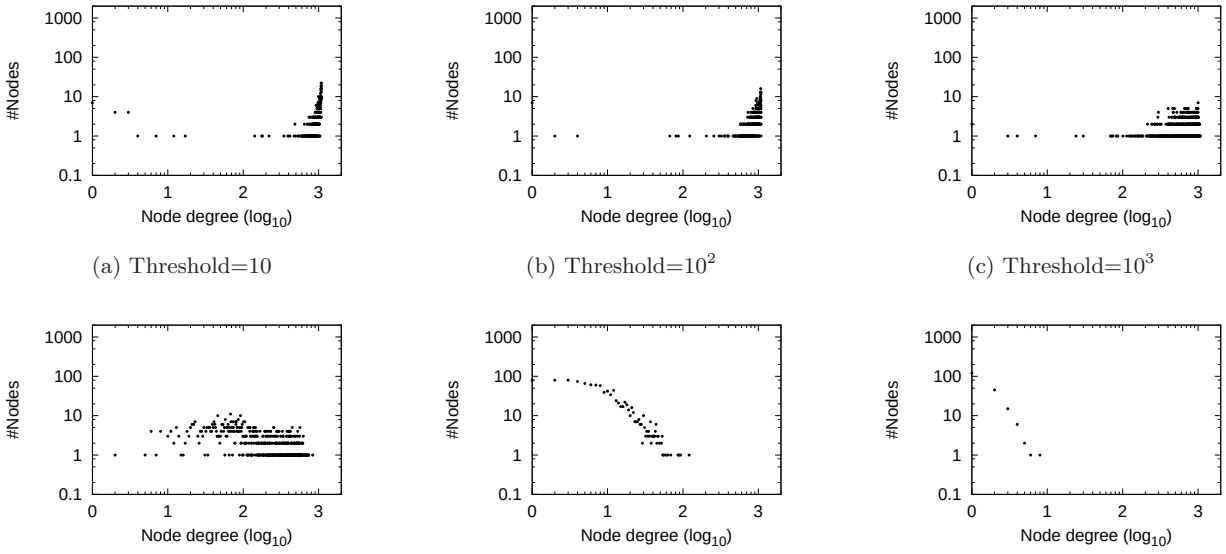
Building graphs. For each extracted spatial communication graph, we set the threshold h to be 10, 10^2, 10^3, 10^4, 10^5, or 10^6, and therefore obtain 6 constrained graphs.

Graph analysis. In the following, we first show how degree distribution changes when the threshold increases, and then show how the size of the largest connected component changes, when the threshold is varied.

Since the 10 sets of constrained graphs obtained from 10

Graph id	1	2	3	4	5	6	7	8	9	10
Avg	872	636	666	750	759	655	609	487	938	892
Max	1093	1073	1075	1062	1060	1007	971	895	1200	1198

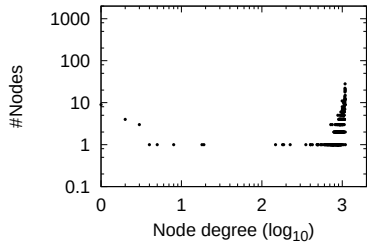
Table 1: Average and maximum node degree in spatial communication graphs



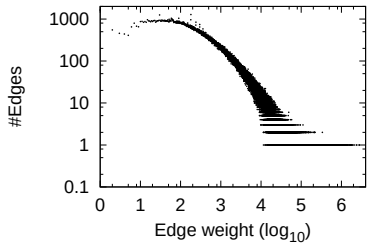
(a) Threshold=10 (b) Threshold=10^2 (c) Threshold=10^3

(d) Threshold=10^4 (e) Threshold=10^5 (f) Threshold=10^6

Figure 2: Node degree distribution of a set of constrained graphs



(a) Degree distribution



(b) Weight distribution

Figure 1: Node degree and edge weight distribution of a spatial communication graph obtained from the D4D dataset

spatial communication graphs show similar characteristics, in the following, we only present the degree distributions of a set of 6 constrained graphs that are obtained from the spatial communication graph that models the first trunk of data.

Figure 2 demonstrates the degree distributions of a set of constrained graphs. When the threshold is 10, there is still a large number of nodes of high degree. As threshold increases, the high-degree nodes gradually become medium-degree nodes, and the medium-degree nodes become low-degree nodes. When the threshold becomes 10^5, the node degree distribution demonstrates a power law behavior. When the threshold increases to 10^6, there are only several low-degree (≤ 10) nodes left.

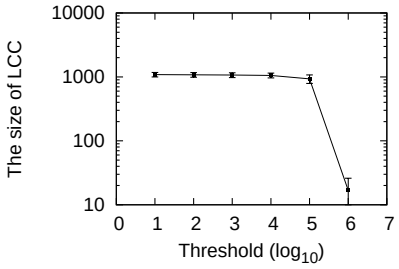


Figure 3: The size of the largest connected components changes with increasing threshold

Figure 3 illustrates the mean and standard variance on the size of the largest connected component (LCC) over all constrained graphs. We make two observations: (1) when the threshold increases from 10 to 10^5, the size of the largest connected component does not change much; however, (2) when the threshold becomes 10^6, the largest connected component is broken into pieces.

4. EXPERIMENTAL RESULTS

In this section, we present an experimental study to investigate how well existing techniques solve the constrained link prediction problem on the D4D dataset.

4.1 Experiment setup

Existing techniques. In this work, we applied three techniques to the constrained link prediction problem: (1) a preferential-attachment based method (referred to as PA), (2) a technique that extends the preferential-attachment based method by considering nodes' distance (referred to as PADist), and (3) a latent node radius based method (referred to as LR).

PA exploits the topological information of graphs to make predictions [2]. Given an incoming node u, the likelihood that a node v will link to u is proportional to $k_v^{-\gamma}$, where k_v is the degree of node v, and γ is a model parameter learned from the training data. In other words, the node of a higher degree always has a higher likelihood to be linked. One might notice that PA does not exploit all the available information. In particular, the spatial information of nodes is ignored in PA.

PADist is a technique that uses both topological and spatial information [3, 10]. The intuition is if two nodes are closer in the geographical space, they are more likely being linked. In particular, the likelihood that an incoming node u forms a link to node v is proportional to $k_v d_{uv}^{-A}$ [10], where d_{uv} is the distance between node u and node v, and $A > 0$ is a model parameter.

LR is a latent parameter model that considers topological as well as spatial information, and uses latent node variables to summarize other potential factors that affect link formation [7]. The general idea of LR includes (1) instead of node popularity/degree and nodes' distance, there might be other factors that affect link formation such as nodes' surroundings (*e.g.,* network structure and node density); and (2) to capture these information, a latent radius is attached to each node to summarize its *spatial reach* (*i.e.,* an incoming node is more likely to link to a node of a larger latent radius). In particular, the likelihood that a node u forms a link to a node v is proportional to $\frac{1}{\alpha}(r_v - d_{uv}) - \frac{k_v}{\gamma}$, where (1) r_v is the latent radius of node v, (2) d_{uv} are the spatial distance between u and v, (3) k_v is the node degree of v, and (4) α, γ are model parameters.

Training/testing data. Given a constrained graph G_h, a pair of training/testing data is constructed as follows. (1) Let u be a node in G_h with node degree larger than 0. We first remove all the links that attach to u, and the resulting graph is the training data. (2) Meanwhile, we mark node u as a testing data, and the removed links connecting u are the ground truth. Therefore, if G_h has n nodes of non-zero degree, we will obtain n pairs of training/testing data.

4.2 Prediction performance

We used two metrics to measure the prediction performance of the applied techniques.

Precision at k. The first metric is precision at k[1]. Let A_k be an edge set that contains the top-k edges by their estimated likelihood, and T be the ground truth edge set that contains the edges in the original graph. Precision at k is calculated by $\frac{|A_k \cap T|}{k}$.

In the experiment, we set k to be 1 for the following reason. Given a test data u, the size of its ground truth set is not fixed. If k is set to a large value for the cases of small ground truth sets, the measurement result is not fair. Since the ground truth sets of test data are never empty, it is always safe to set k to be 1.

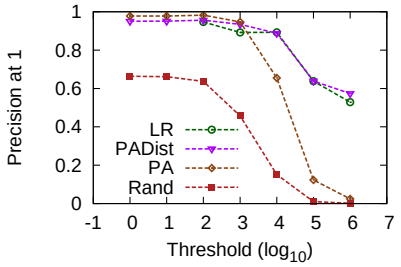


Figure 4: Precision of existing techniques

The performance of the applied techniques on precision at 1 is presented in Figure 4. Note that the performance of a random algorithm, referred to as Rand, is illustrated for reference. For each testing data, Rand randomly picks an edge, and mark this edge as the one of the highest likelihood. Three observations can be made: (1) all the techniques show a decreasing performance when the threshold increases; (2) when the threshold ranges from 1 to 10^3, all the techniques demonstrate similar performance; and (3) when the threshold goes beyond 10^3, the performance of PA rapidly decreases; however, PADist and LR still have a similar performance. When the threshold is no larger than 10^3, PA works well because there are a large number of high-degree nodes such that when PA picks the edge of the largest degree as the top-likelihood edge, it is very likely that PA picks the right edge. However, when the threshold goes beyond 10^3, the number of high-degree nodes rapidly decreases, and the performance of PA decreases rapidly as well. In contrast, PADist and LR take spatial information into consideration. Even when constrained graphs are very sparse at threshold 10^6, their precision is no less than 50%.

Normalized discounted cumulative gain. The second metric is normalized discounted cumulative gain, referred to as nDCG [6]. Let R_u be a ranked list of edges sorted by their estimated likelihood and T be the ground truth edge set. For the rank i edge $e_i \in R_u$, the credit of e_i is defined by

$$credit(e_i) = \frac{\delta_i}{\log_2 i}$$

where if $e_i \in T$, $\delta_i = 1$; otherwise, $\delta_i = 0$. Moreover, we define $nDCG = \frac{1}{Z}[\delta_1 + \sum_{i=2}^{|R_u|} credit(e_i)]$, where Z is a normalization coefficient defined by $Z = 1 + \sum_{i=2}^{|T|} \frac{1}{\log_2 i}$. Intuitively, a ranked edge list R_u has higher nDCG, if there are more edges in T having ranks higher than those edges that are not in R_u.

Figure 5 presents the nDCG performance of the three techniques. Note that the performance of a random algorithm, referred to as Rand, is demonstrated for reference. For each testing data, Rand randomly generates ranks for all possible edges. We make four observations: (1) all the three tech-

[1]http://en.wikipedia.org/wiki/Precision_(information_retrieval)

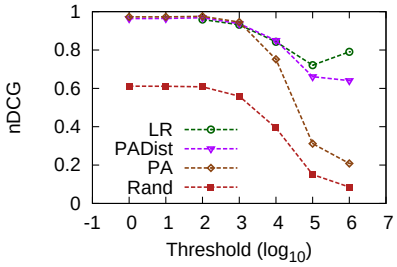


Figure 5: Normalized discounted cumulative gain

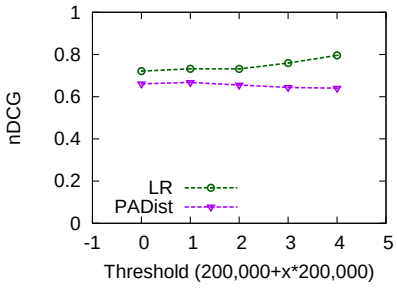


Figure 6: The nDCG performance of PADist and LR methods on the constrained graphs with threshold ranging from $200,000$ to $1,000,000$

niques have a high nDCG, when threshold ranges from 1 to 10^3; (2) the performance of PA rapidly drops as the threshold goes beyond 10^3; (3) as the threshold increases from 10^4 to 10^5, the nDCG performance of PADist drops faster than that of LR; and (4) when the threshold increases from 10^5 to 10^6, the nDCG performance of PADist keeps falling; however, we observe an increasing nDCG for LR. Similar to the precision results, the nDCG performance of PA rapidly decreases after threshold goes beyond 10^3, and a rapid decrease on the number of high-degree nodes causes the drop of performance. PADist and LR consider spatial information such that even when the number of high-degree nodes largely decreases, their nDCG values is still higher than 0.6. Moreover, as shown in Figure 6, we observe an increasing nDCG for LR, when the threshold increases from 10^5 to 10^6. As the threshold is approaching 10^6, there might be some latent factors that dominate link formation instead of node degree and nodes' distance. LR is able to capture these latent factors such that when the threshold is 10^6, LR performs 10% better than PADist in case of nDCG.

4.3 Summary

In this section, we conducted an experimental study to investigate how existing techniques solve the constrained link prediction problem. We applied three techniques: PA, PADist, and LR, and measured their prediction performance on two metrics: precision at k and nDCG. We observe that (1) on both precision and nDCG, all the three techniques perform well on low-threshold constrained graphs; (2) in case of precision, PA's performance drops rapidly, when the threshold goes beyond 10^3, while the performance of PADist and LR drops gradually; and (3) in case of nDCG, while PADist still performs better than PA since it considers spatial information, LR obtains the best performance by capturing latent factors that affect link formation.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we studied the constrained link prediction problem on the D4D dataset. We conducted an experimental study to explore how existing techniques perform when applied to the constrained link prediction problem. In particular, three techniques were evaluated: a method based on preferential attachment model (referred to as PA), a method considering both preferential attachment and spatial information (referred to as PADist), and a latent parameter method that captures preferential attachment, spatial information, and latent factors that affect link formation (referred to as LR). In case of precision, LR and PADist have similar performance, and both of them perform better than PA. In case of nDCG, PADist performs better than PA, and LR has the best performance.

Our future work is summarized as follows.

- First, we are going to extend the latent radius model to support more general applications. This evaluation is a first step towards our long-term goal of understanding the latent factors (outside space and network connectivity) that shape network flow in such communication networks. The extensions to the latent radius approach (LR) include the following directions. (1) One extension is to directly model edge weights. This extension is useful for traffic analysis and prediction. (2) Another extension is to model temporal evolution of the traffic flow. This extension allows multi-resolution temporal analysis, as opposed to fixed (*e.g.,* 2 weeks) periods. (3) The third extension is to model multigraphs (there are parallel edges between a pair of nodes). This extension can be applied on data cleaning: communication records in the D4D dataset can be considered as multi-edges among antennas, and recovering the missing destination in the records can be formulated as a link prediction problem on multigraphs.
- Second, we are going to explore the interpretation of the learned latent radii. We plan to correlate the latent radii with external information such as transportation networks and population density. We would seek to interpret the interleaved factors modeled by the estimated latent radii. We can then explore network design problems, including (1) what is an optimal positioning of new antennas (*e.g.,* to increase the network traffic or to satisfy certain conditions in the latent radii space)? and (2) how does external information affect such decisions?

6. REFERENCES

[1] L. Akoglu, M. McGlohon, and C. Faloutsos. RTM: Laws and a recursive generator for weighted time-evolving graphs. In *Proceedings of ICDM'08*, pages 701–706, 2008.
[2] A. Barabási and R. Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.
[3] F. Cerina, V. De Leo, M. Barthelemy, and A. Chessa. Spatial correlations in attribute communities. *PloS one*, 7(5):e37507, 2012.
[4] M. D. Chinn and R. W. Fairlie. Ict use in the developing world: An analysis of differences in computer and internet penetration. *Review of International Economics*, 18(1), 2010.
[5] P. Expert, T. S. Evans, V. D. Blondel, and R. Lambiotte. Uncovering space-independent communities in spatial

networks. *Proceedings of the National Academy of Sciences*, 108(19):7663–7668, 2011.

[6] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002.

[7] N. D. Larusso, B. E. Ruttenberg, and A. K. Singh. A latent parameter node-centric model for spatial networks. *CoRR*, abs/1210.4246, 2012.

[8] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proceedings of SIGKDD'05*, pages 177–187, 2005.

[9] M. McGlohon, L. Akoglu, and C. Faloutsos. Weighted graphs and disconnected components. In *Proceedings of SIGKDD'08*, pages 524–532, 2008.

[10] S. Yook, H. Jeong, and A. Barabási. Modeling the internet's large-scale topology. *Proceedings of the National Academy of Sciences*, 99(21):13382–13386, 2002.

Interactive Visualization of Cellphone Network Data Using D3: The Case of Ivory Coast

Maria Virginia Rodriguez[1], Veena Mendiratta[3], Benedict Lim[1], Derek Doran[2], Diego Klabjan[1]
[1]Northwestern University, [2] University of Connecticut, [3]Bell Labs, Alcatel-Lucent
Contact Author: veena.mendiratta@alcatel-lucent.com

1. Introduction and Motivation

The ability to implement a plethora of data mining algorithms on large data sets to gain valuable information and insights empowers today's decision makers with decision support tools previously unavailable. Historic patterns, relationships as well as predictions are some of the valuable insights that can be garnered. Even though a primary focus of analyzing mobile phone datasets focus on revealing structural properties with the calling network (zang07, belik10, hidalgo08, nanavai08, ozgul11, onnela07, doran12), researchers also focused on using the datasets to capture communication and mobility patterns within a society. Recent advances in visualizing such complex patterns incorporate the social relationships among users (onnella07) integrated with their spatial positions (shen08). Techniques have also been proposed to visualize human mobility across a large city (kwan12), and to integrate sensor and mobile phone data for the purpose of infrastructure planning (pattath06). More recent advances integrate algorithmic techniques to intelligently collapse edges in network visualizations in order to remove ambiguity (luo12).

The contribution of this paper is to demonstrate the effectiveness of a popular data visualization language, D3, in representing cellphone traffic data. Visualization of cellphone data in current literature has so far been of a static nature, representing volume and network data in the static form such as heat maps, density graphs, etc. Latest work at Vincent Blondel's Research Group on Large Graphs and networks effectively showcases cellphone data as an animated time series displaying a heat map describing network traffic over time (blondel12). This allows for user interaction with the data set to customize visualizations, allowing the analyst to experiment with different methods of conceptualizing the data, and paving the way for future optimization or predictive analytics. We recognize the value of descriptive analytics in the analytics process, and in this paper aim to build on the interactive cellphone visualizations to create our own unique visualizations that will allow analysts to better explore the interactions between different regions in Ivory Coast.

Moreover, the visualizations were created in D3 (Data-Drive-Documents) which is a Javascript visualization library. This paper demonstrates the effectiveness of working in a D3 environment to create dynamic and interactive dashboards for visualizing temporal cellphone network datasets.

The paper is structured as follows: In the second section, we briefly describe the dataset. In the third section, we explain the three visualizations that were created for this paper: The first visualization is at the regional level: showing the volume of calls within each region over time. In the second visualization, we develop the visualization at the user level, describing the geographic dispersion of where users make phone calls. In the third visualization, we explain the visualization of the interaction of users between regions.

2. Description of Data

The data was provided by Orange in conjunction with the "Data for Development" (D4D) open data challenge. The data used in this analysis pertained to the phone calls made by 500,000 users from December 2011 to April 2012. Each call record is characterized by the id of the user

and date, hour and location from where the call was made.

For the purposes of this analysis and the dataset, Ivory Coast is divided into 255 sous-prefectures. The analysis in this paper takes into account the activity of only 237 of them since the Orange network antennas only cover 237 sous-prefectures. The majority of the sous-prefectures without antennas are located in the north of the country in areas with little development.

3. General Analysis of Data Visualizations

There are three main visualizations that can be accessed via http://www.klabjan.dynresmanagement.com/datavisualization/d4d.html. In this section, we describe the visualizations in detail, and how the visualizations can be used to supplement certain types of analysis, such as those of infrastructure planning.

3.1 First Visualization: Call Volume Heat Map

Description:
This interactive visualization in Figure 1 represents the number of calls from each sous-prefecture on a weekly basis in the form of a heat map. The user can cycle through 20 weeks of data to look at the temporal evolution of the call volumes. Each region on the Ivory Coast map is color coded so that the regions are indicated by a linear range of colors, where regions with the darkest colors have the highest number of calls, and the regions with the lightest colors have the least number of calls.

In addition to visually presenting the relative magnitudes of number of calls between regions, when a cursor is placed over each sous-prefecture, the dashboard also displays the total number of calls for that particular sous-prefecture. When a sous-prefecture is clicked on, the map is zoomed in and it is possible to see the locations of each of the antennas in that region.

Potential Uses for Analytics:
Analysts can use the visualizations to detect changes in relative call patterns between regions over time and link it with current events happening in Ivory Coast. Cellphone activity levels could be predictors of impending conflicts, or a way for the government to monitor economic and social activity in different regions. Visualizing cellphone activity levels can empower the government with an inexpensive way to monitor activities across the country and make appropriate policy adjustments to boost economic growth.

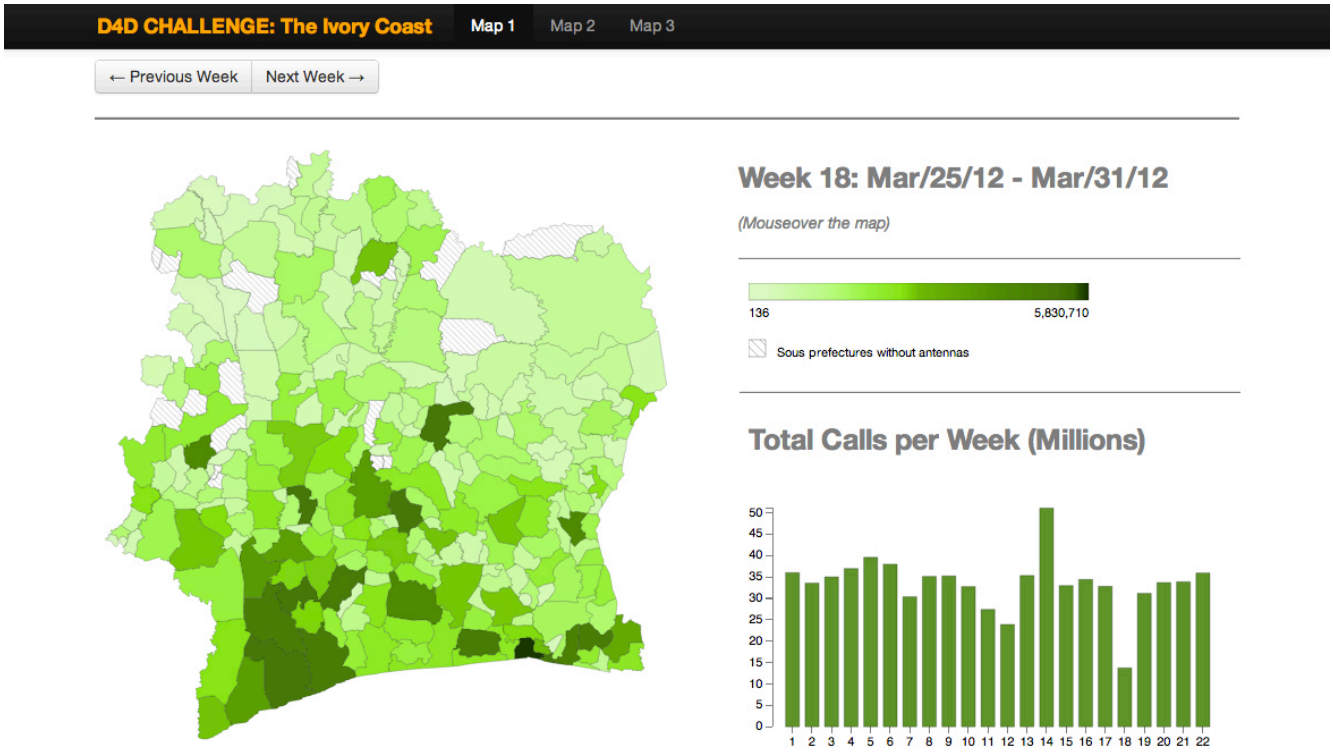


FIGURE 1: Call Volume Heat Map Visualization

3.2 Second Visualization: User Movement Graph

Description:
We focus on the subset of users who made calls from two separate sous-prefectures. The visualization in Figure 2 shows which other sous-prefectures users who make calls one sous-prefecture also make calls from. A mouse over a particular region would show lines to the other cities users physically made calls from, and the circles denote the total number of calls made by this subset of users in that sous-prefecture.

Potential Uses for Analytics:
One of the national priorities of the Ivory Coast is infrastructure development. User trajectories based on the mobile calling patterns show the movement of users between various locations. Trajectories that occur on a regular basis – for example, daily weekday pattern between 2 locations indicating commuting for work – can be used for planning of roads and public transportation. Other, more diverse mobility patterns show the variation of the mobility of the users by sub prefecture. This variation can provide a measure of the relative isolation (or "outgoingness") of the population of a sub prefecture which can be used for the allocation of development projects and for the development of longer range infrastructure projects.

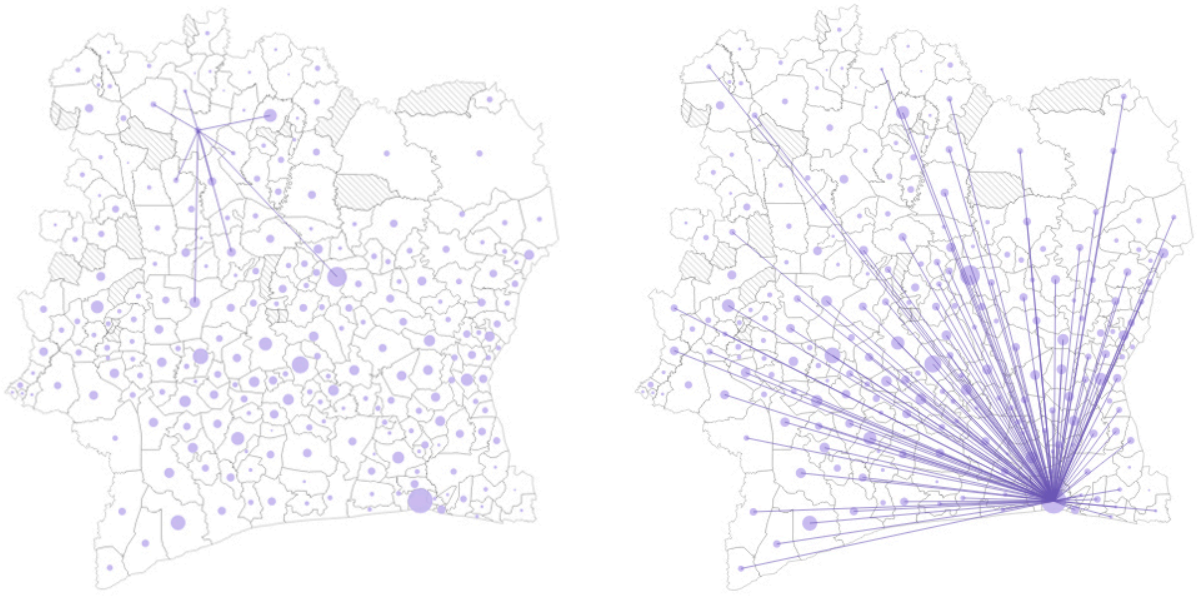


FIGURE 2: User Movement Graph

3.3 Third Visualization: Isolation Analysis

Description:

The visualization in Figure 3 represents the percentage of users within each sous-prefecture that only made calls within that sous-prefecture by displaying a circle of proportional size. When a cursor is placed over the region, the dashboard also displays the percentage and actual number of users that only made calls within that sous-prefecture.

Potential Uses for Analytics:

This analysis can be used to create a measure of isolation of the users in a sous-prefecture where the proportion of users that made calls from only one sous-prefecture is defined as the isolation index. Being able to identify sous-prefectures that have high isolation indexes in rural regions is extremely important. Improvements in communication and connectivity with other regions can act as a catalyst to open up these isolated regions to new technology, access to more advanced healthcare as well as to business opportunities. Sub-prefectures with a high isolation index would also imply that the residents are finding employment close to their place of residence, which is a source of demographic data for the government.

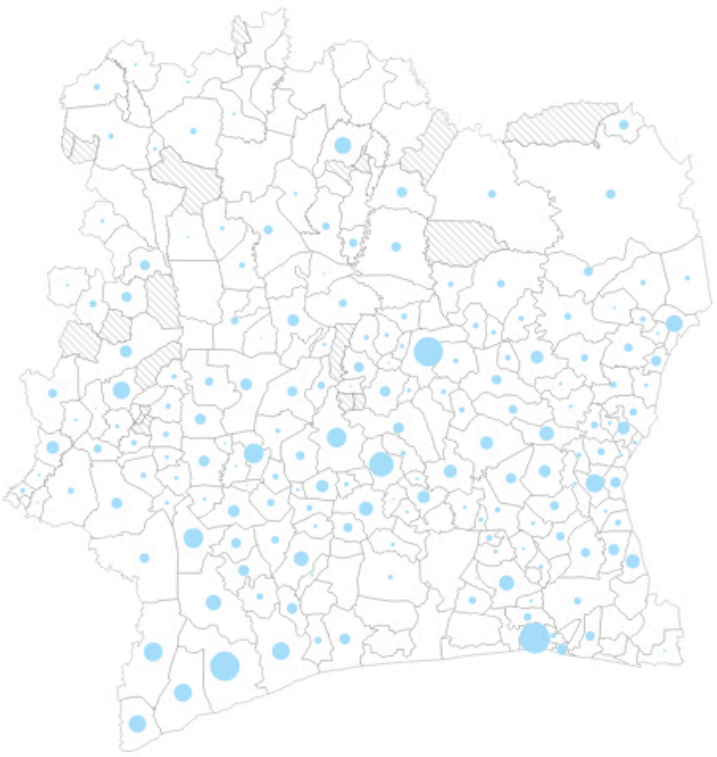


FIGURE 3: Isolation Analysis

4. Conclusion

In conclusion, we showed how cellphone data can be creatively manipulated with a powerful visualization tool like D3 and its potential uses for analytics. Call location data is a rich source of data for purposes such as infrastructure planning, however the analysis can be taken one step further if we can combine this data with data on user personal networks for future research.

One example of a synergy between call location and personal network data would be an analysis dedicated to finding groups of users who are physically isolated from geo-location data and also are in clusters of strong clique formations. Without geographic data, one could interpret users in this group as having strong community ties and high amounts of social capital. However, the physical isolation of this group of users would mean that these users are in fact not actively building connections and brokering relationships elsewhere in the country, and the government could invest in building roads or improving public transportation to help open up the region and improve their connectedness with the rest of the country – This would empower the government to be able to efficiently target and divert resources to regions with the most need.

5. Bibliography

[belik10] V. Belik, T. Geisel, and D. Brockmann, "Human movements and the spread of infectious diseases," in *Proc. of 1st Work- shop on the Analysis of Mobile Phone Datasets and Networks*, 2010, pp. 44–46

[ozgul11] Fatih Ozgul and Ahmet Celik and Claus Atzenbeck and Nadir Gergin. "Investigating Terrorist Attacks Using CDR Data: A Case Study". *Counterterrorism and Open Source Intelligence*, LNSN Volume 2, 2011, pp. 343-354

[nanavati08] A. Nanavati, R. Singh, D. Chakraborty, K. Dasgupta, S. Mukherjea, G. Das, S. Gurumurthy, and A. Joshi, "Analyz- ing the Structure and Evolution of Massive Telecom Graphs," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 5, 2008

[doran12] Derek Doran and Veena Mendiratta and Chitra Phadke and Huseyin Uzulangilou. "The Importance of Outlier Relationships in Mobile Call Graphs", Proc. of IEEE Intl. Conference in Machine Learning and Applications, 2012.

[blondel12] Blondel, Vincent. "Visualization of Cote D'Ivoire Cellphone Volume." Geofast Cote D'Ivoire. Research Group on Large Graphs and Networks, n.d. Web. 15 Feb. 2013.

[hidalgo08] C. Hidalgo and C. Rodriguez-Sickert, "The dynamics of a mobile phone network," *Physica A: Statistical and Theoreti- cal Physics*, no. 387, pp. 3017–3024, 2008

[onnela07] J.-P. Onnela, J. Saramaki, J. Hyvonen, G. Szabo, D. Lazer, K. Kaski, J. Kertesz, and A.-L. Barabasi, "Structure and tie strengths in mobile communication networks," *Proceedings of the National Academy of Sciences of the United States*, vol. 104, pp. 7332–7336, 2007

[zang07] H. Zang and J. Bolot, "Mining Call and Mobility Data to Improve Paging Efficiency in Cellular Networks," in *Proc. of 13th ACM International Conference on Mobile Computing and Networking*, 2007.

[kwan12] Matthew Kwan. "Visualization and analysis of mobile phone location data". PhD Dissertation, RMIT University, 2012.

[shen08] Zeqian Shen and Kwan-Liu Ma. "MobiVis: A Visualization System for Exploring Mobile Data". Proc. of IEEE Pacific Visualization Symposium, 2008, pp. 175-182

[luo12] Sheng-Jie Luo and Chun-Liang Liu and Bing-Yu Chen and Kwan-Liu Ma. "Ambiguity-Free Edge-Bundling for Interactive Graph Visualization", *IEEE Transactions on Visualization and Computer Graphics*, Vol. 18, no. 5, 2012

[pattath06] Avin Pattath and Brian Bue and Yun Jang and David Ebert and Xuan Zhong and Aaron Ault and Edward Coyle. "Visualization and Analysis of Network and Sensor Data on Mobile Device". Proc. of IEEE Symposium on Visual Analytics and Technology, 2006

D4D Challenge Submission

NVizABLE: A Web-Based Network Visualization Interface

Jennifer Smith
The Pennsylvania State University
Department of Geography
Pennsylvania, United States
jms1186@psu.edu

Joshua Stevens
The Pennsylvania State University
Department of Geography
Pennsylvania, United States
josh.stevens@psu.edu

Muhammed Y. Idris
The Pennsylvania State University
Department of Political Science
Pennsylvania, United States
muhammedy.idris@gmail.com

Abstract—The ultimate goal of the D4D challenge was to conduct research which "contribute[s] to the socioeconomic development and well-being of populations" [1]. The data provided consisted of over 2.5 billion communication records between 5 million anonymous users. Such high resolution data, while having many benefits, can be at times cumbersome and costly to analyze. As such we focused our efforts in developing a tool which assist analysts, through leveraging analytical and technical features, to explore and identify relationships within the data for their research; which we hope in line with the objectives of D4D and Group's Orange, will "relate to an objective and improved quality of life for all" [1]. We present a new web-based application – the Network Visualization and Big Data Learning Environment (NVizABLE) [2]. This application depicts network graphs in an exploratory, interactive, and web-based interface. This system will allow users to identify network characteristics through ad hoc exploration rather than pre-determined calculation. In this paper we introduce the tool as well as our motivation and development techniques.

I. Introduction

Data driven research is fundamental to any study which aims to contribute to the betterment of society. Without any rigorous empirics, governments, organizations, and analysts cannot objectively answer fundamental research questions such as what is the effect of X policy on Y outcome. The higher the resolution data the more robust our understanding of these phenomena can be. However, higher resolution data does not come without its own limitations. "Big data" or the collection of large or complex datasets can make analysis difficult and costly. Moreover, it may be unclear as to whether or not the associations or patterns identified are substantive or spurious. As such a new field in visual analytics has focused its efforts on the development of tools which allow analysts to leverage their analytical skills coupled with the technical advantages of modern computer systems to better make sense of large data sets.

We present one such tool to depict cell phone communications data, provided by the D4D challenge from Orange in the Ivory Coast, as network graphs in an exploratory, interactive interface. This system will allow researchers to identify network characteristics of cell phone communication data through ad hoc exploration rather than pre-determined calculation. This approach will provide analysts with the ability to generate hypotheses and reason about network relationships without extensive knowledge in network science or graph theory. Lastly, from this approach, network-based methodologies may become accessible to a wider variety of disciplines.

We dedicate the rest of this paper to walk D4D organizers as well as potential researchers step-by-step through our research project. In the next section we turn to data acquisition and preparation, followed by the NVizABLE tool. Lastly, as the tool we are presenting is a prototype we conclude with future development work to be completed in the coming months.

II. Data Acquisition and Preparation

A. *Network Data*

Network data were obtained from the Data for Development challenge (D4D) (held by the Orange Group), consisting of mobile phone data in the Ivory Coast. This dataset contains 2.5 billion records of phone calls and SMS communication among five million users spanning December 1, 2011 to April 28, 2012. To protect privacy, the Orange Group anonymized individuals. Achieving greater levels of privacy through the aggregation methods used by Orange is not a hindrance to fine-grained analysis, as such detailed information only exists with extreme uncertainty in this region. As [3] notes, mobile phone users in Côte d'Ivoire commonly share or loan their devices to others at a cost, thus attributing any particular usage to an individual is questionable at best. Furthermore, the Orange Group removed customers who were new or left the company in the 150 day period. The matching incoming and outgoing calls were paired to eliminate redundancy. This dataset contains the duration and total number of calls between antennas for each hour. Any call that extended beyond the hour period was placed within

D4D Challenge Submission

the hour the call originated in. Additional attributes include antenna ID and the corresponding geographic location. In light of advancements in technology, big data is currently being produced at an extremely high rate. This dataset affords the authors a unique challenge in developing a network geovisual analytics interface that can present the data in a coherent manner and allow users to explore and find new spatial and statistical patterns among masses of data for human behavior.

B. *D4D Cleanup and Preparation*

The original data exists as a tab separated values file (TSV) as shown in Figure 2.

```
2012-04-28 23:00:00 1236 786 2 96
2012-04-28 23:00:00 1236 804 1 539
2012-04-28 23:00:00 1236 867 3 1778
2012-04-28 23:00:00 1236 939 1 1
2012-04-28 23:00:00 1236 1020 6 108
2012-04-28 23:00:00 1236 1065 1 1047
2012-04-28 23:00:00 1236 1191 1 67
2012-04-28 23:00:00 1236 1236 18 2212
2012-04-28 23:00:00 1237 323 1 636
2012-04-28 23:00:00 1237 710 1 252
```

Figure 2. Screenshot of the D4D data in the TSV format.

The data are first passed through a Bash script that performs the following tasks:

1. The sed program is called to convert tabs and spaces to commas, remove special characters, and output a subset of records that do not result in a network with more than 500 edges. This produced a file of containing 971 records.
2. The new file is sent to a Python program that creates the network by connecting all source and target nodes. Characteristics such as call length, number, neighbors, and others are added to each node as additional attributes.
3. Once the network is created, more analysis is performed to calculate statistics for the entire network. These include average degree, average shortest path length, and clustering coefficients for both the entire network and individual nodes.
4. Three JavaScript Object Notation (JSON) files are then created. The first contains the network structure and node attributes, the second contains basic characteristics of the network, and the third contains explicitly geographic information about the network.
5. These files are exported and subsequently uploaded to the web server. Note that additional JSON files can be created as needed for other components of the application.

An example of the JSON file is shown in Figure 3 below.

```
{
  "directed": false, "graph": [],
    "nodes": [
    {"neighbors": 20, "source": "620", "target": "844", "duration": "103", "number": "3"},
    {"neighbors": 9, "source": "708", "target": "1023", "duration": "73", "number": "1"},
    {"neighbors": 11, "source": "1144", "target": "1144", "duration": "21", "number": "2",
    ]
}
```

Figure 3. Screenshot of an example JSON file.

The data cleanup and preparation stages are visually outlined in the figure below.

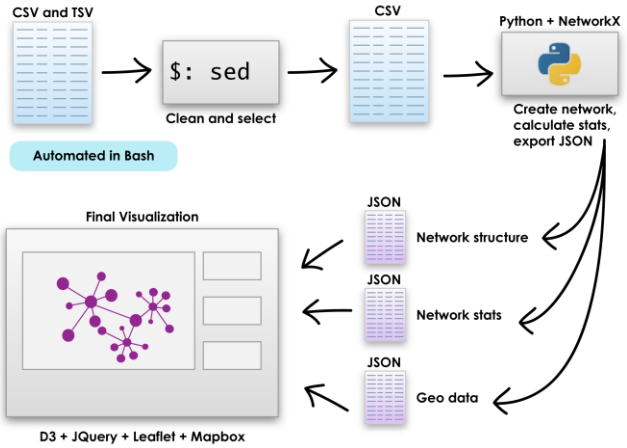


Figure 4. Data Flow Diagram of NVizABLE cleanup and preparation.

C. *Creating an Interactive Interface: NVizABLE*

Upon completing the previous steps, a web page reads the data in using the Data-driven Documents (D3) and JQuery libraries [12]. After the JSON is read, it is stored as an object that is used to create the interactivity of the application. D3 draws and animates the network while JQuery handles the asynchronous data updates, pop-ups, and text-replacement for the information windows.

Data-Driven Documents (D3) [5] is "an embedded domain-specific language for transforming the document object model based on data. With D3, designers selectively bind input data to arbitrary document elements, applying dynamic transforms to both generate and modify content" [6, p. 2302]. Of importance to processing and performance, D3 also allows developers to bypass extraneous computation (such as only transforming selected attributes). A map view displays the geographic distribution of the network and is built with Leaflet (a JavaScript library for interactive maps) and map tiles designed in Mapbox

D4D Challenge Submission

III. NVIZABLE

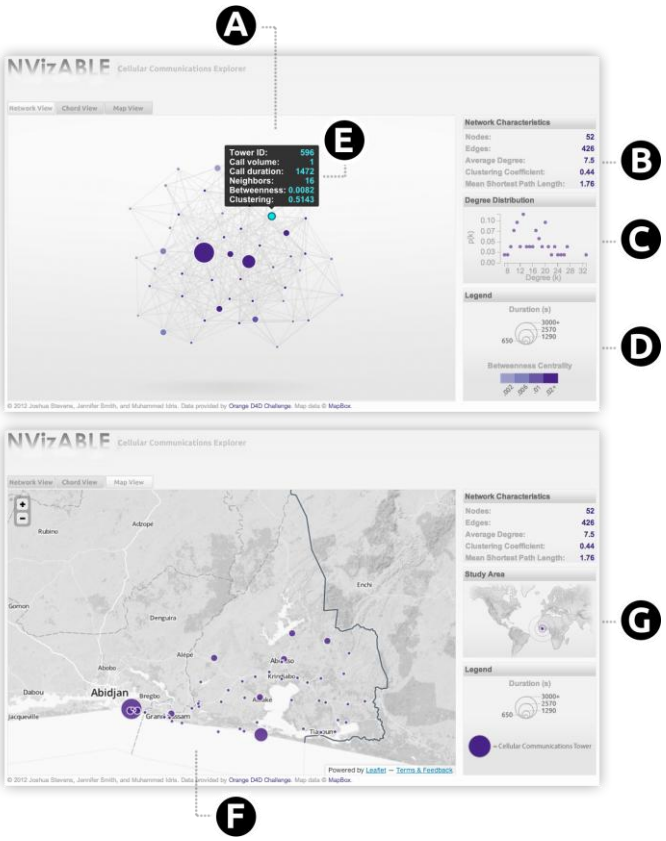


Figure 5. NVizABLE User Interface (*In Network View Tab: [*A: Main network frame, B: network characteristics frame, C: degree distribution frame, D: network legend, E: node information from mouse hovering] *In Map View Tab: [*F: Map frame, G: Study area overview inset])

NVizABLE offers basic network functionality allowing the user to interact and view the network within the interface through the following procedures:

Within the Network View tab:

1. Network Layout: NVizABLE supports a standard network layout view that allows users to see the overall network structure and connections among nodes (i.e. cell towers). NVizABLE employs a force-directed network view wherein the nodes are arranged by predefined gravities that dictate how close or far apart nodes may settle near their neighbors. Originally users were able to refine node positions and rotate the network further, though we found that this functionality impeded users' ability to keep track of node relationships. Due to the animated nature of force-directed graphs, it also became difficult to select individual nodes to activate their pop-ups. For these reasons, the current implementation of NViZABLE maintains the network arrangement once node positions are calculated. If a new arrangement is desired, users can simply reload the website.
2. Find the duration of calls for each cell tower and the corresponding betweenness centrality. These are visually encoded by the size and color of each node, respectively. The betweenness centrality is indicative of the overall importance of the node in the network. The embedded legend in the lower right-hand corner provides users with a constant reference to this information.
3. The degree distribution frame shows the probability for any of the included degrees, or number of neighbors for a node, to occur.
4. Network characteristics are found in the top right panel of the main network view. These characteristics provide overall statistics of the network including the total number of nodes, the total number of edges (or links), the average degree of the network, the overall clustering coefficient, and the mean shortest path length.
5. Hover over a node, allowing users to see the statistics that have been calculated for the particular node. In addition, the basic information of the node is displayed including the tower ID, number of calls, and total call duration within one hour, as well as network data statistics for the number of neighbors, betweenness centrality, and the clustering coefficient.

Within the Map View tab:

1. Interactively zoom in and out of the region and pan around the map area to explore the spatial patterns and layout at the desired level of detail.
2. In the map view tab, hovering over a node displays the basic attribute information of the node including the tower ID, number of calls, and total call duration within one hour.
3. The network characteristics view remains unchanged and available for the user to view as found on the network view tab.
4. Overview map to show the study region/location
5. The embedded legend in the lower right-hand corner graphically explains the visually encoded node information to the users.

Currently, the chord view tab is disabled due to the time constraints for finishing the project. This network view will be completed in future renditions of NVizABLE. The chord view displays the relationships between objects and provides mouseover filtering within the D3 environment [4].

IV. CONCLUSION AND FUTURE WORK

While NVizABLE is fully deployed on the web, we hope to extend the tool's functionalities based upon feedback through user evaluation as well as goals we have previously identified. Future work is hoped to include:

D4D Challenge Submission

1. Linked views for coordinated brushing between and across components. "Brushing is a collection of dynamic methods for viewing multidimensional data. The effect of an operation appears simultaneously" [7, p. 127] across all views. As with [8], we hope to "focus on the probing, brushing, and linking of data in order to help analysts refine their hypotheses. These methods emphasize the interaction between human cognition and computation through dynamically linked statistical graphs and geographical representations of the data" [p. 207].
2. Inclusion of a parallel coordinate plot to visualize multivariate attributes such as node statistics and data attributes allowing users to visually compare multivariate and single values between nodes.
3. Implement the interactive chord diagram view to represent network relationships [9].
4. Allowing users to select and filter data based on different parameters is a necessary step to free up display space and allows users to find important patterns and relationships.
5. Developing a scalable visualization that reduces clutter for large datasets. Traditionally, other data transforming methods have been used, including: "for large numbers of links or nodes, aggregation, for large numbers of time periods, averaging, and for detecting changes, thresholding and exception reporting" [5, p. 16]. New visualization methods allow developers to enhance network visibility without transforming or compromising the original data structure. For instance, creating a node-link tree structure to allow for representation of big data (volume) in the network may support this goal. NAViGaTOR [10], a program limited to biological data successfully implemented collapsible network groups to simplify the structure of the overall network, allowing for better scalability in light of big data. Within NVizABLE, this structure would allow the user to interactively click on node and expand subgraphs so that the original messy structure does not overwhelm the display space and new meaningful relationships can be discovered.
6. Incorporate more geographic analysis capabilities and link the graphics and statistics between the network and geographic views. Ultimately, a side-by-side comparison may prove to be especially useful for users to visually discern and compare patterns of both structures that may not have been possible before.
7. With advancements in technology, NVizABLE should be able to support real-time data to help analysts in making decisions and finding insights in time-sensitive situations.
8. With streaming real-time and temporal data, it will also be important to incorporate techniques that allow users to view the changing network structure over time. This may include the use of a time slider or animation.
9. Allow users to input their own formatted data directly in NVizABLE (more usable and accessible to the public).
10. Develop a scalable system to "tightly integrate state-of-the-art automatic data analysis methods with interactive visualization techniques and be integrated smoothly into custom-designed processes for the exploration and analysis of complex information spaces." [11, p. 252].
11. Incorporate techniques for depicting uncertainty in the data/network structure.

NVizABLE is a new visual analytics tool which depicta network graphs in an exploratory, interactive interface. The main goal of NVizABLE was to develop a general tool that could synthesize the heterogeneous Orange cell phone communications data to aid researchers in addressing complex social issues to better the living situation of individuals in the Ivory Coast specifically, and others in developing countries generally. We believe our project also provides analysts a methodology for parsing the Orange cell phone communications data into usable network data.

NVizABLE also has implications for the study of network data generally. It is our hope that this tool will assist users from many domains who have both large or small datasets to find meaningful patterns and insights upon using the interface. By deploying a web-based interface, the program is accessible to all users with an Internet connection. By incorporating evaluation approaches from conception to development, we have iteratively refined and addressed user concerns and needs from a variety of disciplines and data types.

REFERENCES

[1] Data for Development, "D4D Orange Challengce " *http://www.d4d.orange.com/home*

[2] J. Stevens, J. Smith, and M. Idris. (2012, December 19). *NVizABLE: Cellular Communications Explorer*. Available: http://www.cartocadabra.com/nvizable/

[3] V. D. Blondel, M. Esch, C. Chan, F. Clerot, P. Deville, H. Etienne., *et al.*, "Data for Development: The D4D Challenge on Mobile Phone Data," Orange Labs, France2012.

[4] M. Bostock. (2012, December 19). *Chord Diagram*. Available: http://bl.ocks.org/4062006

[5] M. Bostock. (2012, December 19). *Data-Driven Documents*. Available: http://d3js.org/

[6] M. Bostock, V. Ogievetsky, and J. Heer, "D3: Data-Driven Documents," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, pp. 2301-2309, 2011.

[7] R. A. Becker and W. S. Cleveland, "Brushing Scatterplots," *Technometrics*, vol. 29, pp. 127-142, 1987.

[8] R. Maciejewski, S. Rudolph, R. Hafen, A. M. Abusalah, M. Yakout, M. Ouzzani, *et al.*, "A Visual Analytics Approach to Understanding Spatiotemporal Hotspots," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, pp. 205-220, 2010.

[9] R. A. Becker, S. G. Eick, and A. R. Wilks, "Visualizing Network Data," *IEEE Transactions of Visualization and Computer Graphics*, vol. 1, pp. 16-28, 1995.

[10] K. R. Brown, D. Otasek, M. Ali, M. J. McGuffin, W. Xie, B. Devani, *et al.*, "NAViGaTOR: Network Analysis, Visualization and Graphing Toronto," *Bioinformatics*, vol. 25, pp. 3327-3329, 2009.

[11] G. Andrienko, N. Andrienko, D. Keim, A. M. MacEachren, and S. Wrobel, "Challenging Problems of Geospatial Visual Analytics," *Journal of Visual Languages and Computing*, vol. 22, pp. 251-256, 2011.

[12] T. j. Foundation. (2012, December 19). *jQuery*. Available: http://jquery.com/

D4D Submission

Mobile Data Delivery through Opportunistic Communications among Cellular Users: A Case Study for the D4D Challenge*

Ying Zhu Chao Zhang Yu Wang
Department of Computer Science, University of North Carolina at Charlotte
9201 University City Blvd., Charlotte, North Carolina, USA
{yzhu17,czhang25,yu.wang}@uncc.edu

ABSTRACT

The appearance of smartphones and increasing popularity of various mobile applications and services have caused the explosion of mobile data traffic. To avoid overloading the cellular networks, different offloading solutions (such as WiFi networks or femtocells) have been proposed and adopted. Recently, offloading cellular traffic through opportunistic communications among mobile phones becomes a new and promising option, due to free cost. In this paper, by using real trace data from the Orange "Data for Development" (D4D) challenge, we investigate the feasibility of delivering data packets among mobile cellular users through opportunistic communications in a large scale network. Our experimental results show that by using social or location properties of mobile users opportunistic routing can indeed complement the traditional cellular network to deliver delay-tolerant data packets among certain portion of cellular users. Such solution is especially cost efficient and beneficial for developing countries, as Ivory Coast.

1. INTRODUCTION

Due to the increasing popularity of mobile applications and services for smartphones, we are currently facing the challenges of mobile data explosion. Based on the most recent Cisco's report [1], mobile data traffic grew 70 percent in 2012 and reached 885 petabytes per month at the end of 2012, which was nearly 12 times the size of the entire Internet in 2000 (75 petabytes per month). Cisco also forecasts that mobile data traffic will surpass 10 exabytes per month in 2017. In addition, the recent advance in machine-to-machine (M2M) communications may potentially add billions of devices into mobile Internet. By the end of 2013, the number of mobile-connected devices will exceed the number of people on earth [1]. However, the current cellular networks do not have enough capacity to support all of the fast-growing mobile data from these devices.

To avoid overloading the cellular networks, different offloading solutions (such as WiFi networks [2–4] or femtocells [5]) have been proposed and adopted. According to Cisco [1], globally, 33 percent of total mobile data traffic was offloaded onto fixed network through WiFi or femtocell in 2012. Without offload, mobile data traffic would have grown 96 percent rather than 70 percent in 2012. Recently, offloading cellular traffic through opportunistic communications [6–8] among mobile phones becomes a new and possible solution. Compared with current WiFi or femtocell solutions, this method uses occasional device-to-device contact opportunities to deliver data rather than using the fixed network infrastructure. The major advantage of this solution is low cost and easy to deploy. However, due to the intermittent connectivity in opportunistic networks, the target data types are limited to delay-tolerant bulk data for non-realtime applications.

In this paper, using real trace data from the Orange "Data for Development" (D4D) challenge [9], we investigate the feasibility of delivering data packets among cellular users through opportunistic communications in a large scale network. Different from the previous studies [6–8] where broadcast traffic from the service provider to all subscriber users are offloaded to opportunistic networks, we focus on data delivery among individual mobile users using opportunistic routing. The released D4D dataset [10] provides anonymized call patterns and mobility data of 5,000 to 50,000 mobile phone users (based on Call Detail Records (CDR) of phone calls and SMS among these users) in Ivory Coast. It is a perfect resource to study the performance of opportunistic routing in a large-scale real network, since the dataset provides fine-quality mobility and location information of a large population of mobile users. Previous study on opportunistic routing [11–19] usually only focus on data delivery in small-scale delay tolerant networks or pocket switch networks with limited number of mobile users in a relevant small region (such as a group of researchers in a conference venue or a group of students on a campus). We believe that this is the first study on data delivery via opportunistic communications in large-scale networks in wide deployment

*This material was prepared for the Data for Development (D4D) Challenge and the third conference on the analysis of mobile phone datasets (NetMob 2013), May 2-3, 2013, MIT, USA. This work was supported in part by the US National Science Foundation (NSF) under Grant No.CNS-0915331 and CNS-1050398.

area. In our experiments, we consider six different opportunistic routing methods and evaluate them under various settings . Our results show that by using social or location properties of mobile users opportunistic routing can indeed achieve certain level of delivery ratios so that it can complement the traditional cellular network to deliver delay-tolerant data packets among active cellular users.

Middle East and Africa are the fastest growing regions of mobile phone market. Based on [1], their monthly mobile data traffic will experience the highest compound annual growth rate of 77 percent around the world between 2012 and 2017. Therefore, we strongly believe that the proposed solution is a cost efficient complement to the existing and growing cellular infrastructure for the developing countries in these regions (such as Ivory Coast). Thus, this could contribute to the socio-economic development and well-being of the Ivory Coast population (and fulfill the goal of D4D challenge).

The remainder of this paper is organized as follows: We briefly review related work in Section 2. We then introduce the D4D dataset and how we select and use the dataset for opportunistic communications in Section 3. In Section 4, we present six different opportunistic routing schemes which we test over the D4D dataset. Simulation results are reported in Section 5. Finally, conclusions are presented in Section 6.

2. RELATED WORK

2.1 Opportunistic Networks

Opportunistic network [20,21], where occasional contact opportunities are used to deliver data, is one of the emerging communication paradigms. In opportunistic networks, communication is challenged by sporadic and intermittent contacts as well as frequent disconnections and reconnections. To handle intermittent connectivity, opportunistic routing methods [11–19] share the same principle, "store and forward": If there is no connection available at a particular time, the current node can store and carry the data until it encounters other nodes. When the node has such a forwarding opportunity, all encountered nodes could be the candidates to relay the data. Therefore, relaying selection and forwarding decision need to be made by the current node based on certain forwarding strategy.

The simplest routing method is *epidemic routing* [11], in which a node forwards copies of message to any nodes it encounters. This flooding-based method can guarantee the best delivery ratio, but suffers from huge message overheads. To reduce the overheads, many methods restrict the number of message replicas in the network to a certain constant (such as in *Spray and Wait* [13]) or just one (such as in *SimBet* [15]) or a small one by only replicating the message when certain condition is met (such as in *delegation forwarding* [14]). We call the methods which allow multiple replicas and those which allow a single replica as multi-copy routing and single-copy routing, respectively.

Forwarding decision (or replicating decision) in opportunistic routing usually relies on certain type of quality metric. The message is only forwarded to a node with higher quality metric. During an encounter, if there are multiple nodes with higher quality metric, only the one with highest quality metric is selected as the relay. Examples include *Fresh* [16] (picking the node which has met the destination more recently), *Greedy-Total* [17] (picking the node with a higher encounter frequency to all other nodes), *MobySpace* [18] (picking the node which has more location similarity with the destination), *SimBet* [15] (picking the node with higher social centrality and more common neighbors with the destination), or *Bubble Rap* [12] (picking the node with higher centrality within certain community structure).

All these opportunistic routings are usually evaluated for data delivery in small-scale delay tolerant networks or pocket switch networks with limited number of mobile users (such as a group of researchers in a conference venue or a group of students on a campus).

2.2 Cellular Traffic Offloading

Cellular traffic offloading with complementary network communication technologies has become an emerging topic in recent years due to the dramatic increasing of mobile traffic load. Current solutions mainly rely on either femtocells and WiFi networks for delivering data originally targeted for cellular networks. Femtocells [5] operate on the same licensed spectrum as the macrocells of cellular networks and can offer better indoor voice and data services by offloading traffic from macrocells. WiFi networks on the other hand work on the unlicensed frequency bands and have also been widely used for offloading from cellular networks [2–4]. For example, major cellular operators, such as Orange, AT&T, T-Moblie, all have deployed their own WiFi networks world wide. Recently, Han et al. [6, 7] proposed the third type of solution: offloading cellular traffic to opportunistic networks formed by mobile cellular users. By studying how to select the initial set of users to push the content to all users in the networks, their proposed simple heuristics can improve the delivery efficiency and offload a large fraction of data from the cellular network. Li et al. [8] then studied the problem of multiple mobile data offloading through opportunistic communications among different data subscribers under resource constraints. Our study also uses opportunistic communications to offload traffic from cellular network, but we focus on offloading peer to peer traffic among mobile users instead of broadcasting traffic from the service provider to all subscriber users (as in [6–8]).

2.3 Cellular Dataset Analysis

The appearance of smartphones equipped with various sensors (especially GPS) and contact/event logs enables pervasive monitoring of mobile user behaviors and mobility. There are several cellular datasets recently collected via smartphone based testbeds: Nokia Data Collection Campaign [22], MIT reality project [23], Nodobo [24], and Context project [25]. These real-life tracing data provide abundant resources to study social, spatial, and temporal characteristics of mobile users in different environments. The D4D datasets [10] are newly released cellular datasets, which complement the existing datasets as the scale of number of users is much larger than those existing ones. This gives us an unique opportunity to study the feasibility of opportunistic routing in large-scale networks.

Table 1: Numbers of users, towers, and contacts in four different settings

Setting	# of users	# of towers	# of encounters
A) subset users within full region	13,436	1,095	617,136
B) subset users within limited region	6,318	496	327,717
C) all users within full region	46,254	1097	6,787,594
D) all users within limited region	21,768	497	3,736,173

(a) traffic load distribution (b) limited region near Abidjan (c) tower distribution in limited region

Figure 1: Illustration of the limited region in Settings B and D: (a) traffic load distribution in the whole nation generated using Geofast site [40] for the first two weeks; (2) the selected region near Abidjan with the heaviest traffic; (3) the detailed cellular tower distribution within the limited region.

Different cellular datasets have been studied by the research community for a wide range of purposes, such as human mobility modeling [26–28], importance place extraction [29–31], mobile recommendation systems [32, 33], urban sensing and planning [34, 35], sociology [36–38], ecology and epidemiology [39]. In this study, we use the D4D to study social or location based opportunistic routing in large-scale cellular networks as a possible offloading solution.

3. D4D DATASETS AND PREPARATION

The released D4D datasets [10] are based on anonymized Call Detail Records (CDR) of phone calls and SMS exchanges between 500,000 Orange mobile users in Ivory Coast between December 1, 2011 and April 28, 2012 (150 days). Among the released four datasets, we mainly use the second one (**SET2**): individual trajectories with high spatial resolution. This dataset contains the access records of antenna (cellular tower) of each mobile user over two-week periods. Such information provides high resolution trajectories for all mobile users. We will use the sequences of visited cellular towers of all users to generate contact encounters among mobile users and location profiles of each mobile user. In the results present in this paper we only use the first two weeks (December 1 to 14, 2011) data for our simulations.

Since D4D datasets do not have direct encounter information between phones via short range communications (such as Bluetooth or WiFi), to support opportunistic communications we assume that two phones can direct communicate to each other if they share the same cellular tower at particular time. Though this assumption may not be true in reality, it gives us an approximated environment for opportunistic communications. All of our experiments (presented in Section 5) are based on the generated encounter databases from SET2.

We will consider four different settings (A-D) for our experiments. Table 1 summaries some statistics of these settings. In term of number of nodes (mobile users), we either use all 500,000 users or a subset of users (around 15,000) in the original SET2. When we pick up the subset of users, we just simply choose the first 15,000 users in our encounter database. Notice that the number of users in our generated encounter database is less than the number of users in original SET2 (such as $46,254 < 50,000$). This shows that there are many mobile users who do not share any cellular towers with other users. The smaller size of user set could accelerate the execution time of our simulations. Notice that the number of encounters is significantly reduced after picking the subset users, though the cellular towers stay the same level. We also have settings where we limit the physical locations of encounters to a small region. As shown in Figure 1(a), the traffic load distribution within Ivory Coast is unbalanced. This figure shows the number of calls (both incoming and outgoing calls) during the first two weeks. Darker color indicates heavier traffic loads. Therefore, when picking up the small region, we choose the region with the heavies traffic load. The longitude and latitude ranges of the region (shown as a tiny blue rectangle in Figure 1(b) around Abidjan) are $[-8.49, -2.69]$ and $[4.41, 10.47]$, respectively. Abidjan is the

economic and former official capital of Ivory Coast and the largest city in the nation. From Table 1 we can see that this region holds a large number of cellular towers and mobile users. Figure 1(c) shows the detailed tower distribution in this region.

4. OPPORTUNISTIC ROUTING PROTOCOLS

To test the feasibility of opportunistic routing among large scale mobile users, we implement six different routing methods which are listed below.

- **Epidemic** [11]: during any encounter, a copy of the message is forwarded to all encountered nodes and the current node still hold a copy of the message.
- **Naive**: during any encounter, the message is always forwarded to the encounter node and the current node will not hold the message after forwarding. If there are multiple nodes during the same encounter, the next hop is randomly picked. It can be treated as a single-copy version of Spray and Wait [13].
- **Fresh** [16]: the message is only forwarded from the current node v_i to the encountered node v_j if v_j has met the destination more recently than v_i does. If there are multiple nodes satisfying such a condition during the same encounter, v_i forwards the message to the one who has met the destination most recently.
- **Destination Frequency** [14]: the message is only forwarded from v_i to v_j if v_j has met the destination more often than v_i does. If there are multiple nodes satisfying such a condition during the encounters, v_i forwards the message to the one who has met the destination most often.
- **Centrality-Based**: the message is only forwarded from v_i to v_j if v_j has higher centrality than v_i does. Here, we simply consider the degree centrality of each node, i.e., how many nodes it has encountered. A node with higher degree centrality is more popular in the network. If there are multiple nodes satisfying the condition during the encounter, v_i forwards the message to the one who has the highest centrality. Similar idea has been used in Greedy-Total [17], SimBet [15] and Bubble Rap [12].
- **Location-Based**: the message is only forwarded from v_i to v_j if v_j has higher location similarity to the destination than v_i does. If there are multiple nodes satisfying the condition during the encounter, v_i forwards the message to the one with the highest similarity. This idea has been used in MobySpace [18] and is based on the obversion that people with similar location profiles (places visited) are likely to meet each other at their common places.

Since most of these methods are quite standard and straightforward to implement, we only introduce the detail about our implementation of location-based method (how to estimate the location similarity between two mobile users).

We treat every cellular tower t_k as one place, where $k = 1, \cdots, N$ and N is the total number of towers. For each mobile user v_i, we first extract its total visit duration and frequency of each tower t_k, represented by $d(v_i, t_k)$ and $f(v_i, t_k)$. Both are normalized between 0 and 1. Then we can estimate the probability that user v_i visits tower t_k using $p(v_i, t_k) = d(v_i, t_k) \times f(v_i, t_k)$. Therefore, for each user v_i, we can have a vector of $P(v_i) = \{p(v_i, t_k) | k = 1, \cdots, N\}$ to represent his *location profile*. Given two mobile users v_i and v_j, their location similarity can be defined as

$$S(v_i, v_j) = |P(v_i) \cdot P(v_j)|_1 = \sum_{k=1}^{N} p(v_i, t_k) \times p(v_j, t_k).$$

Notice that here we only consider the total visit duration and frequency of each user for a particular tower. It is possible to consider more detailed visit pattern, such as time-dependent visit patterns: visit duration and frequency in certain time period (morning, afternoon or night). However, our simulation results show that such refinement does not lead to sufficient improvement.

5. PERFORMANCE EVALUATIONS

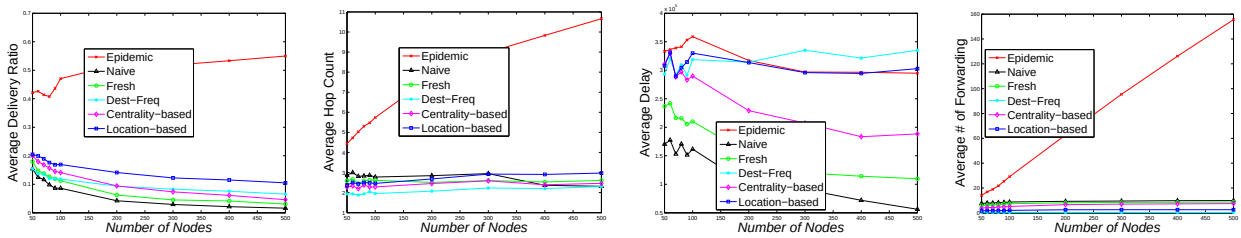
In this section, we present our experimental results on performances of all opportunistic routing methods over the contact database we generated from the D4D dataset as described in Section 3.

In all experiments, we compare each algorithm using the following routing metrics.

- **Delivery ratio:** the average percentage of successfully delivered messages from the sources to the destinations.
- **Hop count:** the average number of hops during each successful delivery from the sources to the destinations.
- **Delay:** the average time duration of successfully delivered messages from the sources to the destinations.
- **Number of forwarding:** the average number of messages forwarding in the network during the whole period.

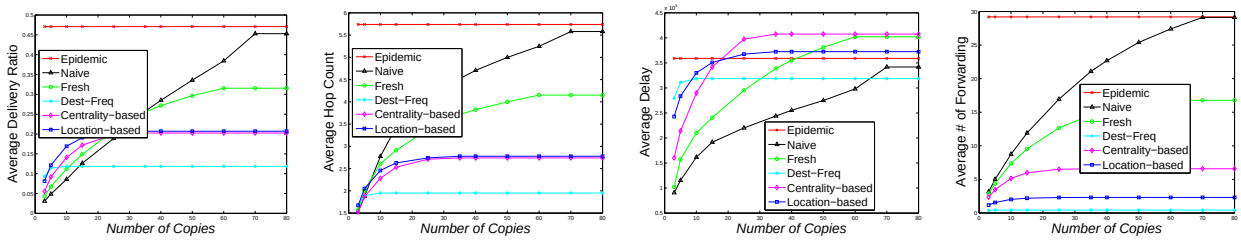
For all experiments, we perform $5,000$ random routing tasks among the selected participators. All results reported here are the average over these tasks.

For each experiment, we pick different number of nodes to participate the opportunistic communications, ranging from 50 to 500. Here we always pick the most active nodes (based on overall centrality) in the user set, since they are better candidates for opportunistic forwarding. We believe that these active users are the major target customers for our proposed traffic offloading scheme via opportunistic communication. For those users who are not active or even isolated in the opportunistic network, the only choice is using the cellular or fixed networks. We did perform some experiments over random chosen users, however the delivery ratios of all routing methods (even Epidemic) are very low (worse than 1 percent).



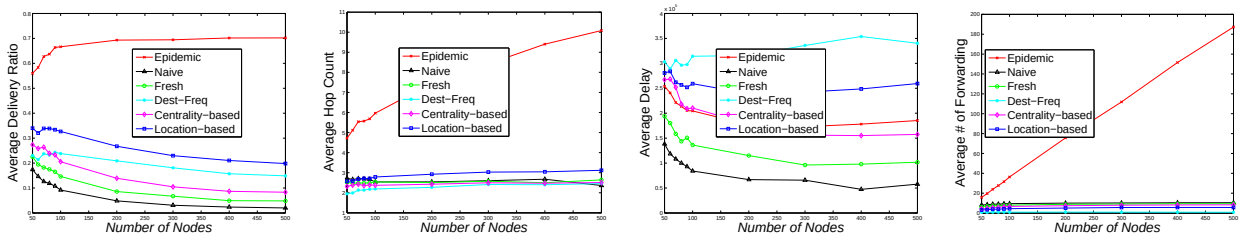
(a) delivery ratio (b) hop count (c) delay (d) # of forwarding

Figure 2: Performance results over Setting A (the number of copies is fixed at 10).



(a) delivery ratio (b) hop count (c) delay (d) # of forwarding

Figure 3: Performance results over Setting A (the number of copies is ranged from 3 to 80, and the number of nodes is fixed at 100).



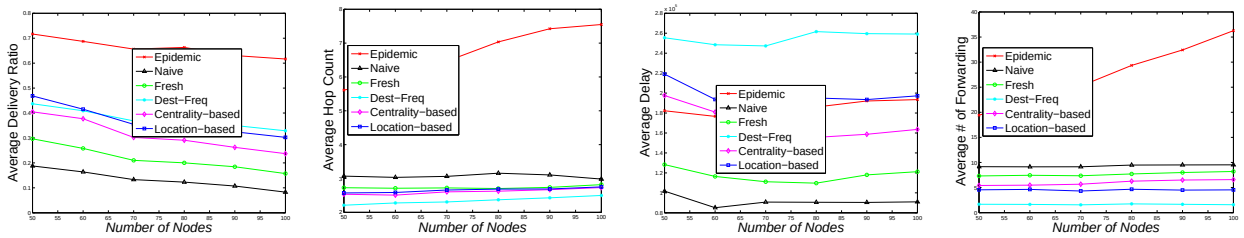
(a) delivery ratio (b) hop count (c) delay (d) # of forwarding

Figure 4: Performance results over Setting B (the number of copies is fixed at 10).

For all opportunistic routing methods except for Epidemic, we allow multiple copies of the same message but limit the number of copies by a small constant. In the default setting, we use 10 as the constant bound.

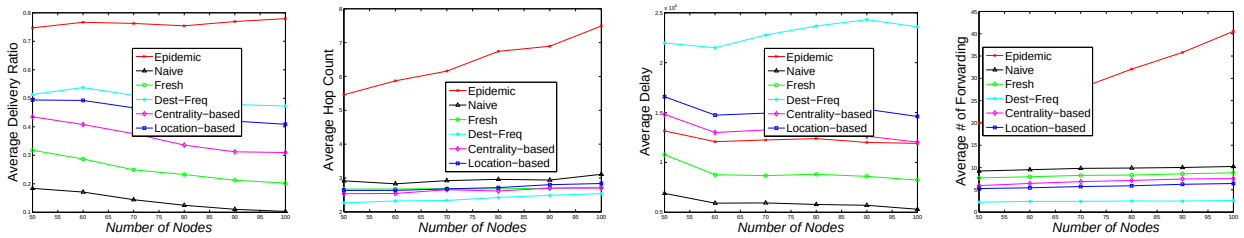
In the first set of simulations, we use Setting A (with around 15,000 selected users and within the full region). Figure 2 illustrate the results. From Figure 2(a), we can see that location-based and centrality-based methods can achieves better delivery ratio than Naive and Fresh methods. This confirms that the understanding and usage of social or location relationships among mobile users is beneficial for making smarter forwarding decision. Notice that that even though Epidemic routing has the best delivery ratio, it costs extremely large amount of forwarding as shown in Figure 2(d). It is also noticeable that the delivery ratio is decreasing as the number of nodes increases. This is reasonable since we always choose the most active nodes as the participators. With more nodes included, more routing tasks are among less active nodes. This again shows that our proposed solution mainly benefits the active mobile users in the network. In terms of hop count and number of forwarding, all opportunistic routing methods are at the similar level except for Epidemic. Notice that for delay since we only consider the successful routes, thus Epidemic usually has the largest delay.

For the same Setting A, we then test the effect of the number of copies in multi-copy opportunistic routing. We fixed with 100 nodes and change the number of copies from 3 to 80. Figure 3 shows the results. It is obvious that with more message copies all methods can achieve higher delivery ratio but increase the number of forwarding too. There is clearly a trade-off between number of copies and forwarding overhead. When the number of copies reaches certain value, the delivery ratio will be stable. Further adding more copies does not help. For different methods, such critical value of copy number may vary.



(a) delivery ratio (b) hop count (c) delay (d) # of forwarding

Figure 5: Performance results over Setting C (the number of copies is fixed at 10).



(a) delivery ratio (b) hop count (c) delay (d) # of forwarding

Figure 6: Performance results over Setting D (the number of copies is fixed at 10).

To test the performance of all methods in a small and dense region, we then test our methods on Setting B, which limits the region around a rectangle region near Abidjan. Compared with the results in the full region (Setting A), all methods can achieve better performances in this setting. This is reasonable since a limited dense network provides more close opportunities for message delivery among mobile users than a larger and sparser network does.

Last, we also perform simulation over the full population of D4D dataset (Settings C and D). Figure 5 and Figure 6 show the results, respectively. Compared with previous results, all methods can achieve better performances too. The reason is still the same that within larger population the selected participants are more active thus lead to better chances for mobile delivery. Once again, better performance can also be achieved in a smaller and denser area.

In summary, via the above simulations over the D4D dataset, we can have the following overall conclusions.

- Epidemic can achieve the highest delivery ratio since it takes every forwarding opportunities and does not have limitation on the number of copies. However, it suffers from the large number of forwarding, especially when the number of nodes is large.
- Location-based, Centrality-based, and Destination Frequency can achieve relevant high delivery ratios while still use reasonable number of forwarding.
- Compared with different settings, all opportunistic routing can achieve better performance when the participants are active users and the physical region is small and dense. This can be shown in Figure 7 which summarizes the average delivery ratios over four different settings under the same parameters.

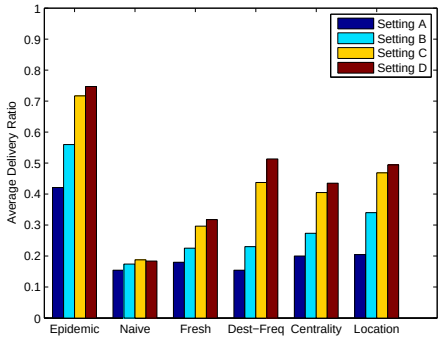


Figure 7: Average deliver ratios over Settings A to D (the number of nodes and the number of copies are 50 and 10, respectively).

6. CONCLUSIONS

In this paper, by leveraging the rich trace data from the Orange D4D challenge, we investigate the feasibility of delivering data packets among mobile cellular users through opportunistic communications. Our experimental results show that by using social or location properties of mobile users opportunistic routing can indeed achieve certain level of deliver ratio, especially among the active mobile users

and within dense region. Therefore, it is possible that such solution can complement the traditional fixed network to deliver delay-tolerant data packets. On the other hand, there are still spaces to further improve the deliver ratio of opportunistic routing in such large scale networks. We will continue investigate new techniques to enhance the performance of opportunistic communications. Finally, we would like to thank the Orange D4D challenge organizers to provide such a great opportunity for us to participate in this event. We hope that Orange and other cellular companies can further release high quality real life datasets to research community.

7. REFERENCES

[1] Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2012-2017 (February 6, 2013). http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.html.

[2] A. Balasubramanian, R. Mahajan, and A. Venkataramani. Augmenting mobile 3G using WiFi. In *ACM MobiSys 2010*, 2010.

[3] K. Lee, I, Rhee, J. Lee, S. Chong, and Y. Yi. Mobile data offloading: how much can WiFi deliver? In *ACM Co-NEXT 10*, 2010.

[4] S. Dimatteo, P. Hui, B. Han, and V.O.K. Li. Cellular traffic offloading through WiFi networks. In *Proceeding of IEEE 8th International Conference on Mobile Ad Hoc and Sensor Systems, MASS 2011*, 2011.

[5] V. Chandrasekhar, J.G. Andrews, and A. Gatherer. Femtocell networks: A survey. *IEEE Communications Magazine*, 46(9):59-67, 2008.

[6] B. Han, P. Hui, V.S.A. Kumar, M.V. Marathe, G. Pei, and A. Srinivasan. Cellular traffic offloading through opportunistic communications: a case study. In *Proceedings of the 5th ACM workshop on Challenged networks (CHANTS '10)*, 2010.

[7] B. Han, P. Hui, V.S.A. Kumar, M.V. Marathe, J. Shao, and A. Srinivasan. Mobile data offloading through opportunistic communications and social participation. *IEEE Transactions on Mobile Computing*, 11(5):821-834, 2012.

[8] Y. Li, G. Su, P. Hui, D. Jin, L. Su, and L. Zeng. Multiple mobile data offloading through delay tolerant networks. In *Proceedings of the 6th ACM workshop on Challenged networks (CHANTS '11)*, 2011.

[9] The Data for Development (D4D) Challenge. http://www.d4d.orange.com.

[10] V.D. Blondel, M. Esch, C. Chan, F. Clerot, P. Deville, E. Huens, F. Morlot, Z. Smoreda, and C. Ziemlicki. Data for development: The D4D challenge on mobile phone data. arXiv.1210.0137v2, Jan 2013.

[11] A. Vahdat and D. Becker. Epidemic routing for partially connected ad hoc networks. Tech. Rep. CS-200006, Duke University, 2000.

[12] P. Hui, J. Crowcroft, and E. Yonek. Bubble Rap: Social-based forwarding in delay tolerant networks. In *Proc. ACM MobiHoc*, 2008.

[13] K T. Spyropoulos, Psounis, and C. Raghavendra. Spray and wait: an efficient routing scheme for intermittently connected mobile networks. In *Proc. of ACM SIGCOMM workshop on DTN (WDTN)*, 2005.

[14] V. Erranmilli, M. Crovella, A. Chaintreau, and C. Diot. Delegation forwarding. In *Proc. of ACM MobiHoc*, 2008.

[15] E.M. Daly and M. Haahr. Social network analysis for routing in disconnected delay-tolerant MANETs. In *Proc. ACM MobiHoc*, 2007.

[16] H. Dubois-Ferriere, M. Grossglauser, and M. Vetterli. Age matters: efficient route discovery in mobile ad hoc networks using encounter ages. In *Proc. of ACM MobiHoc*, 2003.

[17] V. Erramilli and M. Crovella. Diversity of forwarding paths in pocket switched networks. In *Proc. of ACM IMC*, 2007.

[18] J. Leguay, T. Friedman, and V. Conan. DTN routing in a mobility pattern space. In *Proc. of ACM SIGCOMM workshop on DTN*, 2005.

[19] L. Zhao, F. Li, C. Zhang, and Y. Wang. Routing with multi-level social groups in mobile opportunistic networks. In *Proc. of IEEE Globecom*, 2012.

[20] P. Hui, A. Chaintreau, J. Scott, R. Gass, J. Crowcroft, and C. Diot. Pocket switched networks and human mobility in conference environments. In *WDTN '05: Proceedings of the 2005 ACM SIGCOMM workshop on Delay-tolerant networking*, pages 244–251, New York, NY, USA, 2005. ACM.

[21] Y. Zhu, B. Xu, X. Shi, and Y. Wang. A survey of social-based routing in delay tolerant networks: Positive and negative social effects. *IEEE Communication Survey and Tutorials*, 15(1):387-401, 2013.

[22] J. K. Laurila, D. Gatica-Perez, I. Aad, J. Blom, O. Bornet, T.-M.-T. Do, O. Dousse, J. Eberle, and M. Miettinen. The mobile data challenge: Big data for mobile computing research. In *Mobile Data Challenge 2012 (by Nokia) Workshop, in conjunction with Int. Conf.. on Pervasive Computing*, Newcastle, June 2012.

[23] MIT Media Lab. Reality mining project. http://reality.media.mit.edu/.

[24] University of Strathclyde. Nodobo data release. http://nodobo.com/release.html.

[25] University of Helsinki. The context project. http://www.cs.helsinki.fi/group/context/#data.

[26] F. Calabrese, G.D. Lorenzo, L. Liu, and C. Ratti. Estimating origin-destination flows using mobile phone location data. *IEEE Pervasive Computing*, 10(4):36-44, 2011.

[27] C. Song, Z. Qu, N. Blumm, and A.-L. Barabasi. Limits of predictability in human mobility. *Science*, 327:1018–1021, 2010.

[28] M.C. Gonzalez, C.A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 453:779–782, 2008.

[29] S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, M. Martonosi, J. Rowland, and A. Varshavsky. Identifying important places in people's lives from cellular network data. In *Proceedings of the 9th international conference on Pervasive computing*, Pervasive'11, pages 133–151, Berlin, Heidelberg, 2011. Springer-Verlag.

[30] J. H. Kang, W. Welbourne, B. Stewart, and G. Borriello. Extracting places from traces of

locations. In *Proceedings of the 2nd ACM international workshop on Wireless mobile applications and services on WLAN hotspots*, WMASH '04, pages 110–118, New York, NY, USA, 2004. ACM.

[31] Y. Ye, Y. Zheng, Y. Chen, J. Feng, and X. Xie. Mining individual life pattern based on location history. In *Proceedings of the 2009 Tenth International Conference on Mobile Data Management: Systems, Services and Middleware*, MDM '09, pages 1–10, Washington, DC, USA, 2009. IEEE Computer Society.

[32] V. Zheng, Y. Zheng, X. Xie, and Q. Yang. Collaborative location and activity recommendations with GPS history data. In *Proc. of WWW' 10*, pages 1029–1038, 2010.

[33] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma. Mining interesting locations and travel sequences from GPS tragectories. In *Proc. of WWW' 09*, pages 791–800, 2009.

[34] R. A. Becker, R. Caceres, K. Hanson, J. M. Loh, S. Urbanek, A. Varshavsky, and C. Volinsky. A tale of one city: Using cellular network data for urban planning. *IEEE Pervasive Computing*, 10:18–26, 2011.

[35] J. Reades, F. Calabrese, A. Sevtsuk, and C. Ratti. Cellular census: Explorations in urban data collection. *IEEE Pervasive Computing*, 6(3):30–38, July 2007.

[36] G. Chittaranjan, J. Blom, and D. Gatica-Perez. Mining large-scale smartphone data for personality studies. *Personal and Ubiquitous Computing*, 2012.

[37] N. Eagle, A. Pentland, and D. Lazer. Inferring Friendship Network Structure by Using Mobile Phone Data. *PNAS*, 106(36), 2009.

[38] W. Dong, B. Lepri, and S. Pentland. Tracking co-evolution of behavior and relationships with mobile phone. *Tsinghua Science and Technology*, 17(2):136-151, 2012.

[39] S. Eubank, H. Guclu, V. S. A. Kumar, M. Marathe, A. Srinivasan, Z. Toroczkai, and N. Wang. Modeling disease outbreaks in realistic urban social networks. *Nature*, 2004.

[40] The Geofast Web Interface. http://www.geofast.net.

EEMC: An Energy-Efficient Mobile Crowdsensing Mechanism by Reusing Call/SMS Connections

Haoyi Xiong [†, ‡], Leye Wang [†, ‡], and Daqing Zhang [†]
[†] Institut Mines-Télécom, Télécom SudParis, CNRS SAMOVAR, UMR 5157, France
[‡] EDITE, Université Pierre et Marie Curie - Paris VI, France
{haoyi.xiong, leye.wang, daqing.zhang}@telecom-sudparis.eu

Abstract—Mobile crowdsensing (MCS) has been proved effective to enable large-scale monitoring applications, especially in developing countries, since it doesn't rely on any sensor network or other infrastructure, but crowdsources the sensing tasks to mobile phone users. However, as MCS consumes the energy of each participatory mobile device and affects user's experiences, it might discourage users to participate in the monitoring tasks. In order to minimize the energy consumption of MCS in each mobile phone, it's necessary to reduce the power consumption caused by at least one of the three main factors: sensing, computation, and data transmission. Besides, the overall energy consumption of a MCS task depends on the number of its participants. Thus, to minimize the total energy cost, MCS should select a minimum set of participants from all qualified users. In this paper, we propose *EEMC*–an energy-efficient mobile crowdsensing mechanism relying on the reuse of wireless connections. *EEMC* reuses the connections of call/SMS to cut off the energy consumption of data transmission for task assignment and sensing result collection. Furthermore, to minimize the total energy cost, *EEMC* estimates the probability of each user to have a new call or SMS in near future, and assigns tasks only to those users who have the highest probability of reusing connections. Evaluation results with D4D dataset show that *EEMC* requires only 43% of energy that was consumed by the traditional schema without connection reuse in order to collect 40 sensing samples of activity recognition from the commerce center of Abidjan city within a delay between 30 to 45 minutes.

Index Terms—Mobile Phone Energy, Mobile Crowdsensing, Crowdsource, Sensors, and Data Transmission

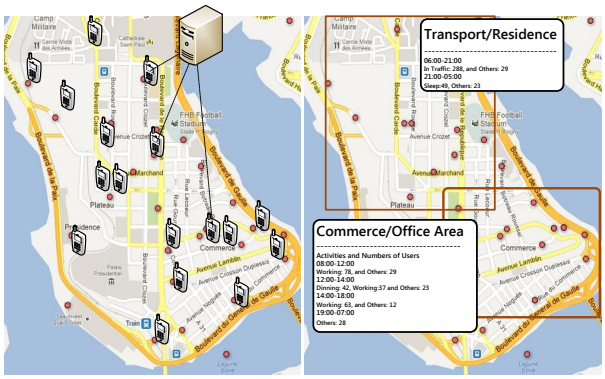
1 Introduction

Mobile Crowdsensing (MCS)—a term coined by *Ganti et al.* [4] has recently spurred lots of interest, and enabled a broad range of applications. MCS applications leverage the sensors equipped by the millions of mobile devices, e.g., Android phones, iPhones or iPads, already "deployed in fields" where people carry these devices in their daily lives. MCS has successfully extended the sensing capacities from single physical space to a community or the whole city, from recognizing individual behavior to understanding the collective behavior of crowds.

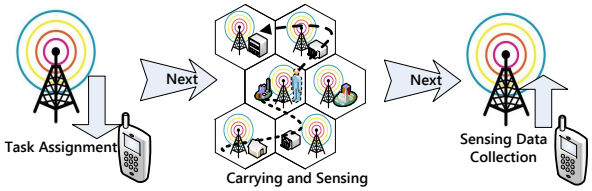
A typical MCS use case of MCS is *Crowds Life Logging* shown in Figure 1. The process shown in Figure 1b outsources the sensing tasks to the mobile phones of crowds, and then collects the sensing results including the human activity, e.g., working, walking and sleeping, of the user as well as the time and location of the activity from participants. As shown in Figure 1a, the long-term crowds life logging might characterize the land use of a location [10], while the real-time crowds life logging in short term can be used to detect large emergency or social events [1] in specific region. This use case shows the sensing capabilities of MCS to enable city-wide collective behavior monitoring without pre-deployment of any static and specific sensor networks. The crowdsourcing of mobile participants provides developing countries, which usually lack the sensor network infrastructure, with a sustainable eco-system for city-wide sensing and monitoring.

Even though mobile crowdsensing is widely studied recently, the effectiveness of MCS is usually constrained by the low participatory rate of mobile phone users. One of the main factors that discourages the participants is the energy consumption on mobile phones. Due to the additional energy consumption of sensing, computation, and data transmission, the MCS applications affect the user's experience in participating the crowdsensing tasks. Therefore, we are motivated to propose an energy-efficient crowdsensing mechanism–*EEMC* to reduce the total energy cost of MCS at the mobile side and also grantee the collection of enough samples within a tolerable time of delay.

Generally, *EEMC* crowdsources the MCS tasks to the mobile phone users each of whom is currently having a phone call or SMS, and probably going to have another call or SMS within a short duration. More specific, it transfers data for task assignments by reusing the connections of current calls or SMS, and collects the sensing results by reusing the connections



(a) Crowds Life Logging at Abidjan City



(b) MCS Process on Mobile Phone

Fig. 1: The Use Case

of calls or SMS in the near future. The reuse of Call/SMS connection cuts off the energy consumption of connection establishment, maintenance and tails from the mobile battery usage of mobile crowdsensing. *EEMC* significantly reduces the total energy cost in data transmission, since the connection establishment, maintenance and tail (e.g., 24 Joules for 3Gs) consumes much more energy than the data transfer of sensory data (e.g., 9 to 30×10^{-3} Joules for 100 bytes via 3G). Besides, it seems that sensory task execution may consume less energy than the cost of connection also. For example, it costs 0.66 Joules to recognize a user's current human activity based on the traces of accelerometer sensors in recent 30 seconds with accuracy higher than 90% [12]. We will present and discuss the detailed measurement results of energy cost comparison in Section 2.

However, selecting the participants for such mobile crowdsensing is a challenging problem, since some mobile phone users who have been assigned with tasks by their previous connections, won't have another phone call or SMS in near future. That means it is required to assign more tasks to collect enough number of samples, if our data transmission totally relies on the reuse of connections. In this paper, we first formulate the problem of task assignment in *EEMC*, then we propose a set of algorithms to address the *EEMC* problem. All the algorithms are evaluated on D4D dataset [3].

The rest of this paper is organized as follow: Section 2 presents the energy consumption of MCS on mobile phone; Section 3 introduces our solution of *EEMC*, and discuss the algorithms of task assignments; Section 5 presents the evaluation result of proposed algorithm; and Section 6 makes concludes this work.

2 Energy Consumption of Mobile Phone in Mobile Crowdsensing

In this section, we mainly investigate the energy consumption of mobile phone for MCS applications.

2.1 Energy Consumption of Data Transmission

Type	Connection (**J**)	Data Transfer (**mJ/byte**)
3G (UTMS)	12.0	0.04-0.16 download 0.09-0.3 upload
SMS (SS7)	2.0	3.0
WIFI	5.0-12.0	0.01
2G (GSM/GPRS)	4.0	0.036

TABLE 1: Energy Cost of Data Transmission: the specific energy consumption depends on the waiting time, buffer size or bandwidth

In Table 1, we will discuss the energy consumption of data transmission, including the cost of connection establishment, data transfer, connection maintenance and tail, by using the network of 2G, 3G, WIFI and SMS (SS7). We take the energy consumption to establish, to maintain and to end a connection into account as "connection" in the table.

Since the payload of data transmission in MCS, including datagrams for both the command word of task assignment and sensory data result, is less than 50 bytes, the data transfer, at most, costs 0.5 mJ (by WIFI) to 0.15 J (by SS7). Such energy consumption can almost be ignored by the cost to make twice connections for task assignment and data collection, i.e., 4 to 24 Joules. Therefore, by using arbitrary communication methods of 3G, GSM, WIFI and SMS (SS7), the cost of connection holds almost all of the energy in data transmission of sensory data for a few bytes. All above measurement and instrumental results are investigated from the work [2], [6], [9]; and interested readers are encouraged to see also in these papers.

Please note that, in the rest of this paper, we mainly focus on the energy consumption of data transmission by using 3G and SMS, due to the low availability of WIFI and the upgrading of 2G network.

2.2 Energy Consumption of Sensors

The power of sensors, including accelerometer, pressure, temperature, microphone and compass sensors, equipped by the mainstream mobile phones are also covered by Table 2. Particularly, we take care of the sensor energy consumption under various frequency and duty cycles settings, so as to succeed different

Sensing Task	Sensors (frequency, duty cycle)	Energy (J)	Total Energy (J)
Human Activity Monitoring	Accelerometer (160Hz, 10%) Microphone (1Hz, 50%) Compass (1Hz, 10%) Pressure (1Hz, 100%)	0.66 0.755 0.015 0.0006	1.43
Environment Monitoring	Temperature (1Hz, 20%) Microphone (1Hz, 20%)	0.0012 0.3	0.3

TABLE 2: Energy Cost of Sensors and Sensing Tasks

sensing tasks, e.g., environmental monitoring and human activity recognition.

The instrumental results listed in Table 2 is measured by work [7], [8], [11]. It shows that the human activity recognition task involves more sensors with higher working frequency and more intensive duty cycles than executing the environmental monitoring task. Therefore, it consumes much more energy than environmental monitoring task as shown in "Total" column. Obviously, based on above two sensing tasks, we can see the energy cost of mobile phone-equipped sensors. Confidently, we can conclude that sensory task execution consumes less energy than the total connection cost of data transmission in two-way.

Composing the energy consumption of two-way data transmission with sensory task execution, we still find that the cost of connection, including the consumption for connection establishment, maintenance and tail, consumes the most of energy through the whole process. For example, for each assigned MCS task, the crowds life logging which relies on the high-energy consumed human activity recognition, costs 4 in 5.43 Joules by data transmission via SMS, or consumes 24 in 25.43 by data transmission via 3G. In summary, if we can cut off the energy of connection cost, then it is possible to reduce the total energy consumption of MCS on mobile phone significantly.

3 Improving Sensory Data Transmission Through Reusing the Connections of Voice Call and SMS

In this section, we mainly introduce the technical issues to transfer data in parallel by reusing the connections of phone call and SMS. Besides, the implementation of mobile crowdsensing based on the reuse of connections is also discussed.

3.1 The Reuse of Connections

The technique to transfer data in parallel of voice call or SMS was illustrated in Figure 2. This technique was thought energy efficient, since the MCS application no longer needs to establish its own wireless connections, but it reuses the existing connections to upload or download data. For example, MCS encodes the sensory data of few bytes into the blank of a SMS. Besides, [5] presents their evaluation of parallel data transfer via voice call on Nokia N95. It shows they can have a 32KB/sec channel without affecting

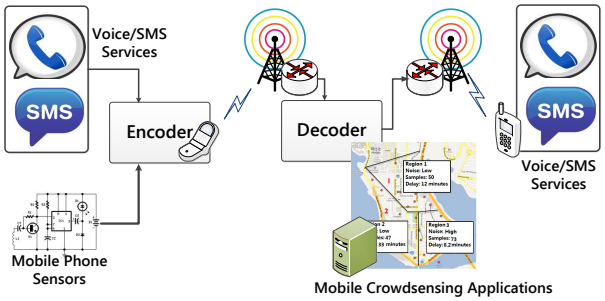


Fig. 2: Reuse of Connections

the quality of voice call; and the data transmission without involving any significant overhead. Finally, to transfer the sensory data of few bytes or a command word of task assignment through a 32KB/sec channel, we believe the data transfer costs zero energy, since this technique, at least, cuts off the energy cost to establish, to maintain and to end a connection.

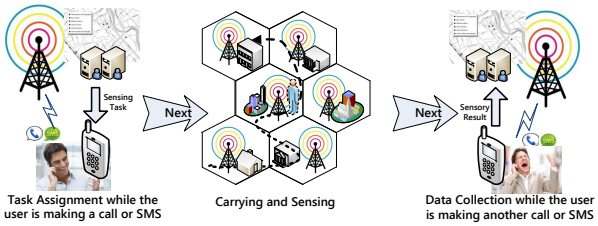


Fig. 3: MCS by Reusing Call/SMS Connections

3.2 Mobile Crowdsensing by Reusing Wireless Connections

As the diagram shown in Figure 3, *EEMC* is designed to reuse the wireless connection of current phone call or SMS to assign task in parallel of the voice or text, and then it will try to collect the sensory result by reusing the connection of another call or SMS in future. Through our investigation in section 2, the cost to make a connection consumes the most of energy of MCS on mobile phone. *Therefore , by reusing the connections of SMS and calls, EEMC is confident to reduce the total energy cost of mobile phone for MCS task significantly.*

4 Task Assignment and Solution

Still, we are hard to identify the users who will have another phone call or text in short time; and

it becomes so difficult to assign task. That means if we assign the tasks to too few guys who have an incoming call during the sensing period, then it may be impossible to collect enough samples within tolerable delay. Or, oppositely, if we assign the task to so many or even arbitrary users , then it occurs huge energy overhead caused by sensory task execution of those users who has been assigned with sensing task but won't have another phone call or text within tolerable delay. Therefore, the core of *EEMC* is indeed the algorithms that minimize the overhead of task assignment as well as guarantees to collect minimum number of samples within maximum tolerable delay.

In this section, we introduce the problem in task assignment, and propose a set of task assignment algorithms, including two baselines and one solution.

4.1 Problem Formulation of Task Assignment

First of all, we introduce the concept of a *sensing task*. It characterizes the location and time to conduct the mobile crowdsensing. Furthermore, it specifies our requirement to a **valid sample**, including the minimum sensing duration and maximum tolerable delay. These two features concern the validity and effectiveness of a sensing result.

Definition 1. Sensing Task is defined as $S = (R, T_o, C_y, D_v, D_m)$, where:

- R is a set of cellular towers which covers the $region$ of this sensing task;
- T_o is the starting time of the sensing task;
- C_y refers to the duty cycle of sensing task meaning that the sensing task S will be assigned during $[T_o, T_o + C_y]$;
- D_v refers to the minimum sensing duration to sense a valid sample, since usually the sensors need work for a while to guarantee the validity of sensing; and
- D_m is the maximum tolerable delay of the sensing task meaning that the sensing task times out after $T_o + C_y + D_m$.

Besides the sensing task, we propose another concept – *monitoring task* to track the progress of task assignment as well as data collection, and to set our expected number of valid samples for a specific sensing task.

Definition 2. Monitoring Task is defined as $M = (S, N_{exp}, A_s, D_c, N_{valid}, S_d,)$ where:

- S identifies a specific sensing task;
- N_{exp} counts the expected number of valid samples;
- A_s refers to a set of mobile phones who were assigned the sensing task S;
- D_c is the set of mobile phones who responses a valid sample within the maximum tolerable delay; and
- N_{valid} counts the number of valid samples respond in real case, i.e., $N_{valid} = min(N_{exp}, |D_c|)$. Due to the nature of connection reuse, we can infer that $A_s \geq N_{valid}$ always.

Since the historical Call/SMS behaviors of a user can predict when she would have a phone call or SMS, therefore we include the historical Call/SMS traces of each user on any region as a part of our task assignment algorithms.

Definition 3. Call Trace is defined as $Trace(u, R) = < t_1, t_2 \dots t_n >$, where:

- u identifies a specific mobile phone, and $u \in U$ the full set of all phones;
- R refers to a sepcific region;
- $< t_1, t_2 \dots t_n >$ means the time series when U made a call or SMS in region R.

Problem Definition. Suppose current time is t. The current monitoring task on region R is $M = (S, N_{exp}, A_s, D_c, N_{valid}, S_d)$, mobile phone u has been connecting with the cell tower of region R for a call or text. The problem is to define a trade-off function

$$f : u \mapsto \{true, false\}$$

to decide whether the platform assigns the sensing task S to u, based on all provided information. The trade-off function f should be optimized to collect enough valid samples as our expectation N_{exp} with the minimized overhead. Therefore, the design of trade-off function should optimally achieve the following optimization process–

- **Prior Criteria:** If $f(u) = true$, then platform assigns u with the task S and $\{u\} \cup A_s \rightarrow A_s$.
- **Posterior Criteria:** If u finally responses a valid sample of S within tolerable delay D_m, then $\{u\} \cup D_c \rightarrow D_c$.
- **Goal:** *minimize* Overhead= $|A_s| - N_{valid}$
- **Constraint:** $|D_c| \geq N_{exp}$

Please understand that, due to the uncertainty of future call event, we may not find the solution for above optimization with the settings of a short delay D_m or a large mount of expected samples N_{exp}.

4.2 Algorithms of Task Assignment

In this section, we implement two baseline algorithms: flooding and first-K, as well as an optimal solution– *Frequency Selection of Human-carrier*.

4.2.1 Flooding

Given a monitoring task M, flooding algorithm assigns task to arbitrary users who connects the cellular tower of region R by call or SMS. However to avoid the bias of sampling, we will not assign task to those users who have been assigned with the same task. Therefore, the trade-off function of flooding algorithm is formulated as Equation 1.

$$f(u) = \begin{cases} true & \text{if } u \notin A_s \text{ of M,} \\ false & \text{else.} \end{cases} \quad (1)$$

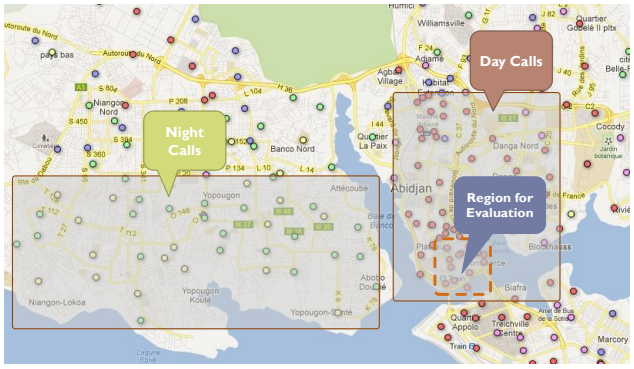


Fig. 4: Spatial Patterns of Call/SMS in Abidjan City

4.2.2 First-K

Being similar with flooding algorithm, given a monitoring task M, first-K algorithm assigns task to first $K * N_{exp}$ users who connects the cellular tower of region R by call or SMS during C_y, and it also refuses to re-assign task to the same user. Therefore, the trade-off function of first-K algorithm is formulated as Equation 2.

$$f(u) = \begin{cases} true & \text{if } |A_s| < k * N_{exp} \text{ and} \\ & u \notin A_s \\ false & \text{else.} \end{cases} \quad (2)$$

4.2.3 Frequency Selection of Human-carrier

This algorithm considers more about the call/SMS behavior of users. It selects the most frequent users in the region R of given monitoring task M. In our research, we select top F users (**Freq**) who have made the most phone calls and SMS at the given region during last two weeks. Our algorithm automatically ignores the users out of **Freq**, and assign task to the first $k * N_{exp}$ users during C_y as first-K. The trade-off function is addressed as Equation 3.

$$f(u) = \begin{cases} true & \text{if } u \in \textbf{Freq}, |A_s| < k * N_{exp} \\ & \text{and } u \notin A_s, \\ false & \text{else.} \end{cases} \quad (3)$$

All above algorithms will be evaluated on D4D dataset in section 5. The result shows *Frequency Selection of Human-carrier* outperforms other two in overhead reduction and sample collection within given delay.

5 Evaluation

In this section, we will present our evaluation of *EEMC* on D4D dataset [3]. Firstly, we will introduce the spatial-temporal patterns of call/SMS in Abidjan city where *EEMC* is evaluated. Then, we mainly stay focused at the performance of proposed task assignment algorithms under various settings of the maximum tolerable delay and expected number of valid samples.

Parameters	Values
Starting Time T_0	2011.12.7 10:00
Duty Cycle C_y	30 minutes
Valid Duration D_v	5 minutes
Maximum Tolerable Delay D_m	5-60 minutes
Expected Valid Samples N_{exp}	20, 30, and 40

TABLE 3: Settings of Experiments

Setting	algorithm	$K * N_{exp}$	Participants
Flooding	*Flooding*	n/a	all users
First-K 2	*First-K*	2	all users
First-K 1.5	*First-K*	1.5	all users
FSHC (1.5, 500)	*FSHC*	1.5	top 500 users

TABLE 4: Settings of Algorithms

5.1 Description of D4D Dataset for Evaluation

Figure 4 illustrates the geographic distribution of cellular towers. The cellular tower in red identifies that the most (85%) of call/SMS happen in the day time– i.e.,06:00-18:00, while the white point represents the cellular tower where 85% phone calls or SMS occur during 00:00-06:00 and 18:00-24:00. Therefore, we label the region with cellular towers as the selection of box "Day Calls" and as "Night Calls".

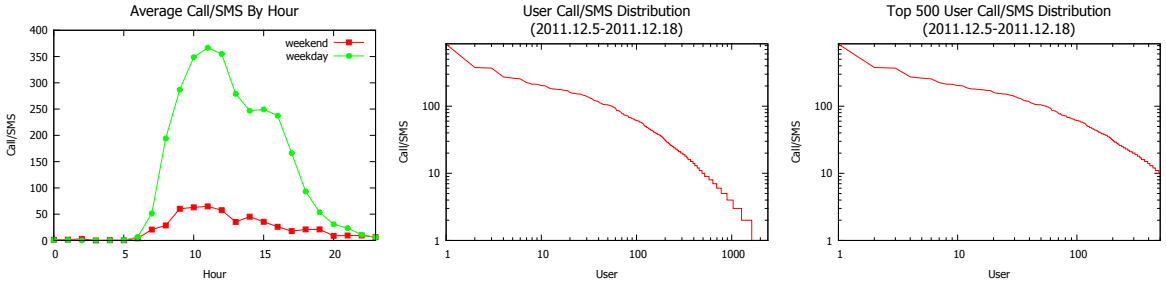
The region for evaluation was selected as the red dash box from "Day Call" area. To understand the regularity of Call/SMS in selected region, we make statistics of its Call/SMS traces as Figure 5. The figure 5a shows that place is very active in the working hours and seems as a commerce area. The number of Call/SMS per each user is illustrated in figure 5b, and the Call/SMS of top 500 users is presented in figure 5c as well. They show the frequency of phone call follows the scale-free law. Top 10 users can almost have 10 or even more Call/SMS per day. It shows the top 500 users can cover the most of Call/SMS in the whole region.

5.2 Experiment Result

We set the parameters of the experiments as Table 3. It sounds that 30 minutes for duty cycle of sensing task (C_y); and 5 minutes for minimum sensing duration (D_v) are appropriate for most sensing task scenarios. As the experiment area has 91 different individual callers during the experiment sensing task duty cycle (2011.12.7 10:00-10:30), it means that at most 91 participants would be assigned the task in our crowdsensing mechanism. Accordingly, we choose 20, 30 and 40 as the N_{exp}, representing about 20% to 50% of the total potential participants, which we believe is sufficient for most crowdsensing tasks.

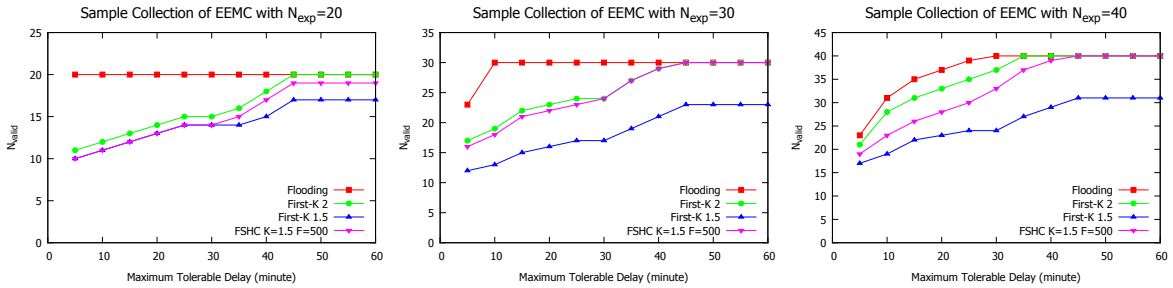
We evaluate *EEMC* on the real traces with a list of algorithms. The settings of these algorithms are shown in table 4. Figure 6 presents the evaluation result of these algorithms.

There are 2 figures for each N_{exp} setting: N_{valid} and $Overhead$. First, let's look at Figure 6c. This 'N_{valid}' figure shows how many valid samples crowd

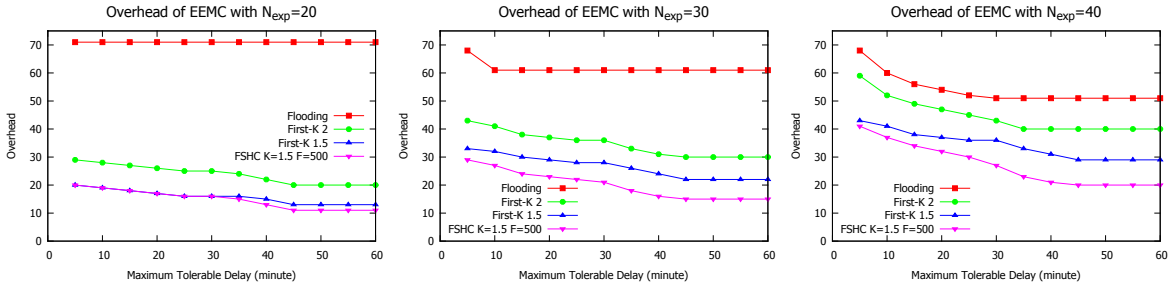


(a) Temporal Patterns of Call/SMS (b) Users vs. Call/SMS (c) Top 500 Users vs. Call/SMS

Fig. 5: Statistics on the Region of Evaluation



(a) N_{valid} with $N_{exp} = 20$ (b) N_{valid} with $N_{exp} = 30$ (c) N_{valid} with $N_{exp} = 40$



(d) *Overhead* with $N_{exp} = 20$ (e) *Overhead* with $N_{exp} = 30$ (f) *Overhead* with $N_{exp} = 40$

Fig. 6: N_{valid} vs. *Overhead* with various N_{exp}

sensing task could receive at different D_m(maximum tolerable delay) when $N_{exp} = 40$. When an algorithm gets N_{exp} valid samples, the crowd sensing task is finished successfully. Undoubtedly, *Flooding* beats the other algorithms for receiving N_{exp} samples in the smallest D_m. *First-K(K=2)* and *FSHC(K=1.5, F=500)* get N_{exp} samples by using a larger D_m that is 10-15 minutes longer than *Flooding*. Meanwhile, *First-K(K=1.5)* doesn't get enough responses even in D_m *=60 minutes*, where *First-K(K=1.5)* still lacks 9 samples to reach N_{exp}.

Second, let's move to Figure 6f. This '*Overhead*' figure shows how many users that are assigned in the crowd sensing task but don't respond a valid sample at different D_m when $N_{exp} = 40$. The smaller *Overhead* is, the less energy is wasted in assigning useless users. *FSHC(K=1.5, F=500)* is the most energy-efficient among these 4 algorithms, which is only 49% and 60% *Overhead* compared to *Flooding* and *First-K(K=2)* when $D_m > 45minutes$. In the previous Figure 6c, we see that when $D_m > 45minutes$, *FSHC(K=1.5, F=500)* receives N_{exp} samples like *Flooding* and *First-K(K=2)*. It means that when D_m is large enough for *FSHC(K=1.5, F=500)* to get N_{exp} samples, the saved energy is significant compared to the other algorithms that also receive N_{exp} results.

On the other hand, it's reasonable that *FSHC(K=1.5, F=500)* is more energy-efficient than *First-K(K=1.5)*. Because both these algorithms assign $1.5 * N_{exp}$(in this case, 60) users and *FSHC(K=1.5, F=500)* can receive more valid samples, it's definite that *FSHC(K=1.5, F=500)* will get less *Overhead* users. More importantly, these results show that by taking user's call frequency into account, simple algorithm like *FSHC* could get a great performance progress over *First-K* (i.e. larger possibility of receiving enough samples in a short maximum tolerable delay while assigning the same amount of participants).

The other figures in Figure 6 are the results of $N_{exp} = 20$ and $N_{exp} = 30$. The tendency is sim-

N_{exp}	Flood	First-K 2	FSHC	Orig(3G)	Orig(SMS)
20	130	57	**43**	509	109
30	130	93	**64**	763	163
40	130	114	**93**	1017	217

TABLE 5: Estimating the Total Energy Consumption (Joules) for *Crowds Life Logging* with maximum tolerable delay $D_m = 45$ minutes

ilar to the condition when $N_{exp} = 40$, which we illustrated previously. However, the 'N_{valid}' result for $N_{exp} = 20$ (Figure 6a) shows that although *FSHC(K=1.5, F=500)* gets more valid samples than *First-K(K=1.5)*, it still does not achieve N_{exp} samples even for $D_m = 60minutes$. This figure discloses the limitation for *FSHC*: whether it will finish the crowd sensing task is heavily dependent on the setting of K and F. Currently, it's hard for us to choose the appropriate K and F that could both receive enough samples in a reasonable short delay to finish a specific crowd sensing task, and produce few $Overhead$ users to save energy.

5.3 Discussion

In terms of energy consumption, the performance of *EEMC* can be converged to the estimation of joules. Table 5 presents the estimation of energy consumption in joules for the *Crowds Life Logging* tasks with maximum tolerable delay $D_m = 45$ minutes. The result is based on the total amount of task assignments shown in figure 6. Orig (3G) and Orig (SMS) refer to the original crowdsensing mechanism based on 3G or SMS data transmission. The energy consumption for Orig (3G) is $12 * 2 + 1.43 = 25.43J$ for each user, while the energy cost of orig (SMS) is $2 * 2 + 1.43 = 5.43J$ for each user. Besides, ideally, they assign task as many as we expect–i.e., assign totally 40 tasks for $N_{exp} = 40$. As the energy cost of *EEMC* for each user is only $1.43J$ through reusing connections, the total energy cost of Flood, First-K and FSHC is obviously less than Orig (3G) and Orig (SMS). Even though *EEMC* will assign more tasks than its expected number of valid samples, it can still effectively reduce the total energy cost. All in all, to collect 40 valid samples within 45 minutes, *EEMC* based on FSHC costs 43% of energy that was consumed by the Orig(SMS), and costs 9% of energy that was consumed by the Orig (3G).

6 Conclusion and Future Work

In this paper, we analyze the energy consumption of mobile phone for MCS task execution. We propose a novel crowdsensing mechanism *EEMC* to reduce the energy consumption of each participants and minimize the total energy cost of crowds. *EEMC* relying on the reuse of connections could extend the time of battery life and enhances the mobile user's experience, since it cuts of the energy consumption to establish, maintain and tail a connection for the data transmission. We use D4D dataset to evaluate *EEMC* and verify its feasibility and energy efficiency. The result shows, *EEMC* can effectively collect sufficient samples of sensing results with ultra low energy consumption. The low energy cost may encourage mobile phone users with higher willingness of participatory. In terms of future work, we plan to release *EEMC* mobile phone client software to Mobile APP store soon. We would like to see how *EEMC* can be used to enable real-world monitoring applications in developing countries.

7 Acknowledgment

This work is supported by the EU FP7 Project MONICA (No. 295222) and EU FP7 Project SOCIETIES (No. 257493).

References

[1] J.P. Bagrow, D. Wang, and A.L. Barabási. Collective response of human populations to large-scale emergencies. *PloS one*, 6(3):e17680, 2011.

[2] N. Balasubramanian, A. Balasubramanian, and A. Venkataramani. Energy consumption in mobile phones: a measurement study and implications for network applications. In *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*, pages 280–293. ACM, 2009.

[3] V.D. Blondel, M. Esch, C. Chan, F. Clerot, P. Deville, E. Huens, F. Morlot, Z. Smoreda, and C. Ziemlicki. Data for development: the d4d challenge on mobile phone data. *arXiv preprint arXiv:1210.0137*, 2012.

[4] R.K. Ganti, F. Ye, and H. Lei. Mobile crowdsensing: Current state and future challenges. *Communications Magazine, IEEE*, 49(11):32–39, 2011.

[5] J.K. Nurminen and J. Nöyränen. Parallel data transfer with voice calls for energy-efficient mobile services. *MobileWireless Middleware, Operating Systems, and Applications*, pages 87–100, 2009.

[6] GP Perrucci, FHP Fitzek, and J. Widmer. Survey on energy consumption entities on the smartphone platform. In *Vehicular Technology Conference (VTC Spring), 2011 IEEE 73rd*, pages 1–6. IEEE, 2011.

[7] B. Priyantha, D. Lymberopoulos, and J. Liu. Littlerock: Enabling energy-efficient continuous sensing on mobile phones. *Pervasive Computing, IEEE*, 10(2):12–15, 2011.

[8] N. Priyantha, D. Lymberopoulos, and J. Liu. Eers: Energy efficient responsive sleeping on mobile phones. In *Workshop on Sensing for App Phones*, 2010.

[9] A. Rice and S. Hay. Decomposing power measurements for mobile devices. In *Pervasive Computing and Communications (PerCom), 2010 IEEE International Conference on*, pages 70–78. IEEE, 2010.

[10] J.L. Toole, M. Ulm, M.C. González, and D. Bauer. Inferring land use from mobile phone activity. In *Proceedings of the ACM SIGKDD International Workshop on Urban Computing*, pages 1–8. ACM, 2012.

[11] Y. Wang, J. Lin, M. Annavaram, Q.A. Jacobson, J. Hong, B. Krishnamachari, and N. Sadeh. A framework of energy efficient mobile sensing for automatic user state recognition. In *Proceedings of the 7th international conference on Mobile systems, applications, and services*, pages 179–192. ACM, 2009.

[12] Z. Yan, V. Subbaraju, D. Chakraborty, A. Misra, and K. Aberer. Energy-efficient continuous activity recognition on mobile phones: An activity-adaptive approach. In *Wearable Computers (ISWC), 2012 16th International Symposium on*, pages 17–24. IEEE, 2012.

The Differing Tribal and Infrastructural Influences on Mobility in Developing and Industrialized Regions

Alexander Amini, Kevin Kung, Chaogui Kang, Stanislav Sobolevsky, Carlo Ratti

SENSEable City Lab,
Massachusetts Institute of Technology,
77 Massachusetts Ave,
Cambridge, Massachusetts 02139, USA

***Abstract*—This study leverages mobile phone data to analyze mobility patterns in developing countries, especially in comparison to more developed countries. Developing regions, such as the Ivory Coast, are marked by a number of factors that may influence mobility, such as less infrastructural coverage and maturity, less economic resources and stability, and in some cases, more cultural and language-based diversity. By comparing mobile phone data collected from the Ivory Coast to similar data collected in Portugal, we are able to highlight differences in mobility patterns—such as differences in likelihood to travel, as well as in the time required to travel—that are relevant to consideration on policy, infrastructure, and economic development. Moreover, our study illustrates how cultural and lingual diversity in developing regions (such as Ivory Coast) can present challenges to mobility models that perform well for less culturally diverse regions.**

I. Introduction

Transportation and communication networks form the fabric of developed nations. The rollout of such infrastructure in developing regions can play a major role in supporting, or deterring, a region's ability to thrive economically and socially. Likewise, citizens' use of these networks can tell us much about the region. Use of communication and transportation networks can tell us how ideas and disease may be spreading, or how to most effectively augment services, such as healthcare and education [1].

Studies of mobile phone data have given us insight on numerous aspects of human mobility [2][3][4][5]. However, these studies tend to focus on regions with the highest mobile phone coverage, which also happens to be in more stable, mature, and developed regions. Thus, the models developed based on this data, are a reflection of these developed regions. However, these highly industrialized and wealthy regions represent less than one-third of the world's population, with the remaining two-thirds living in developing and poor regions. These developing regions are facing the most rapid demographic and economic shifts worldwide, and are in great need of such models to help inform policy makers, urban planners, and service providers. Yet, little work has been done to assess the appropriateness of models conceptualized for industrialized regions for use in developing regions.

The Data4Developement (D4D) dataset was collected from cell towers in the Ivory Coast and released for research purposes, so that developing regions could also be analyzed. With over 60 distinct tribes [18], Ivory Coast boasts rich cultural and linguistic diversity, in addition to its rapid urbanization. These contrasting social interactions offer researchers a unique opportunity to understand the communication and mobility patterns and needs of a developing nation during key phases of its transformation.

Our goal was to leverage the contrast between mobility data from Ivory Coast and a more industrialized nation (Portugal) in order to assess the ability of human mobility models developed for industrialized regions to accurately model developing regions.

We started with a comparison of the bulk characteristics of human mobility in the two countries: Ivory Coast and Portugal. We computed metrics such as the probability density function of agglomerated jump sizes in migration and the radii of gyration. We then delved into an analysis of the commuting patterns between the two countries.

While differences in metrics such as likelihood of migration and mean migration distance could easily be explained by differences in infrastructural coverage and maturity, we also assessed regional partitioning.

To better understand the regional partitions, we leveraged a suite of community detection algorithms. By applying these algorithms to both Ivory Coast and Portugal, we immediately started to see surprising differences in the fundamental structure of the mobility networks. For example, the official administrative boundaries of Ivory Coast were not well aligned with the communities detected, unlike communities detected in Portugal. Given the cultural and linguistic diversity of Ivory Coast, we also investigated the alignment of detected communities with tribal regions and identified striking improvements.

To assess the implications of these structural differences, we tested human mobility models, such as the Radiation Model [5] and the Gravity Model of Migration [7]. Both models predict mobility fluxes across regional partitions, such as city, county, or state boundaries. We found that the communities that respected the tribal and cultural partitioning of the region enabled significantly better mobility modeling than those formed by administrative boundaries.

Our findings shed new light on the applicability of metrics and models conceptualized for industrialized regions, to developing regions. Our results demonstrate the importance of considering cultural and linguistic diversity in the construction of new models to address the challenges of developing regions. The insights gained from our study have important applications to policymaking, urban planning, and the services deployments that are transforming Ivory Coast and many other developing countries.

In the following sections, we provide additional details on the data used in this study, the results derived, and the conclusions drawn.

II. Data Description

We used five datasets to assess and compare the human mobility patterns in the Ivory Coast and Portugal. The first two datasets, D1 and D2, were provided by Orange telecom as SET1 and SET2 respectively, via the Data for Development (D4D) Challenge [8]. Both datasets are based on anonymized Call Detail Records (CDRs) of 2.5 billion calls and SMS exchanges between 5 million users December 1, 2011 until April 28, 2012 (150 days).

SET1 provided the number and duration of all calls between any pair of antenna, aggregated hour. Calls spanning multiple time slots were considered to be in the time slot they started in. Communication between Orange customers and customers of other providers were removed.

SET2 contains consecutive call activities of each subscriber over the study period. Each record in this dataset represents a single connection to an antenna and contains the following fields: timestamp, anonymized ID of the user, and the antenna ID they connected to. To further anonymize this data, the original dataset was subsampled to the calls of 50K randomly sampled individuals for each of 2-week periods in the dataset. The geographical positions of the antenna for D1 and D2 were also provided and are illustrated in Figure 1. Records without antenna IDs were removed; 107 antennae had no calls and 128 antennae had no population movements.

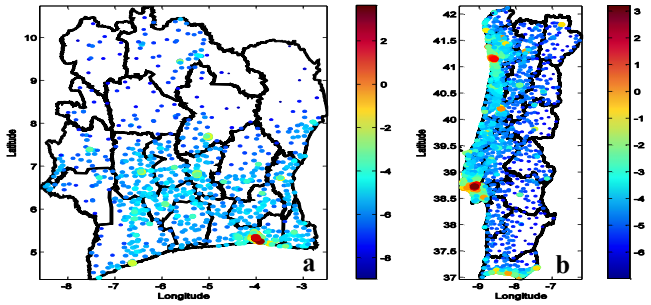


Figure 1 Geographical locations of mobile phone antennas used in this study for Ivory Coast (a) and Portugal (b). The size and color of the antennas are logarithmically mapped based on the density of antennae in that region.

The third dataset, D3, was also provided by Orange from 6511 antennae distributed across Portugal as shown in Figure 1. D3 included the same fields as D2, for 400 million anonymized CDRs from 2 million users, for the time period of January 1, 2006, to December 31, 2007.

Datasets D4 and D5 provided a high-resolution population density data for Ivory Coast [30] and Portugal [31], respectively. To map the population data to the antennae, we created a Voronoi tessellation [19] of each country based on the antennae location. For the 12 locations that had 2-3 antennas in a single location, those 2-3 antennas were collapsed into a single Voronoi cell. Each antenna was assigned the total population within the corresponding Voronoi cell. Figure 2 provides a logarithmic scale population map using the data from D4 and D5. The population maps were created as an interpolation of the populations assigned to each antenna.

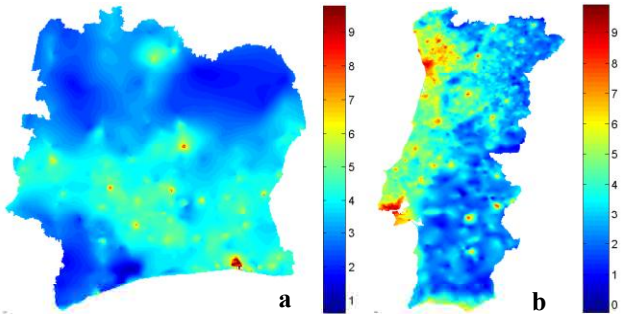


Figure 2: Population map of Ivory Coast (a) and Portugal (b). Populations are mapped on a logarithmic scale.

III. Collective Mobility Patterns

We first performed a bulk mobility pattern analysis based on D2 and D3 by plotting the probability density function $P(\Delta r)$ of the individual travel distances (or jump sizes) Δr in a trace of agglomerated de-identified callers over a period of two weeks, for Ivory Coast (left plot, solid black line) and Portugal (right plot, solid black line), as shown in Figure 3. We found that the distributions were qualitatively similar to each other except that at the administrative level, the distributions are much more scattered than those observed in Portugal, suggesting greater regional variance. We attempted to fit the density function to a truncated power law of the form $P(\Delta r) = (\Delta r + \Delta r_0)^{-\beta} \exp(-\Delta r/\kappa)$, as described in [1], where Δr_0, β, and κ are the fit constants. While the two distributions had similar cutoff distance ($\kappa_{Portugal} = 106 \pm 10$ km; $\kappa_{Ivory} = 122 \pm 5$ km), the two countries have slightly different power law coefficients ($\beta_{Portugal} = 1.37 \pm 0.06$; $\beta_{Ivory} = 1.62 \pm 0.03$).

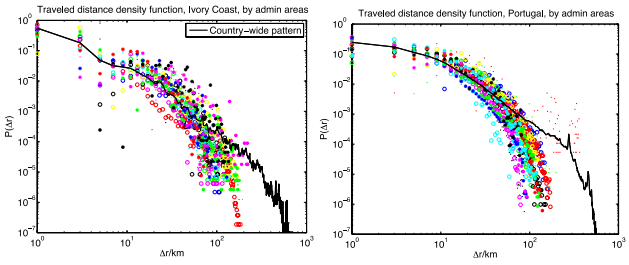


Figure 3 Probability density functions of agglomerated jump sizes in migration, as defined earlier in the text as $P(\Delta r)$, plotted for Ivory Coast (left) and Portugal (right). The solid black line indicates the country-wide pattern. Different administrative regions are identified by different scatter marker types and colors. Data taken from sets D2 and D3 as defined above.

We also investigated whether there are regional differences in the mobility pattern. In both cases of Ivory Coast and Portugal we identified the first level administrative boundaries as the highest country-defined level of partitioning. For Ivory Coast these are called "régions", while in Portugal they are referred to as "districts". We partitioned the mobility data by the different level-one administrative regions and overlaid the same density functions specific for each administrative region on the same plots above. Different administrative regions are identified by different scatter marker types and colors. We observed the same truncated power law behavior across the different regions, but the Ivory Coast regions exhibited significantly greater diversity than similarly defined regions in Portugal.

Another important metric for assessing mobility patterns is the radius of gyration for the different callers, as defined by the mean squared variance of the center of mass of each user's trajectory. We computed the radius of gyration using the same method described in [2] and constructed the probability density functions in the same manner as described above in Figure 3, we observe similar qualitative behaviors.

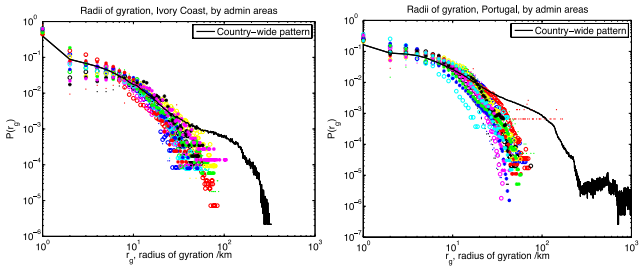


Figure 4 Probability density functions of radii of gyration, defined as the mean squared variance from the center of mass of the trajectory of each user, plotted for Ivory Coast (left) and Portugal (right). The solid black line indicates the country-wide pattern. Different administrative regions are identified by different scatter marker types and colors. The plots illustrate that both countries follow the same paradigm outlined in [2] and strengthens the validity of comparison between the two countries as we do later. Data taken from sets D2 and D3 as defined above.

The distributions are plotted in Figure 4 and show that the bulk mobility data from Ivory Coast adheres well within this the scale-free framework proposed in [2]. The similarity in the bulk mobility characteristics between Ivory Coast and Portugal serves to strengthen the argument that we can make valid comparisons between the two datasets, as described in the sections below.

IV. Commuting Patterns

Daily commuting patterns are a critical component of any region's mobility requirements. Displacement is defined as movement from one cell tower to another cell tower between two consecutive calls, and is a key marker for assessing mobility. We computed displacement over 40 minute windows for Ivory Coast and Portugal. To focus on daily commuting patterns, we excluded data collected during weekends, and computed the fraction of inter-call events that were accompanied by displacements in each 40-minute window. We averaged the fraction of displacement for each 40-minute window across 45 weekdays to get a 24-hour temporal profile of the probability of displacement during a workday.

The first and probably the most significant difference is the absolute difference in the probability of displacement. We observe that in Portugal, in a given period, people are much more mobile compared to their counterparts in Ivory Coast.

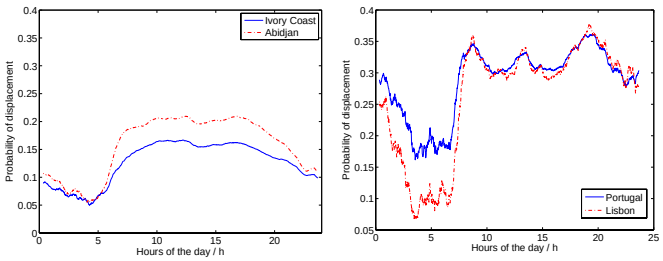


Figure 5 Probability of displacement in Ivory Coast (left) and Portugal (right), with the nation-wide data in solid blue lines. We also compared displacement levels between the respective capital cities, Abidjan and Lisbon, as shown in dashed red lines.

Both countries exhibit a commuting pattern; there is a sharp rise in the probability of displacement around 7-9 a.m. The evening decline is not as sharp, suggesting that people leave work at different times in the evening.

Significant quantitative differences between the countries can also been seen throughout the day. In Portugal, people in Lisbon and across the nation exhibited similar likelihood to commute during the busiest hours. However, a significantly higher percentage of people in Abidjan were mobile than across the nation. Additionally, while displacement levels in Abidjan and across Ivory Coast were similar during the lowest period (4-7 a.m.). Displacement for the same period is significantly higher for Portugal than for Lisbon, and is likely an indicator of more significant numbers of suburban commuters in Portugal than in Ivory Coast.

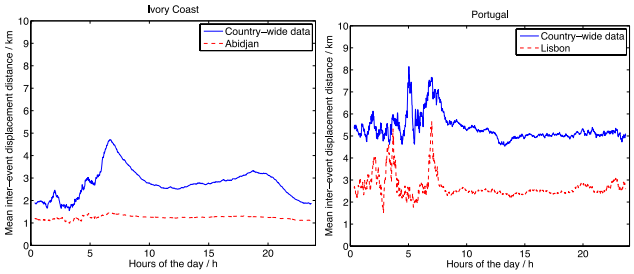


Figure 6: Comparison of mean inter-event migration distance in Ivory Coast (left) versus Portugal (right), for country-wide data (solid blue lines) and the capital city (red dashed lines). Data taken from sets D2 and D3 as defined above.

Figure 6 provides a comparison of the mean migration distances between the 2 countries for the same period. Here again we note that the average distance traveled is significantly less in Ivory Coast and its capital city, than in Portugal. In the country-wide data, we observe a sharp increase in the mean inter-event displacement distance near the morning peak commute (around 5-9 a.m.) in both Ivory Coast and Portugal. However, we also note a difference in the spike of distance encountered in Lisbon during morning commute does not occur in Abidjan. This difference may be indicative of people both living and working in close proximity in Abidjan, as opposed to commuting in from outside or across the city as is often the case in developed regions with more comprehensive public transport facilities.

We then decided to examine the country-specific commuting pattern more closely by looking at how the distance commuted may affect the daily behaviors. To do this, we put all the observed distances traveled into specific bins (0-1 km, 1-5 km, 5-10 km, 10-20 km, and 20-50 km). We then computed the daily temporal profile of the probability of displacement for each bin for the two countries, as shown in Figure 7.

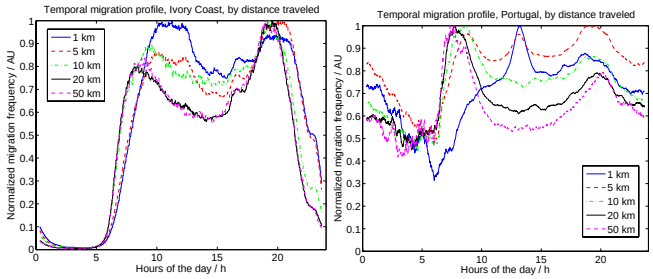


Figure 7 Temporal migration profiles in a typical workday for Ivory Coast (left) and Portugal (right) for displacements of different distances: 0-1 km (blue solid lines), 1-5 km (red dashed lines), 5-10 km (green dash-dotted lines), 10-20 km (black solid lines), and 20-50 km (magenta dashed lines). Data taken from sets D2 and D3 as defined above.

By looking at the figures closely, we can draw many interesting and relevant insights regarding commuting. Firstly, let us focus on Ivory Coast (left plot). We see that regardless of the distances commuted, the temporal profiles show a bimodal pattern: a morning peak (around 9 a.m.) as well as an evening peak (around 7 p.m.), as expected from a typical commuting pattern. Between the two peaks, there is a valley which deepens as the distances commuted increases. This also makes intuitive sense, as people who live far away from their work place are likely only going to make the long commute twice a day, and there are only rare occasions during the work day where such a migration is required. Whereas, for shorter distances traveled (such as the solid blue line or the dashed red line), this valley is less prominent. This difference likely is explained by the fact that during the day, people at work may make frequent short trips, such as to visit their clients, to replenish their inventories, etc.

If we compare this pattern in Ivory Coast with that in Portugal (right plot), we observe further differences. First, we note that while the bimodal behavior is largely preserved in Portugal, it is less prominent for shorter commuting distances (in particular, in the 0-1 km bin, as shown in the solid blue line).

In fact, in bins associated with shorter displacement distances, we also observe a third peak around midday (13:10 for the 0-1 km bin and 13:09 for the 1-5 km bin, respectively), which is absent in Ivory Coast and in the bins with longer distances traveled in the Portugal data. This third peak likely represents people going for lunch, and it is only observed in short-distance displacements, as it is rare for people to travel for long distances for lunch when they are at work. The absence of this feature in the Ivory Coast data can be accounted for by the fact that in Portugal, workers are required not allowed to work for more than five consecutive hours and are subsequently forced to a break for at least one hour in the middle of the workday, which gives them more chance to travel further for lunch or other reasons [32].

The third interesting observation is that depending on the distances traveled, the bimodal peaks seem to shift in time, and this pattern is especially noticeable in the Ivory Coast data, where we see a shift towards earlier commuting for people making longer travels. In order to quantify this pattern further for both countries, we computed the peak migration time as a function of distance traveled, as show in Figure **8** for the morning (left plot) and evening (right plot) of both Ivory Coast and Portugal.

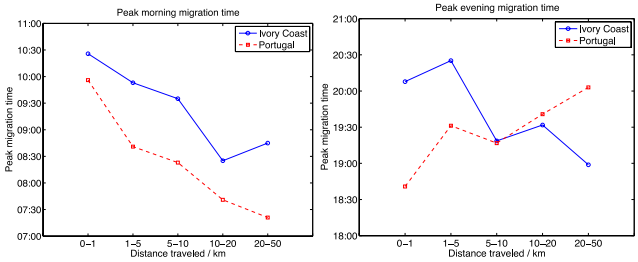


Figure 8 The estimated peak migration times for morning (left) and evening (right) as a function of distance traveled for the country-wide data for Ivory Coast and Portugal. Data taken from sets D 2 and D 3 as defined above.

As we see in the peak morning migration time on the left plot, there is a negative trend between the peak morning migration time and the distance traveled: the further the distance, the earlier the peak migration time. This makes intuitive sense, as in order to get to work on time, a person living further from his/her work place needs to set off from home earlier, both due to the long distance and due to the possibility of traffic. Assuming a typical 8-hour workday, we observe that the peak migration occurs later for Portugal, which may imply that the average commuting time at a given distance is shorter than Ivory Coast. This may make sense if we assume that Portugal has more options for travel compared to Ivory Coast (in addition to cars and shared buses)—such as commuter rails, metro trains, etc. As shown in Table 1 below, both countries demonstrate a statistically significant negative correlation between peak morning commute time and the distance traveled, as evaluated by the Spearman's rank correlation test [22] (both ρ, the correlation coefficient, as well as P, the p-value, are given).

On the other hand, the pattern for the peak evening commute time (right plot of Figure 8) is more ambiguous. While there seems to be a negative correlation in Ivory Coast (i.e. people who live further from work go home earlier), the correlation seems to be positive in Portugal (i.e. people who live further from work go home later). One potential explanation is that due to the limited number of long-distance transportation options for people in Ivory Coast, commuters are forced to travel according the availability of the transportation systems. However, the relatively strong transportation system, and urban infrastructure of Portugal gives commuters a wider variety of options to travel home when they please.

Country	Morning commute	Evening commute
Ivory Coast	$\rho = -0.9$; $P < 0.02$	$\rho = -0.8$; $P < 0.05$
Portugal	$\rho = -1.0$; $P < 10^{-6}$	$\rho = 0.9$; $P < 0.02$

Table 1 Spearman's correlation test for peak commute time as a function of distance traveled

The mobility metrics discussed in this section highlight a number of challenges with respect to mobility in Ivory Coast. Lower likelihood of travel and mean migration distances may be indicative of limited or poor quality transportation infrastructure. However, they may also point to more systemic and fundamental problems with how services are rolled out across communities and regions. To further investigate this potential, we used community detection algorithms to probe social structure.

V. Community Analysis

Large networks, such as the telecommunications or transportation networks of a nation, often exhibit community structure, i.e., the organization of vertices into clusters with many edges joining vertices of the same cluster and comparatively few edges joining vertices of different clusters. Identifying the community structure in such networks has many applications, such as better placement and provisioning of services.

Network modularity [23] is a measure of the strength of the division of a network into clusters. Networks with high modularity have dense connections between nodes within clusters, and sparse connections between nodes in different clusters. Modularity is computed as the fraction of edges that fall within a cluster, minus the expected such fraction if the edges were distributed at random. The value of modularity lies in the range [-1 1], and is positive if the edges within groups exceeds the number expected on the basis of chance.

We computed network modularity as:

$$Q = \frac{1}{2m}\sum_{ij}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\sigma(c_i, c_j) \qquad (1)$$

where A_{ij} is the weight of the link from i to j, k_i is the sum of the weights from node i, c_i is the community that node i was assigned to, $m = \frac{1}{2}\sum_{ij} A_{ij}$, and $\sigma(c_i, c_j)$ is 1 if $c_i = c_j$ and 0 otherwise.

High modularity in mobility networks may point to an efficient organization of residences, employment, and services all in close proximity, or it may point to restrictive policies or infrastructures that limit free movement across communities. We were interested in the community structure of developing nations, such as Ivory Coast, especially in comparison to more developed nations, such as Portugal.

We used D1 and D3 to identify the community structure of antennae within the mobile phone networks of Ivory Coast and Portugal. We set nodes to the locations of each cell tower, and edges to the sum duration of all calls between these cell towers. We tested the following Community Detection algorithms: Lovain [28], Le Martelot [26], Newman [27], Infomap [29], and a new method of community detection suggested in [20]. We computed the modularity of community structures identified by each of these methods. The method described in [20] provided the highest modularity, and was subsequently chosen to be used for this part of the study.

An especially interesting difference in the communities identified for Ivory Coast and those identified for Portugal was the similarity between identified communities and the official administrative boundaries of the nations. While the communities identified for Portugal exhibited high similarity with the 20 official administrative boundaries (districts), this was not the case for the Ivory Coast's 19 official administrative boundaries (régions).

Figure 9 graphically compares the communities identified (in color) with their first level administrative boundaries (outlined in black). We also tested the similarity quantitatively by computing the clustering coefficient between the administrative boundaries of each nation with the communities detected. As shown in Table 2, communities identified for Portugal show significantly higher similarity (as much as 46% higher clustering coefficients) to administrative boundaries, in comparison to that of the Ivory Coast.

While this significant difference in community and official boundary alignments may be attributable to the layout of infrastructure along official boundaries, we began to question whether there might be more fundamental differences. Previous studies have shown that other factors, such as geographical features, can play an important role in how communities are formed and services are sought [24][25]. However, little has been done to investigate the direct impact of culture and language on human mobility.

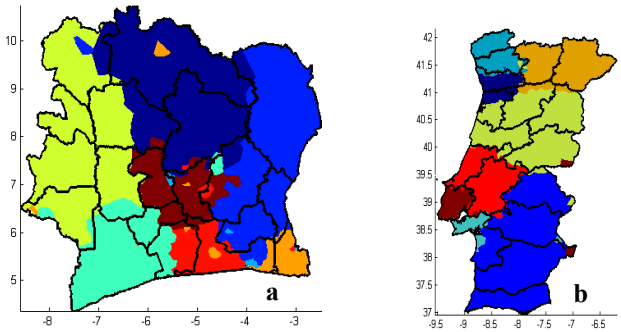


Figure 9 Community Detection of Ivory Coast (a) and Portugal (b). Generated communities are represented by different colors, while black lines represent the first level administrative boundaries of each country.

Clustering Index	Ivory Coast	Portugal	Percent Increase
Wallace [11]	0.368314	0.68835	46.49296
Adjusted Rand [12]	0.311090	0.57217	45.62926
Jaccard [13]	0.259467	0.45858	43.41893
Fowlkes-Mallows[14]	0.414960	0.63117	34.25502
Melia-Heckerman[15]	0.435379	0.62461	30.29574
Hubert [10]	0.650485	0.80197	18.88880
Larsen [16]	0.495615	0.57454	13.73665
Rand [12]	0.825242	0.90098	8.406486

Table 2 Comparison of different clustering coefficients to compare similarity between generated communities and the respective administrative boundaries.

Ivory Coast represents an especially interesting context to investigate cultural and linguistic influences on mobility within a single nation. The Ivory Coast is a nation made up of more than 60 distinct tribes, classified into 5 principle regions [18]. The official language is French, although many of the local languages are widely used, including Baoulé, Dioula, Dan, Anyin and Cebaara Senufo, and an estimated 65 languages are spoken in the country.

Intuitively, these cultural and linguistic differences are likely to influence mobility patterns in the region. However, it is also known that as regions become more urbanized, cultural ways are often blended or lost altogether. Portugal represents an interesting context for the latter, as Portuguese is the single national language of Portugal, and any tribal boundaries pre-date Roman times.

In the next section, our goal was to understand if tribal boundaries would be evident in the community and mobility patterns of a region.

VI. Tribal Community Analysis

Since the communities detected in Section IV exhibited low similarity to administrative boundaries, we modified the community detection approach to use the tribal boundaries as the Level 1 boundaries. We then ran a hierarchical community detection using the Louvain method [28] in each of these Level 1 tribal partitions. The Louvain method provided the closest final number of partitions to the administratively defined subprefectures of the algorithms that we tested, and was thus chosen as the least biased algorithm to use when generating these communities and comparing them to the administrative subprefectures. By doing so, we were able to generate sub-tribal communities while also conserving the physical shape of each tribal region. Figure 10a shows the official prefecture and subprefecture boundaries, while Figure 10b shows the tribal and sub-tribal communities that were created using this approach.

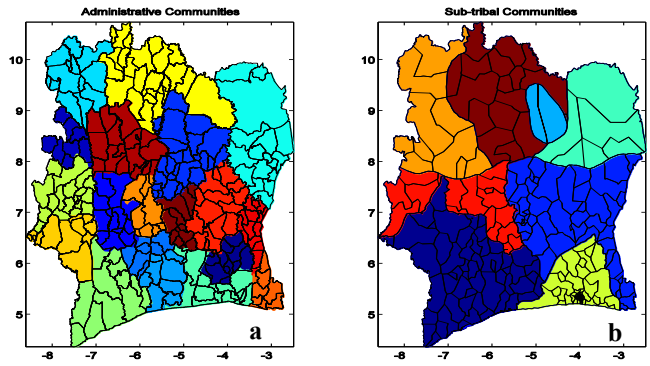


Figure 10 Comparison between the administrative subprefecture (a) boundaries and the generated sub-tribal (b) boundaries.

	Administrative Subprefectures	Sub-tribal Communities
Network Modularity	0.6158	0.6548
Population (IQR [21])	5.46E+04	8.38E+04
Area km^2 (IQR [21])	1.06E+03	1.82E+03
Partitions	255	217
Hubert Index [10]	0.8110	

Table 3 Statistical Comparison of Administrative subprefectures to the generated sub-tribal communities

Table 3 shows key metrics for the Administrative Communities and the Tribal Communities. The number of communities identified was similar and also present similar population and area distributions; however, the actual communities themselves show a relatively low similarity index.

As a first measure of impact of tribes on mobility, we computed the mobility network for each community.

Figure 11 provides a plot of the mobility network with each node representing a community, and each edge colored to reflect the number of migrations between the connected nodes in the mobility network. The intra-tribal community mobility network is plotted separately from the inter-tribal community mobility network, in order to facilitate comparison. The intra-tribal network plots only those edges between communities in the same tribe. The inter-tribal network plots only edges between communities of differing tribes.

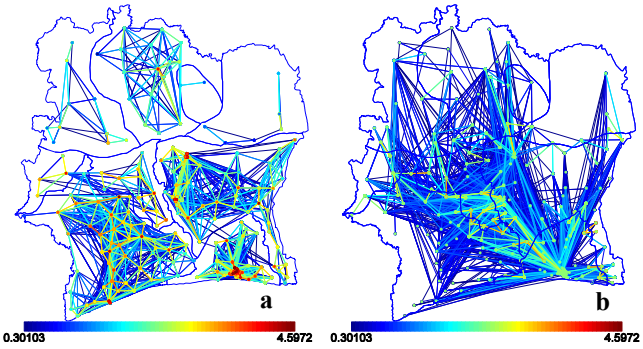


Figure 11 Intra (a) and Inter (b) tribal migrations over the entire dataset. The color of the edges is logarithmically mapped based on the total flux of migrations between the two nodes.

This diagram provides a first insight into tribal influences on mobility. Note that the number of intra-tribal migrations (as indicated by the color coding of edges) dwarfs the number of inter-tribal migrations. Additionally, the inter-tribal migrations are largely dominated by connections to the largest city, Abidjan.

We quantified the strength of these tribal ties by computing the network modularity of the sub-tribal communities versus that of the administrative boundaries.

The network modularity [23] of the sub-tribal communities was 0.6548, in comparison to a network modularity of 0.6158 for the administrative boundaries, showing again that mobility patterns have a stronger connection in a sub-tribal country partitioning compared to that of an administrative partitioning.

We also investigated whether the use of these sub-tribal communities would provide an advantage in modeling the mobility network of the Ivory Coast. The Radiation Model [5] was recently proposed as a parameter free mobility model in which individuals move and interact based on the population density of the source and destination regions, and

that of the surrounding regions. Using the Radiation Model, the average flux between two regions i and j is:

$$T_{ij} = T_i \frac{m_i n_j}{(m_i+s_{ij})(m_i+n_j+s_{ij})} \qquad (2)$$

where i and j are locations with populations' m_i and n_j respectively, at distance r_{ij} from each other, with s_{ij} representing the total population in the circle of radius r_{ij} centered at i (excluding the source and destination population). T_i signifies the total flux that originates from region i.

We tested the accuracy of the Radiation Model on both the administrative and sub-tribal communities of Ivory Coast. We computed the Radiation model using dataset D2, and specifically modeled the migrations between communities. We plotted the ground truth flux (x-axis) versus the predicted flux (y-axis) for both sets of partitions in Figure 12.

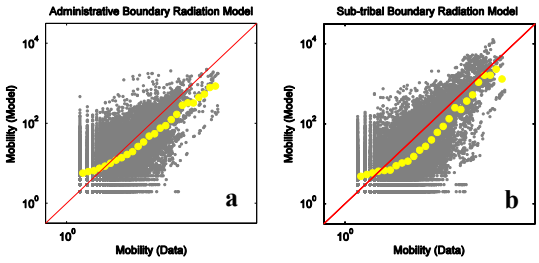


Figure 12 Testing the Radiation Model using call networks aggregated on an administrative level (a) and sub-tribal boundaries (b).Yellow data points represent the mean value in each bin of the scatter data, while the red lines represents y=x.

We also quantified the high accuracy of mobility model under sub-tribal communities (Figure 12b) by computing the Mean Absolute Percentage Error (MAPE) [8] for each partitioning. MAPE values range from [0, ∞) and are an accepted measure of the error a model presents against a given set of truth data.

We compute MAPE values according to:

$$M = \frac{100\%}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right| \qquad (3)$$

where A_t is the actual value, F_t is the forecasted value, and n is the number of data points.

Figure 13 shows the normalized MAPE comparison for the sub-tribal and administrative boundaries, and demonstrates that the MAPE for predictions made via administrative boundaries ranged from 20% to 50% higher than the MAPE of the sub-tribal communities. The plots illustrate the higher accuracy (lower MAPE) produced when sub-tribal communities were used, as opposed to administrative boundaries. This would indicate that, in terms of mobility patterns of Ivory Coast, it is more efficient to model mobility on a partition that accounts for tribal, cultural, and lingual differences in groups of people, as opposed to the administratively defined country partition.

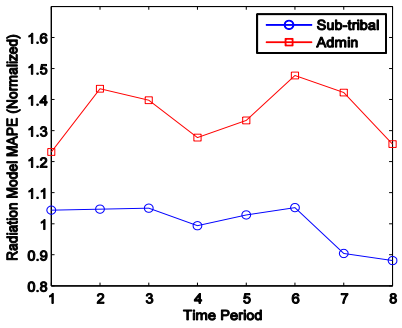


Figure 13 MAPE of the radiation model when using sub-tribal and administrative boundaries for each of the two week datasets given in D2.

We also partitioned the mobility model predictions according to intra-tribal and inter-tribal flux in order to quantify the strength of the connectivity of the tribes. Figure 14 illustrates the comparison of intra- to inter- tribal flux and the accuracy of the model predicting these fluxes, and supports the dominant pattern of intra-tribal migrations over inter-tribal migration. Quantitatively, the MAPE for the inter-tribal mobility was 11.3% higher than the MAPE of the intra-tribal mobility predictions. Again, the fact that the Radiation model produces more accurate results for migrations within a single tribe compared to those between tribes suggests that the tribes themselves are playing a key role in the overall improved accuracy of the model.

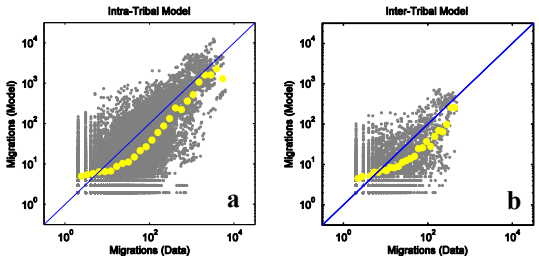


Figure 14 Testing the Radiation Model accuracy segregating intra (a) and inter (b) tribal migrations. Yellow data points represent the mean value in each bin of the scatter data, while the blue lines represents y=x.

In order to further explore the relationship between tribal and administrative boundaries we decided to apply an alternative approach for modeling human mobility and interaction, the Gravity Model [7]. The Gravity Model is formulated on Newton's Law of gravity, and predicts flux between a source and destination based on the populations of the source and destination, and the distance between the source and destination. More specifically, according to the gravity model, the average flux migrations from regions i to j is:

$$T_{ij} = \frac{m_i^{\alpha} n_j^{\beta}}{r_{ij}^{\gamma}} \qquad (4)$$

where i and j are locations with populations m_i and n_j respectively at a distance of r_{ij} from each other. α, β, and γ are adjustable parameters chosen to fit the data.

We found the Gravity Model did not improve the accuracy for either the administrative or the sub-tribal communities of Ivory Coast. Figure 15 compares the predicted probability of migration versus the ground truth migration, for both the Gravity Model and Radiation Model. As a more direct comparison of accuracy, Figure 16 provides the error (MAPE) for the models plotted in Figure 15. For the Ivory Coast, these figures show the higher accuracy (i.e., lower error) of Radiation Model for both administrative and sub-tribal communities, and it shows that using the Radiation Model with sub-tribal communities provides the highest accuracy (i.e., lowest MAPE).

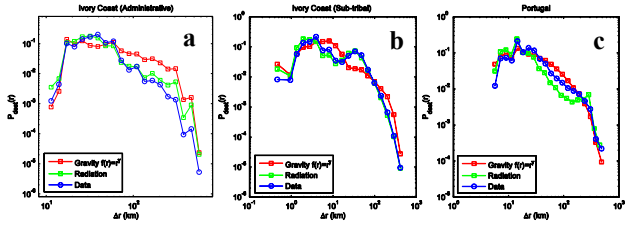


Figure 15 Comparison of the migration models for Ivory Coast's administrative (a) and sub-tribal (b) boundaries, as well as in Portugal's administrative partitioning (c). Each signal represents the probability of migration to a location 'r' kilometers away from the originating location.

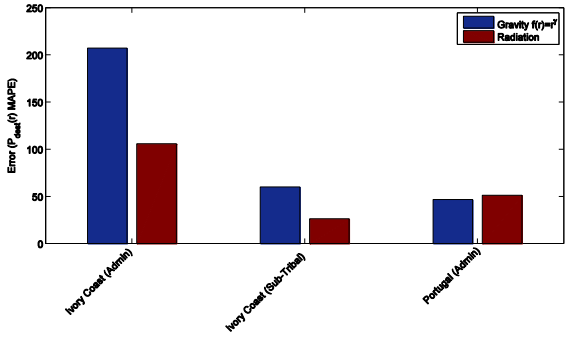


Figure 16 The error (MAPE) of each model applied to each region of Figure 15 when predicting the probability of migrations that occur at a distance of r kilometers from their originating location.

Interestingly, we can also see in Figure 15 that the Portuguese administrative municipality boundaries perform well for both the Radiation and Gravity Models. This may be indicative that the municipal boundaries were designed to align with cultural and social communities or that cultural and social communities have adapted to fit administrative boundaries.

However, we believe a more likely explanation is the growing homogeneity of language and culture that comes with maturing industrialization and urbanization. This is reflected in the predominance of Portuguese as the national language in Portugal, compared to the more than 60 local languages spoken in Ivory Coast.

There are several important implications from these findings.

1. Models of mobility, migration, and interaction that are conceptualized in mature, industrialized, and urbanized regions may not directly map to developing regions with more pronounced cultural and linguistic differences. Such models need to better account for these differences.
2. If administrative boundaries are drawn and services placed based on models that do not accurately reflect these influencers, results could include inefficiencies, leading to inequality of services (e.g., longer or less accessible commutes), and potentially discrimination and alienation of segments of the population.
3. Additional analysis could provide better insights on how cultural and linguistic impacts how we move and interact, how cultural and linguistic diversity evolve as urbanization increases, and how urbanization might better support diversity and inclusion, instead of just increasing homogeneity.

VII. Concluding Remarks

Africa is a continent that has been shaped by human migration over tens of thousands of years. Indeed, migrations within and beyond the African borders have recently been shown as influencing all civilizations as we know them. However, until recently, there has been a dearth of data on the forms and patterns of migration within the nations of Africa. Moreover, much of the mobility research is based on theories that have emerged from highly industrialized nations.

Our study has demonstrated that these conceptions are not necessarily applicable in the African context. We have made these differences clear by comparing our findings in Ivory Coast to one such industrialized nation, Portugal. For example, we have shown that the probability of displacement during normal commuting hours in Portugal is often nearly double that of Ivory Coast for the same time of day. Similarly, average distances traveled by commuters in Portugal is nearly double that of commuters in Ivory Coast.

While differences in the likelihood of travel and average distance travel can be attributed to quantitative differences in infrastructural support for mobility this already strongly affects the whole mobility picture leading to a number of quantitative dissimilarities. Our study shows evidence of more fundamental differences in infrastructural support for

mobility, such as tribal, cultural, and lingual differences. In addition, we demonstrate that the similarity between administrative boundaries and communities detected in mobile phone data is markedly lower in Ivory Coast than in Portugal.

By identifying the tribal influence on mobility in the Ivory Coast, we were able to illuminate further differences in mobility patterns. For example, we were able to show intra-tribal migrations were much more frequent than that of inter-tribal migrations over the same distance. Taking this into account by exploiting our tribally aligned communities drastically improves modeling of human mobility in Ivory Coast. We validated this higher accuracy by computing the Mean Absolute Percentage Error (MAPE) across all data points for both models, and found a 20% to 50% higher error for the administrative boundaries. We also validated our results by computing the distribution of migrations by distance migrated and found that by using this sub-tribal method of modeling human mobility we were able to improve the accuracy of the models so drastically, that the Ivory Coast performed even better than its developed country-counterpart, Portugal.

We are excited by the findings of this study and plan to further validate our findings by comparing to other developing and developed regions.

VIII. Acknowledgements

The authors wish to thank Orange Telecom for providing the datasets used in this study. We would also like to thank all the members of the SENSEable City Consortium for their support throughout the study with special thanks to Yuji Yoshimura, Riccardo Campari, Michael Szell, and Tao Pei.

IX. References

[1] Robertson C, Sawford K, Daniel S L, Nelson T A, Stephen C, 2010, "Mobile phone–based infectious disease surveillance system, Sri Lanka" Emerging Infectious Diseases 16(10), 1524

[2] González M C, Hidalgo C A, Barabási A L, 2008, "Understanding individual human mobility patterns" *Nature* 435 (5) 779-782

[3] Ratti C, Pulselli R, Williams S, Frenchman D, 2006, "Mobile Landscapes: using location data from cell phones for urban analysis" *Environment and Planning B: Planning and Design* 31(3) 453-473

[4] Reades J, Calabrese, F, Ratti C, 2009, "Eigenplaces: analysing cities using the space-time structure of the mobile phone network" *Environment and Planning B: Planning and Design* 36(5) 824-836

[5] Simini F, González M C, Maritan A, Barabási A L, "A universal model for mobility and migration patterns" *Nature* 484(5) 96–100

[6] Calabrese F, Di Lorenzo G, Liu L, Ratti C, 2011, "Estimating origin-destination flows using mobile phone location data" *IEEE Pervasive Computing* 10(4) 36-44

[7] Rodrigue, Jean-Paul. The geography of transport systems. Routledge, 2009.

[8] Swanson D A, Tayman J, Bryan T M, 2011, "MAPE-R: a rescaled measure of accuracy for cross-sectional subnational population forecasts" *Journal of Population Research* 28 (2) 255-243

[9] Vincent Blondel, Markus Esch, Connie Chan, Fabrice Clerot, Pierre Deville, Etienne Huens, Frederic Morlot, Zbigniew Smoreda, Cezary Ziemlicki. "Data for Development: The D4D Orange Challenge on Mobile Phone Data." 2012-2013.

[10] Hubert L & Arabie P, "Comparing Partitions", 1985, *Journal of Classification* 2 (1) 193-218

[11] Wallace, D L. "Comment." *Journal of the American Statistical Association* 78.383 (1983): 569-576.

[12] Rand W M, 1971, "Objective criteria for the evaluation of clustering methods", *Journal of the American Statistical Association* 66 (336) 846-850

[13] Jaccard P, 1901, "Étude comparative de la distribution florale dans une portion des Alpes et des Jura", *Bulletin de la Société Vaudoise des Sciences Naturelles* 37 547–579

[14] Fowlkes E B & Colin L M, 1983, "A method for comparing two hierarchical clusterings", *Journal of the American statistical association* 78 (383) 553-569

[15] Meilă, Marina, and David Heckerman. "An experimental comparison of several clustering and initialization methods." Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc., 1998.

[16] Larsen B, & Aone C, 1999, "Fast and effective text mining using linear-time document clustering" *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* 16-22

[17] Meilă M, 2007, "Comparing clusterings—an information based distance" *Journal of Multivariate Analysis* 98 (5) 873-895

[18] *Côte d'Ivoire*. 2013. Encyclopedia Britannica Academic Edition Inc. <http://www.britannica.com/EBchecked/topic/139651/Cote-dIvoire>.

[19] Voronoï G, 1908, "Nouvelles applications des paramètres continus à la théorie des formes quadratiques. Deuxième mémoire. Recherches sur les parallélloèdres primitifs." *Journal für die reine und angewandte Mathematik (Crelles Journal)* 1908 (134) 198-287

[20] Sobolevsky S, Szell M, Couronn T, Smoredab Z, Claxtonc R, Ratti C. "Regional delineation using networks of human interactions: the cases of France, UK, Italy, Portugal, Belgium and Singapore." *(Publication in Process)* (2012).

[21] Zwillinger D, & Kokoska S, 1999 *CRC standard probability and statistics tables and formulae* (CRC Press)

[22] Jerrold H Z, 1972, "Significance testing of the Spearman rank correlation coefficient." *Journal of the American Statistical Association* 67 (339) 578-580

[23] Newman, M E J, 2006, "Modularity and community structure in networks" *Proceedings of the National Academy of Sciences* 103 (23) 8577-8582

[24] Ratti C, Sobolevsky S, Calabrese F, Andris C, Reades J, Mauro M, Claxton R, Strogatz S H, 2010, "Redrawing the Map of Great Britain from a Network of Human Interactions" *PLoS ONE* 5(12): e14248. doi:10.1371/journal.pone.0014248

[25] Onnela J P, Arbesman S, González M C, Barabási A L, Christakis N A, 2011, "Geographic Constraints on Social Network Groups" *PLoS ONE* 6(4): e16939. doi:10.1371/journal.pone.0016939

[26] Le Martelot E, & Hankin C, 2011, "Multi-scale community detection using stability as optimisation criterion in a greedy algorithm" *Proceedings of the 2011 International Conference on Knowledge Discovery and Information Retrieval (KDIR 2011)* 216-225

[27] Newman M E J, 2006, "Finding community structure in networks using the eigenvectors of matrices" *Physical review E* 74 (3) 036104

[28] Blondel V D, Guillaume J L, Lambiotte R, Lefebvre E, 2008, "Fast unfolding of communities in large networks" *Journal of Statistical Mechanics: Theory and Experiment* 2008(10) P10008 doi:10.1088/1742-5468/2008/10/P10008

[29] Lancichinetti A, & Fortunato S, 2009, "Community detection algorithms: A comparative analysis" *Physical Review E*, 80(5) 056117. doi:10.1103/PhysRevE.80.056117

[30] *AfriPop: Cote D'Ivoire.* n.d. AfriPop. <http://www.clas.ufl.edu/users/atatem/index_files/CIV.htm>.

[31] *Portugal*. n.d. Portal do Instituto Nacional de Estatística. <http://www.ine.pt/>

[32] Working Times in Portugal. AngloInfo. http://portugal.angloinfo.com/working/eu-factsheets-working/working-time/

Approaching the Limits of Predictability in Human Mobility: A Study of 500,000 Mobile Phone Users in Cote d'Ivoire after the 2011 Civil War

Xin Lu

Flowminder Foundation, 17177 Stockholm, Sweden
Department of Public Health Sciences, Karolinska Institutet, 17177 Stockholm, Sweden
Department of Sociology at Stockholm University, 17177 Stockholm, Sweden
Department of Information Systems and Management at National University of Defense Technology, 410073 Changsha, China
e-mail: xin.lu@ki.se

Erik Wetter

Flowminder Foundation, 17177 Stockholm, Sweden
Department of Management and Organization at Stockholm School of Economics, 11383 Stockholm, Sweden
e-mail: erik.wetter@hhs.se

Nita Bharti

Department of Biology, Center for Infectious Disease Dynamics and Huck Institute of Life Sciences, Penn State University, University Park, PA 16801, USA
e-mail: nbharti@princeton.edu

Andy Tatem

Geography and Environment, University of Southampton, SO17 1BJ, UK
e-mail: andy.tatem@gmail.com

and

Linus Bengtsson

Flowminder Foundation, 17177 Stockholm, Sweden
Department of Public Health Sciences, Karolinska Institutet, 17177 Stockholm, Sweden
e-mail: bengtssonlinus@gmail.com

Abstract: Studies of mobility patterns and prediction of individual movements are important in many research fields, such as mobile computing, epidemic modeling, urban planning and disaster response. In this study we analyze travel patterns of 500,000 individuals from Cote d'Ivoire using mobile phone call record data. By measuring the uncertainties of movements using entropy, considering both the frequency and temporal correlation of individual trajectories, we find that the theoretical maximum predictability is as high as 88%. To verify whether such a theoretical limit can be approached, we implement a series of Markov chain (MC) based models to predict the actual locations visited by each user. Results show that MC models can produce next-location estimates with an accuracy of 90% and

that the first-order MC model provides equally high prediction accuracy as models of higher orders. The correlation of entropy and predictability with prediction accuracy are very high at -0.85 and 0.80 respectively. Our results indicate that human mobility is highly dependent on historical behaviors, and that the maximum predictability is not only a fundamental theoretical limit for potential predictive power, but also an approachable target for actual prediction accuracy.

Keywords and phrases: mobile phone, Cote d'Ivoire, predictability, entropy, Markov chain, prediction, accuracy.

Received February 2013.

1. Introduction

Studies of mobility patterns and prediction of individual mobility trajectories are important in many research fields, such as mobile computing, epidemic modeling, traffic planning and disaster response [15, 4, 10]. Real-time locations visited by individuals are typically collected through mobile devices equipped with global-positioning system (GPS) capability, mobile phone cell towers, or wireless local area network (WLAN) access points.

Various methods have been proposed to forecast trajectory movements, such as Markov chain (MC) models [17, 14], neural networks [13], Bayesian networks [1], and finite automaton [16]. It has been shown that the prediction accuracy may vary according to the algorithm and the context from which the location data come. For example, in an evaluation of next cell prediction based on more than 6000 users on Dartmouth's campus-wide Wi-Fi wireless network, it was found that the best predictor (the $\mathcal{O}(2)$-MC model) had an accuracy of about 65-72% [20]. In another study where mobility traces of six researchers and GPS-locations of 175 individuals were used, the prediction accuracy was shown to be in the range of 70% to 95% with an $\mathcal{O}(2)$-MC model [11, 24, 8]. Low prediction accuracy was also found in certain circumstances. For example, in an evaluation of MC models for pedestrian-movement prediction, the accuracy was found to be as low as 2%, 45% and 74.4% for the $\mathcal{O}(1)$-MC model, hidden-Markov model, and the mixed MC model, respectively [2].

Due to limited data access, the above studies investigated small numbers of individuals or special populations, and thereby the study conclusions and feasibility of the proposed new predictive algorithms can hardly be generalized to the general population.In addition, since the prediction accuracy may be constrained by the type of location data studied, it has not been clear how well these algorithms perform versus the best possible algorithm that could theoretically be constructed; i.e., what is, for the given data type, the best possible accuracy achievable and how well do the predictive algorithm perform versus such a benchmark? The highest potential accuracy of predictability, termed "maximum predictability" ($\Pi^{\max}$), is defined by the entropy of information in a person's trajectory (frequency, sequence of location visits, etc.). $\Pi^{\max}$ can be calculated by solving a limiting case of Fano's inequality (a relation derived

from calculation of the decrease in information in a noisy information channel) [12, 7, 6].

By measuring Π^{max}, Song et al showed, using a mobile phone dataset of 50,000 users in a high-income country, that there is a 93% potential predictability in user mobility, despite large differences in travel patterns such as travel distances[19]. Under much more extreme conditions and in a low-income setting, Lu et al analyzed a complete mobile phone dataset of 2.9 million anonymous subscribers after the earthquake in Haiti in 2010, and found that despite massive population movements and increased travel distances following the earthquake, the predictability of population's movements remained as high as 85%, showing fundamental regularity in human movement behavior [15]. These findings are promising for the design and improvement of predictive algorithms, however, these studies did not show how much of the potential predictability is achievable in practice.

In this study we aim to fill this gap in knowledge by measuring the maximum predictability and performance of actual prediction algorithms on a mobile phone data set of 500,000 users from Cote d'Ivoire (CIV), West Africa. We also give an overview of population mobility patterns during the data collection period, which took place after the 2011 civil war. We find that the maximum predictability and regularity in mobility in CIV is on the same high level as found in studies in Haiti and Europe [15, 19]. The evaluation of practical predictive algorithms reveals that the maximum predictability can be approached with MC-based models. Interestingly, higher order MC models in this mobile phone data set did not generate improved prediction accuracy as compared to the first order MC model.

2. Materials and Methods

2.1. The Mobile Phone Dataset

Mobility data was provided by the telecom company Orange and derived from call detail records (CDR) from a random sample of 500,000 anonymous Orange mobile phone subscribers in CIV, during December 1, 2011 to April 28, 2012. The user's location was provided as the location of the subprefecture (sous-préfecture in French) of the mobile phone tower through which the call was routed. CIV is composed of 19 regions, which are further divided into 255 subprefectures (237 of these subprefectures have at least one tower, see Fig. 1). The original CDR contains approximately five million users (1/4 of the total population of CIV) [22, 3]. Detailed description of the data can be found in [5].

The number of Orange subscriptions per person varies considerably throughout the country, as does the overall population density in CIV. For example, the region Lagunes which includes the economic capital Abidjan, is the home of 25% of the CIV population and is the most frequently visited location for 43% percent of the mobile phone subscribers in this dataset (see Fig. 2A). The distribution of the number of location updates (calls and SMS) follows a log-normal

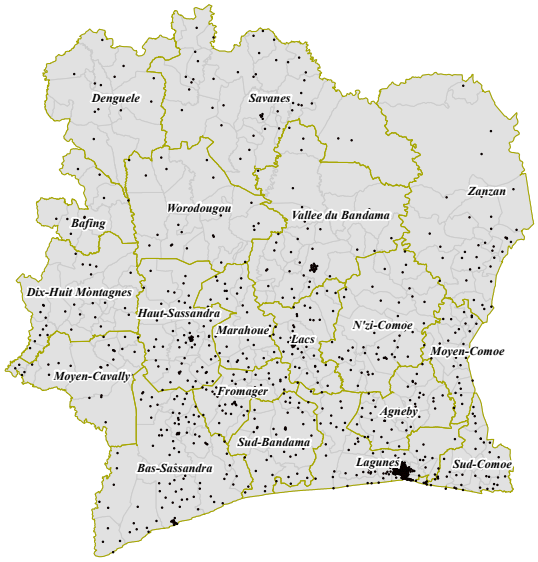


FIG 1. *Administrative map of Cote d'Ivoire and distribution of cell phone towers*

distribution (Fig. 2B), with 81.3% having between 100 and 2000 location updates. Seventy-seven percent of users had at least one location update per day during two thirds of the data collection period (Fig. 2C). In addition to the calling activities, there was also a high heterogeneity in the movement patterns. While sixteen percent of subscribers were only found in one subprefecture during the period, a few users were registered in more than 50 subprefectures (Fig. 2D).

2.2. Measures of Mobility

We use the average travel distance, $\bar{D}$, and the radius of gyration of the trajectory, r_g, to measure the mobility property of individuals. Specifically, Let $M_i = \{m_1, m_2, ..., m_n\}$ be the sequence of observed location updates for person i during the data period. Then $\bar{D}$ and r_g are defined by:

$$\bar{D}(i) = \sum_{j=2}^{n} |m_j - m_{j-1}| \tag{1}$$

and $r_g(i) = \sqrt{\frac{1}{n}\sum_{j=1}^{n} |m_j - \bar{m}|^2}$, where $|m_j - m_{j-1}|$ is the distance between location m_j and m_{j-1}, and $\bar{m} = \frac{1}{n}\sum_j m_j$ is the center of mass of the trajectory [9].

The radius of gyration is rather different from the average travel distance. Someone who moves in a comparatively confined space will have a small radius of gyration even though he or she covers a large distance. On the other hand, r_g

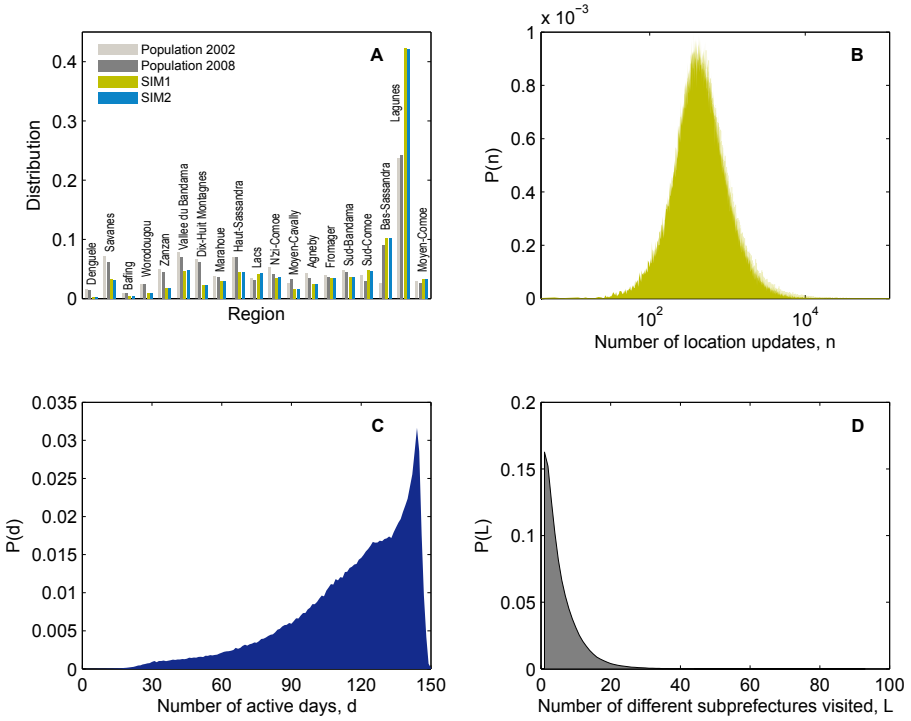


FIG 2. *Characteristics of the mobile phone users. (A) The proportion of users in each region compared to the population. Population data from 2002 is obtained from the AfriPop project [21], and the 2008 comes from data made by UN OCHA and CNTIG [22]. SIM1: the number of users who made their first calls in this region; SIM2: the number of users who appeared for the majority of their time in this region. We use SIM1 and SIM2 to approximate the number of residential mobile phone users in each region. (B) The distribution of number of observations for each user during the data collection period. Note that the x-axis is logged. (C) The number of active days on which each user had made at least one call. (D) The distribution on the number of different subprefectures visited by each user.*

can be larger than $\bar{D}$ if someone travels with small steps but in a fixed direction or in a large circle. Note that in the dataset we only know the location of each individual by subprefecture, consequently, the centroid of each subprefecture is used to approximate the coordinates of individuals. Such an approximation would introduce imprecision for the measure of travel distances, but still provides useful information when comparing mobility between users as those who traveled many subprefectures will have larger r_g and $\bar{D}$ than those who stay most of their time in one or two subprefectures.

2.3. Measures of Entropy and Predictability

We are primarily interested in the stable, long-term patterns of population mobility behavior as opposed to short-term movements. Here we therefore focus on entropy and predictability analysis of day-to-day movement of individuals. Let $X_i = \{x_1, x_2, ..., x_T\}$ be the sequence of daily locations for person i during the data period of T days. x_j is the last observed location ID of person i on day j, otherwise we mark x_j "unknown". The uncertainty (or disorderness) of the trajectories can be measured by entropy. Larger entropy indicates greater disorder, and consequently reduce predictability of an individual's movements.

Entropy: Following notation in [19] we measure: (i) the random entropy, $S_i^{rand} = \log_2 L_i$, capturing the predictability of each user by assuming that the person's whereabouts are uniformly distributed among L_i distinct locations in X_i; (ii) the temporal-uncorrelated entropy, $S_i^{unc} = -\sum_{k=1}^{L_i} p_k \log_2 p_k$, where p_k is the frequency at which the person visited the k^{th} location among the L_i distinct locations. S_i^{unc} takes into account the number of different locations visited as well as the proportion of times i spent at each location, decreasing the uncertainty of the trajectory, and; (iii) the true-entropy, $S_i^{real} = -\sum_{X'_i \subset X_i} P(X'_i)\log_2[P(X'_i)]$, where $P(X'_i)$ is the probability of finding a subsequence X'_i in X_i, considering both spatial and temporal patterns.

Predictability: Given the entropy E for an individual i, Fano's inequality gives an upper limit for the predictability of i, i.e., the level of accuracy the best possible predictive algorithm can achieve:

$$\Pi_i \leq \Pi_i^{\text{Fano}}(E, L_i) \tag{2}$$

where Π_i^{Fano} is given by

$$E = H(\Pi_i^{\text{Fano}}) + (1 - \Pi_i^{\text{Fano}})\log_2(L_i - 1) \tag{3}$$

and

$$H(\Pi_i^{\text{Fano}}) = -\Pi_i^{\text{Fano}}\log_2(\Pi_i^{\text{Fano}}) - (1 - \Pi_i^{\text{Fano}})\log_2(1 - \Pi_i^{\text{Fano}}) \tag{4}$$

Let $\Pi_i^{\text{rand}} = \Pi_i^{\text{Fano}}(S_i^{\text{rand}}, L_i)$, $\Pi_i^{\text{unc}} = \Pi_i^{\text{Fano}}(S_i^{\text{unc}}, L_i)$ and $\Pi_i^{\text{max}} = \Pi_i^{\text{Fano}}(S_i^{\text{real}}, L_i)$, since $S_i^{\text{rand}} \geq S_i^{\text{unc}} \geq S_i^{\text{real}}$, it is true that $\Pi_i^{\text{max}} \geq \Pi_i^{\text{unc}} \geq \Pi_i^{\text{rand}}$. Comparing between these three predictability measurements provides the ability to investigate how the spatial distribution and temporal correlations in the individual's trajectories improve potential predictive power. Since Π_i^{max} provides the best possible predictive power (because it utilize the maximum information from S_i^{real}) we refer to it in this paper as the "maximum predictability".

2.4. Prediction Algorithms

Predicting a user's next location using Markov chain models: To investigate how close we can get to achieving Π with actual prediction algorithms we implement several variants of Markov chain (MC) based models.

In an MC-based model, the trajectory of each individual is modeled as a Markov chain of order n , which assumes that the movement of individuals between the L_i locations is a process with limited memory in the sense that the future location is visited depending only on the previous n visited location, i.e., $P(X_i^{t+1} = x^{t+1}|X_i^t = x^t, X_i^{t-1} = x^{t-1}, ..., X_i^1 = x^1) = P(X_i^{t+1} = x^{t+1}|X_i^t = x^t, X_i^{t-1} = x^{t-1}, ..., X_i^{t-n+1} = x^{t-n+1})$, where X_i^t is a random variable representing the location for individual i at time t .

Given the previous n locations $X_i^t = x_i^t, X_i^{t-1} = x_i^{t-1}, ..., X_i^{t-n+1} = x_i^{t-n+1}$, the prediction is then determined by the transition matrix, P , choosing the destination location $x^{pre}(1 \leq pre \leq L_i)$ which maximize the probability:

$P(X_i^{t+1} = x^{pre}|X_i^t = x_i^t, X_i^{t-1} = x_i^{t-1}, ..., X_i^{t-n+1} = x_i^{t-n+1})$
$= \max_{k=1}^{L_i}\{P(X_i^{t+1} = x^k|X_i^t = x^t, X_i^{t-1} = x^{t-1}, ..., X_i^{t-n+1} = x^{t-n+1})\}$

Increases of the order n in the Markov chain do not necessarily lead to improvement in the prediction accuracy. However, to investigate the correlation of predictive powers with the length of historical locations considered, we vary n from 1 to 7 (one day to one week). If predictions for a higher ordered MC(n) model did not exist (i.e., the order of the previous n locations is unique in history), the prediction from a lower ordered model, MC($n-1$), was used.

The performance of each model is evaluated by the accuracy, γ , which is the proportion of accurate predictions from all predictions made:

$$\gamma = \frac{\text{number of correct predictions}}{\text{total number of predictions}}. \tag{5}$$

Users who were not active on a specific day were excluded from the prediction.

Next place prediction using historical frequency data: For comparison we implement a simple algorithm predicting the next location based on the most visited location in the historical trajectory: $P(x^{pre}) = \max_{k=1}^{L_i}\{p_k|X_i^t = x_i^t, X_i^{t-1} = x_i^{t-1}, ..., X_i^1 = x_i^1)\}$, where p_k is defined the same as in S^{unc}. As no temporal correlation is considered in this algorithm, we refer it as MC(0).

Using the MC models, we repeatedly update the transition matrices and the visiting frequency for each user when new locations are observed in the trajectory. We predict for each user the most likely location s/he would visit on each day based on all the historical information, i.e., for each day t, the transition matrices and visiting frequency are constructed based on the trajectory from day 1 to day $t-1$.

3. Results

3.1. Overview of Mobility and Aggregated Flows

The absolute change in the number of subscribers in each region is dominated by the changes in the region of Lagunes, where Abidjan is situated (Fig. 3A). Seven-day cyclical patterns (workday-weekend cycles) are clearly visible for several regions, e.g. Lagunes and Sud-Comoe, but other more complex trends are also evident. An irregular change of population flow was observed near the end

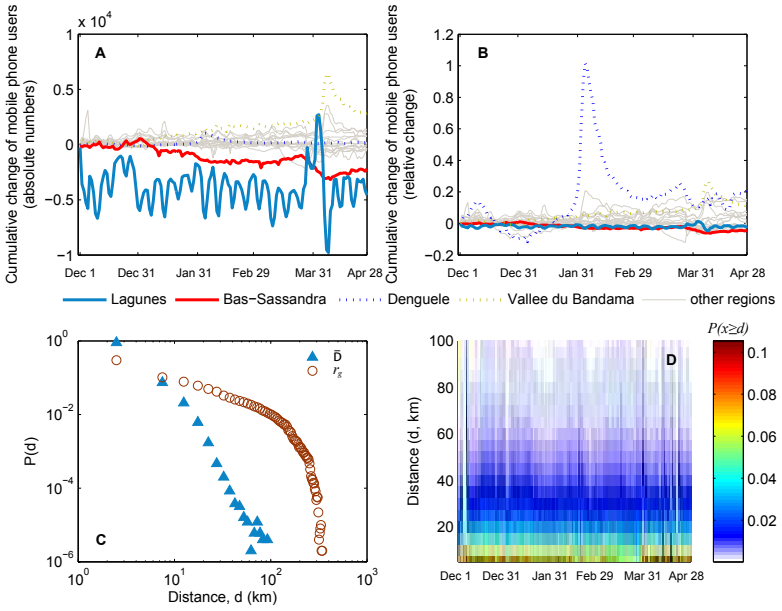


FIG 3. *Overview of population movements: (A) Shows the cumulative change in number of users in each region. (B) Shows the same data as in panel (A) but changes are shown in proportion to the number of users in each region at the beginning of the period. (C) Gives the distribution of average travel distance $\bar{D}$ and the radius of gyration, r_g. (D) Shows the cumulative probability distribution of average daily travel distance over the 150 days.*

of March and early April when the numbers of users rapidly increased in Abidjan then decreased a few days later (potentially partly related to Easter). In addition, Bas-Sassandra, in the southwest experienced a decrease during large parts of this period.

In relative terms (Fig. 3B), several regions showed considerable change over the period, dominated by Denguele, which however had a small change in absolute terms (see Fig. 3A). As we see from Fig. 3C, both the distribution of $\bar{D}$ and r_g obey a skewed decay over increasing traveling distances. While the movement of the vast majority of users were confined within an area of 10km, a few users traveled on average as far as 100 to 300 km (see also Fig. 3D where the distribution of daily average travel distances is shown). Note that the radius of gyration here is calculated from location data on the level of the sub-prefecture and thus exclude short movements.

3.2. Regularity and Potential Predictability

We now focus on the regularity of the daily observed trajectories of the users by allocating the last observed location (subprefecture) to each user's trajectory. To avoid the illusion of high predictability stemming from users with many

unknown locations, or from users who never traveled to other locations, we include users who visited at least two subprefectures, and were observable for more than 120 days in the period (208,288 users).

The distributions of S^{rand}, S^{unc}, and S^{real} are presented in Fig. 4A. We can see that, consistent with findings from previous studies, the disorderness of visited locations is greatly reduced if we consider both the spatial and temporal correlation of the visiting sequences. The median value of S^{rand} is 2.0, indicating that if we assume that individuals randomly choose a location to visit the next day, a typical individual could be found in any of $2^{2.0} \approx 4$ locations. On the other hand, if we utilize information contained in the frequency and sequence order of the trajectory of individuals, the uncertainty in a typical individual's whereabouts reduces to $2^{S^{\rm unc}} = 2^{0.91} \approx 1.88$ and $2^{S^{real}} = 2^{0.71} \approx 1.64$, in less than two locations.

Not surprisingly, the reduced uncertainty leads to increased maximum predictability, as shown in Fig. 4B. If information only available is the number of distinct visited locations, L_i, the accuracy of any predictive algorithm cannot exceed 0.35. With the additional information on frequency and temporal correlation, the average predictability increases to $< \Pi^{\rm unc} > \approx 0.84$, and $< \Pi^{\rm max} > \approx 0.88$, respectively.

In Fig. 4C, we have investigated the correlation between the radius of gyration and the average predictability, $< \Pi >$. There is a steady decrease of $< \Pi^{\rm rand} >$ and $< \Pi^{\rm unc} >$ when r_g increases (measured based on the centroid of each subprefecture). On the other hand, $< \Pi^{\rm max} >$ stays around 0.85 for a wide range of $r_g \in [20, 300]$. This finding is consistent with previous studies, revealing the independence of predictability on traveling distance in human mobility [15, 19]. However, we have also examined other travel distance measurements. Increases in average travel distances ($\bar{D}$) causes a slightly decreasing predictability. $< \Pi^{\rm max} >$ ranges from 0.9 to 0.7 when $\bar{D}$ increases from 1 to 20 km, and stays around 0.63 0.68 for $\bar{D} \in [20, 70]$. However, interestingly, predictability decreases considerably with increasing number of distinct locations visited. From Fig. 4E, we can see that the average predictability $< \Pi^{\rm unc} >$ and $< \Pi^{\rm max} >$ decays linearly with the number of different visited locations.

3.3. Prediction accuracy based on Markov-chain models

The predictability analysis in the previous section reveals that, by combining information on frequency with temporal correlation of the trajectory, the theoretical upper bound of prediction accuracy can be as high as 0.88. However, the highest prediction accuracy that can be achieved with properly designed predictive algorithms is not yet clear. In this section, we predict the location of users on each day, by considering all previous trajectory data points with the MC(n) models as described above. The accuracy of these models is presented in Fig. 5A and shows accuracies of more than 0.8 for almost all days. The accuracy of MC-based models ($< \gamma > \approx 0.90$), MC(1) toMC(7), produce substantially higher accuracies than the estimation method based only on frequency information,

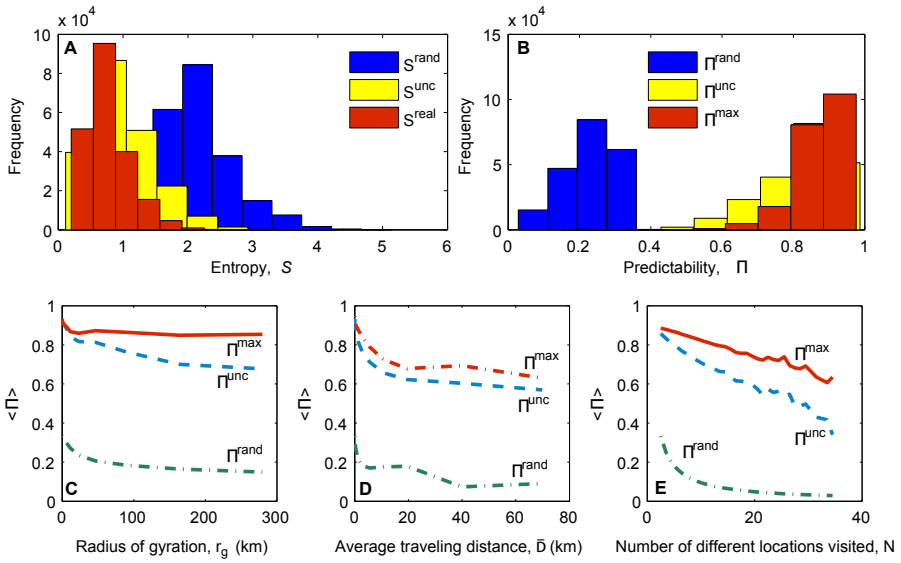


FIG 4. *Entropy and predictability analysis based on trajectory of visited subprefectures. (A) Shows the frequency distribution of* S^{rand}, S^{unc} *and* S^{real}. *(B) Shows the frequency distribution of* Π^{rand}, Π^{unc} *and* Π^{real}. *(C) Shows the correlation between radius of gyration and* Π^{max}. *(D) Shows the correlation between average travel distance* $\bar{D}$ *and* Π. *(E) Shows the correlation between the number of different locations visited and* Π.

i.e., MC(0) ($< \gamma > \approx 0.85$).

There is however little difference between the performance of the MC-based models with different orders. At the beginning of the period when historical information is limited, the accuracy of MC(1) is slightly higher than the other models, however this difference becomes very small when the historical trajectory is over 100 data points.

Another interesting finding from Fig. 5A, is that the MC-based models performs more robustly than MC(0), for example, during the later period of the data, there is a sharp decrease in the accuracy of MC(0) (from 0.88 to 0.77), while the accuracy of MC-based models sustains a much smaller decrease, from 0.92 to 0.87. Irruptions of decreased accuracy from MC(0) model indicate that people moved abnormally from regular patterns, the sustainability of MC-based models reveals that such abnormalities can be captured partly by considering the temporal correlation of visiting sequences in the trajectories.

The increase of $< \gamma >$ over the observation time is not very apparent from Fig. 5A, as $< \gamma >$ is calculated based on a mixture of users with long and short known historical trajectories. To investigate the effect of trajectory length on the performance of algorithms, we have removed, for each user, the unknown locations and calculated the average prediction accuracy for users with valid historical trajectories of the same length L_{hist} . The result is shown in Fig.

5B. We can see that the accuracy of MC(0) approaches relative stability after around 15 historical data points. For a wide range of $L_{hsit} \in [15, 120]$, $< \gamma >$ stays steady around 0.85, indicating that the visiting behavior on frequency is relatively stable over time for users with valid historical trajectories of this range. On the other hand, there is a steady increase of $< \gamma >$ for the MC-based models. When the available historical trajectories that contain more than 100 data points the average prediction accuracy climbs above 0.9.

The performance of MC-based models indicates that, while the predictability of a typical user is $\Pi^{\max} = 0.88$, which gives an upper bound for the accuracy of any predictive algorithm when the trajectory is stable, the MC-based models are able to produce estimates as high as 0.90, even higher than the theoretical upper limit. Possible reasons for why the practical algorithm can produce accuracies higher than $\Pi^{\max}$ could be, first, that the trajectory data contains only one data point for each day, which means that the maximum length of the trajectory can only be 150 and the movement pattern of individuals may have not reached stability with this length; secondly, we use the lower order MC$(n-1)$ model to generate estimates when the transition probability for MC(n) does not exist, which may have added random fluctuations to the prediction accuracy.

MC-models considering higher orders (longer correlations of previous locations) do not necessarily improve prediction accuracy. For example, for trajectories with the same historical length, the performance of the MC(4) model always produce less precise predictions compared to other MC-models (Fig. 5B). This finding is consistent with previous studies, in which the MC$(n > 2)$ models was found to not bring important improvement at the cost of a significant overhead in terms of computation and space for the learning and storing of the mobility model [20, 8]. It is worth noting that a large part of the predictive power of the studied prediction algorithms is due to the fact that many individuals spent a substantial time in his/her top visited locations. For example, users who visited four distinct subprefectures, still spent almost 80% of their time in their most visited locations (Fig. 5C).

3.4. Entropy, predictability and prediction accuracy

The evaluation of predictive algorithms above reveals that, for this data, the maximum predictability $\Pi^{\max}$ can be achieved with a first-order Markov chain model (MC(1)). In this section, we investigate whether the individual predictability, $\Pi_i^{\max}$, is correlated with the accuracy in predicting all the locations when the trajectory increases from 1 to T . We measure the individual prediction accuracy ($< \gamma_i >$) by the proportion of accurate predictions over all days for each individual (days without any location data are excluded).

First, we check the correlation between prediction accuracy and the disorderness in the trajectory, i.e., S^{real}. As we can see from Fig. 6A, $< \gamma_i >$ is highly correlated with the trajectory's entropy, the larger the entropy, the lower the prediction accuracy. The correlation coefficient between S^{real} and $< \gamma_i >$ is as high as -0.849, with $p < 0.000$. Second, we investigate the correlation between

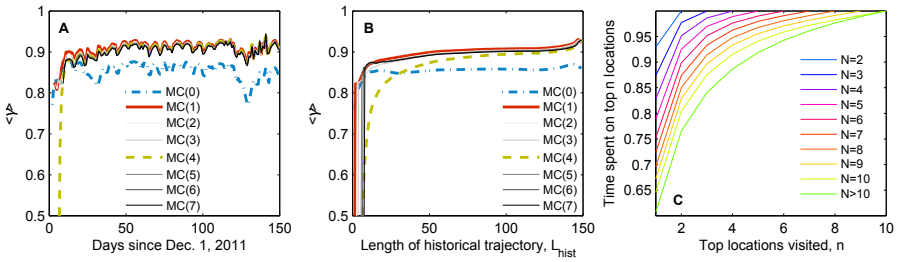


FIG 5. *Visiting behavior and prediction accuracy. (A) Shows the proportion of accurate predictions for each day based on historical data (users who were not active on a day are excluded in the prediction). In (B), the accuracy of predictive algorithms increases with the length of historical trajectories. (C) The fraction of time users spent in the top n visited subprefectures. The subscribers are divided into 10 groups based on the number of distinct locations they had visited (N).*

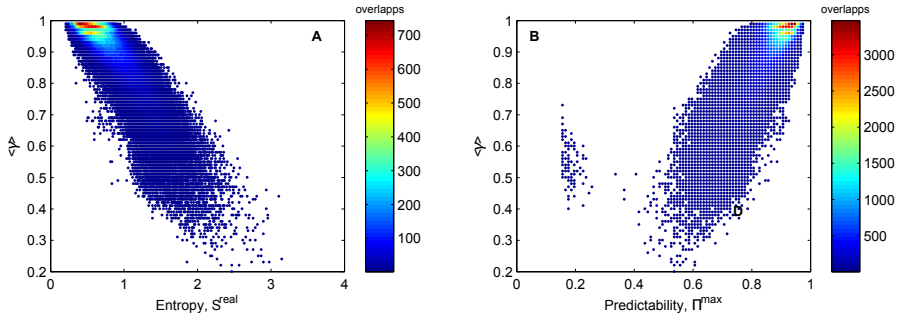


FIG 6. *Correlation between entropy, predictability and prediction accuracy. Data points are aggregated into intervals of equal lengths. (A) Shows the correlation between entropy and prediction accuracy. (B) Shows the correlation between predictability and prediction accuracy.*

prediction accuracy and the maximum predictability, i.e., $\Pi_i^{\max}$. Not surprisingly, $\Pi_i^{\max}$ also correlates highly with $< \gamma_i >$, with a correlation coefficient of 0.802, $p < 0.000$ (Fig. 6B).

The high correlation between predictability and prediction accuracy of the MC(1) model reveals that, as a measurement for disorderness and potential predictability, S^{real} and $\Pi_i^{\max}$ capture the theoretical limits for the predictive analysis of human movement behaviors, and provide an *approachable* upper bound of predictive power for this type of mobility data. More broadly, the approach used here provides an important strategy to evaluate and guide the design and improvement of mobility prediction algorithms.

4. Conclusion

In summary, by investigating the movement of 500,000 mobile phone users during the post-civil war period in Cote d'Ivoire, we have found a potential predictability in user mobility as high as 88% in this West-African lower-middle income setting. The finding of high predictability is consistent with two previous studies which investigated the mobility patterns of mobile phone users in very different settings, one in a high-income country with stable social conditions [19] and one in a low-income country following an extreme natural disasters [15] . By applying MC-based estimate algorithms, we found that the first order MC model (MC(1)) is able to produce an average predictive accuracy of 90%.

This paper is, to the best of our knowledge, the first to investigate both the predictability and the practical algorithms that aim to approach the limiting accuracy on a massive mobile phone data set. Our results not only show that the predictability of human mobility is high, but also show that this high predictability is achievable for daily population movement predictions. These findings indicate that the movement of human behaviors is far from random, and the individuals' movements are highly influenced by their historical behavior. With a deep understanding of individuals' travel patterns, public policy decision making, such as humanitarian relief, urban planning, traffic design, etc., may be significantly improved.

One would perhaps assume that Markov chains of second or seventh order would produce next-location estimates with higher accuracy, as aggregated flows based on mobile phone data frequently show weekly cycles, see e.g., Fig. 3A, and [15, 4, 19]. However, our evaluations on the MC(n) models show that this information is not necessary in this setting. This can be due to the fact that many people in Cote d'Ivoire do not have a two-day weekend, and that unplanned journeys are less common in resource-limited settings, as may travel in general [18, 23] . The trajectories used for prediction contains only the last observed location on each day, which may also bring difficulties for the time series to reach stability, and limit the performance of MC models with higher orders. Nevertheless, we believe that the evaluation of predictive performance on a daily basis is most practical for the long-term investigation of population movements. For the purpose of this study, we have only included the Markov chain based models in the evaluation of predictive performance of algorithms. Future studies may want to compare other predictive algorithms, such as mixed MC models, neural networks and finite automaton, and to evaluate the feasibility of predicting aggregated population movements with individual-based travel behavior models. For the study of achievability of predictability, stable trajectories with long historical length are needed in future investigation.

5. Acknowledgement

The authors would like to thank the operator France Telecom-Orange and the "Data for Development" committee for sharing the mobile phone dataset and organizing the D4D challenge.

References

[1] Akoush, S. and Sameh, A. Mobile user movement prediction using bayesian learning for neural networks. In *Proceedings of the 2007 international conference on Wireless communications and mobile computing* 191-196. ACM.

[2] Asahara, A., Maruyama, K., Sato, A. and Seto, K. (2011). Pedestrian-movement prediction based on mixed Markov-chain model.

[3] Bank, T. W. (2011). Population Total in Cote d'Ivoire. http://data.worldbank.org/country/cote-divoire [accessed Febuary 13, 2013].

[4] Bengtsson, L., Lu, X., Thorson, A., Garfield, R. and von Schreeb, J. (2011). Improved Response to Disasters and Outbreaks by Tracking Population Movements with Mobile Phone Network Data: A Post-Earthquake Geospatial Study in Haiti. *Plos Medicine* **8**.

[5] Blondel, V. D., Esch, M., Chan, C., Clerot, F., Deville, P., Huens, E., Morlot, F., Smoreda, Z. and Ziemlicki, C. (2013). Data for Development: the D4D Challenge on Mobile Phone Data. *arXiv:1210.0137v2*.

[6] Brabazon, A. and O'Neill, M. (2008). *Natural computing in computational finance* **1**. Springer.

[7] Fano, R. M. (1961). Transmission of information: A statistical theory of communications. *American Journal of Physics* **29** 793-794.

[8] Gambs, S., Killijian, M.-O. and del Prado Cortez, M. N. Next place prediction using mobility Markov chains. In *Proceedings of the First Workshop on Measurement, Privacy, and Mobility* 3. ACM.

[9] Gonzalez, M. C., Hidalgo, C. A. and Barabasi, A. L. (2008). Understanding individual human mobility patterns. *Nature* **453** 779-782.

[10] Kenett, D. Y. and Portugali, J. (2012). Population movement under extreme events. *Proceedings of the National Academy of Sciences* **109** 11472-11473.

[11] Killijian, M.-O., Roy, M. and Trédan, G. Beyond San Francisco cabs: Building a*-lity mining dataset. In *Proceedings of the Workshop on the Analysis of Mobile Phone Networks* 75-78.

[12] Kontoyiannis, I., Algoet, P. H., Suhov, Y. M. and Wyner, A. J. (1998). Nonparametric entropy estimation for stationary processes and random fields, with applications to English text. *Information Theory, IEEE Transactions on* **44** 1319-1327.

[13] Liou, S.-C. and Lu, H.-C. Applied neural network for location prediction and resources reservation scheme in wireless networks. In *Communication Technology Proceedings, 2003. ICCT 2003. International Conference on* **2** 958-961. IEEE.

[14] Liu, G. and Maguire Jr, G. (1996). A class of mobile motion prediction algorithms for wireless mobile computing and communication. *Mobile Networks and Applications* **1** 113-121.

[15] Lu, X., Bengtsson, L. and Holme, P. (2012). Predictability of population displacement after the 2010 Haiti earthquake. *Proc Natl Acad Sci U S*

A **109** 11576-81.

[16] Petzold, J., Bagci, F., Trumler, W. and Ungerer, T. (2003). Global and Local Context Prediction.

[17] Ross, S. M. (2009). *Introduction to probability models*. Academic press.

[18] Rubio, A., Frias-Martinez, V., Frias-Martinez, E. and Oliver, N. Human mobility in advanced and developing economies: A comparative analysis. In *AAAI Spring Symposia Artificial Intelligence for Development, AI-D, Stanford, USA*.

[19] Song, C. M., Qu, Z. H., Blumm, N. and Barabasi, A. L. (2010). Limits of Predictability in Human Mobility. *Science* **327** 1018-1021.

[20] Song, L., Kotz, D., Jain, R. and He, X. (2006). Evaluating next-cell predictors with extensive Wi-Fi mobility data. *Mobile Computing, IEEE Transactions on* **5** 1633-1649.

[21] Tatem, A. and Linard, C. (2010). High Spatial Resolution Data on Persons Per Grid Square in Cote d' Ivoire. http://www.clas.ufl.edu/users/atatem/index_files/CIV.htm [accessed Febuary 13, 2013].

[22] UN OCHA Côte d'Ivoire and Le Comité National de Télédétection et d'Information Gégraphique (CNTIG), (2011). Common and Fundamental Operational Datasets Registry. http://cod.humanitarianresponse.info/fr/country-region/c%C3%B4te-divoire [accessed Febuary 13, 2013].

[23] Wesolowski, A., Eagle, N., Noor, A. M., Snow, R. W. and Buckee, C. O. (2013). The impact of biases in mobile phone ownership on estimates of human mobility. *Journal of the Royal Society Interface* **10**.

[24] Zheng, Y., Li, Q., Chen, Y., Xie, X. and Ma, W.-Y. Understanding mobility based on GPS data. In *Proceedings of the 10th international conference on Ubiquitous computing* 312-321. ACM.

Identification and Characterization of Human Behavior Patterns from Mobile Phone Data

Pavlos Paraskevopoulos
University of Trento, Italy,
Telecom Italia - SKIL
p.paraskevopoulos@unitn.it

Thanh-Cong Dinh
University of Trento, Italy,
Telecom Italia - SKIL
thanhcong.dinh@unitn.it

Zolzaya Dashdorj
University of Trento, Italy,
Telecom Italia - SKIL,
Fondazione Bruno Kessler
dashdorj@disi.unitn.it

Themis Palpanas
University of Trento, Italy,
themis@disi.unitn.eu

Luciano Serafini
Fondazione Bruno Kessler
serafini@fbk.eu

ABSTRACT

The availability of datasets coming from the telecommunications industry, and specifically those relevant to the use of mobile phones, are helping to conduct studies on patterns that appear at large scales, and better understand social behaviors. This study aims to develop methods for enabling the extraction and characterization of normal behavior patterns, and the identification of exceptional, or divergent behaviors. We study the call activity and mobility patterns, classify the observed behaviors that exhibit similar characteristics, and we analyze and characterize the anomalous behaviors. Moreover, we link the identified behaviors to important events (e.g., national and religious holidays) that took place in the same time period, and examine the interplay between the behaviors we observe and the nature of these events. The results of our work could be used for early identification of exceptional situations, monitoring the effects of important events in large areas, urban and transportation planning, and others.

1. INTRODUCTION

Starting with the assumption that important events affect the behavioral patterns of a significant number of people in such a way that these changes are reflected in their use of the mobile telephony, this study aims to develop methods for enabling the extraction, analysis, and evaluation of quantitative and qualitative information about the calling and mobility behavior patterns of users. We focus on the characterization of normal behavior patterns, the identification of exceptional, or divergent behaviors, the characterization of such behaviors (e.g., offering explanations for these behaviors), and the prediction of similar patterns. Examples of the situations we are interested in are national and religious holidays, as well as major events of local interest (e.g., sports events).

In order to achieve the above goals, we need to complement the information present in the D4D datasets with contextual information that describes the environment and context in which a user is making a phone call, and that can provide an additional set of feature for the characterization of the user behavior. Information about the context of a call can be extracted from other sources on the web, like event databases, weather forecast database, etc.

In this work, we study the call activity and mobility patterns, classify the observed behaviors that exhibit similar characteristics, and we analyze and characterize the anomalous behaviors. The results of our work can be used for early identification of exceptional situations, monitoring the effects of important events in large areas, urban and transportation planning, and others.

This paper is organized as follows. The next section briefly discusses related works. Section 3 describes D4D datasets and methods for preprocessing the data. We present our core analysis methodology in Section 4. Some experiments are conducted in Section 5. Finally, we summarize our work in Section 6.

2. RELATED WORK

Calls placed from mobile phone devices are traced in logs which can serve as an indication to understand personal and social behaviors. Researchers in the areas of behavioral and social science are interested in examining mobile phone data to characterize and to understand real-life phenomena [9, 5, 7, 15, 10], including individual traits, as well as human mobility patterns [1, 2], communication and interaction patterns[2, 11, 14].

Dynamics of call activity: Candia et al. [4] proposes an approach to understand the dynamics of individual calling activities, which could carry implications on social networks. The author analyzed calling activities of different groups of users; (some people rarely used a mobile phone, others used it more often). The cumulative distribution of consecutive calls made by each user is measured within each group and the result explains that the subsequent time of consecutive calls is useful to discover some characteristic values for the behaviors. For example, peaks occur near noon and late evening. The fraction of active traveling population

and average distance of travel are almost stable during the day. This approach can be applied for detecting anomalous events.

Moreover, a number of interesting approaches propose to analyze mobile phone data to understand personal movement patterns, in particular individual tracking and monitoring [13, 3, 16], and behavioral routines [6].

Human mobility: Furletti et al. [8] extract user profiles from mobile phone data. The authors analyze moving human behaviors which correspond to specific human profiles (such as commuter, resident, in-transit, tourist), inferred by profile assumptions. A classification technique based on neural networks, called self organizing map, is used to classify users by similar profiles that have temporal constraints based on their temporal distributions. The result shows that the percentage of residents was compatible with the customer statistics provided by the Telecom operator, and short-ranged temporal profiles like commuter and in-transit are significantly different and distinguishable from the profiles with a larger extent like resident. The authors tested their approach on a case study in the city of Pisa (Italy). The data consists of around 7.8 million call records during the period of one month. They identified a peak which was caused by the reporting of an earthquake news. The authors highlighted the necessity to align temporal call distributions with a series of high level observations concerning events and other contextual information coming from different data-sources, in order to have more specific interpretation of the phenomena.

Phithakkitnukoon et al. [12] analyze the correlation of geographic areas and human activity patterns (i.e., sequence of daily activities). pYsearch (Python APIs for Y! search services) is used in order to extract the points of interest from a map. The points of interest are annotated with activities like eating, recreational, shopping and entertainment. The Bayes theorem is then used to classify the areas into a crisp distribution map of activities. Identifying the work location as a frequent stop during the day from the trajectories of individuals, it derives the mobility choices of users towards daily activity patterns. The stop extractions are done in the same way as in [2]. The study shows that the people who have same work profiles have strongly similar daily activity patterns. But this similarity is reduced when the distance of work profile location of the people are increased. Due to the limitation of heterogeneity of activities in this paper, the result includes some strange behaviors, like shopping during the night in the shopping area, which cannot be explained by the ground truth of activities.

Anomaly detection: Candia et al. [4] propose a simple approach to detect exceptional situations on the basis of anomalies from the call patterns in a certain region. The approach partitions the area using Voronoi regions centered on the cell-towers, and computes the call pattern in the "normal situation". These patterns are compared with the actual data and anomalies are detected with the use of the percolation method.

Mobility patterns: In [2] the author analyze the mobility traces of groups of users with the objective of extracting standard mobility patterns for people in special events. In particular this work presents an analysis of anonymized traces from the Boston metropolitan area during a number of selected events that happened in the city. They indeed demonstrate that people who live close to an event are preferentially interested in those events.

Social response to events: The social response to events, and behavior changes in particular, have been studied by J.P.Bagrow et al. [1]. The authors explored societal response to external perturbations like bombing, plane crash, earthquake, blackout, concert, and festival, in order to identify real-time changes in communication and mobility patterns. The results show that from a quantitative aspect, behavioral changes under extreme conditions are radically increased right after the emergency events occur and they have long term impacts.

Crowd mobility: Calabrese et al. in [2] characterize the relationship between events and its attendees, more specifically of their home area. The consecutive calls are measured in the same manner as in [4], in order to determine the stop duration of the trajectories. Given an event, for each cell-tower of the grid, the count of people who are attending that event, and whose home location does not fall inside that cell-tower, describes the attendance of events in geo-space. Most of the people attending one type of event are most probably not attending other types of event and people who live close to an event are preferentially attracted by it. As a consequence, the approach could partly predict starting locations of people who are coming to the future events. This could be useful to determine anomalies and additional travel demands for the capacity planning considering the type of an event. Conversely, knowing event interests of people helps to detect the event. But estimating the actual number of attendees and validating the models is still an open problem due to the presence of noise in the ground truth data. So, it derives to other issues like refining mobility patterns belonging to the events which occurred in the similar region at a closer time, and distinguishing home locations for people who live in the same location where events are organized.

3. DATA DESCRIPTION AND PREPROCESSING

D4D provided four datasets, each one having different features, giving us the possibility to try more that one techniques on the available data. We now discuss the characteristics of these datasets, as well as some necessary preprocessing that we performed before applying our techniques.

3.1 Description of Datasets

The first dataset provided to us describes the aggregated communication between cell-towers. The second and the third datasets refer to mobility traces, having fine and coarse resolution data, respectively. The fourth dataset contains data about the communication between the users, creating sub-graphs. All the datasets contain data for the whole country of Ivory Coast and were collected from December 2011 to April 2012 (five-month period).

We concentrated mainly on the first two datasets. The first one consists of 175.645.538 rows, each one having a record for each available column, while the second consists

of 55.319.911 rows.

In preprocessing the data we observed that the volume of missing data was rather large, which made it difficult to make accurate predictions or connections with the events, during the subsequent analysis phases. This problem was created due to some technical problems, and as a result led to the loss of the origin, or the destination cell-tower id. The missing cell-towers were recorded as '-1'. More precisely, just for the first dataset, the amount of missing data was so big that for each cell-tower we had an average of 143.162 records, while at the same time the number of records for cell-tower '-1' were 1.846.084.

At this point we have to mention that even though the cell-tower ids range from 1 to 1238, there are some ids that don't belong to a cell-tower. Furthermore, there are some cell-towers that do not have any record during the whole five-month period. As a result we have just 1214 cell-towers with records plus one, the cell-tower '-1' that represents the missing data.

3.1.1 Aggregate Communication Between Cell-Towers Data

This dataset contains data about the number of calls and their total duration. The data was grouped by their origin and their destination cell-tower. Furthermore the dataset contains timestamps about the time that the calls were initialized, but not the time that they were terminated.

3.1.2 Mobility Traces: Fine Resolution Data

This dataset provides us with the cell-tower ids that some specific users connected to for a predefined period and with the timestamps for each connection. The users that were "tracked" for the construction of this dataset were a sample that was changing every two weeks and was chosen every time at random. The id for each user is unique during this two-week period but after two weeks it is assigned to another user. This reduced the resolution of the data, but was necessary in order to protect the privacy of the users.

3.1.3 Events Data

In order to collect some interesting events that took part during the five-months period covered by our sample, we used the Google Search Engine and we manually extracted the most important events related to Ivory Coast. Examples of such events are public holidays, important festivals, sport events, concert shows, and news that could change the activity of a user.

The extracted events refer only to the time period between the beginning of December 2011 and the end of April 2012. These events are listed in Table 1, and include events of both both regional and national importance.

3.2 Preprocessing of Datasets

The datasets were structured in such a way that an immediate analysis was not possible in order to arrive to clear conclusions about the changes of the calling activity. Before starting the development of our methods, we had to manipulate the data in a way that we would keep just the most useful (for us) data and turn them in a more usable form. In the following part of this section, we describe these preprocessing steps.

3.2.1 Useful Variables

Two of the methods described in this paper used the first dataset, which has only two types of values. The first is the hourly number of calls for each cell-tower, and the second is the total duration of these calls. Due to the volume of the data, we decided to aggregate the 24 hourly values that each cell-tower has for each day into a single daily value. Even though this aggregation leads to some information loss, it allows for an initial fast analysis, which can subsequently be refined, by using the hourly data values, for the cases where we detect some abnormal behavior.

We note that many cell-towers did not contain 24 values for every day in the dataset (due to the missing data problem we discussed earlier). Moreover, some cell-towers did not have values for each day of the period that the available dataset was produced, but this did not cause a problem for our analysis.

Apart from the two variables provided in the dataset, we derived and used a third variable that helped us to perform our analysis. This variable is the "duration per call" (dpc) that can be extracted by the division of the daily duration of calls by the number of calls, for each cell-tower. The values for this variable were calculated according to Equation 1.

$$dpc_{i,j} = \frac{number\_of\_calls_{i,j}}{total\_duration_{i,j}}, i, j \in N \quad (1)$$

3.2.2 Normalizing the Data

There as some cell-towers that are in urban areas and some others that are in areas that don't have many citizens. This has as a result that the first group of cell-towers have a continuously high activity, with respect to both the number of calls and their duration. Furthermore there are some days, like public holidays, that have more calls than the days when there is not any special event. These two factors do not allow us to cluster the data because the days or the cell-tower that have this overhead would always be reported as outliers.

In order to eliminate this problem we normalize the data by using z-normalization. In statistics, the z-normalization ensures that all elements of the input vector are transformed into the output vector whose mean-μ is 0, while the standard deviation-σ (and variance) is 1. For this transformation, we used Equation 2.

$$x'_{i,j} = \frac{x_{i,j} - \mu_j}{\sigma_j}, i, j \in N \quad (2)$$

We normalize the values in two ways. First we normalize by day in order to have normalized data with respect to each individual day. This can be achieved by finding the mean value and the standard deviation for each day and then compute the new value according to each day. This kind of normalization helps us to identify patterns across

Date	Location	Event	Event type
Dec 25, 2011	Ivory Coast	Christmas Day	public holiday
Jan 01, 2012	Ivory Coast	New Year's Day	public holiday
Feb 05, 2012	Ivory Coast	Day after the Prophet's Birthday (Maouioud)	public holiday
Feb 13, 2012	Ivory Coast	Post African Cup of Nations Recovery	public holiday
Apr 09, 2012	Ivory Coast	Easter Monday	public holiday
Feb 05, 2012	Ivory Coast	Mouloud	public holiday
Feb 22, 2012	Ivory Coast	Ash Wednesday	public festival
Jan 14, 2012	Ivory Coast	Arbeen Iman Hussain	public festival
Jan 8, 2012	Ivory Coast	Baptism of the Losd Jesus	public festival
Mar 25- Apr 1, 2012	Bouake	Carnaval	public festival
Apr 1- May 1, 2012	Ivory Coast	Fete du Dipri	public festival
Apr 6, 2012	Ivory Coast	Good Friday	public festival
Feb 9, 2012	Ivory Coast	Mawlid an Nabi (Shia)	public festival
Feb 4, 2012	Ivory Coast	Mawlid an Nabi (Sunni)	public festival
Feb 5, 2012	Ivory Coast	Yam	public festival
Dec 7, 2011	Ivory Coast	Anniversary of the death of Felix Houphouet Boigny	public festival
Apr 13-14, 2012	Abidjan	Assine Fashion Days in Cote D'Ivoire	show concert
Apr 1-4, 2012	Yamoussoukro	Education international 22nd congress	conference meeting
Apr 25, 2012	Sakre	Violence attack in Sakre	emergency event
Dec 17-18, 2011	Yale	Violence	emergency event
Jan 7, 2012	Abidjan	Hilary Clinton's visit	news event
Jan 7-8, 2012	Abidjan	Kofi Annan's visit	news event
Mar 12-13, 2012	Abidjan	Election of National Assembly President and Prime Minister	news event
Dec 11, 2011	Abidjan	New parliament election	news event
Jan 30, 2012 19-20	Ivory Coast	ACNF 2012 match vs Angola	sport
Jan 26, 2012 20-21	IvoryCoast	ACNF 2012 match vs Burkino Faso	sport
Jan 22, 2012 17-18	IvoryCoast	ACNF 2012 match vs Sudan	sport
Feb 4, 2012 20-21	IvoryCoast	ACNF 2012 match vs Equatorial Gulnea	sport
Feb 8, 2012 20-21	IvoryCoast	ACNF 2012 match vs Mall	sport
Feb 12, 2012 20:30-21:30	IvoryCoast	ACNF 2012 final match vs Zambia	sport

Table 1: List of important regional and national events in Ivory Coast, for the time period between December 2011 and April 2012.

days. In this case we use Equation 2, where i is the cell id and j is the day.

In addition, we normalize by the cell-tower, using each individual cell-tower's mean value and standard deviation. This action helps us to identify patterns for the cell-towers. In this case we use Equation 1 again, but in contrast to the previous case, i is the day and j is the cell id.

3.2.3 Creating sequences of movement

In order to develop the third method described in this paper, we used the second dataset. The original data do not allow us to analyze movements of customer. Therefore, we need to preprocess the data using the following steps:

- Step 1: Grouping the close cell-towers to avoid spatial gaps.
- Step 2: Creating sequences of movement.
- Step 3: Data cleansing to correct the "lost" cell-tower named '-1'.

In Step 1, we determine the customers' trajectories. The trajectory $Traj$ of an arbitrary customer c is represented by an array of places with their associated time-stamp. The identifier of a place (Id) is the identifier of the cell-tower where the customer has network connection. Let p_i be the place i-th and t_i is its associated time-stamp, where $i \in [1, n_c]$, n_c is the maximum number of places which customer c has visited. We have:

$$Traj_c = \{(p_1, t_1) \rightarrow (p_2, t_2) \rightarrow ... \rightarrow (p_{n_c}, t_{n_c})\}$$

Next, for each trajectory, we group pairs of two consecutive places whose Harversine distance is within 500 meters. The pseudocode for this process is shown in Algorithm 1.

Algorithm 1 Group close cell-towers by Haversine distance

```
1: procedure GroupByHaversine(Traj_c)
2:     p_i ← Traj_c.p_1                          ▷ Obtain the first place
3:     while i < n_c do
4:         p_{i+1} ← Traj_c.p_{i+1}
5:         if Harversine(p_i,p_{i+1}) < 500 ∧ p_i ≠ p_{i+1} then
6:             loc ← p_i p_{i+1} ▷ Create representative location
7:             Traj_c.p_i ← loc
8:             Traj_c.p_{i+1} ← loc
9:         p_i ← p_{i+1}
10:        i ← i + 1
```

For Step 2, we split each trajectory, which was obtained from Step 1, into 140 sub-trajectories according to the 140 days of the dataset's observation period. Then, we create daily sequences of movements using those sub-trajectories. A sequence of movements is a 24-element array. In each element, there is an ordered list of locations where the customer visited during the hour that this element represents. Note that, a sequence has its identifier which made of the sample's Ids, that of the customer c and the day d in format yyyy-MM-dd. Thus, we have:

$$Seq_{s,c,d} = \{(< 0 >, l_0), (< 1 >, l_1), ..., (< 23 >, l_{23})\}$$

The i-th element is described by $(< i >, l_i)$ where i is the hour and l_i is the list of locations that the customer visited within hour i and $i+1$. For example, considering the following $SubTraj$ which is a sub-trajectory obtained from sample number 0 and the customer 1:

$SubTraj$ = {(264,'2011-12-06 16:59:00'), (264,'2011-12-06 21:00:00'), (264,'2011-12-06 22:36:00')}

Its sequence is obtained as below:

$Seq_{0,1,2011-12-06}$ = {(<0>),..., (<16>,264),..., (<21>,264), (<22>,264), (<23>)}

Finally, for each customer, Step 3 simply replaces the cell-tower '-1' in sequences of movements of that customer by the most frequent visited-cell-tower in the same hour range.

4. PROBLEM DESCRIPTION

In this work, we concentrate on three problems. First, we identify and investigate the anomalous behavior that some cell-towers exhibit, and examine the reasons that could cause such a behavior. The second problem that we tackle is to characterize the way that a social event affects the calling activity of a region, or of the entire country in general. Finally, we analyze the social response to some major events, and investigate how different events affect the mobility of users.

4.1 Analysis of Anomalous Behavior

In this part we present a method that helps us to analyze the calling activity of the entire country of Ivory Coast. The method that we present allows us to normalize data, creating clusters and extracting usage patterns. Furthermore, we can identify activities in specific regions or even in the entire country that are not normal.

4.1.1 Identifying Outliers

After having normalized the data we have to compare the values with respect to the day or the cell-tower. In order to compare these values we calculate the mean and the standard deviation for each cell-tower or for each day, depending on the analysis that we intend to do. Furthermore, we calculate the difference of each point from the mean. If we have a point A that is much farther away from the mean than a point B that is the closest to A and between A and the mean, then the point A is marked as an outlier. Such an example is the plot depicted in Figure 1.

In this figure, we have an analysis of the data points clustered by day after having normalized by the day. As we can see there are two days, the day 66 and 67, that have some cell-towers whose points are much farther from the mean than the rest of the points, creating a gap between them and the rest of the cluster. By analyzing these outliers, and after having set a threshold[1] of '3.5', we found that the cell-towers that have these values, never had such a calling activity during the rest of the period that covers our dataset.

[1]We set this threshold manually after observing the data.

The algorithm that implements our method is shown in Algorithm 2.

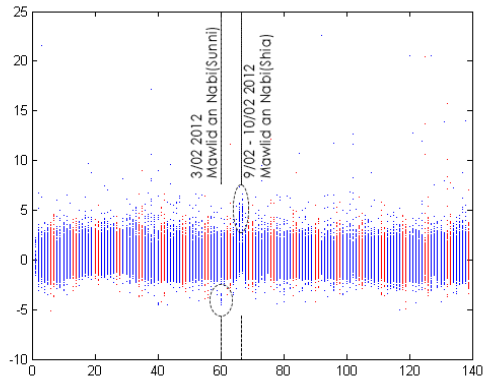


Figure 1: Daily plot for dpc normalized by day.

Algorithm 2 Grouping By Distances

```
1: procedure GroupByDistances(threshold)
2:     x'_{i,j} ← NormalizedValues
3:     for i = 1 → MaxID do        ▷ The MaxID is either the maximum ID of the cell-tower or the maximum ID for the days, depending on the analysis we intend to do. In most of the cases the day.
4:         Distances ← allthedistances        ▷ between each point that belongs to i and i's mean
5:         array ← sortthedataaccordingtothedistances
6:     for i = 1 → MaxID do
7:         for all points ∈ i do
8:             if distance ≥ closerpoint then        ▷ the point that is closest and between the examined point and the mean
9:                 while (NotTheEnd) do
10:                    while (NoPointWithGreaterDistance) do
11:                        checkNextPoint()
12:                        if it is close to the previous point then        ▷ they propably form a sub-cluster return point ▷ as a possible outlier
```

4.1.2 Identifying Outliers Using the Standard Deviation

A second method that we used to identify the outliers is the comparison of the standard deviations. This method can be mainly applied on the daily values because each day has more or less the same features. More specifically each day has (almost) the same number of values and each value derives from a cell-tower that is every day at the same longitude and latitude. The only difference is that if an event is local then it will be hard to detect using the normalization by cell-tower. This results in the creation of datasets that have some steady main features plus some features that change and allow us to analyze them.

Following this method, we look at the standard deviation for each day and we compare it with the standard deviations of the 12 adjacent days, 6 before the day under examination and 6 after. This helps us to draw a conclusion on whether the calling pattern is more or less the same for this day as it should be, in respect to the period that we analyze. In case the standard deviation is not similar to the majority of the compared days, we can come to the conclusion that some event, such as a public holiday, has taken place.

The pseudocode for this technique is shown in Algorithm 3.

Algorithm 3 Grouping By The Standard Deviation

```
procedure GroupByStd(threshold)
2:     x'_{i,j} ← NormalizedValues
       for j = 1 → MaxID do        ▷ The MaxID is either the maximum ID of the cell-tower or the maximum ID for the days, depending on the analysis we intend to do. In most of the cases the day.
4:         Difference ← 0
           Similar ← 0
6:         for k = (j − 6) → j + 13 do ▷ Compare with the 6 previous days and the 6 days after
               if i ≠ j then
8:                 if (σ[i] ≠ (threshold ∗ σ[j]))||(σ[i] ≤ (threshold ∗ σ[j])) then
                       Difference ← Difference + 1
10:                else
                       similar ← similar + 1
12:        if Difference ≥ 6 then return i
```

4.1.3 Correlated Abnormal Behaviors

We have already analyzed the cases that a value is an outlier for a cell-tower, or for a day. The problem that rises is the importance and the weight that the value has in general. If, for example, cell-tower 1 has for one day 100 calls and for the next day again 100, this could be possibly a normal pattern according to the cell-tower whose values remain more or less stable. What happens, though, if the values for all the cell-towers apart from this one are increased during the second day? This means that cell-tower 1 is an outlier. If we perform only a normalization with respect to the cell tower this outlier could be lost.

In order to avoid this situation, we have to correlate the two normalized values. This can be achieved by the subtraction of the two normalized values, and then look for outliers in this new space. This correlation can be achieved by using Equation 3.

$$weight_{i,j} = x'_{i,j} - x'_{j,i}, i, j \in N \quad (3)$$

4.2 Analysis of Social Response to Events

Here, we are interested in analyzing call volume changes at each cell-tower during events (e.g., public holidays, festivals etc.), which cover the entire population or the population of a specific region. The goal is to discover similar behaviors and the regularities of the behavioral patterns in related to events. This enables us to understand and characterization normal behaviors, and identify exceptional, or divergent behaviors. We assume that social responses to events are similar at the scale of cell-tower area to that of the country.

For example, a comparison of social responses to an event in two cell-towers is shown in Figure 2. The behaviors have a similar shape, pointing to the fact that during a public holiday people are in celebration and relaxing mood, which leads to reduced call volumes. A day after the holiday, the daily activities pick up again and the call volumes increased.

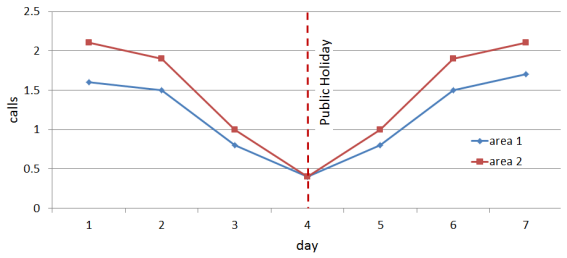


Figure 2: The behavioral patterns (social response) to a public holiday in the vicinity of the epicenter before (-3 < D_{event} < 0), during (-1 < D_{event} < 1), after (1 < D_{event} < 3) the holiday in two cell-towers. The dashed line describes a day-D_{event} when the event has been started.

To infer this type of information about social response, we analyze the call volume changes in the vicinity of the epicenter of an event before/during/after this event, in order to identify and extract behavioral patterns (social responses). As an event, we consider a public holiday which covers the entire population of the country.

To achieve this, we first annotate the information about public events that are described in Table 1 to daily call records at each cell. This event information is the context in which call volume is decayed or grown and it is described as a pair of location-time <l,t>. The location of the public events is Ivory Coast. Second, we extract call volume changes in the vicinity of the epicenter before/during/after the given event in order to extract behavioral patterns for the event from each cell. Third, we cluster these behavioral patterns to events by the call volume changes in order to find a similar pattern among all cell-towers.

4.2.1 *Annotate the information about public events to daily call records*

In order to prepare the daily call records, we group call records by day at each cell and normalize the daily call volume at each cell using the z-normalization procedure presented in Section 4.1. The equation to normalize the call volumes at each cell is described in Equation 4.

$$v_i' = \frac{v_i - \mu}{\sigma}, i \in N \tag{4}$$

1. v is a call volume
2. v' is a normalized call volume
3. μ is a mean value of call volume
4. σ is a standard deviation of call volume
5. i is a day parameter (i=1,2,3..,140)

Now, we need to annotate the information about events to the daily call records using the date. We create a data table that contains daily call records and event id in the following format.

```
CREATE TABLE DAILY_RECORDS (
date DATE,
originating_ant INTEGER,
total_duration_voice_calls INTEGER,
nb_voice_calls INTEGER,
event_id INTEGER
);
```

The following is an example of the resulting daily call records:

```
2011-12-05 2 175122 633 5
2011-12-06 3 3069220 310 2
2011-12-07 4 5665196 306 1
2011-12-08 5 5606249 303 0
```

4.2.2 *Behavioral Pattern Extraction For Public Events*

We analyze the call volume changes before/during/after the events. The behavioral patterns that occur in response to an event are represented as time-series of call activities in a given day-D_{event}, when the event has occurred, and in addition for d_1 days before (e.g., -3 < D_{event} < -1), during d_2 days (e.g., -1 < D_{event} < 1) and after d_3 days (e.g., 1 < D_{event} < 3). The pseudocode of behavioral pattern extraction is described in Algorithm 4. The algorithm returns the time-series of call volumes in certain periods in a given cell.

Algorithm 4 Extraction of behavioral pattern algorithm

```
1: function EXTRACT_PATTERN(D_event, d1, d2, d3, cell − tower)
2:     call_records ← daily_records(cell − tower)
3:     for d = 1 → 140 do
4:         if call_records[d, 1] = D_event then
5:             t ← 0                                  ▷ before the event
6:             for j = (D_event − d1) → D_event do
7:                 bpattern[t] ← call_records[j, 2]
8:                 t = t + 1
                                                      ▷ during the event
9:             for j = D_event → (D_event + d2) do
10:                bpattern[t] ← call_records[j, 2]
11:                t = t + 1
                                                      ▷ after the event
12:            for j = (D_event + d2) → (D_event + d3) do
13:                bpattern[t] ← call_records[j, 2]
14:                t = t + 1
       return bpattern
```

4.2.3 *Clustering Behavioral Patterns*

We cluster these behavioral patterns using hierarchical agglomerative clustering techniques in order to understand the most similar behaviors. To measure the dissimilarity between sets of call volumes, we use the Euclidean distance metric as described in Equation 5.

$$\mathrm{d}(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{5}$$

1. x is a set of call volume 1
2. y is a set of call volume 2

Based on the dissimilarity between sets of call volumes, we cluster the pairs using average linkage clustering, a method for estimating the distance between clusters in a hierarchical cluster analysis. It specifies the distance between two clusters is computed as the average distance between objects from the first cluster and objects from the second cluster as presented in Equation 6.

$$D(C1, C2) = \frac{1}{N_{c1} N_{c2}} \sum_{i=1}^{N_{c1}} \sum_{j=1}^{N_{c2}} d(c1_i, c2_i) \quad (6)$$

1. $C1$ is cluster 1
2. $C2$ is cluster 2
3. N_{c1} is a number of values in cluster 1
4. N_{c2} is a number of values in cluster 2
5. $d(c1_i, c2_i)$ is the distance of cluster 1 and cluster 2

This way, we characterize the social responses to the events. Supervised classification and clustering can be used (e.g., k-means, k-nearest neighbor, SVM etc.).

4.3 Analysis of human mobility

To give an overview of the datasets, Figure 3 illustrates the total number of displacements, i.e., users changing a cell-tower, for all customers over time. There are ten time series that represent the entire dataset, where "series1" and "series10" are the first and the last consecutive two-week periods, respectively. Note that the beginning time of the first period is Monday 2011-12-05.

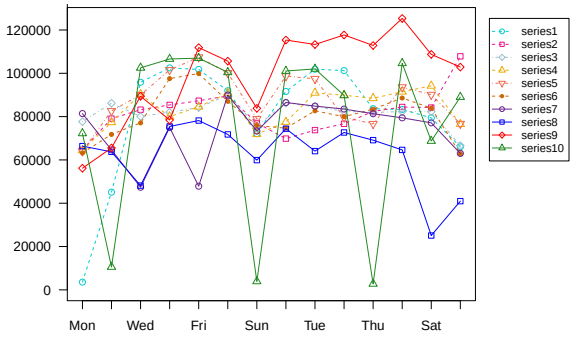


Figure 3: Total number of displacements

As shown in Figure 3, the time series in "series8" dips below other series, while coming to "series9", it keeps moving up until the beginning of "series10", it suddenly fluctuates. The interesting point here is that these strange behaviors of displacements might have correlation to the events happening during the same periods, i.e, Easter Monday on 2012-04-09.

In this section, we describe a method to analyze the correlation between movements of customers and the upcoming events. Thus, enabling us to detect in advance the locations (i.e., cell-towers) that are potentially related to some event. The proposed method consists of two steps: i) Inferring the cell-towers that show abnormal behavior in the number of distinct customers; ii) For each cell-tower obtained from the previous step, detecting potential cell-towers that may contain events.

In the first step, for each cell-tower, we count the number of distinct customers, who had network connection within the coverage of the cell-tower,considering a specific day d. Then, we use the sigma approach to extract all cell-towers that fall out of the following range R_d as outliers:

$$R_d = [\mu_d - 1.5 \times \sigma_d, \mu_d + 1.5 \times \sigma_d]$$

Where μ_d and σ_d is the mean and the standard deviation of number of distinct customers of all cell-towers in day d , respectively.

Before going to the second step, we infer the home location of each customer by estimating the most frequent cell-tower where that customer stays during night hours (from 6PM to 8AM). Once the home locations of the customers are obtained, we use Algorithm 5 to detect the cell-towers that may have events among the cell-towers obtained from the previous step. Note that the sequence database S is obtained from Section 3.2.3, C and H are the lists of abnormal cells and home locations, and d is the day that cell-towers in C have abnormal behavior.

Algorithm 5 Dectecting potential cell-towers

procedure DETECTEVENT($S, C, H, d, minsup$)
  for all $c \in C$ **do**
    for $i \leftarrow 0, 23$ **do**
      $cnt \leftarrow \text{count}(S, d, i, c)$ ▷ Count the number of distinct customers have visited cell-tower c in hour i on the day d in database S
      $\mu \leftarrow \text{getMean}(S, d, i, c)$ ▷ Calculate the average number of distinct customers concerning the days before d
      $\sigma \leftarrow \text{getStd}(S, d, i, c)$
      $lb \leftarrow \mu - 1.5 \times \sigma$ ▷ Calculate lower bound
      $ub \leftarrow \mu + 1.5 \times \sigma$ ▷ Calculate upper bound
      if $cnt > ub$ **then**
        $support \leftarrow \text{getSupport}(S, d, i, c, H)$ ▷ Calculate relative support level concerning only the sequences of customers who have visited cell-tower c in hour i of day d and their home location is not c
        if $support < minsup$ **then**
          $out \leftarrow c, d, i$
      if $cnt < lb$ **then**
        $support \leftarrow 1 -$ getSupport(S, d, i, c) ▷ Considering home location
        if $support < minsup$ **then**
          $out \leftarrow c, d, i$

5. EXPERIMENTAL EVALUATION

5.1 Social response to events

In this section, we characterize the social responses to a public event, Easter Monday (9 April, 2012). We extracted the

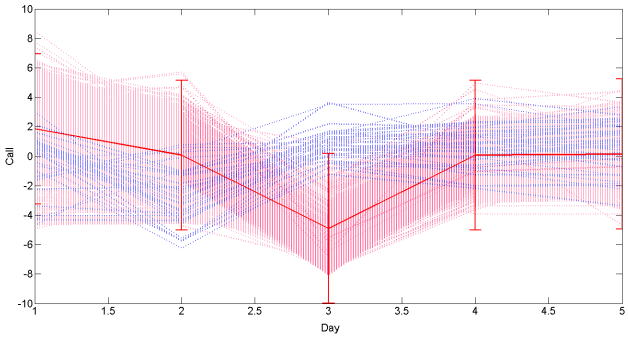


Figure 4: The call volume changes in the vicinity of the epicenter before (-2 < D_{event} < 0), during (0 < D_{event} < 1), after (1 < D_{event} < 3) Easter Monday. Pink and blue color represent daily call volumes in each cell-tower, red color represents for the μ (mean) $\pm$ 2*σ (standard deviation).

daily call volumes at each cell-tower in the vicinity of the epicenter before (-2 < D_{event} < 0), during (0 < D_{event} < 1), after (1 < D_{event} < 3) Easter Monday as described in Figure 4. The figure shows that behavioral pattern for the Easter Monday is v-shaped and can fit inside 2 standard deviations. Before (-2 < D_{event} < 0) and during (0 < D_{event} < 1) the event, the call volumes are reduced to the lowest activity levels, suggesting that people are having a rest during public holiday and the urge to communicate is the weakest. After (1 < D_{event} < 3) the event, the call volumes increase again to reach the normal activity levels. WE observe that some call volumes start decaying before the event. The reason could be that there are some areas where people do activities during the holiday such as concert place, relaxing place, eating place so on. Further, we investigate the call volumes by hour for these abnormal behaviors in those cell-towers.

However, the call volumes before/during the event (-1 < D_{event} < 1) are obviously divided into groups. Figure 4 shows that call volumes (described in blue color) during the events are conversely increased. To cluster these behavioral patterns (-2 < D_{event} < 1) , we create a matrix that contains the call volumes in certain times at each cell-tower. The dataset can be seen as a collection of cell-tower vectors that contains the call volumes in certain times. We calculate the euclidean distance between all possible pairs of vectors to find dissimilarities. Then, we perform hierarchical clustering, starting from the two cell-towers that has the maximum distance as a cluster in first stage. In next stage, the two nearest clusters are merged into a new cluster. The process is repeated until the whole data set is agglomerated into one single cluster as represented in Figure 6. Figure 5 shows the classification of behavioral patterns before/during/after the event. After the clustering, we identified the most two groups; call volumes are increased during the event that covers 39 cell-towers out of 1214 cell-towers (total numbers of cell-towers which are identified from the experiment dataset) and call volumes are decreased during the event that covers 1173 cell-towers out of 1214 cell-towers.

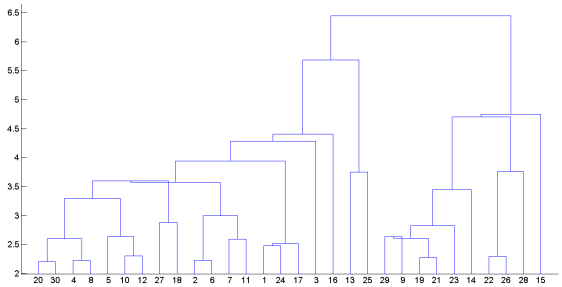


Figure 6: Groupings of behavioral patterns

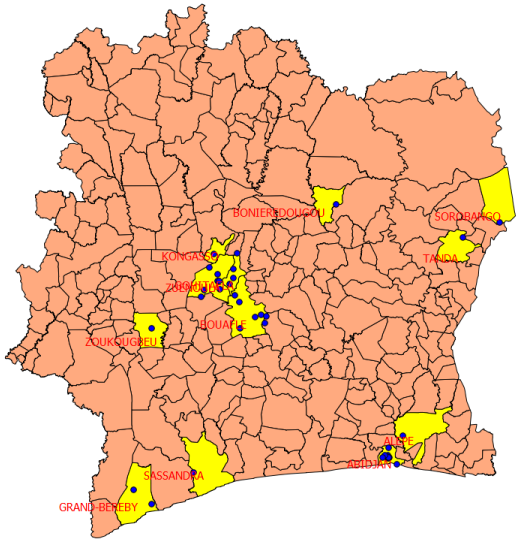
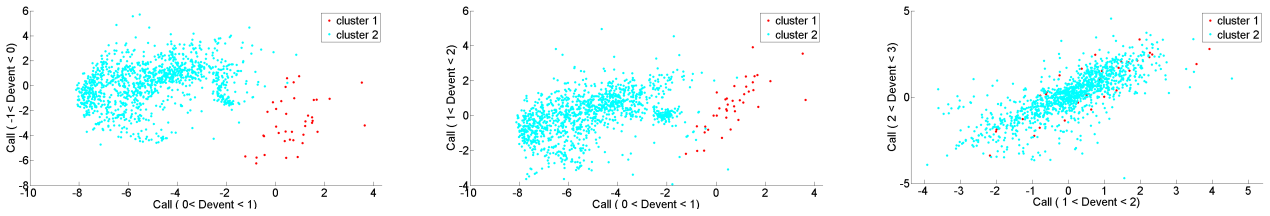


Figure 7: Location of the cell-towers where the call volumes are increased during the event (-1 < D_{event} < 1)

The location of the cell-towers where the call volumes are increased during the event are shown in Figure 7. The call volume changes are concentrated in the center of the country (Daloa, Zuenoula, Bouafle) which is the main cocoa-growing region and south east of the country (Abidjan) which is the capital of the country. This could explain that during the Easter Monday, people work in the field as in working day at the cocoa-growing region, and other shops or restaurants are operated at the capital. Due to the fact that the event information is limited in web site for experiment period, we are not able to provide more accurate information to these call volume changes. Further, we can investigate the mobility patterns of users to check if there are movement from other cities to these area.

Similarly, we analyzed the social responses to all public holidays which are described in Table 1. The characterization of social responses to each public holiday is represented in Table 2. A day before the public holiday, Post African Cup of Nations Recovery there was the final match that Ivory Coast vs Zambia for 2012 African Cup of Nations. This affects to



(a) A day before the event (-1 < D_{event} < 0)

(b) During the event (0 < D_{event} < 1)

(c) A day after the event (1 < D_{event} < 2)

Figure 5: Classification of behavioral patterns for the period (-1 < D_{event} < 2)

behavioral patterns during the public holiday, Post African Cup of Nations Recovery. Further, we investigate the mean pattern at each cell to investigate the magnitude of the call volume changes during the events. Also, the behavioral patterns during New Year's Day are overlapped with the public festival, New Year on 31 December. Some cells are not active during the events

5.2 Anomalous behaviors

In this section we are going to present some results after having applied on the first dataset the method described in section 4.1.

The first step is to compute the duration per call (dpc) that each cell-tower has. The next step is to normalize the data by day and by the cell-tower, as described in 3.2.2. After the normalization step we are going to have six values, two for each initial variable, the daily number of calls (nb), the daily duration of the calls (dur) and the dpc for each cell-tower.

Having normalized the data we cluster it in the way described in section 4.1.1. This has as result to identify behaviors that are not normal. Such type of behaviors we can see in Figure 1. In this figure we can see the dpc normalized by each day values. At the x axis we have the day id and at the y axis we have the normalized value of the dpc by day for each day. With red color we can see the weekends and with blue color the weekdays. This difference at the colors makes it easier for us to achieve even a visualized comparison between the values. As we can see almost all the days follow the same pattern, having values that are in a small range plus some values that are (unique) outliers. Investigating these outliers, we found that are mostly the same cell-towers. This allows us to consider it as a normal pattern for these specific cell-towers and we don't analyze them more.

Although most of the days in the sample follow the same pattern, there are some days like the days with ids 60, 66 and 67 that have a sub-cluster of outliers. Investigating these outliers we found out that, for the days 66 and 67, these cell-towers are only 36 and they are close to each other. Furthermore, we found out that the calling activity referring to the dpc for these cell-towers is unique for these days and they don't have such an activity for the rest of the five-month period. For the day 60 we have the same conclusions as we did with the days 66 and 67 with the difference that the outliers are negatives.

Finally we came to the conclusion that specific events or actions change the calling activity for these days for these specific cell-towers. You can see the cell-towers that are outliers for the days 66,67 and the day 60 in Figures 8 and 9 respectively.

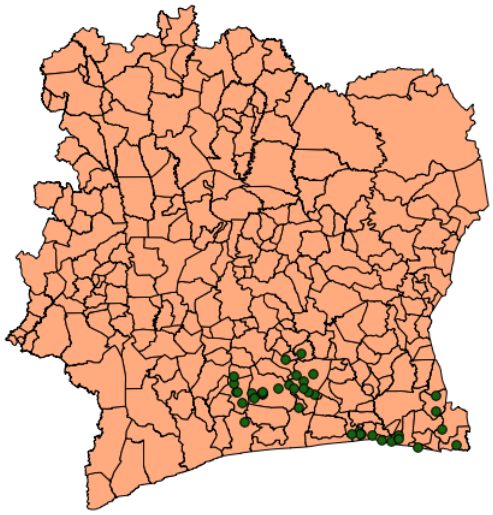


Figure 8: Cell-Towers Positive Outliers for 9-10 of February 2012

One more fact that evaluated our conclusions is the analysis that we did for the number of the calls for each day when this number was normalized by the day. In order to achieve this type of analysis we used the method described in section 4.1.2. By analyzing these values we found out that just a single event could cause a difference on the calling activity for just a region(when it is a local event) or even for the whole country.

Such an example is depicted in Figure 10 where we can see the difference of the calling activity for the whole country

Public holiday	Pattern extraction period	Cluster count	Cell-towers at each cluster	not active cell-towers
Christmas Day	(-1 < D_{event} < 1)	3	[1],[572],[503]	138
New Year's Day	(-2 < D_{event} < 1)	4	[656],[8],[2],[411]	137
Day after the Prophet's Birthday (Maouioud)	(-1 < D_{event} < 1)	4	[943], [115], [2], [1]	153
Post African Cup of Nations Recovery	(-2 < D_{event} < 2)	2	[475],[548]	191
Easter Monday	(-1 < D_{event} < 1)	2	[39], [1173]	2

Table 2: Public holiday classifications

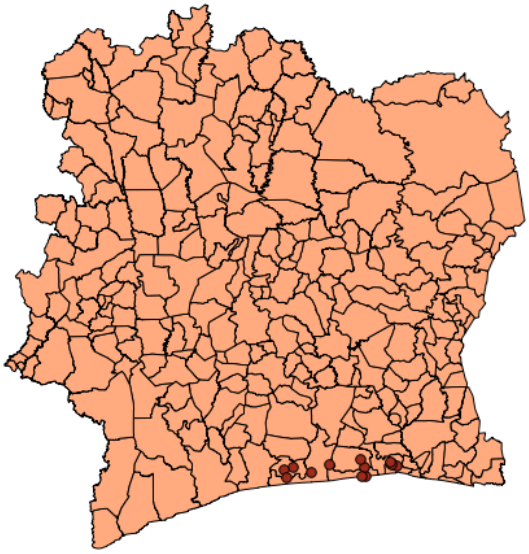


Figure 9: Cell-Towers Negative Outliers for 3 of February 2012

during the Christmas and the new year event, the easter, the days that we have unique events such as festivals and the rest of the days. We depict with blue color the weekdays and with red color the weekends.

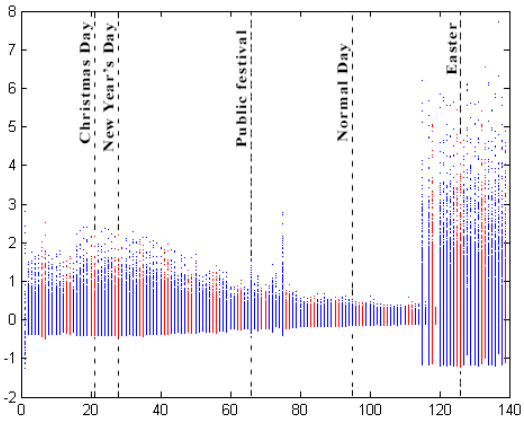


Figure 10: Number of calls for each day (normalized by day)

Finally we evaluate the method described in 4.1.3 by analyzing the duration of the calls for each cell-tower while we have normalized it both by the day and the cell-tower. Subtracting the second value from the first we get a result that for cell-towers with ids 731 to 750 the weights for the weekends are mostly clustered at the positive values while the weights for the weekdays are clustered to the negative values. This makes it clear that these cell-towers have specific patterns for the weekdays and the weekends. In Figure 11 we have this analysis visualized. Again with blue color we have the weekdays and with red color the weekends. In x axis we have the cell-tower id and in y axis we have the normalized daily duration of calls.

5.3 Human mobility

Using the method described in section 4.3, we found the cell-towers that have abnormal behavior in movements of customers. As in Fig.12, there are significant signs of augment of movement of customers in the Christmas day, the day after the Prophet's Birthday (Maouioud) and the New Year holiday. While the movement of customers dramatically decreases before Carnaval.

We choose the New Year day and Carnaval day as the interesting days for experimenting Algorithm 5. The poten-

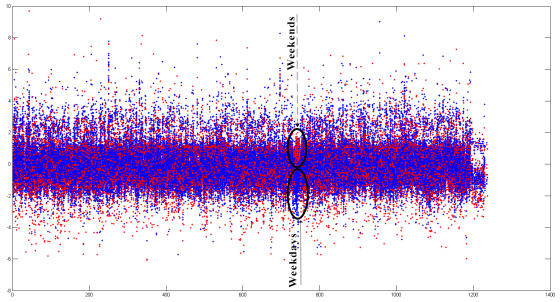


Figure 11: Weights For The Correlation of The Two Types of Normalized Values For The Duration (for each cell-tower)

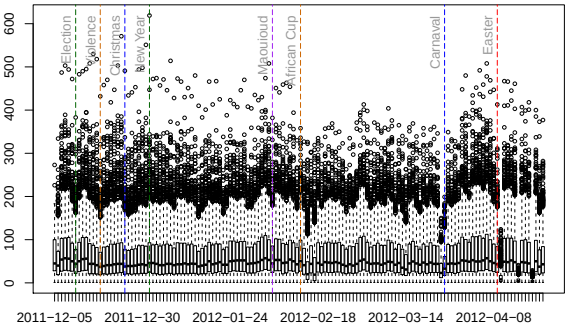


Figure 12: Result of inferring abnormal cell-towers. Points represent cell-towers and y-axis is the number of distinct customers visited the cell-towers

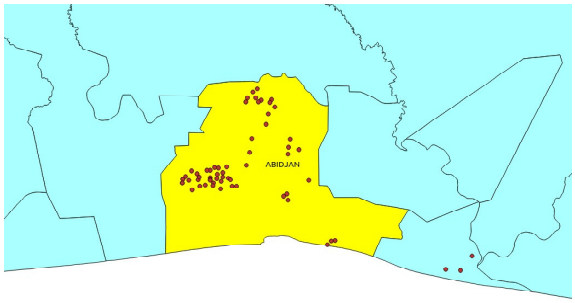


Figure 13: Result of detecting event's location. Points represent cell-towers and the day is 2012-01-01

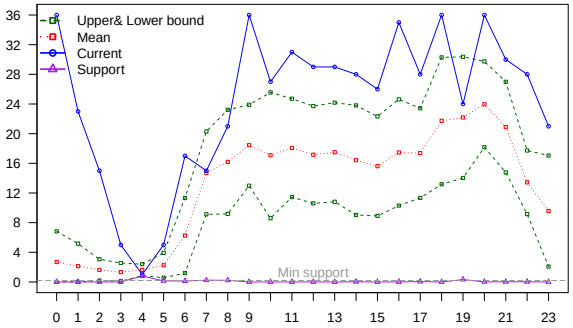


Figure 14: Number of distinct customers who visited cell-tower's Id '728', 2012-01-01, x-axis is hours. Minimum support is 0.2

tial cell-towers that have events on the New Year day, are shown in Fig. 13. Almost cell-towers are located in Abidjan, which is the largest city of Ivory Coast. Among these cell-towers, we investigate one random cell-tower to illustrates Algorithm 5. As shown in Fig. 16, we can infer that there may have special events during period from 0AM to 3AM and 9AM to 11PM at the location cover by cell-tower's Id '728'. During these events, there is significant increase in movement of customers arriving this location.

In another context, Fig. 15 plots the cell-towers that potentially have events among the abnormal cell-towers in the day before Carnaval event day. The cell-towers are mainly located in Bouake and Abidjan. We also investigate one random cell-tower as shown in Fig. 16. There may have special events during period from 4PM to 6PM of the day 24th March, 2012 at coverage of cell-tower's Id '642'. During this event, the significant customers, who visited this cell-tower, may move to other locations or may not use their mobile phone as usual.

6. CONCLUSION AND DISCUSSION

In this paper, we presented three methods for identifying patterns and outliers in the behavior and mobility of mobile phone users, by analyzing the data recorded by cell-towers. Our methods can be used to predict events and actions that are possible to happen if some specific circumstances exist, for example, to predict activities when we have dry-periods, or important events, such as festivals and public holidays.

We are currently extending our techniques in order to become more targeted. The goal is to anble a more detailed analysis, focusing on particular regions of the country, or examining in detail the patterns that occur in finer time scales (e.g., by analyzing the data at the hourly level).

7. REFERENCES

[1] J. P. Bagrow, D. Wang, and A.-L. Barabási. Collective response of human populations to large-scale emergencies. *CoRR*, abs/1106.0560, 2011.

[2] F. Calabrese, P. F. C., G. Di Lorenzo, L. Liu, and C. Ratti. The geography of taste: analyzing cell-phone mobility and social events. In *the Proc. of the 8th international conference on Pervasive Computing*, Pervasive'10, pages 22–37, Berlin, Heidelberg, 2010. Springer-Verlag.

[3] F. Calabrese, M. Colonna, P. Lovisolo, D. Parata, and C. Ratti. Real-time urban monitoring using cell phones: A case study in rome. *Trans. Intell. Transport. Sys.*, 12(1):141–151, Mar. 2011.

[4] J. Candia, M. C. Gonzalez, P. Wang, T. Schoenharl,

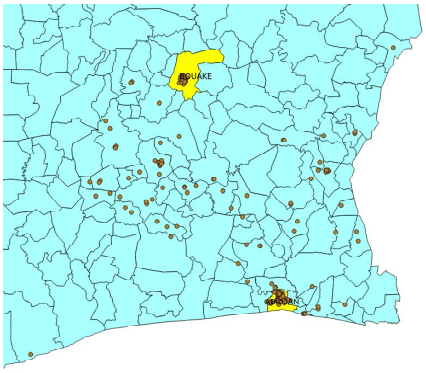


Figure 15: Result of detecting event's location. Points represent cell-towers and the day is 2012-03-24

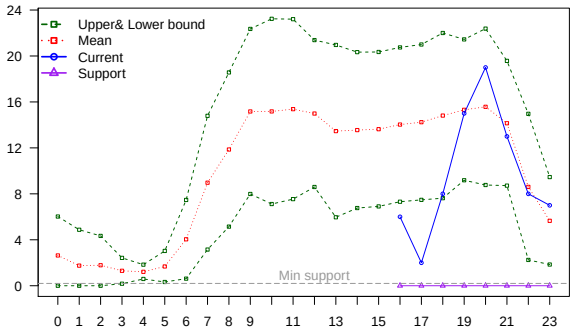


Figure 16: Number of distinct customers who visited cell-tower's Id '642', 2012-03-24. Minimum support is 0.2

G. Madey, and A. L. Barabasi. Uncovering individual and collective human dynamics from mobile phone records, 2007.

[5] D. Choujaa and N. Dulay. Tracme: Temporal activity recognition using mobile phone data. In *Embedded and Ubiquitous Computing, 2008. EUC '08. IEEE/IFIP International Conference on*, volume 1, pages 119 –126, dec. 2008.

[6] N. Eagle and A. (Sandy) Pentland. Reality mining: sensing complex social systems. *Personal Ubiquitous Comput.*, 10(4):255–268, Mar. 2006.

[7] D. Fox. Location-based activity recognition. In *the Proc. of the 30th Conf. on Advances in Artificial Intelligence*, pages 51–51, 2007.

[8] B. Furletti, L. Gabrielli, C. Renso, and S. Rinzivillo. Identifying users profiles from mobile calls habits. In *Proceedings of the ACM SIGKDD International Workshop on Urban Computing*, UrbComp '12, pages 17–24, New York, NY, USA, 2012. ACM.

[9] M. Kwan, C. Arrowsmith, and W. Cartwright. Visualizing population movements within a region, 2011.

[10] M. A. Muhammad Awais Azam, Laurissa Tokarchuk. Human behavior detection using gsm location patterns and bluetooth proximity data. *The 4th international conference on mobile ubiquitous computing, systems, services and technologies-UBICOMM 2010*, pages 428–433, 10 2010.

[11] J. P. Onnela, J. Saramaki, J. Hyvonen, G. Szabo, D. Lazer, K. Kaski, J. Kertesz, and A. L. Barabasi. Structure and tie strengths in mobile communication networks, 2006.

[12] S. Phithakkitnukoon, T. Horanont, G. Di Lorenzo, R. Shibasaki, and C. Ratti. Activity-aware map: identifying human daily activity pattern using mobile phone data. In *the Proc. of the 1st Intl. Conf. Human Behavior Understanding*, pages 14–25, 2010.

[13] C. Ratti, A. Sevtsuk, S. Huang, and R. Pailer. Mobile landscapes: Graz in real time. In G. Gartner, W. E. Cartwright, and M. P. Peterson, editors, *Location Based Services and TeleCartography*, Lecture Notes in Geoinformation and Cartography, pages 433–444. Springer, 2007.

[14] C. Ratti, S. Sobolevsky, F. Calabrese, C. Andris, J. Reades, M. Martino, R. Claxton, and S. H. Strogatz. Redrawing the map of great britain from a network of human interactions. *PLoS ONE*, 5(12):e14248, 12 2010.

[15] T. Sohn, A. Varshavsky, A. LaMarca, M. Y. Chen, T. Choudhury, I. Smith, S. Consolvo, J. Hightower, W. G. Griswold, and E. de Lara. Mobility detection using everyday gsm traces. In *8th International Conference on Ubiquitous Computing (UbiComp)*, Irvine, CA, September 2006.

[16] J. Steenbruggen, M. Borzacchiello, P. Nijkamp, and H. Scholten. Mobile phone data from gsm networks for traffic parameter and urban spatial pattern assessment: a review of applications and opportunities. *GeoJournal*, pages 1–21, 2011.

Egocentric and population-density patterns of cellphone communication in Ivory Coast

Paul Schmitt, Morgan Vigil, Mariya Zheleva, and Elizabeth M. Belding
Department of Computer Science University of California, Santa Barbara
{pschmitt, mvigil, mariya, ebelding}@cs.ucsb.edu

Abstract

Traffic analysis of mobile and Internet networks helps researchers understand people's behavior and needs in the context of these networks. Such analysis is an important facet of both the initial design as well as the iterative improvement of applications that leverage such networks. In developing countries where the population is predominantly rural, mobile communications with their high affordability and intuitive interface, are the first communication technology introduced. Thus, the analysis of usage patterns of mobile networks is of great importance, as it facilitates better understanding of people's interaction with technology and their specific technological needs. We approach the D4D challenge as a preliminary analysis on network usage patterns focusing particularly on usage in Rural areas. We also analyze persistence trends of individual's social groups in this mobile network. Based on our results, we provide a discussion of possible practical applications that can leverage mobile networks.

1 Introduction

Mobile networks have revolutionized the way people communicate in the developing world and serve as a platform for enhancement of many aspects of people's day-to-day life. Applications that use underlying mobile networks span from health care [5, 6, 11, 15] and education [18, 1, 14] to agriculture [7, 17, 16] and mobile banking [12]. Multiple successful projects in Africa have spurred from observing people's behavior in mobile or social networks. Johnson et. al., after analysing facebook traffic, design a system to facilitate local content sharing within remote rural communities [9]. [12] describes a system called mPesa that enables transfer of money in the form of airtime in rural Kenya. The design of this system was inspired by analysis of mobile network usage in Kenya, which indicated that people tend to transfer airtime between one another as a means for payment or financial support. Follow up studies on the adoption of mPesa in Kenya show that theft decreased, as people no longer needed to carry cash.

Such projects are of critical importance to introducing new services and enhancing the wellbeing of people in under-serviced areas. At the same time, special attention should be paid in the design process of these systems to make sure that they meet an actual need in the community. Analysis of large scale datasets generated by the targeted communities naturally facilitates the identification of actual community needs.

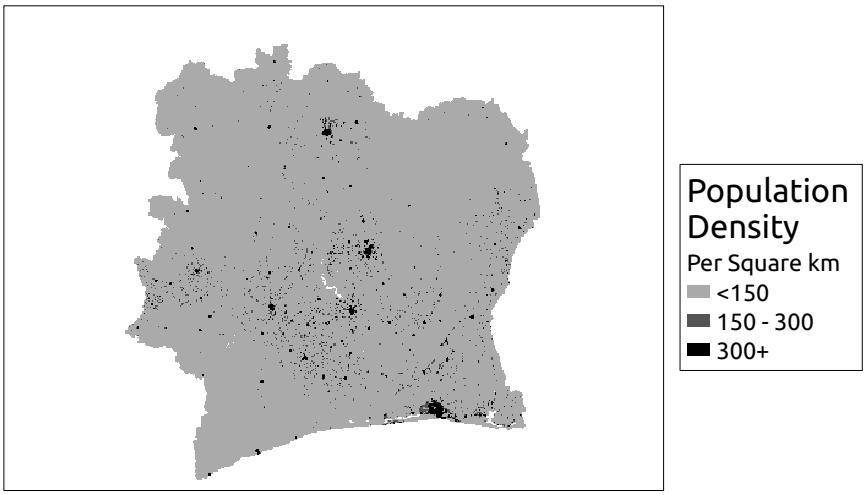
Due to the prevalence of mobile communication technologies in sub-Saharan Africa [13], it is particularly important to understand how information is exchanged over these wireless infrastructures. As communication patterns emerge, it becomes possible to improve current wireless infrastructures and develop new technologies and systems that effectively leverage existing infrastructure. We are particularly concerned with how communication patterns might inform healthcare systems for development. We approach Orange's datasets on mobile call patterns in Ivory Coast with this end in mind.

Our previous work introduces VillageCell [4], a comprehensive connectivity solution designed for rural areas. VillageCell includes low cost free-to-air cellular coverage and exploits locality of interest to facilitate effective use of limited gateway bandwidth. VillageCell has been deployed in rural Macha, Zambia. By understanding how mobile communication flows with respect to population density, we begin to understand the feasibility of a similar connectivity solution for rural Ivory Coast. Additionally, we investigate how mobile communication flows socially. Social graph patterns that emerge provide a preliminary understanding of how social relays might be exploited for information broadcast. We discuss how this can be useful to the development of a healthcare system that uses cellular network technology.

2 Methodology

To facilitate extensive analysis of mobile network traces over different subsets of the Ivory Coast population, we preprocess our data. We now describe the preprocessing techniques and software we use. We also define a set of metrics we utilize in

Figure 1: Population density of Ivory Coast.



the process of evaluation.

2.1 Datasets

Our analysis focuses primarily on information in Set 1 and Set 4 of the Orange Datasets. Dataset 1 represents mobility data aggregated on an hourly basis for ten weeks from December 05, 2011 to April 22, 2012 and includes data about the number of calls and call duration between antenna pairs. In addition, we use the antenna location data provided by the ANT_POS dataset. This set provides data that maps an antenna ID to its corresponding latitudinal and longitudinal coordinates. Set 4 includes ego-centric social graphs that describe up to second order neighbors of 5000 users traced over the entire period. We use this dataset to analyze persistence of social groups that a mobile subscriber communicates with.

As recommended by Orange, we use data from the AfriPop project. The data consist of high resolution population density distribution information in ESRI Float format. We use this data to calculate and associate population density with antenna locations given the ANT_POS dataset.

2.2 Antenna classification

We employ the new European Union typology of "predominantly Rural", "Intermediate", and "predominantly Urban" areas. This typology is a modification of the Organisation for Economic Co-operation and Development (OECD) methodology that seeks to minimize distortions caused by large variations in the area of local administrative units[2]. Using the new OECD method, rural local administrative units are defined as areas with a population density below 150 inhabitants per km^2 applied to grid cells of size 1 km^2. Likewise, urban local administrative units are defined as areas with population density of at least 300 inhabitants per km^2 applied to grid cells of size 1 km^2. The $1km^2$ cell size provides fine granularity, which makes the OECD method equally applicable to countries outside the European Union. See Table 1 for the classifications per km^2 we employ. Note that our classification of "Suburban" directly corresponds to [2]'s classification of "Intermediate".

We utilize the population density information contained in the AfriPop data set and use Quantum GIS to project it as a raster layer. The AfriPop data includes population density information formatted as the number of people per 100 square meters for Ivory Coast. We then re-sample the density data at a lower resolution creating a grid of 2 km squares assigning population density for each as the mean density values of the AfriPop data bounded by the new grid. Each square is assigned to one of the population density categories using the OECD typology. This allows us to create the population density map shown in Figure 1. The different shades indicate the different density classifications as defined by OECD - Rural, Suburban, and Urban. As shown, the majority of land area in Ivory Coast is classified as Rural.

Table 1: OECD population density classifications

Density per km^2	Classification
0-149	Rural
150-299	Suburban
300+	Urban

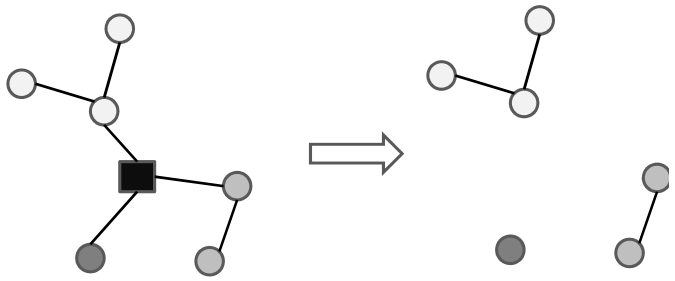


Figure 2: The effect of removing the ego (depicted with a square) from the ego-centric social graph.

Interestingly, World Bank [3] population statistics show that the rural population represents 51.2 percent of the total population in the country while the urban population accounts for 48.8 percent. Our grid assigns population density at a resolution suitable for associating antennas with the underlying population statistics.

2.3 Ego-centric graphs analysis

We examine the ego-centric social graphs dataset to determine persistence of social groups for each ego over time. We also analyze the likelihood that one or few nodes – *top nodes*, persist over time in an ego-centric graph. We hypothesize that such persistent nodes can be used as information relays in an egocentric graph. Our analysis indicates that such persistent nodes indeed exist. Their feasibility as information relays, however, needs further analysis that requires richer datasets providing information about frequency and duration of phone calls as well as physical location of the communicating parties.

In order to extract the separate social groups of an ego, we remove the ego node from each ego-centric social graph (Figure 2) and analyze the connected components that remain after the ego is removed. Each connected component corresponds to one social group. Note that in the text we use the terms *connected component* and *social group* interchangeably.

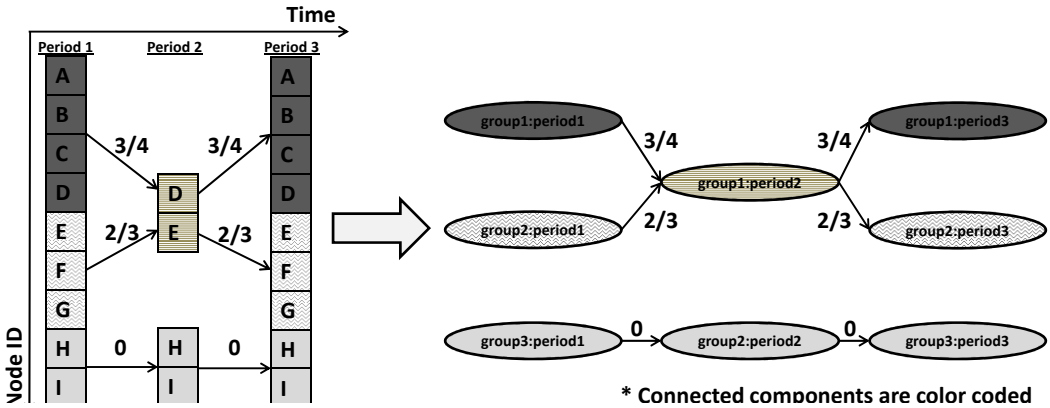


Figure 3: Building a persistence graph.

After extracting the connected components we evaluate the persistence of these components over time. A connected component is 100% persistent over two consecutive periods if the nodes in this connected component are the same in the two periods. For this evaluation we define a *persistence graph* $G = (N, E, W)$ with N nodes, E edges and W weights assigned to each edge. Each node in G is a connected component labeled with the period, to which it belongs. An edge exists between two connected components if they overlap in consecutive periods. The weight assigned to each edge is the *Jaccard similarity*, J, between the connected components. J for two sets A and B can be calculated as follows:

$$J = \frac{A \cap B}{A \cup B} \tag{1}$$

The Jaccard similarity changes between 0 and 1, where 0 indicates no overlap and 1 indicates full overlap.

Figure 3 presents an example of building the persistence graph for a single ego over three consecutive periods. The left-hand side of the picture presents the set of neighbors in each of the three periods. The social groups comprised by these neighbors are color-coded. The right-hand side of the picture presents the resulting persistence graph. Edges exist only between connected components that overlap fully or partially in consecutive periods. There is no edge between connected components that persist over non-consecutive periods (e.g. there is no edge between node "group1:period1" and node "group1:period3").

Our persistence analysis is based on the described persistence graphs and consists of two parts. First, we analyze the in- and out-degree distribution of the nodes in the persistence graph. We note that if the social groups of an ego persist over time, all the nodes in the persistence graph should have in- and out-degrees of either 0 if the node belongs to the first or last period, or 1, if the node is in the intermediate periods. In cases where social groups do not persist, nodes can have a degree of 0 if the corresponding social group does not re-appear in following periods. Nodes can also have in- and out-degrees larger than 1 if social groups merge or split in consecutive periods.

Further we attempt to quantify the level to which social groups overlap by considering the weights of the edges in the persistence graphs. As detailed earlier, edges are drawn between nodes that overlap fully or partially in consecutive time periods. The weights assigned to these edges are the Jaccard similarity between the nodes connected by these edges. For each transition between period t and period $t+1$ we find the normalized Jaccard similarity $\hat{JS}^{(t,t+1)}$ between these periods: that is the sum of edge weights $W_i^{(t,t+1)}$ divided by the number of edges $|E^{(t,t+1)}|$ between the two periods.

$$\hat{JS}^{(t,t+1)} = \frac{\sum_{i=1}^{|E^{(t,t+1)}|} W_i^{(t,t+1)}}{|E^{(t,t+1)}|} \tag{2}$$

We then find the average Jaccard similarity for the entire persistence graph by summing the normalized Jaccard similarities and dividing this sum by the number of period transitions K.

$$\bar{JS} = \frac{\sum_{j=1}^{K} \hat{JS}_j^{(t,t+1)}}{K} \tag{3}$$

Informally, the higher the average Jaccard similarity, the more persistent the social graphs of an ego are over time.

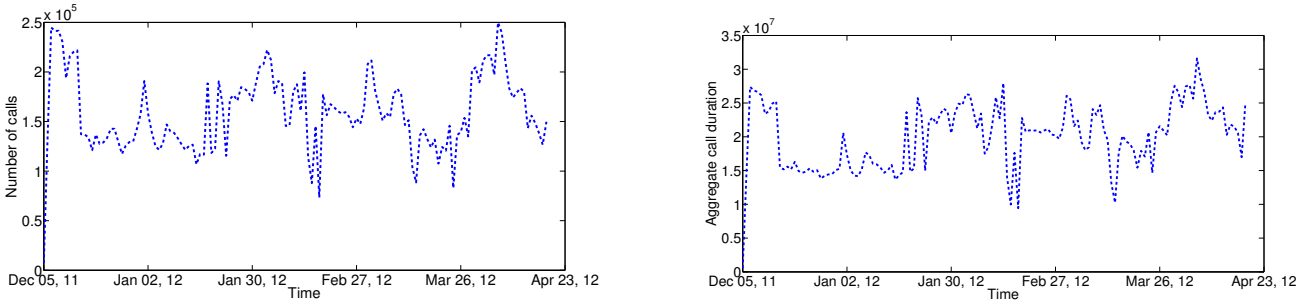
We present our results for social groups persistence in Section 3.6.

3 Evaluation

We begin our analysis by investigating temporal trends in mobile communication in general and across areas of different population density types. We expect to observe regular temporal trends along weekly and monthly intervals, with Rural areas having temporal trends distinctive from those of Urban areas. Progressing from temporal trends, we explore trends related to population density. Again, we expect to see differences in call duration and call frequency based on the population density of the sender. Next, we seek observable relationships between the distance between sending and receiving antennas and call duration and frequency. Finally, we examine patterns in social groups. We hope to observe consistency in social groups over time.

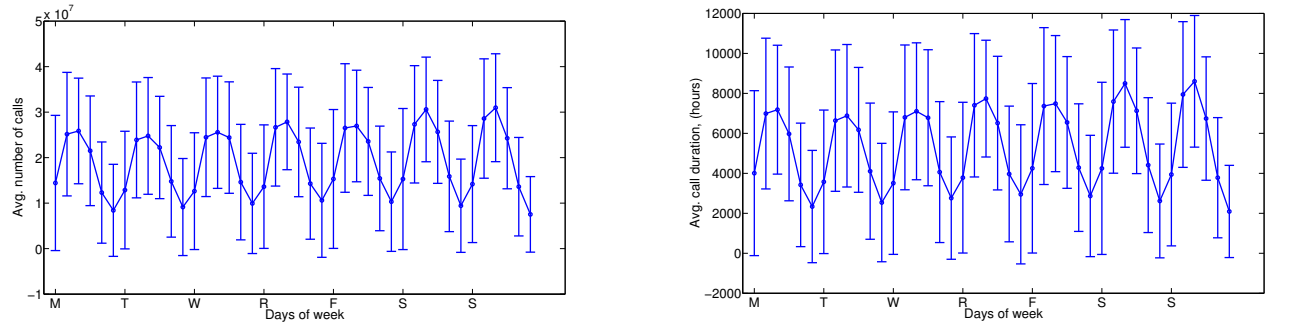
3.1 Usage patterns over time

We evaluate the cellular network activity patterns over the entire capture period. In Figures 4(a) and 4(b) we plot aggregate number of calls and call duration per day. As the figure shows, there is no distinctive call pattern on a weekly or monthly basis; instead subscriber activity seems to be widely correlated with events in the country. We hypothesize that the peaks from the beginning of the period coincide with the weeks before and after the parliamentary elections on



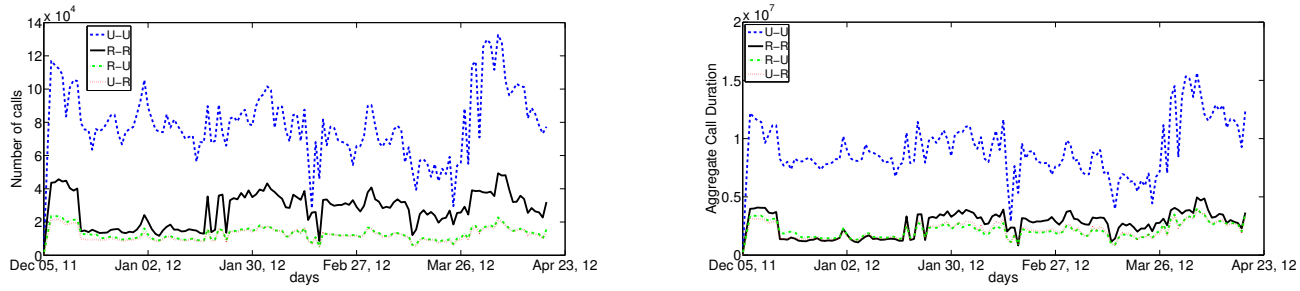
(a) (b)

Figure 4: (a) Number of calls and (b) call duration over time.



(a) (b)

Figure 5: (a) Number of calls and (b) call duration over time.



(a) (b)

Figure 6: (a) Number of calls and (b) call duration over time.

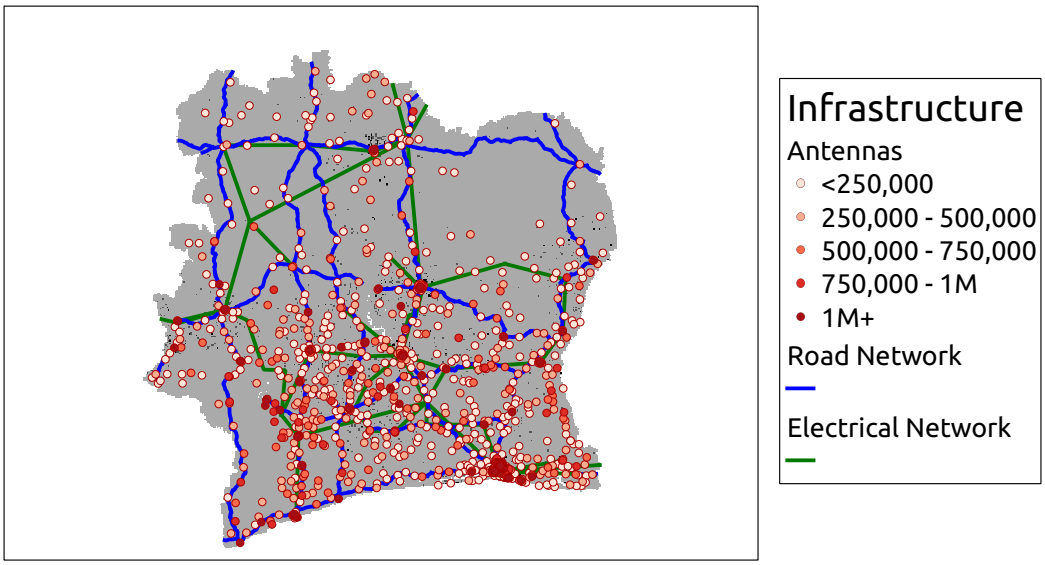


Figure 7: Antenna Usage

December 11th, 2011, while the second peak is most likely traffic around the New Years Eve. The increased utilization from the end of March and April is likely associated with the military coup in Mali and the associated ECOWAS[1] summit that took place in Abijan, Ivory Coast. Such irregular usage pattern is very different than what had been observed in cellular network traces from the western world [10].

The lack of weekly pattern is further confirmed by Figure 5. We average the number of calls and call duration over the entire capture period in a one week window. Each point on the plot presents an average over four hours over all occurrences of each day of the week (that is the first data point from the graphs presents the average number of calls and call duration for the hours from Midnight to 4 AM for all Mondays in the capture period). The figure clearly presents diurnal pattern of network activity with slight increase over the weekend, however, the standard deviation of this graph is very high, indicating that the network activity varies dramatically over the observed period.

Next we analyze whether the calling patterns of rural areas differ from those in urban areas within Ivory Coast. In Figure 6(a) and 6(b) we plot the aggregate number of calls and call duration for each day of the observed period. We analyze four categories of calls depending on the source and destination antenna type: Urban to Urban (U-U), Rural to Rural (R-R), Urban to Rural (U-R) and Rural to Urban (R-U). As the figures show, calling patterns for all four categories follow similar trends, where the number of calls and the aggregate call duration between Urban antennas is about three times higher than between Rural antennas. We also note that while the number of Rural to Rural calls is larger than the number of Rural to Urban and Urban to Rural, the aggregate call duration for these three categories is the same. This result indicates that while calls between Rural residents occur more often, they are shorter in comparison to calls between Urban and Rural residents.

3.2 Antenna activity map

We seek patterns of mobile communication flow in Ivory Coast by associating antennas with their geographical location and the population density of their location. The resultant mapping of antennas to location can be seen in Figure 7. We associate each antenna with the underlying population density of their location in order to assign the Rural, Suburban, or Urban typology. Antennas are shaded based on the total number of outbound calls they originate throughout the entire sampling period with darker colors signifying busier antennas. It is evident that antennas are densely clustered in urban

[1] http://www.ecowas.int/

locations while more sparsely located in predominantly rural regions. We also find that high activity antennas are often located along major transportation corridors.

Table 2: Antenna Density Classifications

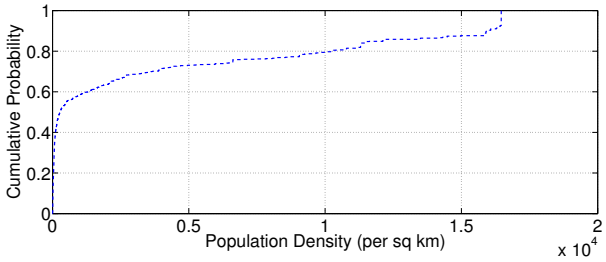
Classification	Antenna Count	Source Calls
Rural	528	146,481,488
Suburban	90	21,529,115
Urban	598	331,630,147
Unknown	15	65,393,926

We join the antenna location and the population density datasets using Quantum GIS and plot all antennas onto a Ivory Coast population density map. This allows us to associate a population density value of the underlying grid with each antenna and assign the OECD typology for each antenna that is provided with geographic information in ANT_POS. Table 2 shows the number of antennas that fall into each of the classifications as well as the total number of calls originated from each antenna type. Of note, we see that relatively few antennas are classified as Suburban. As the antenna location dataset is not fully complete we do not associate any density information for those antennas that do not correlate to a square on the grid. Such antennas are classified as "Unknown" when processing call records.

3.3 Population density

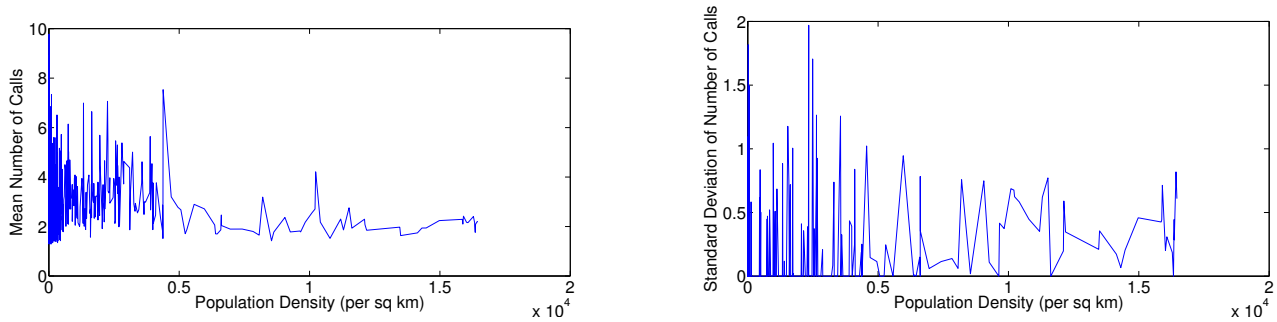
In terms of data density, Figure 8 shows there are observably more records for antenna pairs involving source antennas with population densities with less than 500 inhabitants per km^2. Likewise, Figure 1 shows that the geographical area of Ivory Coast largerly consists of sparsely populated areas regions. This leads us to examine the distribution of Set 1 in terms of population density. Additionally, the distribution shown in Figure 8 demonstrates a clear dichotomy between densely populated regions and sparsely populated regions. This leads us to classify antennas into one of three population density categories: Rural, Suburban, and Urban. In Section 3.5, we then classify each identifiable recorded connection in Set 1 based on its directionality: Rural to Rural, Rural to Suburban, Rural to Urban, Suburban to Suburban, Suburban to Rural, Suburban to Urban, Urban to Urban, Urban to Suburban, or Urban to Rural.

Figure 8: Distribution of population density (per sq km) associated with source antenna.



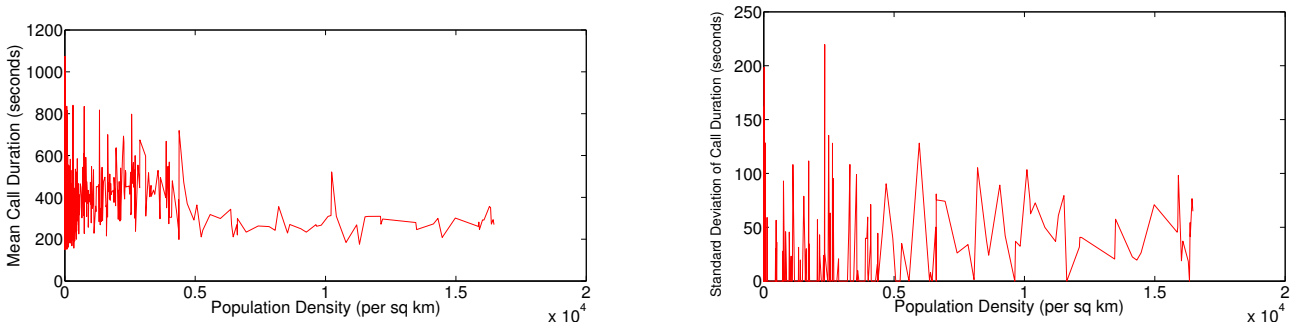
As evidenced by Figure 1, population density in Ivory Coast varies between Rural and Urban areas. We explore the relationship between the population density of a sending antenna and the average number of outbound calls associated with the antenna. Because of the predominant use of "Calling Party Pays" (CPP) policy in sub-Saharan Africa, we focus on the number of outbound calls sent from an antenna rather than incoming calls received by an antenna [13][8]. According to Figure 8, it appears that a large cluster of data points occur in population densities below 500 inhabitants per km^2 and a smaller cluster occurs in population densities above 15,000 inhabitants per km^2. In order to normalize for this, we calculate the mean number of calls and the mean call duration for population densities per km^2 in Figure 9(a) and Figure 10(a). The mean values in Figure 9(a) illustrate the tendency for the outbound number of calls to decrease as population density associated with the sending antenna increases. Likewise, Figure 10(a) illustrates the tendency for the call duration to decrease as population density associated with the sending antenna increases. Due to the CPP policy, we anticipated a larger mean number of outbound calls and mean call duration from antennas associated with the highest population density, which is associated with Ivory Coast's financial district and center of commerce. In order to explore why the mean number of outbound calls and call duration were greater for lower population densities, we plot

the standard deviation of the number of outbound calls and call duration in Figure 9(a) and Figure 10(b). We observed more variation in the standard deviation values associated with lower population densities. We attribute the variation of standard deviation of the number of outbound calls and call duration for lower populaiton densities to be indicative of sparse antenna placement. For instance, many of the lower population densities associated with low standard deviations of number of outbound calls and call duration only have one or zero antennas associated with them. This makes it more difficult to ascertain a "normal" call pattern for areas of low population density, even though most of the geographical area has low population density values (see Figure 1). However, Figure 7 may demonstrate why we observe such erratic mean number of outbound calls and mean call duration in areas of low population density. As can be seen in Figure 7, there are several antennas associated with low population densities that have a high number of outbound calls associated them. In Section 4 we discuss how infrastructure can be inferred based on population density and call frequency associated with antennas.



(a) (b)

Figure 9: (a) Population density vs. mean number of outbound calls and (b) Population density vs. standard deviation of number of outbound calls



(a) (b)

Figure 10: (a) Population density vs. mean call duration and (b) Population density vs. standard deviation of call duration.

3.4 Mean call duration as a function of distance

We investigate the relationship between call distance and the average duration of calls. We calculate the distance in kilometers between all pairs of antennas with known geographic location using the Haversine formula [19] and inputting a mean earth radius of 6,372.80 km. We group connection distances into the nearest 10 km in order to calculate aggregate statistics for each group.

Next we calculate the mean call duration for each of the distance groups. We process Set 1 to find distance information for each record and associate call duration and call counts to the associated distance. Records that include antennas for which we do not have location information are ignored. The impact of distance between source and destination antenna on mean call duration is seen in Figure 11. In general, we observe an increase in average call durations as connection distance increases. We hypothesize that the reason for such a pattern is the calling parties have fewer opportunities for

in-person interactions due to the geographic distance. Lastly, note that with relatively fewer call records for distances greater than 500 km, more noise is introduced into the graph. Given more call records we expect that the relationship trend between distance and average call duration would hold.

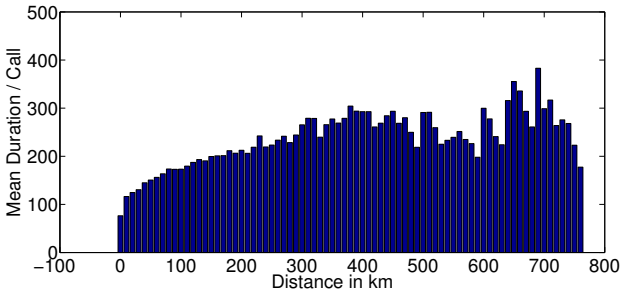


Figure 11: Antenna distance vs. mean call duration.

3.5 Call typology classifications

We investigate the potential correlation between population density and calling patterns by associating antennas with known locations to the corresponding local population density. This process yields antennas denoted as Rural, Suburban, Urban, or Unknown for the antennas which have no geographic location. We process Set 1 to classify call records by each typology source and destination pair in order to investigate potential communication patterns. In this analysis we do not consider records for antennas with no geographic data or records without valid antenna IDs. As seen in Figure 12, the majority of the identifiable connections are classified as Urban to Urban connections. This is followed by 20% of connections classified as Rural to Rural. Connections classified as Rural to Urban or Urban to Rural each account for 9% of connections.

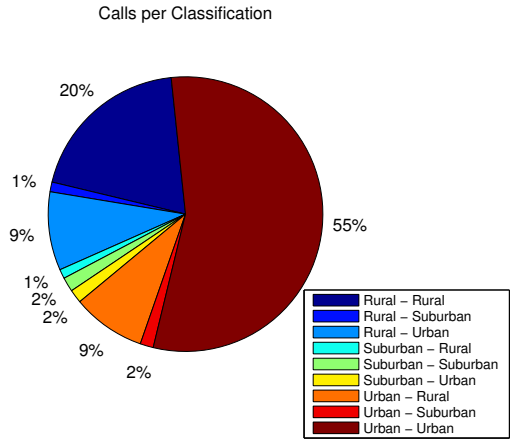


Figure 12: Classification of communication between antenna pairs.

Next, we search for differences in mean call duration across the connection classifications with results shown in Figure 13. We find that the two call classifications with the longest mean call duration are Urban to Rural and Rural to Urban. An observable phenomenon is that calls confined to the same source and destination density type are noticeably shorter on average compared to calls between mixed pairs. Given our prior finding of the relationship between call distance and average duration we posit that the majority of calls that do not cross classification boundaries are confined to a smaller geographic region. For instance, we believe Urban to Urban calls are more likely to be sourced from and destined for the same urban area. Lastly, an interesting observation is that calls originating from Urban antennas generally have a longer duration for any destination type. We believe this is due to the common policy of "Calling Party Pays" and higher buying power of individuals who reside in urban areas.

This trend leads us to look at the average distance between connecting antennas associated with each connection density classification type. Figure 14 shows the relationship between the average call distance for each connection classification type

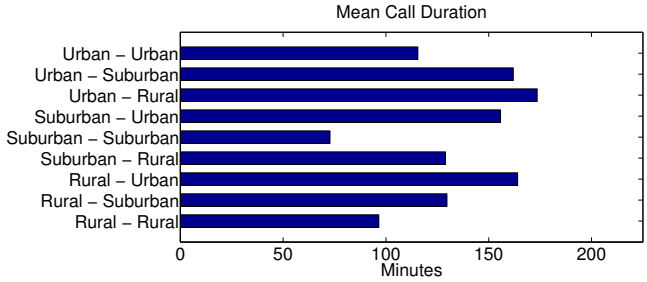


Figure 13: Call durations for classified connections.

and the average call duration. We see that the longest average distance between connecting antennas occurs between Rural to Urban and Urban to Rural connections. The shortest average distance occurs between similar-to-similar connections such as Rural to Rural, Urban to Urban, and Suburban to Suburban. As we would expect based on Figure 11, we see classifications associated with longer average distances between connecting antennas also associated with longer average call duration.

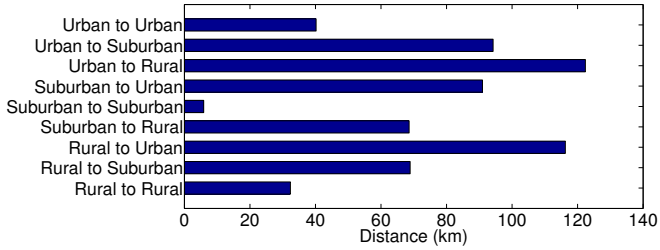


Figure 14: Mean distance for classified connections.

We investigate associating call patterns and population density for calls that are sourced and destined for the same antenna ID. This is motivated by the observation in Figure 12 that roughly 75% of all observed calls are categorized as Rural to Rural or Urban to Urban. We process Set 1 and identify records that include two valid antenna IDs where the source and destination match. We find in Table 3 that 57% of all Rural to Rural calls are both sourced from and destined for the same antenna. We posit that this is due to fewer available antennas to associate with in predominantly rural areas. Furthermore, the cellular coverage provided by a single antenna in rural settings is typically larger, which means that a higher proportion of local users are associated with the same antenna. Calls between users in the same general vicinity in a Rural area are likely to involve only one antenna. Interestingly, Urban connections sourced from and destined for the same antenna represent 23% of all Urban to Urban calls. We believe that the higher density and smaller cell range of Urban antennas provides more diverse antenna association possibilities for users.

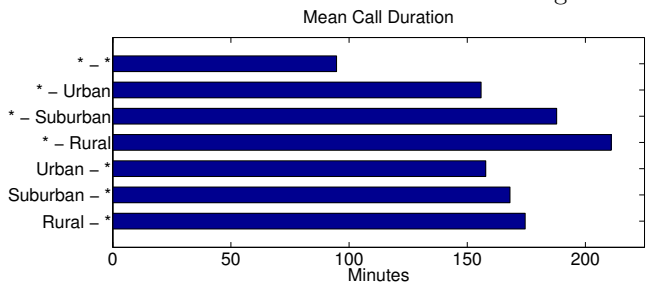
Table 3: Percent of calls made between same source and destination antenna

Classification	Percentage of calls
Rural to Rural	57%
Suburban to Suburban	88%
Urban to Urban	23%

Our final analysis is focused on antennas for which we have no population density information. These antennas include those that are not provided with geographic coordinates in the ANT_POS data set as well as those identified in Set 1 with an invalid ('-1') antenna ID. We classify call records from Set 1 where at least one antenna in the connection is a part of the "Unknown" antenna classification and gather statistics. Figure 15 illustrates that calls between two unclassified antennas are typically shorter than those in which one side of the connection is "known." Given our prior observation that mean call durations are noticeably shorter when the source and destination antenna classification is not mixed we believe

that calls between two unclassified antennas remain within the same density classification, though unknown. Analysis of records in which one of the involved antennas is classified as known shows that calls involving a Rural antenna for at least one half of the connection are longer on average than other types. Also of note is that calls with one half of the connection known are more similar to patterns associated with mixed classification calls than those remaining within the same classification.

Figure 15: Mean call durations for connections including unclassified antennas



3.6 Egocentric graphs

We now examine the ego-centric social graphs provided in dataset 4. Our analysis focuses on persistence of social groups with which individual egos communicate. We regard this analysis as preliminary work on identifying persistent neighbors within one's social network, who can serve as reliable information relays.

First we provide high level analysis of the average number of social groups with which each ego communicates over the entire capture period from December, 2011 to April, 2012. For this analysis we sum the number of connected components that appear in each two-week period and divide this sum by the number of capture periods (i.e. 10). Figure 16(a) plots a CDF of the average number of connected components for each ego. While the average number of components across egos spans from 1 to 10, the majority of egos – 68%, have between 2 and 5 connected components on average. Further, we examine how the number of connected components deviates for each ego. Figure 16(b) plots a CDF of the standard deviation of the number of connected components per ego over the observed period. Almost half of the egos (47%) have standard deviation of less than 1, while 96% of all the egos have standard deviation of less than 4. This indicates that the number of connected components in an ego-centric graph remains relatively constant over time.

Next we analyze the persistence of these social groups over time. First, we look at the in- and out-degree distribution of nodes in the persistence graphs. As detailed in Section 2.3, a node in period t would have in- or out-degree of 0 if it belongs to the first or last observed period or if it does not overlap with any node from the preceding $(t-1)$ or the following $(t+1)$ period. Nodes would have in- and out-degree of exactly 1 if they persist over time and degree larger than 1 if they split or merge over consecutive periods.

We calculate that out of all the nodes in all persistence graphs, 9.49% belong to the first period (i.e. have in-degree of 0) and 8.93% belong to the last period (i.e. have out-degree of 0). At the same time Figure 17(a) indicates that in nearly 60% of the cases nodes have in- or out-degree of 0. This means that about 50% of all the social groups that we observe did not occur in the preceding and following periods. 40% of the nodes have in- or out-degree of 1, which means that these 40% of the social groups persisted in consecutive periods. Only about 3% of the cases have in- or out-degree larger than 1, which means that social groups rarely split or merge over consecutive periods.

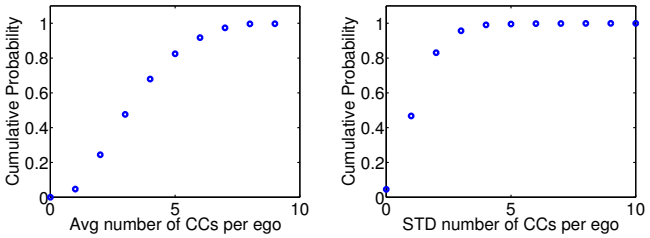
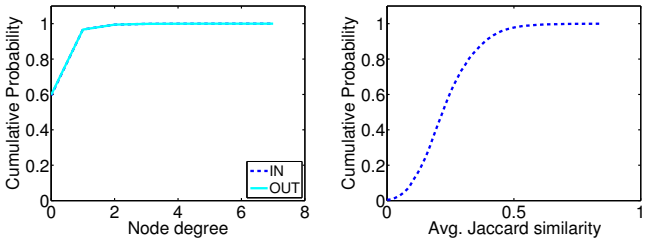


Figure 16: (a) The number of connected components per ego and (b) the standard deviation of the number of connected components per ego over the observed period.



(a) (b)

Figure 17: (a) The in- and out-degree of nodes in all persistence graphs and (b) the average Jaccard similarity for each persistence graph.

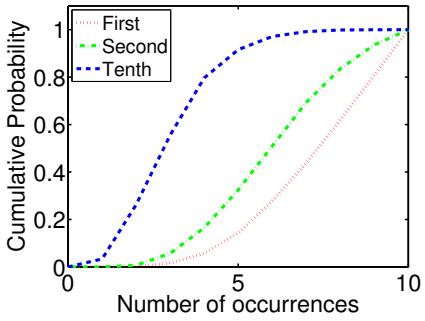


Figure 18: Number of occurrences of the first, the second and tenth most frequent neighbor.

This result indicates an important quality of the observed ego-centric social graphs: there are two distinctive types of social groups with which an ego communicates – (i) those that likely occur only once (in- and out-degree is 0) and (ii) those that likely persist over time and strictly correspond to one social group from the preceding and one social group from the following period. The former group can be associated with one-time calls, for example calling to schedule a doctors appointment, while the latter can be associated with calls recurring over time, such as these between relatives and friends who stay in touch.

We continue our evaluation of social groups persistence by analyzing the weight of edges (representing the similarity) of social groups in consecutive periods. We leverage the average Jaccard similarity metric as defined in Section 2.3; the closer this similarity is to 1, the larger the overlap between social groups in consecutive periods. Figure 17(b) plots a CDF of the average Jaccard similarity for the 5000 ego-centric graphs. The median of this CDF is only 0.22, which means that on average the overlap of social groups over time is relatively small – about 22%.

Finally, we evaluate the frequency of occurrence of the neighbor that appears most often in the social network of an ego. For this evaluation we count in how many of the ten observed periods does each neighbor appear. We then sort the neighbors in decreasing order of appearance frequency. We compare the first, second and tenth most frequent neighbors to determine if there are groups of neighbors that appear more often and what would be a typical size of such groups.

Figure 18 presents our results. The median value for the first top neighbor is 8, while for the second and the tenth top neighbor it decreases to 6 and 3, respectively. These results indicate high persistence of at least one neighbor in the social graph. At the same time, a group of two most persistent neighbors would appear ten times in only 6.8% of the cases, which indicates that a group of most persistent neighbors would typically have very few members.

4 Discussion and conclusion

Our analysis of the Orange mobile network dataset indicates that the usage patterns in Ivory Coast differ drastically from the typical cellphone usage in countries from the western world [10]. Due to the lack of weekly or monthly utilization pattern, we hypothesize that the network utilization is not shaped by people's daily routines, rather peaks in utilization coincide with events in the country.

96% of the territory of Ivory Coast can be categorized as Rural based on the population density [3]. 51.2% of the country population lives in Rural areas while 48.8% lives in Urban areas. At the same time the number of antennas deployed in Rural areas is slightly lower than these deployed in Urban. About a fourth of all the network activity is

initiated by Rural areas, a half of the network activity comes from Urban residents and the remaining fourth is a mix of Suburban and traffic that cannot be identified (due to lack of antenna ID information). This lower activity of Rural compared to Urban residents can be attributed to one of two factors: (i) the population coverage by mobile phone networks is lower in Rural than in Urban areas and (ii) people in Rural areas have lower purchasing power to afford to use mobile communication services. We observe higher number of conducted calls in all scenarios where the mobile call originators and terminators are within close proximity, which indicates high locality of interest in mobile phone communications in Ivory Coast. These results make a strong case for the need of an alternative solution for local voice communication such as [4].

In our evaluation of the relationship between population density and mobile phone usage in Section 3.3, we were surprised to find erratic usage corresponding to very sparsely populated areas, with mean values oscillating between ten and two mean calls per hour. This was in contrast to the more consistent mean number of outbound calls associated with higher population densities. We attribute this to two related factors: sparse antenna placement in sparsely populated areas and antenna placement coinciding with transportation infrastructure. Figure 7 illustrates the coincidence of antennas and the major road system. Even though they are located in regions with low population density, antennas placed alongside major roads transmit a higher number of outbound calls than antennas placed in sparsely populated areas away from infrastructure. Although 42% of recorded antennas are associated with areas classified as Rural, over 96% of square kilometers comprising Ivory Coast's surface area is classified as Rural. With a disproportionate few antennas representing areas of low population density and those antennas behaving in observably different manners based on proximity to infrastructure, it is unsurprising that call duration and frequency are so erratic when observed across areas of low population densities.

We leverage the ego-centric social graphs dataset to analyze the persistence of social groups with which an individual communicates. Our results indicate that two types of social groups exist in the provided ego-centric graphs: (i) social groups that likely occur once and can be attributed to communication activity such as scheduling an appointment, and (ii) social groups that occur persistently over time, which can be associated with regular communication with other subscribers. While the overlap of such persisting social group over time is not very large, there are one to two individuals for each subscriber that persist over the observed period.

Our first hand experience in rural Zambia indicates that often health care initiatives are jeopardized by the lack of reliable information channel between health care providers and targeted individuals. Thus, advanced mechanisms for information dissemination in the context of health care will help significantly improve these services in rural areas. The trends we discover in social groups persistence can serve as a basis for development of algorithms for selection of information relays in egocentric social networks. We hope that the knowledge obtained from such analysis can be further incorporated in information dissemination mechanisms in cases where the ego has limited or no access to a cellphone. We note, however, that while these results are encouraging, further analysis of social groups is needed. Such analysis should focus on social trends in Rural areas specifically and needs to incorporate more information related to individuals' location as well as direction, frequency and duration of calls.

References

[1] Dr math – remote math tutoring using mxit in south africa. `http://www.elearning-africa.com/eLA_Newsportal/mixing-it-with-dr-math-mobile-tutoring-on-demand/`.

[2] European Commission urban-rural typology. `http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/Urban-rural_typology`. Accessed: 09/02/2013.

[3] Rural population in cote d'ivoire. `http://www.tradingeconomics.com/cote-d-ivoire/rural-population-wb-data.html`. Accessed: 03/02/2013.

[4] A. Anand, V. Pejovic, E. M. Belding, and D. L. Johnson. VillageCell: Cost effective cellular connectivity in rural areas. ICTD, Atlanta, Georgia, March, 2012.

[5] R. Anderson, E. Blantz, D. Lubinski, E. O'Rourke, M. Summer, and K. Yousoufian. Smart connect: last mile data connectivity for rural health facilities. In *NSDR*, San Francisco, CA, 2010.

[6] R. Chaudhri, G. Borriello, and W. Thies. FoneAstra: making mobile phones smarter. In *NSDR*, San Francisco, CA, 2010.

[7] M. de Bruijn, F. B. Nyamnjoh, and I. Brinkman. Mobile phones: The New Talking Drums of Everyday Africa. 2009.

[8] J. Donner. The rules of beeping: exchanging messages using missed calls on mobile phones in sub-saharan africa. *International Communications Association*, 2005.

[9] D. L. Johnson, V. Pejovic, E. M. Belding, and G. van Stam. Villageshare: facilitating content generation and sharing in rural networks. In *Proceedings of the 2nd ACM Symposium on Computing for Development*, ACM DEV '12, pages 7:1–7:10, New York, NY, USA, 2012. ACM.

[10] G. Krings, M. Karsai, S. Bernhardsson, V. Blondel, and J. Saramki. Effects of time window size and placement on the structure of an aggregated communication network. In *EPJ Data Science*, May 2012.

[11] A. Kumar, J. Chen, M. Paik, and L. Subramanian. ELMR: Efficient Lightweight Mobile Records. In *MobiHeld*, Barcelona, Spain, 2009.

[12] I. Mbiti and D. N. Weil. Mobile banking: The impact of m-pesa in kenya. Working Paper 17129, National Bureau of Economic Research, June 2011.

[13] M. Minges. Mobile cellular communications in the southern african region. *Telecommunications Policy*, 23(7):585–593, 1999.

[14] S. K. Neil Patel and T. S. Parikh. An asymmetric communications platform for knowledge sharing using cheap mobile phones. In *ACM Symposium on User Interface Software and Technology (UIST)*, Santa Barbara, CA, October 2011.

[15] M. Paik, J. Chen, and L. Subramanian. Epothecary: cost-effective drug pedigree tracking and authentication using mobile phones. In *MobiHeld*, Barcelona, Spain, 2009.

[16] T. Parikh, N. Patel, and Y. Schwartzman. A survey of information systems reaching small producers in global agricultural value chains. In *International Conference on Information and Communication Technologies and Development*, Bangalore, India, December 2007.

[17] N. Patel, D. Chittamuru, A. Jain, P. Dave, and T. S. Parikh. Avaaj otalo: a field study of an interactive voice forum for small farmers in rural India. In *CHI*, Atlanta, GA, 2010.

[18] A. Reda, S. Panjwani, and E. Cutrell. Hyke: a low-cost remote attendance tracking system for developing regions. In *NSDR*, Bethesda, MD, 2011.

[19] C. C. Robusto. The cosine-haversine formula. *The American Mathematical Monthly*, 64(1):38–40, 1957.

Multi-perspective analysis of D4D fine resolution data

Gennady Andrienko, Natalia Andrienko, and Georg Fuchs

Fraunhofer Institute IAIS, Schloss Birlighoven, 53757 Sankt Augustin, Germany

1 INTRODUCTION

Nowadays, huge amounts of movement data describing changes of spatial positions of discrete mobile objects are collected by means of contemporary tracking technologies such as GPS, RFID, and positions within mobile phone call records. Automatically collected movement data are semantically poor as they basically consist of object identifiers, coordinates in space, and time stamps. Despite that, valuable information about the objects and their movement behavior, as well as about the space and time in which they move can be gained even from such basic movement data by means of analysis [1].

Movement can be viewed as consisting of continuous paths in space and time [2], also called trajectories, or as a composition of various spatial events [3]. Movement data can be aggregated in space, enabling identification of interesting places and studying their activity characteristics, and by time intervals, enabling similarity analysis of situations comprising different time intervals as well as detection of extraordinary events.

In this study, we consider the D4D fine resolution call data records (CDR) set of Ivory Coast [4] from multiple perspectives. To set the scope we first evaluate the properties of the data that restrict potentially applicable movement data analysis methods (Section 2). A first analysis step is to study spatio-temporal patterns of calling activities at multiple resolutions of time. To this end we apply spatio-temporal aggregations by antennas, counting number of calls per day (Section 3) and per hour (Section 4). To further identify different kinds of activity neighborhoods and to study their spatial distribution we then characterize antennas by feature vectors of hourly activities within a week and cluster them by similarity of the feature vectors (Section 5). In order to identify peak events – i.e., time intervals during which extraordinarily large number of people made calls in one location simultaneously - we compare time series comprising counts of distinct phone users per time interval and antenna (Section 6). This procedure allows us to identify large-scale events that, possibly, happened in the country. We use trajectories of mobile phone subscribers for reconstructing flows between major towns and between activity regions of the country (Section 7). Finally, we make an attempt at semantic interpretation of individuals' personal places, such as home and work locations, based on these user trajectories (Section 8). We conclude this paper with a short discussion on the results and possible directions for further work.

2 EVALUATING DATA PROPERTIES

The provided data set comprises a total of 55,319,911 CDRs distributed over 10 individual chunks of between 4.8 and 6.5 million records, each corresponding to a set two-week time intervals. Of these, 47,190,414 CDRs are associated with one of the 1,214 antennas and thus be referenced the corresponding antenna's geographic coordinates. CDR temporal references are given with minute accuracy (i.e., seconds were suppressed) ranging from December 5, 2011 till April 22, 2012. Aggregation of geo-referenced calls by days (Figure 1) shows that some days (e.g. March 24, 2012) have much less number of calls than neighboring days. This observation suggests that quite many call activities are missing in the database, especially in April 2012. In addition, 8,129,497 calls refer to unknown antennas (id=-1), with maximal count 166,621 calls on April 1, 2012. Because these CDRs could not be geo-located and thus not related to other calls originating from the same location they were ignored during data import.

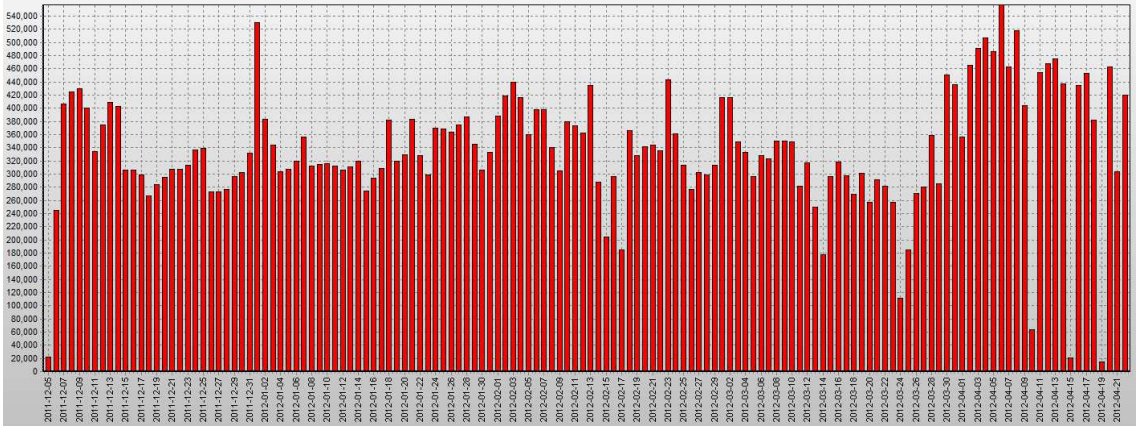


Figure 1. Daily counts of calls[1].

The figure also suggests obvious call peak patterns at New Year, Easter, and, to some extent, at Christmas 2011. Other peaks correspond to public holidays like The Day after the Prophet's Birthday (Sunday, February 5, 2012) and Post African

[1] Readers may access high-resolution versions of the figures contained in the paper at http://geoanalytics.net/and/d4d2013

Cup of Nations Recovery (Monday, February 13,2012)[2]. Wikipedia[3] suggests that religion in Ivory Coast remains very heterogeneous, with Islam (almost all Sunni Muslims) and Christianity (mostly Roman Catholic) being the major religions. Muslims dominate the north, while Christians dominate the south. Unfortunately, the amount of data available for the northern part of the country does not allow comparison of patterns in respect to religious holidays.

A considerable constraint in terms of mobility pattern analysis and semantic interpretation (Sections 7 and 8, respectively) arises from the anonymization procedure applied to the data [4]. Each of the 10 data chunks is a subset of 50,000 distinct mobile phone subscribers tracked over 2 weeks. User IDs associated with each CDR are obviously not real, traceable customer IDs but rather consecutive integer numbers. And while a given user ID is unique with respect to one data chunk, integers are reused (i.e., the counter was reset) between different chunks. This means that it is not possible to analyze movement patterns or flows over periods exceeding two weeks, or generally cover time intervals distributed over multiple chunks (compare Figure 2).

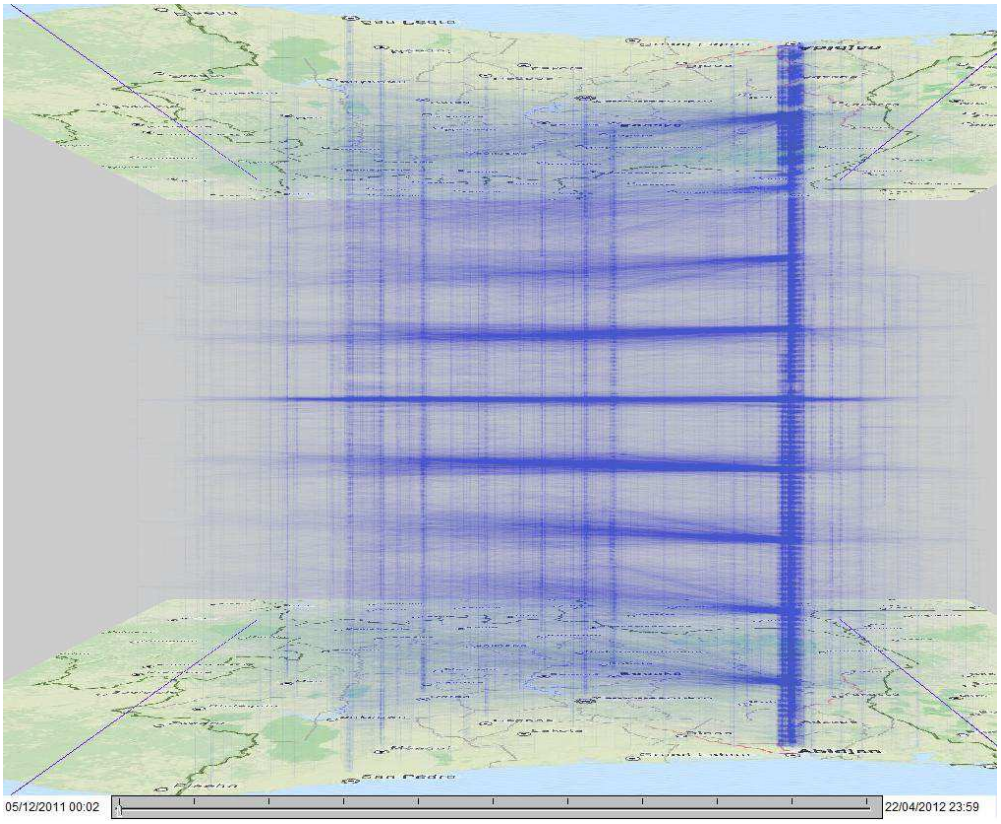


Figure 2. Space-Time Cube displaying the full 20-week data set of CDRs integrated into trajectories (sequence of calls with the same user id) with time increasing from bottom to top of the cube. Besides expected daily cycles e.g. in the area of Abidjan one can spot missing days (near the top), and very clearly the distinct pattern of bi-weekly "false trip" movement caused by re-assigning user IDs to different mobile phone users in other parts of the country between data chunks.

Moreover, a check for repeated combinations of user ID and time stamp produced 5,225,989 pairs that occurred 12,861,168 times in the database. The duplicates have been removed. This operation thus reduced the number of geo-referenced CDRs in the database by about 25%.

3 ASSESSING DAILY AGGREGATES FOR ANTENNAS

We have aggregated the remaining CDRs by antennas and days, producing daily time series of calls for each of the 1,214 antennas. Figure 3 presents an overview of their statistical properties. The upper part of the image shows the call counts' running average line (in bold) and dynamics by deciles (grey bands, min = 0, 10%, 20%, …, 90%, max = 5584) over time. Vertical lines correspond to weeks. The lower part of the image uses segmented bars to represent distribution of antennas categorized by their daily call counts. The darkest blue denotes absence of any calls for at those antennas; blue colors correspond to intervals from 1 to 50 calls per day, yellow represents 50 to 100 calls, orange and reds – more than 100 calls.

[2] Public holidays in Ivory Coast in 2012: http://www.asaralo.com/index.php?option=com_content&view=article&id=2367:public-holidays-in-cote-divoire&catid=160:african-public-holiday&Itemid=2598

[3] http://en.wikipedia.org/wiki/Ivory_Coast#Religion

We can make the following general observations:

- Too few data records on Dec 5, 2011 even though CDR time stamps for that day cover the entire 24h period.
- Gradual increase of counts of antennas without activity (0 calls per day) from Dec 6, 2011 till March 27, 2012.
- Several days with missing data on many antennas (March 29, April 1, April 10, April 15, April 19).
- Absence of typical weekly patterns with different amounts of calls at working days and weekends.

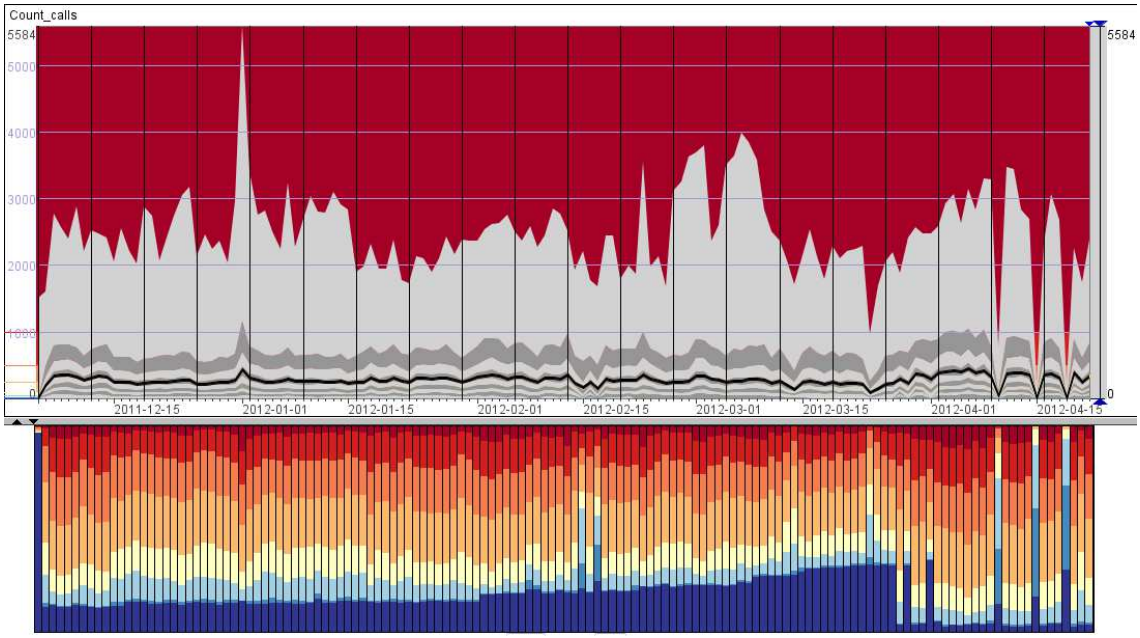


Figure 3. Top: dynamics of deciles of counts of call per antenna distributions. Bottom: daily proportions of antennas with N calls in intervals of 0 (darkest blue), 1..10, 10..50, 50..100 (yellow), 100..200, 200..500, 500..1000, and more than 1000 (darkest red) per day. Note that in the upper image, corresponding interval boundaries are indicated in the scale to the left.

These general observations do not reflect the geographic distribution of patterns. To take the geography into account, we represent the call counts on maps by mosaic diagrams. A diagram consists of a pixel grid with each pixel representing one day's call count by color, using the same color coding as in Figure 3. The pixels are arranged in 2D as in a calendar sheet: columns correspond to days of week (from Monday to Sunday, from left to right) and rows correspond to weeks (from 1 to 20, from top to bottom). Figure 4 shows the entire country, Figure 5 a close-up of the region of the towns Abidjan and Abobo. The large consecutive sections of dark blue colors in many diagrams suggest that the data contain systematically missing portions. In particular, data are completely unavailable 12-14 weeks for many antennas in the northern part of the country, and for more than 16 weeks in the southern part of Abidjan.

Another observation is that all columns in the diagrams look quite similar. This is very different from mobile phone usage patterns observable in Europe and the USA where weekends differ strongly from working days in terms of calling counts. There, calls from the downtown areas of large cities are quite rare on Saturdays and Sundays in comparison to weekdays. We cannot find such patterns in the D4D data set. This suggests that the life style and temporal organization of economic activities in Ivory Coast differ significantly from those cultural regions. Therefore a straightforward application of analysis methods developed primarily for European countries is not valid.

One more complexity of the data is caused by the data sampling and anonymization procedures [4]. For each two-week period, a subset of 50,000 customers has been selected. It is not guaranteed that the subsets represent population samples with similar demographic and economical characteristics. Indeed, clustering days by feature vectors comprising counts of calls at each antenna, followed by assigning colors to clusters by similarity [5] clearly demonstrates the dissimilarity of patterns in consecutive two-weeks periods (Figure 6). Additionally, this display also does not give any evidence of differences between week days and weekends.

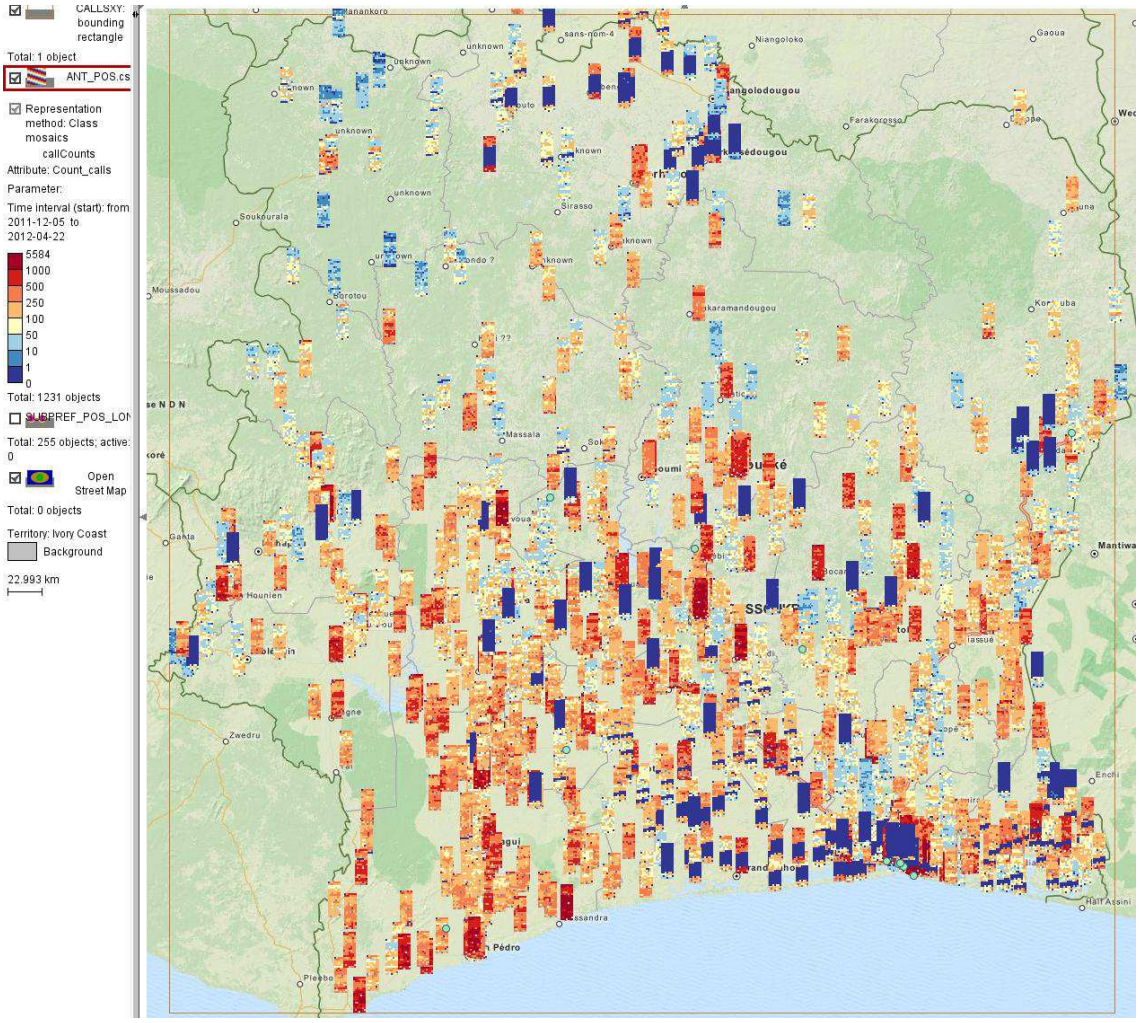


Figure 4. Mosaic (segmented) diagrams show counts of calls for all antennas in the whole country. Counts are represented by colored segments ranging from blue (0 calls) through yellow (50..100 calls) to red (more than 1,000 calls). Diagram rows correspond to weeks (top to bottom – from week 1 to week 20) and columns to days of week (left to right: from Monday to Sunday).

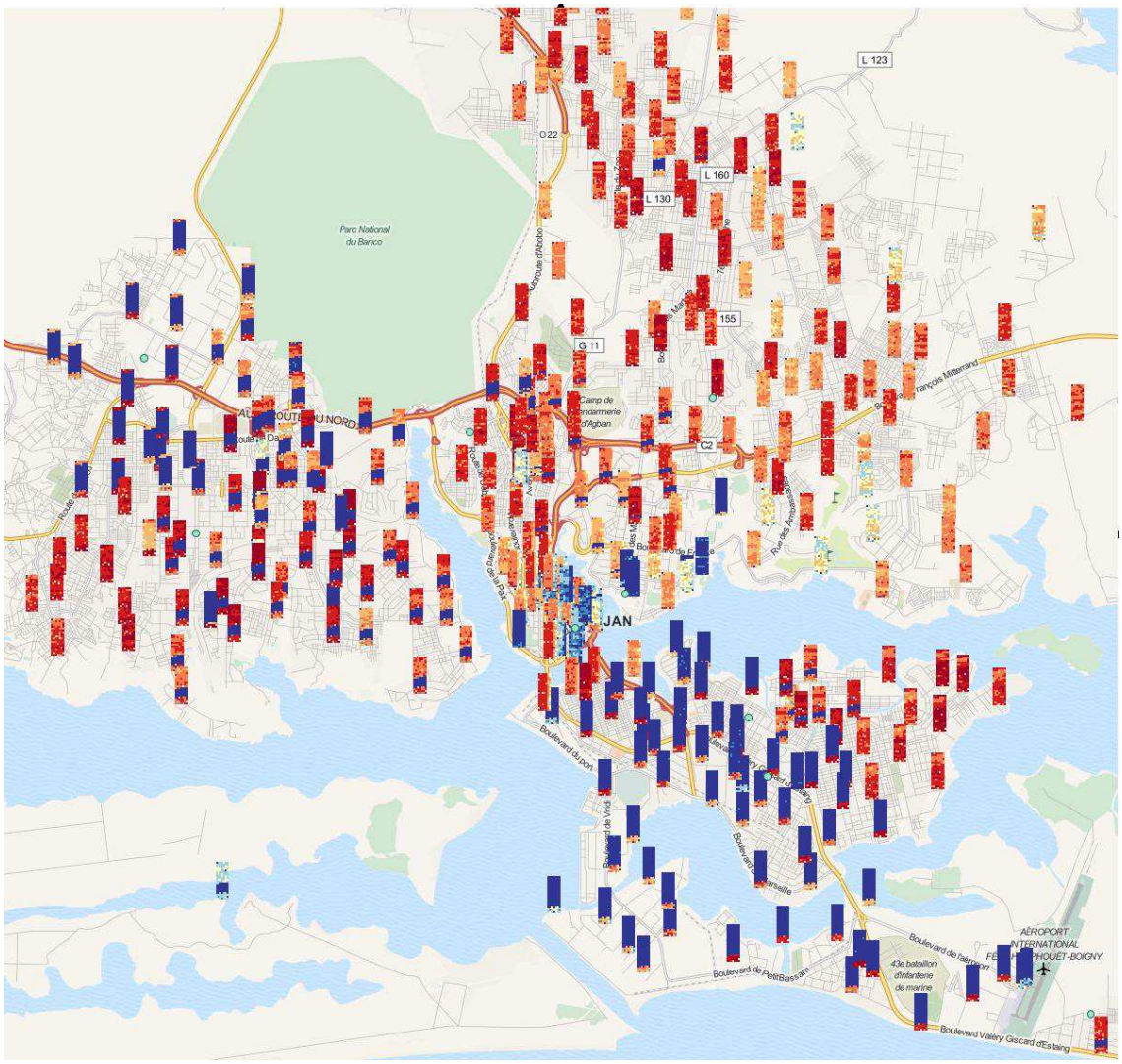


Figure 5. Close-up view of the region of the towns Abidjan and Abobo. The mosaic diagrams are encoded in the same way as in Figure 4 and use the same color coding.

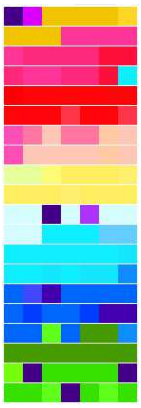


Figure 6. Similarity of situations during 7 days x20 weeks, represented by assigning colors to segments of the diagram according to the cluster the corresponding day belongs to.

4 ANALYZING HOURLY AGGREGATE PATTERNS FOR ANTENNAS

Taking into account the properties of the data, we decided to aggregate calls by antennas for hours of day and days of week, irrespectively of calendar dates. Figure 7 shows mosaic diagram maps for two locations, the country's capital (Yamoussoukro) and a port town (San Pedro). Like in Figures 4 and 5, the diagrams consist of segments representing call counts by colors, from dark blue (no calls) through yellow (50-100 calls per hour) to red. The segments of each diagram are arranged by days of week (Monday to Sunday from left to right) and by hours of day (from 0:00 on top to 23:00 at bottom).

One can see different temporal signatures of calling activities. Thus, in some antennas calls are more frequent at evening times, some have uniform distribution of call counts during daytime hours, while yet others have similar distributions at morning and evening times etc. However, the total amounts of calls differ significantly from one antenna to another, thus making direct comparison and grouping quite difficult.

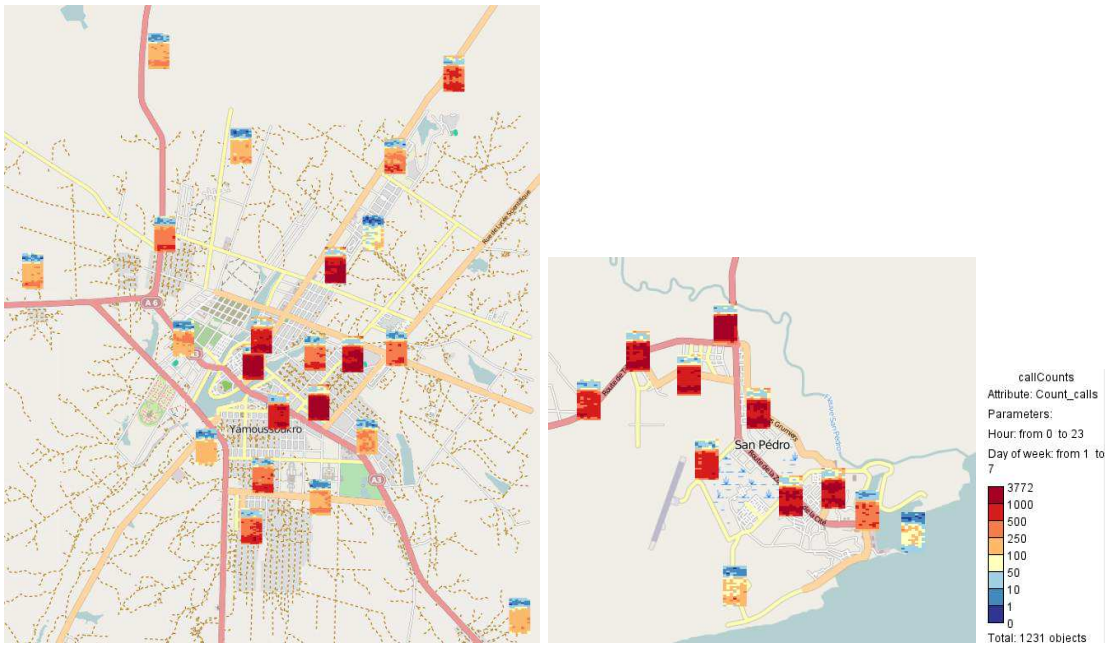


Figure 7. Mosaic diagrams show hourly absolute counts of calls for 7 days of week (by columns, from Monday to Sunday) and 24 hours of day (from 0:00 to 23:00) in Yamoussoukro and San Pedro.

To compensate for different amounts of calls at different antennas, we have applied normalization to each time series by its own mean and standard deviation values, see Figure 8. The resulting images convincingly demonstrate that there exist distinct patterns of hourly calling activities at different antennas. Moreover, these patterns tend to be clustered in geographical space. For example, almost all antennas in the outskirts of Yamoussoukro are characterized by dominant evening call pattern, while in the city centre calls are distributed uniformly during day. There are only few evidences of different calling activity patterns on Saturdays and Sundays (i.e., in the two rightmost columns of the diagrams) in comparison to working days. One such example can be found in the southern part of San Pedro, and some others in the southern part of Yamoussoukro.

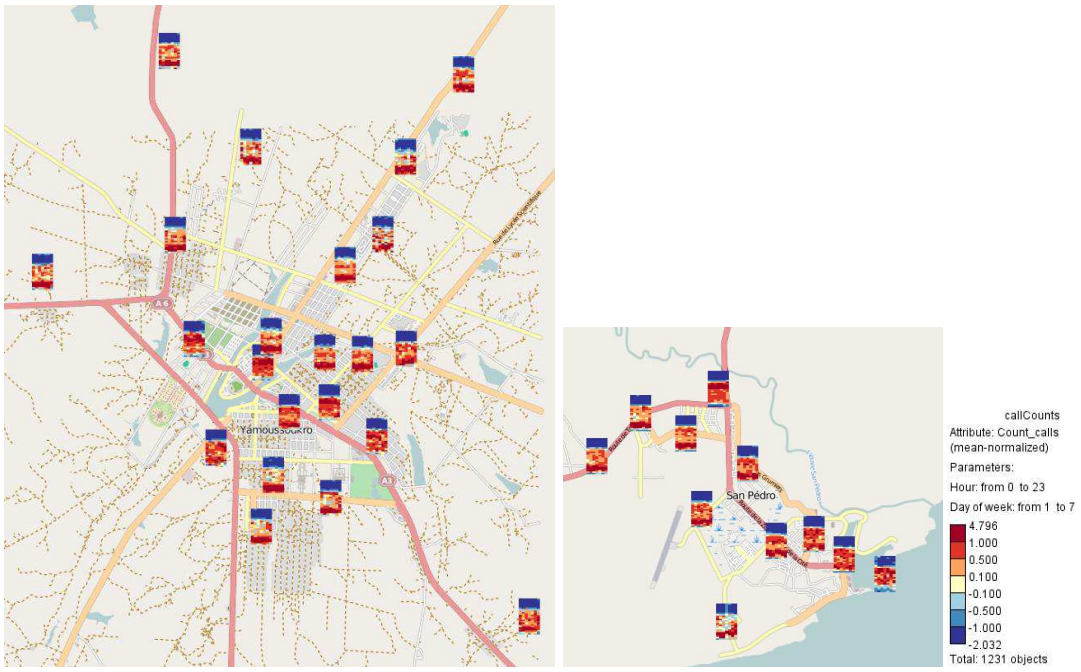


Figure 8. Similarly to Figure 7, mosaic diagrams show hourly show counts of calls for 7 days of week (by columns, from Monday to Sunday) and 24 hours of day (from 0:00 to 23:00) normalized by average count per antenna in Yamoussoukro and San Pedro.

5 CLUSTERING ANTENNAS BY SIMILARITY OF HOURLY AGGREGATE PATTERNS

Visual inspection and comparison of mosaic diagrams has limited applicability. We can perform it for selected cities and regions, but can't apply systematically for the whole country. Instead, we can apply clustering of antennas according to mean-normalized hourly activity profiles over week. We've used *k*-Means and varied the desired number of clusters from 5 to 15, the most interpretable results have been obtained with N=7. Lower number of clusters mixes several behaviors, while large counts extract small clusters with too specific behaviors.

The results are presented in Figure 9. Seven time graphs show profiles of the 7 clusters for 7 days of week. Centroids of the clusters have been projected onto the 2d plane by Sammons mapping [6] (middle left). Following the ideas of [7], colors have been assigned to the clusters according to these 2D positions, thus reflecting relative cluster similarities. The representative feature vector of the cluster centroids are presented by mosaic diagrams (middle right, days of week in columns, hours of day in rows, similarly to Figures 7 and 8), with their placement again corresponding to the respective centroid's Sammons projection. Using these visualizations, we can suggest some interpretations to the clusters:

Cluster 1: High calling activity in the evenings, irrespective of the day of week. Such a profile is typical for residential districts with a high proportion of employed population.
Cluster 2: Uniform calling activity during the day, with some increase in the morning on Monday, Wednesday, Friday and Saturday.
Cluster 3: High calling activity in the evenings, medium activity in mornings, and decreased activity in the middle of the day (except Sundays)
Cluster 4: High calling activity during working hours (except Sundays), with extremes in mornings. Such a profile is typical for business districts.
Cluster 5: Very low calling activity, with only small differences between day and night. This is quite typical for unpopulated areas and for antennas masked (in terms of call handling) by neighboring antennas.
Cluster 6: Similar to cluster 3, however with a less prominent evening pattern but more prominent morning pattern, and increased activity on Saturdays and Sundays.
Cluster 7: Similar to clusters 3 and 6, but with decreased activity on Sundays.

Our experience of analyzing mobile phone usage data in different countries suggests that cluster 1 corresponds to residential districts with high proportion of regularly employed population, in other words, people having fixed out-of-home work schedules, and that cluster 4 represents business districts. We guess that cluster 2 either represents regions with a mix of residential and business land use, or businesses with irregular schedules. Major transportation corridors (main roads, railways) can be characterized by similar temporal patterns, too. Clusters 3, 6 and 7 may represent mostly residential areas with partly employed population, or population with flexible work schedule.

The three maps at the bottom of Figure 9 show, from left to right, the spatial distribution of the clusters for the whole country, its southern part, and the city of Abidjan, respectively. We can observe that our possible interpretations correspond to geographical patterns.

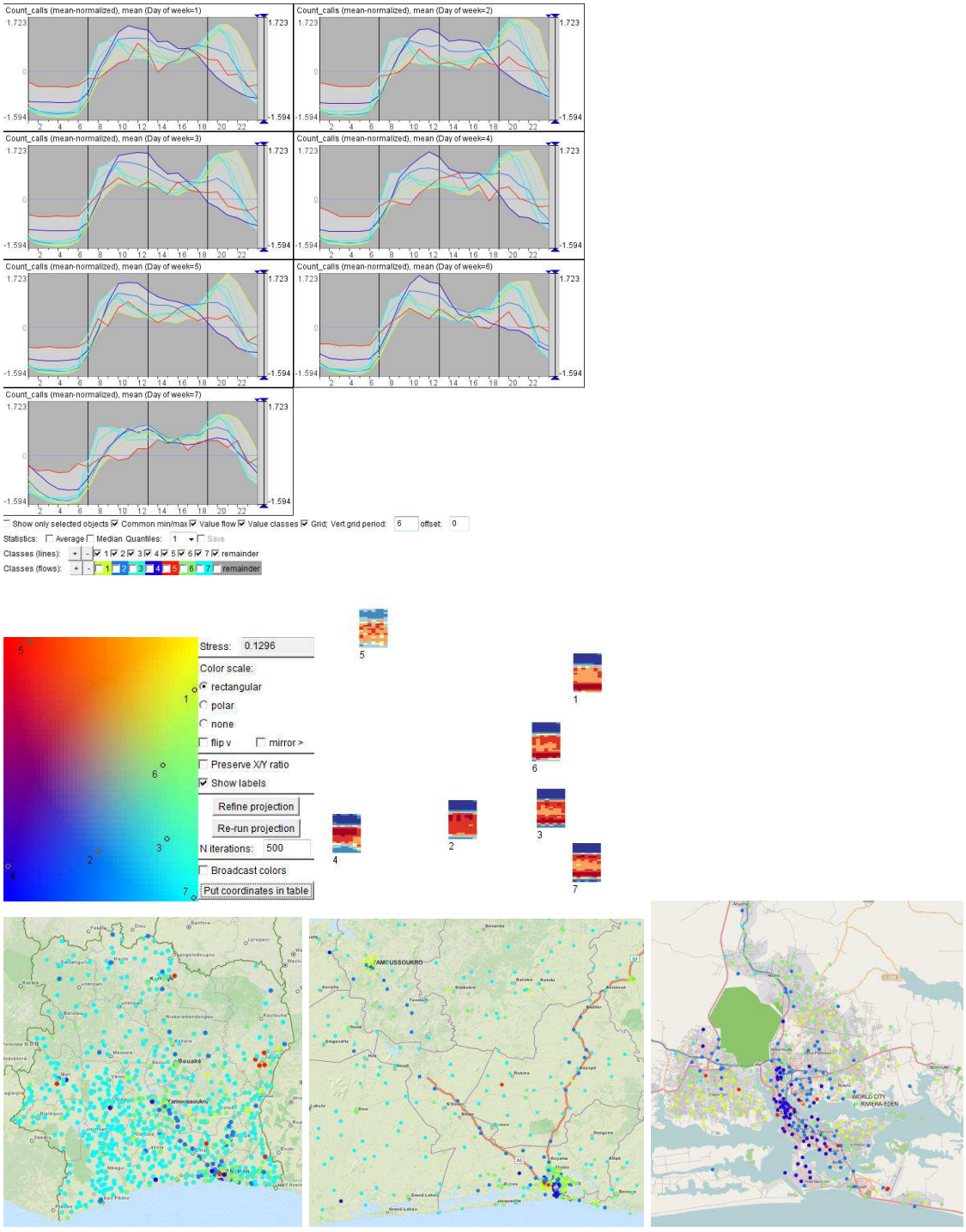


Figure 9. Normalized temporal signatures of antennas are used for defining 7 clusters by *k*-Means. Time graphs in the upper panel show profiles of these clusters during 7 days of week. Colors are assigned to the clusters according to positions of cluster centroids in Sammons mapping (middle-left). Representative activity profiles for the clusters are shown by 2D mosaic diagrams in the middle-right. The maps at the bottom show spatial distributions of the clusters for the whole country (left), south-west part (center) and the region around Abidjan (right).

6 PEAK DETECTION FROM HOURLY TIME SERIES AT ANTENNA LEVEL

Besides examining regular, everyday-life activity patterns we further want to detect interesting events that attracted many people. For this purpose, we need to count the number of different people per antenna cell and time unit (rather than the total number of calls / CDRs as used in the previous sections). It should be noted again that data have been provided in 2-weeks

portions with repeated user IDs across the different portions, therefore limiting time intervals eligible for such analysis in this particular data set due to the inability to distinguish users between data chunks.

We focus our further analysis on trajectories (sequences of positions) of different users during last two weeks of the data set. This is the only period that contains rather complete geographic coverage, see Section 3 for details. For each distinct antenna we have computed hourly counts of distinct user IDs active at this antenna. These counts roughly represent the presence of people in antenna cells. If a person made several calls from the same antenna, we assume that he did not move away between the calls. It should be noted that this assumption may be incorrect in some cases, in that people may transition out of an antenna's cell and back without making a call at another antenna in the meantime.

Figure 10 (top) shows a time graph with a selection of time lines corresponding to antennas. Straight horizontal lines on April 10, April 15 and April 19 correspond to missing data that we already identified earlier in Figures 1 and 3. To find unusual concentrations of people at antennas, we have searched for peaks of averaged presence magnitudes exceeding 20 distinct peoples over a sliding, 3 hours time window [8]. The appropriate parameters for magnitude threshold and time window width have been defined using a sensitivity analysis procedure as suggested by [9]. In particular, the time graph in Figure 10 (top) only shows lines for those antennas that exhibit at least one such peak event. The horizontal event bar immediately below the time graph shows the counts of events over time. The 2D periodic event bar in Figure 10 (bottom left) shows counts of peak events per 24 hours of day (columns) and 14 days of two weeks (rows). The map (bottom-center) and space-time cube (bottom-right) show spatial and spatio-temporal distributions of peak events.

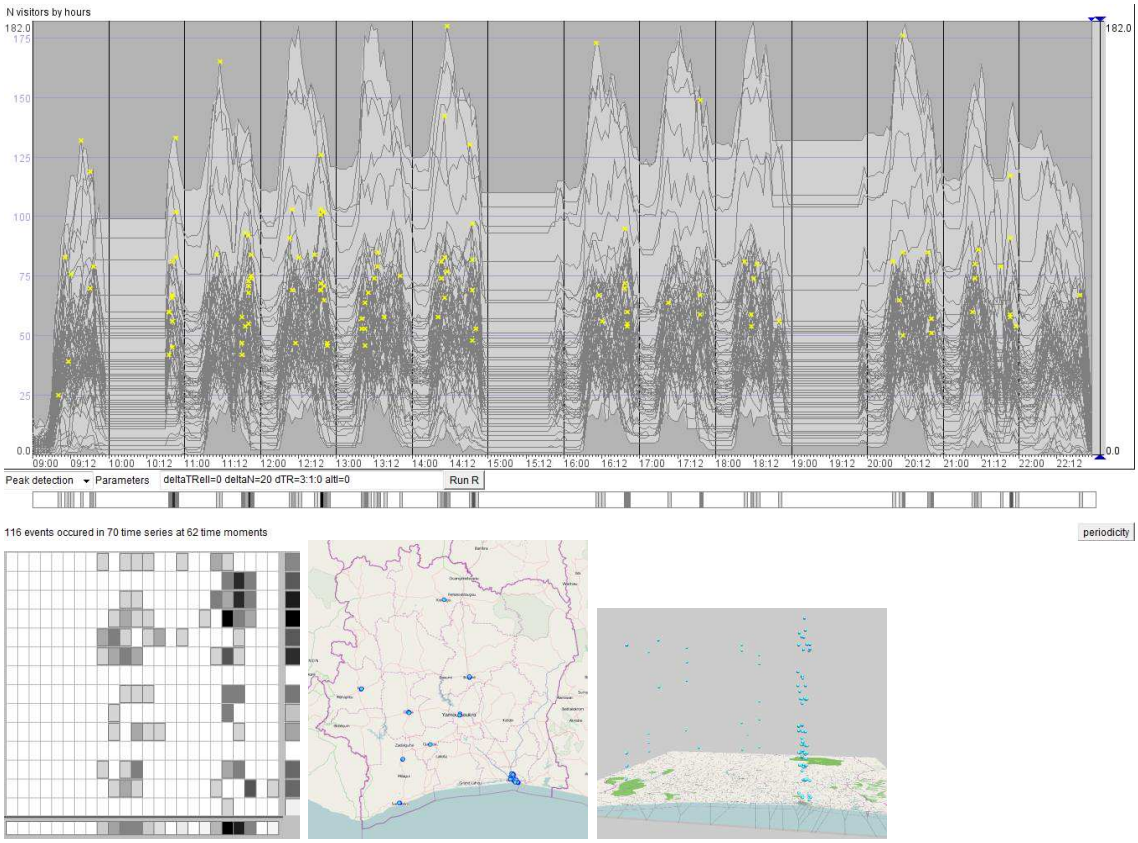


Figure 10. The time graph at the top shows time series of counts of mobile phone users grouped by antennas, at 1 hour resolution. Peaks with magnitude of at least 20 users over 3 hour intervals are marked by yellow crosses. Counts of peaks are shown in 2d periodic event bar at the bottom-left. Positions of peak events are shown on the map of the country in the bottom-center map and in the space-time cube at bottom-right.

We can observe that peak events are frequent in the middle of the day and early in the evening. There are only few exceptions. Thus, several peak events happened during the 15:00 – 16:00h interval on Monday and Fridays of the 1st week, and late in the evening of Saturday of the 2nd week. By clicking the corresponding segment of the periodic event bar, we select the corresponding antennas and time series (see Figure 11). We can see that these peaks happened in 4 different towns in different parts of the country. The time series profiles for those regions indicate that these peaks are rather unusual. We guess that some kind of connected public events happened simultaneously in these regions.

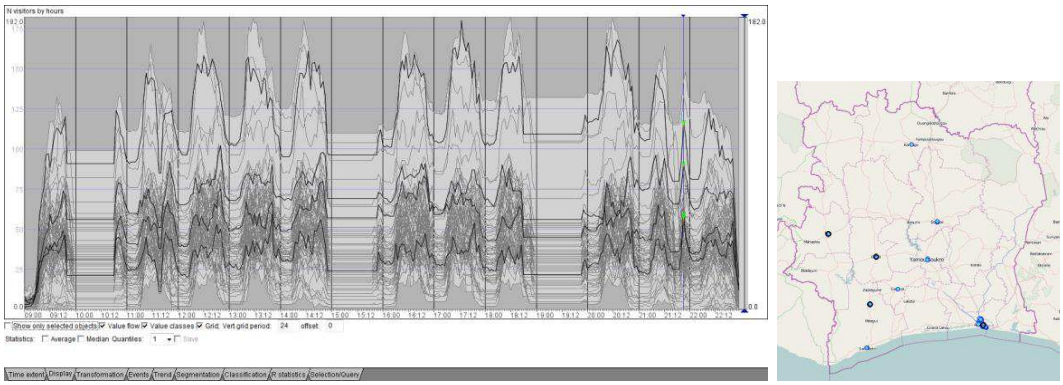


Figure 11. Peaks that happened at 21:00 on the 2nd week's Saturday and their containing time series are highlighted in the time graph (left). Simultaneously, their positions are marked on the map (right).

It is interesting to relate the magnitude of peaks with the maximal values of the time series. We found two extreme cases of time series with peaks of more than 20 people contained in time series with maximum (peak) values of about 40 but average daily values of only about 10..15 people (Figure 12). Both events happened in Abidjan. Probably, some local events happened at about 10:00 on Monday and at 21:00 of Thursday in these locations.

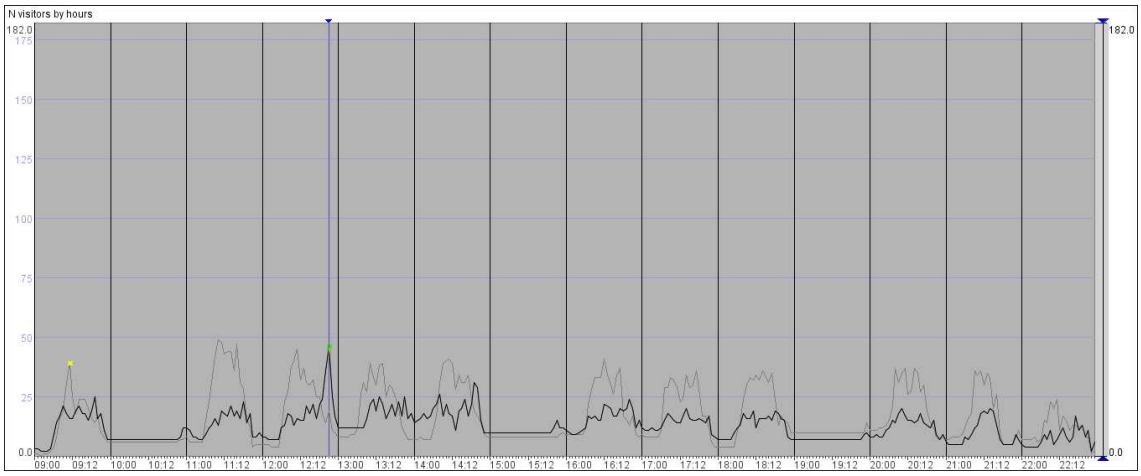


Figure 12. Peaks on Monday morning (yellow cross) and Thursday evening (green cross) are shown on top of two time series with otherwise usually low presence of calling activities. Both peaks have happened in Abidjan.

We found that peak events happened in almost all major towns of the country. To get a flavor of mobility of mobile phone users in Ivory Coast, we outlined areas around the peak events and then calculated counts of direct transitions between these locations, see Figure 13. The thickness of the arrows reflects the magnitude of flows between the corresponding places during the two-week period. This map shows us the strength of connections between locations of different activities.

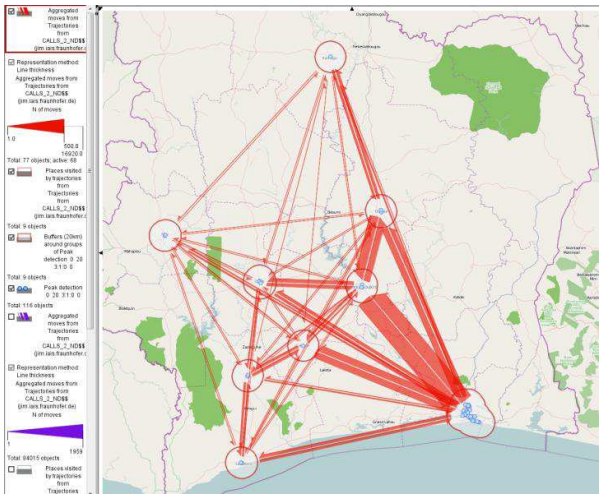


Figure 13. Flows between regions that correspond to peaks in people presence.

7 ANALYSIS OF FLOWS

To explore the mobility flows more systematically, we have applied a method for generalization and summarization of trajectories [10] to the phone user trajectories over last two weeks. The method extracts so-called characteristic points of trajectories, aggregates these points in space with a desired resolution, and finally uses the medoids of the resulting spatial clusters as seeds for generating a Voronoi tessellation of the territory. The method simplifies trajectories while minimizing their distortion with respect to the corresponding original, full-detail version.

Figure 14 (left) shows the original trajectories rendered with high level of transparency (about 99%). This representation gives us a hint about major flows, but does not allow quantifying them. Figure 14 (right) shows the flows between aggregated regions as well as the accumulated counts of distinct users recorded in each region during the two-week period.

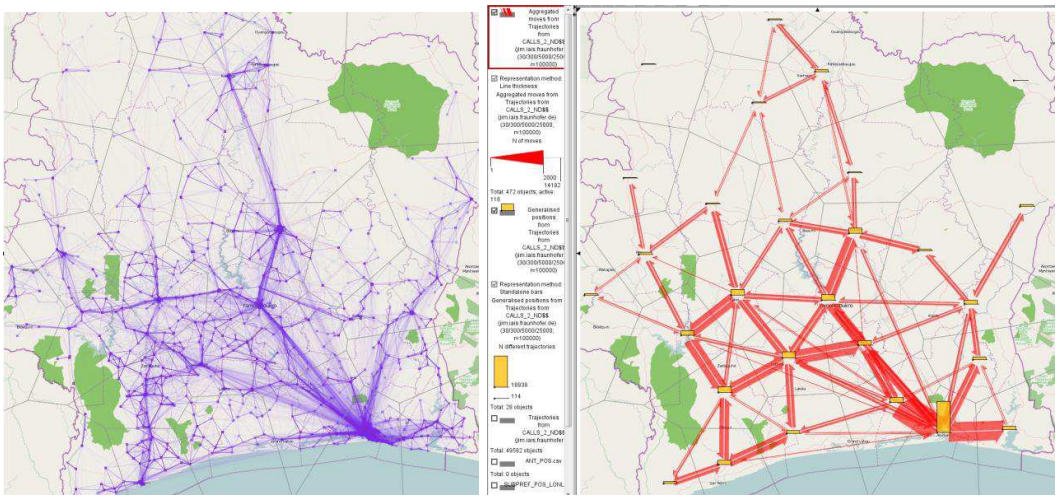


Figure 14. All trajectories during last two weeks drawn as accumulation of semi-transparent lines (left). Trajectories are summarized by 28 aggregated regions (Voronoi polygons) of approximately 100km radius. Flows between regions are represented by red arrows with flow magnitudes encoded in the arrow width. Counts of mobile phone owners registered in each area are shown by yellow bars.

We can observe the consistency between the flow maps in Figures 13 and 14, respectively. However, the latter map uncovers more structural details. In particular, we can see a branch connecting Abidjan with the mid-eastern region of the country. There are only relatively few direct connections between Abidjan and Yamoussoukro, and fewer still between these two and towns in the northern part of the country. This indicates that despite the existences of several local airports, people mostly use ground transportation and make phone calls / send SMS during their lengthy trips. By contrast, air travel typically manifests itself as long-distance flows since the mobile phone is switched of or out of range during flight with no calls at intermediate antennas.

Further analysis (omitted here for space / time constraints) could allow us to identify temporal patterns of flows and assess usual travel times between different locations. We could also find frequent sequences of visited regions and assess the dynamics of such trips.

8 SEMANTIC ANALYSIS OF PERSONAL PLACES

To identify routine trips of people and to obtain interpretations of their personal places, we have applied the procedure proposed in [11] to a small subset of trajectories that are characterized by large numbers of calls in different locations. We have used a sample of the data consisting of 86 trajectories recorded during the last two weeks of the data period and with bounding rectangle diagonals exceeding 5km. The total number of call records in this sample is 133,029. First, we have identified stops as sequences of consecutive calls that occurred within 30 minutes and a rectangular region of less than 500m diagonal. Using these parameters extracted 7,149 stops. The stops have then been clustered by means of the density-based clustering method Optics [12], separately for each trajectory. Parameters have been chosen to group points having at least 5 neighbors within 500m distance. Noise points not grouped into any cluster (1,300 points in total, or about 19% of the set) have been excluded from subsequent analysis as they are assumed to represent infrequently visited locations. For each cluster the counts of calls have been aggregated for every hour of the day. This resulted in time series comprised of 24 one-hour intervals assigned to each cluster.

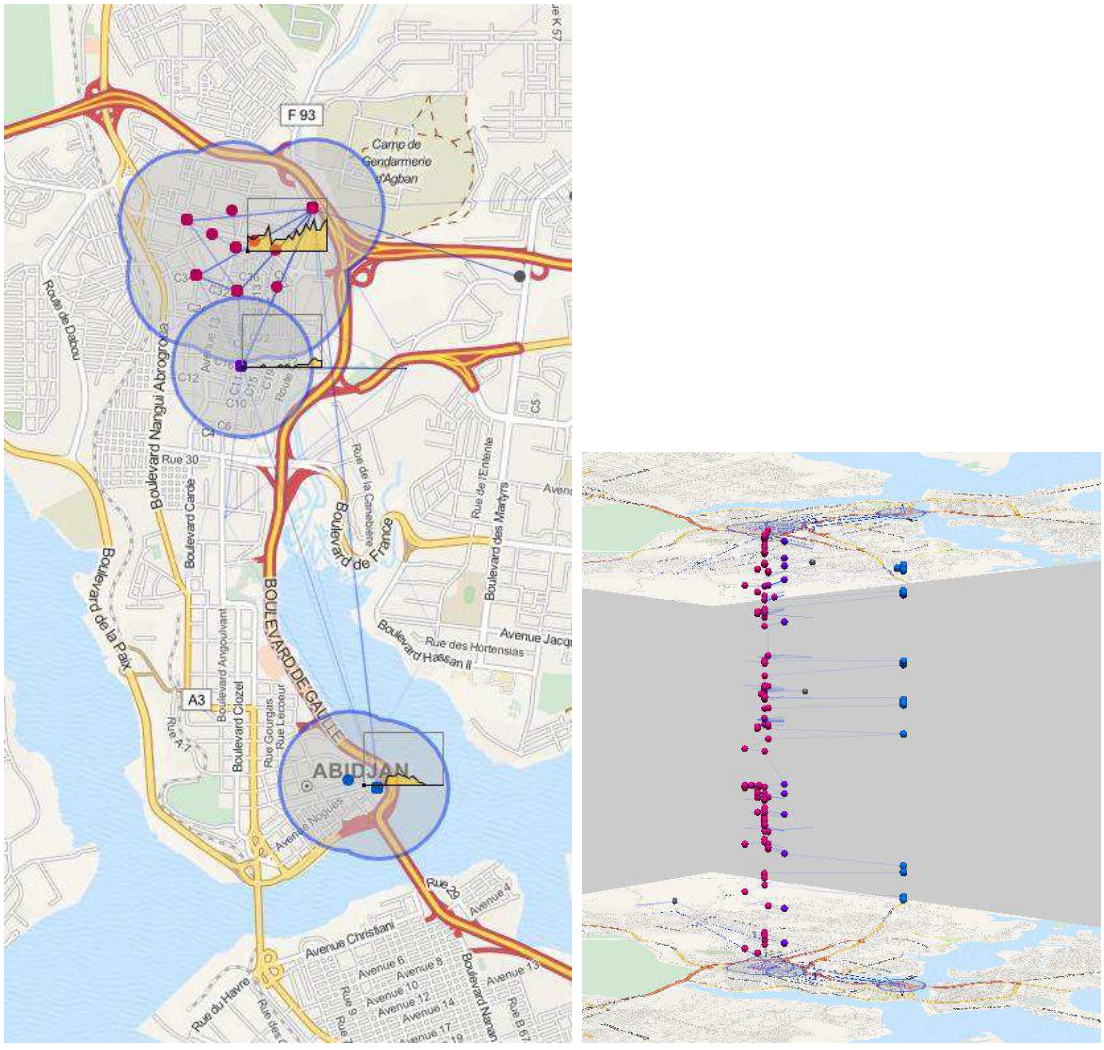


Figure 15. Individual locations of repeated activities are shown by 500m buffer polygons for subscriber #548709. Hourly temporal signatures (according to hours of day) are shown by time flow diagrams. Spatio-temporal positions of calls are shown in the space-time cube. Red dots represent home-based calls, blue dots correspond to the person's work place, and prurple to the primary location of her evening activities. Gray dots in the space-time cube represent irregular activities.

Figure 15 shows routine activity locations for a single person, id #548709, in space and time. The "blue" place to the south was attended only during day time. Most probably, this is the work place of this person. We guess that she has regular work with fixed working times. The "purple" place in the middle is visited less frequently and only during evening times. We

guess this is a place of repeated social activities of the person. Finally, the location in the north is characterized by activities at any times, including night times (but less during the day). We interpret this place as the person's home.

Interpretation of semantically meaningful personal places can be automated. For example, we can compute similarities of all 24-hour feature vectors to a selected vector corresponding to "work" activities, see Figure 16. This map shows locations and temporal signatures of places that can be interpreted as work locations for different mobile phone owners. Partial dissimilarity of the temporal profiles suggests different working hours. For example, some persons seem to be less active at lunch times. Spatial clustering of several people's "work" places suggests concentrations of business locations in the city.

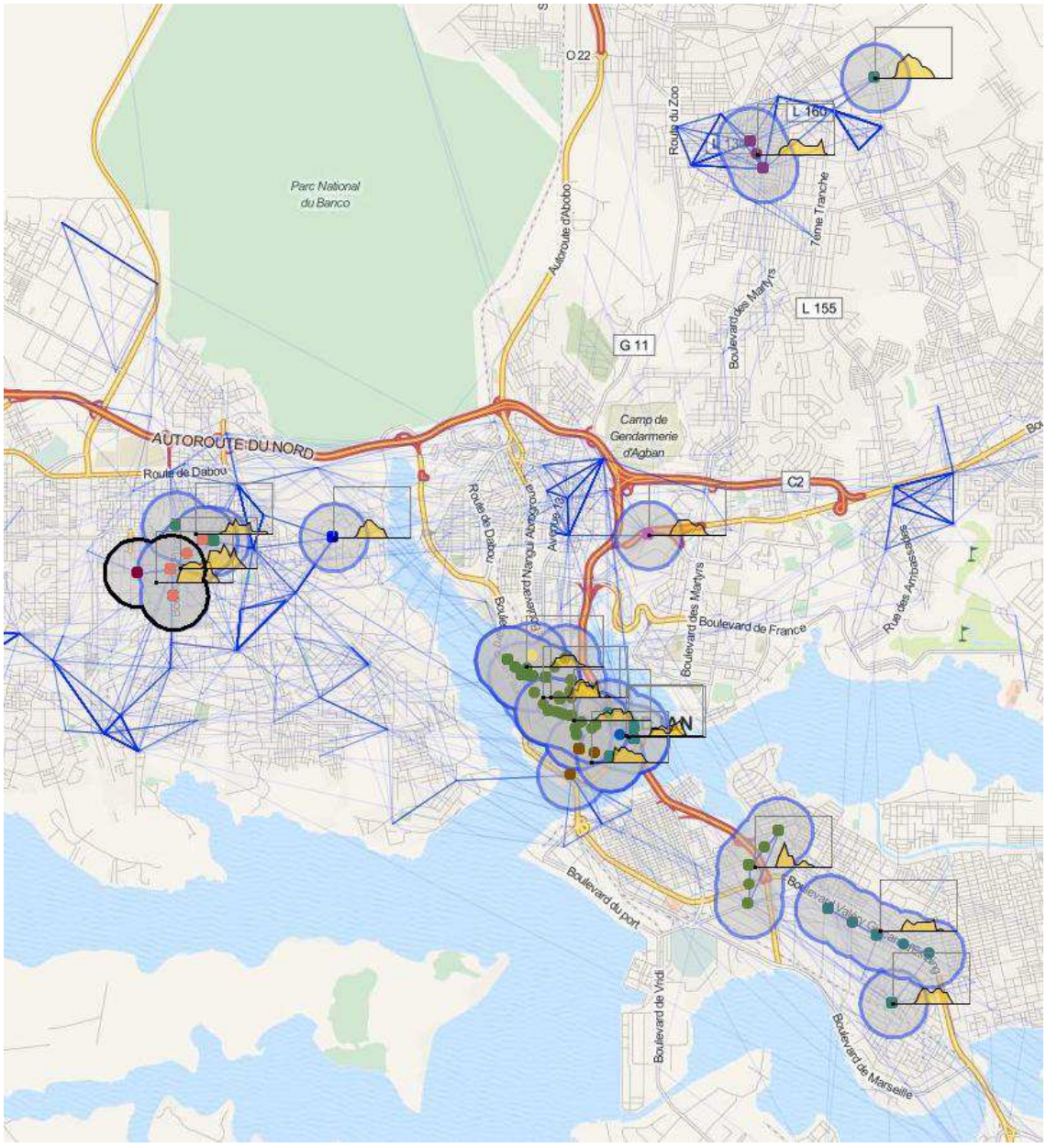


Figure 16. Locations from other trajectories characterized by temporal profiles similar to that was identified as work in Figure 15.

A better quality of semantic interpretation could be achieved if CDRs included times and positions of both the beginnings and ends of calls. In this case it becomes possible to distinguish stationary calls from calls on move, and to estimate movement speeds during the latter.

By applying the described procedure systematically to all subsets of the data and matching routine activity locations of persons in different subsets, it might further be possible to link partial trajectory corresponding to the same person in different data chunks (see Section 2). However, such re-integration may be harmful in terms of personal privacy [13].

9 CONCLUSIONS

In this paper, we report on analysis results of a medium-size set of call data records referring to antenna positions. The analysis was performed with V-Analytics – a research prototype integrating visual analytics techniques for spatial, temporal and spatio-temporal data that our group develops since the mid-90s [14]. We considered the data from multiple perspectives, including views on locations of varying resolution, time intervals of different length and hierarchical organization, and trajectories. We detected a number of interesting patterns that could facilitate a variety of applications, including

- Reconstructing demographic information (to replace expensive and difficult to organize census studies)
- Reconstructing patterns of mobility (to enhance transportation studies)
- Identifying places of important activities (for improving land use and infrastructure)
- Identifying events (for improving safety and security)
- Detecting social networks (for marketing purposes)

While in some cases we considered the complete data set, we had to restrict parts of our analysis to the last two weeks of the provided data due to undesirable properties (namely, missing, incomplete or duplicate data records for several key regions for a large portion of the time period). However, most of the applied techniques scale (or can be scaled up conceptually) for much larger data sets. Some kinds of analysis that we planned to perform were simply impossible due to the data fragmentation into chunks with duplicate user IDs. For example, we were not able to build predictive models of people's presence and mobility [15], as data for longer time periods are needed. We also did not search for interaction patterns between people and did not try to detect social networks.

Limitation caused by data quality could be relaxed by joining the provided data set with data from publicly available sources such as Flickr and Twitter in future work. Textual aggregates of activity records could greatly facilitate understanding and deeper semantic interpretation of the data.

10 REFERENCES

1. G. Andrienko, N. Andrienko, P. Bak, D. Keim, S. Kisilevich, S. Wrobel. A conceptual framework and taxonomy of techniques for analyzing movement. Journal of Visual Languages and Computing, 2011, 22(3), pp.213-232
2. T.Hägerstrand. What about people in regional science? Papers, Regional Science Association, 24, 1970, pp. 7-21.
3. G.Andrienko, N.Andrienko, M.Heurich. An event-based conceptual model for context-aware movement analysis. International Journal Geographical Information Science, 2011, 25(9), pp.1347-1370
4. V.Blondel, M.Esch, C.Chan, F.Clerot, P.Deville, E.Huens, F.Morlot, Z.Smoreda, C.Ziemlicki. Data for development: the D4D Orange challenge on mobile phone data. http://arxiv.org/abs/1210.0137
5. G.Andrienko, N.Andrienko, P.Bak, S.Bremm, D.Keim, T.von Landesberger, C.Pölitz, T.Schreck. A Framework for Using Self-Organizing Maps to Analyze Spatio-Temporal Patterns, Exemplified by Analysis of Mobile Phone Usage. Journal of Location Based Services, 2010, v.4 (3/4), pp. 200-221
6. J.W.Sammon. A nonlinear mapping for data structure analysis. IEEE Transactions on Computers, 1969, v.18, pp. 401–409
7. G.Andrienko, N.Andrienko, S.Bremm, T.Schreck, T.von Landesberger, P.Bak, D.Keim. Space-in-Time and Time-in-Space Self-Organizing Maps for Exploring Spatiotemporal Patterns. Computer Graphics Forum, 2010, v.29 (3), pp. 913-922
8. G.Andrienko, N.Andrienko, M.Mladenov, M.Mock, C. Pölitz. Discovering Bits of Place Histories from People's Activity Traces. In IEEE Visual Analytics Science and Technology (VAST 2010), Proceedings, IEEE Computer Society Press, pp.59-66
9. G.Andrienko, N.Andrienko, M.Mladenov, M.Mock, C. Pölitz. Identifying Place Histories from Activity Traces with an Eye to Parameter Impact. IEEE Transactions on Visualization and Computer Graphics, 2012, v.18 (5), pp.675-688
10. N.Andrienko, G.Andrienko. Spatial Generalization and Aggregation of Massive Movement Data. IEEE Transactions on Visualization and Computer Graphics, 2011, v.17 (2), pp.205-219
11. G.Andrienko, N.Andrienko, A.-M.Oteanu Raimond, J.Symanzik, DC.Ziemlicki. Towards Extracting Semantics from Movement Data by Visual Analytics Approaches. NetMob'2013 (submitted)
12. M.Ester, H.-P.Kriegel, J.Sander, X.Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. Second International Conference on Knowledge Discovery and Data Mining, Portland, Oregon, 1996, pp.226-231
13. G.Andrienko, N.Andrienko. Privacy Issues in Geospatial Visual Analytics. 8th Symposium on Location-Based Services (LBS 2011), Proceedings, Springer, pp.239-246
14. N.Andrienko, G.Andrienko. Visual Analytics of Movement: an Overview of Methods, Tools, and Procedures. Information Visualization, 2013, v.12(1), pp.3-24
15. N.Andrienko, G.Andrienko. A Visual Analytics Framework for Spatio-temporal Analysis and Modelling. Data Mining and Knowledge Discovery, 2013 (accepted). DOI:10.1007/s10618-012-0285-7

AllAboard: a system for exploring urban mobility and optimizing public transport using cellphone data

Michele Berlingerio Francesco Calabrese Giusy Di Lorenzo
Rahul Nair Fabio Pinelli Marco Luca Sbodio
IBM Research, Dublin, Ireland
{mberling,fcalabre,giusydil,rahul.nair,fabiopin,marco.sbodio}@ie.ibm.com

ABSTRACT

The deep penetration of mobile phones offers cities the ability to opportunistically monitor citizens' interactions and use data-driven insights to better plan and manage services. In this context, transit operators can leverage pervasive mobile sensing to better match observed demand for travel with their service offerings. With large scale data on mobility patterns, operators can move away from the costly and resource intensive four-step transportation planning processes prevalent in the West, to a more data-centric view, that places the instrumented user at the center of development. In this framework, using mobile phone data to perform transit analysis and optimization represents a new frontier with significant societal impact, as urban mobility services can be improved. In this paper, we present AllAboard, an interactive system to optimize the public transit network using mobile phone data with the goal to improve ridership and user satisfaction. The system is structured on a modular architecture, currently including three core modules: *Mobility Miner*, *Optimizer*, *User Classifier*; these are implemented as components within an extensible framework, and other components can easily be added in the future. Although our main contribution is the overall system, and the conjunction of the features of the different modules, we also show some experimental evaluation conducted on real data made available in the context of the Orange D4D Challenge. Our results show the power of mobile phone data based optimization of the transit network on real data, and open the way for further future analyses and applications.

1. INTRODUCTION

Coupled with rapid urbanization of the developing world, the deep penetration of the mobile phone offers cities the ability to opportunistically monitor citizens' interactions and use data-driven insights to better plan and manage services.

Within the transportation sector, operators can leverage opportunistic sensing to better match observed demand for travel with their service offerings. With large scale data on mobility patterns, operators can move away from the costly and resource intensive four-step transportation planning processes prevalent in the West, to a more data-centric view, that places the instrumented user at the center of development. In this framework, using mobile phone data to perform transit analysis and optimization represents a new frontier with significant societal impact, as urban mobility services can be improved.

In the context of developing countries, this kind of optimization based on user-centric crowd-sourced data can play an even more important role. From an urban mobility perspective, cities in sub-Saharan Africa, with the exception of Lagos, have seen organized public transport deteriorate over the past few decades [25]. The large entities have been inefficient to operators (cost per passenger-kilometer) and to users as well (fares). An ad-hoc fleet of informal operators have filled this void. In Abidjan, Ivory Coast, the 539 large buses of SOTRA, the public agency, are complimented by 5,000 mini-buses, and roughly 11,000 shared taxis. The mini-buses and taxis operate with little overhead and very little regulation. Kumar et al. [25] suggests that the sector is regulated in principle, but permits are issued upon request. While there may be a process through which routes are allocated [23], the mobility outcomes depending on the informal sector lead to several problems.

The informal urban transport sector is largely self-regulated. It is rigid in terms of routes and terminals with little regard for passenger demand. Mini-buses and other forms of group transport account for 50% of traffic along certain corridors, that could be better served with higher capacity vehicles that are expensive for small private operators to invest in. The informal nature of the sector also leads to a poor safety record, with African cities having the highest fatality rates globally [21]. The unregulated services also have a poor environmental record and low fuel efficiency.

If transit agencies could have an effective tool to quantify the travel demand, as well as recommendations on how to best design the transit network, cities would be able to better support travelers' mobility demand through a regulated and efficient public transport system.

Classical survey-based methods to estimate travel demand have been so far very limitedly used in developing countries, given their high costs. Accurately assessing travel demand through mobile phones offers key benefits and offers a true alternative. Given the high level of penetration, it provides samples that are several order of magnitudes larger that manual surveys. It could potentially be less biased by survey sampling limitations, embracing a wider classes of users that

.

are represented proportionally with their own statistical significance. It is faster than surveys, as mobile phone data can be gathered instantaneously. It allows for dynamic assessment of the travel demand, as the data follows a streaming paradigm, with the potential to make urban services more responsive. Lastly, as demonstrated in this paper, it offers the possibility of leveraging mobility patterns that are more complex than the simple traditional origin-destination (O/D) pairs, allowing for a finer grain modelling.

In this paper, we present an interactive system to optimize the public transit network using mobile phone data with the goal to improve ridership and user satisfaction.

The system consists of three modules that provide a richer characterization of urban movement and recommend optimal expansion of public transit offerings. First, methods to determine travel and activity patterns are developed that use call detail records to establish major activity locations of users and derive frequent sequences. Second, an optimization component employees the activity patterns and O/D flows to determine optimal design of new transit services. The use of frequent activity patterns in this manner represents an improvement over traditional methods purely based on O/D flows, since it allows for user travel preferences to be more accurately reflected in the optimal design. The model takes into account the entire urban network, as opposed to local methods which focus on one route at the time. The aim of the model is to optimize the user experience in terms of reduced travel and wait times. A third component determines how different categories of users, inferred from mobile phone data, are served by public transport. This component helps agencies in quantifying the impact of their services on different market segments and identify potential differences.

A live interactive online application[1] presents the results of each component visually. This allows transit operators, unfamiliar with the new mobile information, to visually explore travel patterns and performance measures of their system. Additionally, strategic planning of transit services requires community interaction and clear presentation of operator justification for service changes or additions. This fosters citizen engagement in the overall planning process and present transparency in decision making.

The main technical novelties provided by our innovative system are:

- it solves a large scale transit optimization, in terms of both number of O/D pairs and number of routes taken into account;
- the data it relies on is entirely crowd sourced, potentially taking into account all mobile phone users;
- the system is able to assess how different classes of users are server by public transport.

The remainder of the paper is organized as follows: Section 2 presents some literature on using mobile phone data for mobility analysis and optimization; Section 3 presents the available data, together with the needed pre-processing to feed our system; Section 4 presents the modular architecture of our AllAboard system; Section 5 presents the core module for mining mobility and trajectories; Section 6 describes the module for transit network optimization; in Section 7, the module for user classification and analysis is presented; we discuss the limitations of our approach in Section 8 and finally conclude in Section 9.

[1] `http://www.dublinked.ie/sandbox/d4d/`

2. MOBILE PHONE DATA FOR MOBILITY ANALYSIS AND OPTIMIZATION

There is an extensive research on modeling travel behavior. The field is dominated by disaggregate econometric models that determine the elasticity of travel demand to changes in supply conditions. Such models can be classified either as trip-based or activity-based models [15]. In trip-based models, mobility is considered only in as a function of socio-demographic characteristics of the traveler, the trip characteristics, and some aggregate measure of trip purpose (typically landuse associated with origin and destination). Activity-based models, in contrast, consider travel to be implied when users engage in a sequence of activities. The aim is to understand the nature, timing, and spatial distribution of choices that lead up to making travel decisions. While trip-based methods are the current state of practice, activity-based models are increasingly being employed by agencies for planning and forecasting [6].

A contrasting approach to identifying human activities and associated travel implications has been through the use of sensor networks. Using technologies such as GPS, RFIDs, WiFis and other electromagnetic sensors, research has focused on learning activities that individuals are engaged in over time and to determine relationships between users and their environments. Pioneering studies that focused on trip prediction and usage [24, 22], have given way to other applications (see, for example [18]). A very different individually-oriented approach, lacking even the need for a GPS unit, was elaborated in the 'Reality Mining' methodology [27, 14]. Their research uses opportunistically collected data from mobile phones in terms of the visible WiFi, Bluetooth, and cellular identifiers to generate a *spatial fingerprint* that is unique to a given place, and then identify recurrent daily patterns in individual behavior. A limitation of this analysis is that the authors only consider common places as 'home' and 'work' without taking other non-work related activities. Such research has reveal key insights in mobility patterns that can be counter intuitive. For example, individual travel behavior that can appear almost random, instead has been found to be repeating and predictable since identifiable routines can be classified [19, 29].

More recently, there has been a major shift away from end-user collection towards the use of massive mobile phone data sets collected by network operators and their partners; these have been shown to have great potential in several areas, from the real-time monitoring of traffic conditions [9] to the statistical modeling of human mobility and the definition of universal laws [19, 29]. These applications are also reinforced by the fact that mobility estimates based on mobile phone network data have been shown to be robust to the substantial biases in phone ownership [32].

In particular, in Song et al. [29] the authors consider the theoretical limits of the predictability of a mobile phone user's location based on their individual spatio-temporal patterns. The authors found that Zipf's law is a suitable approximation of an individual's cumulative place-frequency distribution, and that the probability of returning to a place is related to the number of times that that same location has been visited in the past. However, this descriptive approach

does not directly bind the findings to the underlying reason for moving, which is obviously going to be an important factor in a transportation model. [1] observe that temporal patterns found at hourly, daily, and weekly scales can be connected not only to specific activities, but also to their geographical distribution across the Tallinn metro region. For example, the further an employee is from an office in the Central Business District, the earlier their daily commute to work begins. The broad implication is that near randomness in individual trajectories of the sort that [4] documents for Hasan Elahi is quite rare, and that most of us follow a fairly regular routine that is structured by the interaction between space and time.

Moreover, there has been strong evidence to suggest that people value the ability to perform – and enjoy – unplanned activities [3]. As a result, some researchers have argued that we should move away from normative terms such as 'home' and 'work' towards the more flexible and indeterminate terminology of 'meaningful places' and 'anchor points' as a better way to think about location [2, 12, 13].

In this paper we want to study a city's individual mobility in order to derive patterns that describe different types of people and their daily mobility patterns. Those patterns include classical O/D flows as well as more complex sequential travels. For this purpose we use a large mobile phone location dataset to monitor human locations over the course a two week time interval. We then map human locations to geographical features of the visited places and use that to characterize the human mobility in terms of spatio-temporal patterns.

We leverage previous work that defined analytics to extract user's nominal home and work location, as well as trips [7]. We then extract frequent travel sequences [17] and evaluate whether frequent activity sequences can also be extracted.

3. THE AVAILABLE DATA

The D4D Orange challenge[2] made available data collected in Cote d'Ivoire over a five-month period, from December 2011 to April 2012. The original dataset contains 2.5 billion records, calls and text messages exchanged between 5 million users. The data released in the challenge contains samples from this original data and contains four separate datasets, each with information on different aspects at varying spatio-temporal resolutions.

This paper focuses on the one of the datasets that describe call activity of 50,000 users chosen randomly in every two week period. Specifically, the data contains the cell phone tower and a timestamp at which the user sent or received a text message (SMS) or a call. The dataset consists of tuples in the form `<UserID, Day, Time, Antenna>`. An axillary dataset that describes the spatial location of antennas was used to map each tuple spatially. From the four released datasets, this information provided the greatest spatial resolution, which was critical for the objectives of this paper. Antennas in the dataset span the entire Cote d'Ivoire.

3.1 Preprocessing

Since the work focuses on optimizing public transit, the data for the city of Abidjan is extracted. With 4.5 million inhabitants, it is the largest city in Cote d'Ivoire. A spatial bounding box defined by `(5.2089,-4.2557)` and `(5.4937,-3.7389)` is considered. Users with at least one reported activity within this area were selected. Figure 1 shows the number of distinct users observed during every two week period. The number of users over the five month period show a dip in March but are consistently around 18,000 distinct users.

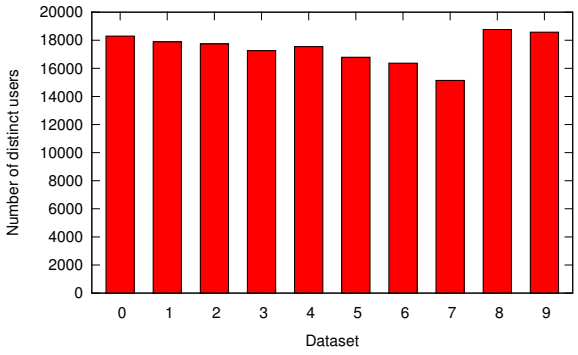


Figure 1: Number of distinct users for each two-week period

The dataset also contained antennas with obfuscated codes which were omitted from the analysis. Additionally, in some occasions multiple identifiers were used to reference the same location. These antennas were combined and treated as one. As a result, 407 antennas were selected within the bounding box for Abidjan.

To evaluate the quality of the dataset, the cumulative distribution of interleaving time between two calls (see Figure 2 (a)) and the cumulative distribution of the number of distinct visited towers (see Figure 2 (b)) were studied. From Figure 2(a) it can be stated that 50% of users make a call every ten minutes, highlighting that users are frequently tracked. However, Figure 2 shows that 50% of the users do not visit more than 6 cell towers on a period of 14 days. This demonstrates that while users are recorded often, their spatial movement is relatively static. Most of the data therefore describes static activity locations rather than mobility patterns.

Hourly profiles of call volumes, plotted in Figure 3, show both a regularly in call profiles (Figure 3(b)) and exceptional days of call activity (Figure 3(a)).

3.2 Public transport data

To model the existing public transportation network, information from SOTRA, the public transport agency, was gathered from their website[3]. Since the schedule information is not geo-referenced, additional data sources such as Open Street Maps (OSM)[4] and other geo-localization toolkits were employed to locate stops across Abidjan. From a set of 301 distinct bus stops, 203 were located in this manner. These stops represent 17 express bus routes, 67 regular bus routes (called the *monbus*), and 1 special route meant for trader with heavy goods (called *marche bus*). This represents the base line SOTRA transit network and is shown in Figure 4.

[2]http://www.d4d.orange.com/

[3]http://sotra.ci

[4]http://www.openstreetmap.org

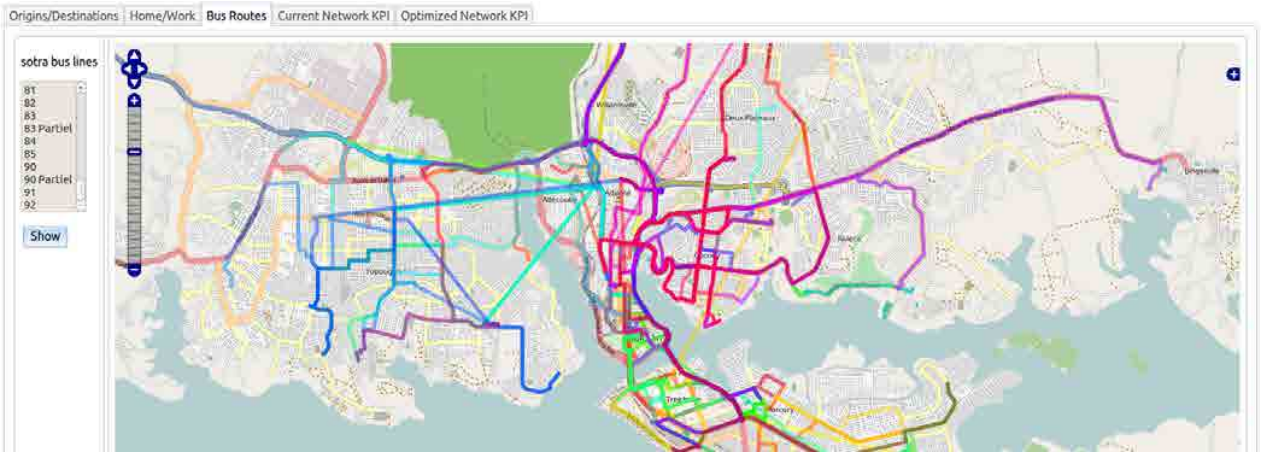


Figure 4: Baseline SOTRA transit network showing 85 routes

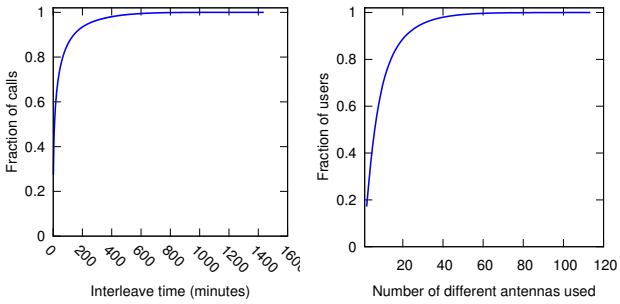


Figure 2: The figures show the cumulative distribution of the interleaving time between two calls (left) and the cumulative distribution of visited cells by user for the two weeks period (right).

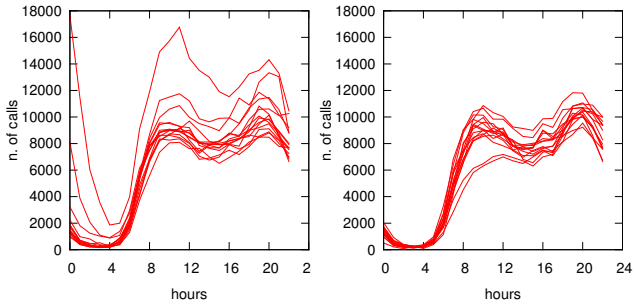


Figure 3: Hourly distribution of number of calls for each day in two different two-week periods

4. THE ALLABOARD PLATFORM

We implemented our system using a modular architecture (see Figure 5). We isolate our data models using an abstract layer that separates algorithm implementations from data stores. The current version of the system includes three core modules: *Mobility Miner*, *Optimizer*, *User Classifier*; these are implemented as components within an extensible framework, and other components can easily be added in the future. Each component provides a lightweight REST service exposing its functionalities. The REST services are also used to implement the AJAX-based Web user interface. The main algorithms for the current three core are described in Sections 5, 7 and 6 respectively.

5. MOBILITY MINING

In this section we present the core concepts implemented in the mobility moning module within AllAboard. The module is able to process mobile phone data and extract information about user stops, trajectories, O/D matrices, flows, frequent sequences, and most likely home and work locations for each user.

In order to achieve this, we need to understand what kind of information can be extracted from the available data, and how to use it.

The methodology to process the data follows the following steps:

1. we extract the location of the stops performed by users
2. based on those, we estimate the O/D matrix flows, used to feed the optimization module presented in Section 6
3. we exploit the data to better understand the mobility of the users, and use the results as additional input to the optimization module. For this, we extract frequent sequential patterns from the sequences of stops.
4. we identify for each user, when possible, most likely home and work location based on both stops and mobile phone activity patterns.
5. we use the above information to assign a semantic label to the extracted frequent sequential patterns in order to characterize the detected sequences.

We now present each step in details.

5.1 Stop detection

As described in Section 3, the available data presents some limitations in terms of both sample rate (i.e., user locations are sparsely distributed over time), and in terms of the coarse spatial resolution (i.e. the localization based on

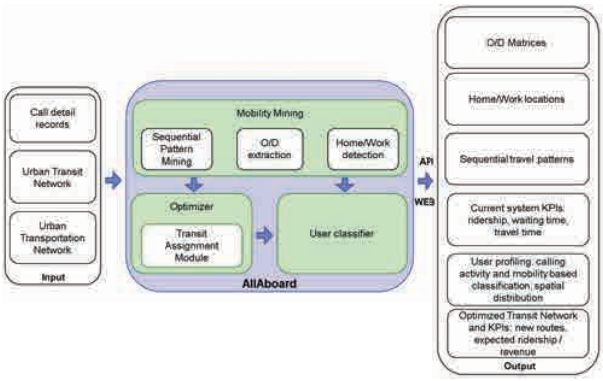


Figure 5: Architecture of the AllAboard platform

mobile phone antennas has a considerable spatial error). For these reasons, it is difficult to capture fine grain movements, but yet it is possible activity patterns suitable for urban planning purposes.

Thus, the focus of this module and the next one on the concept of stops more than movements. In order to describe the entire process let us first introduce some concepts:

DEFINITION 1 (HISTORICAL ACTIVITY). *The historical activity H of a user is defined as an ordered sequence of timestamped locations extracted from mobile phone activity $H = \langle a_1, a_2 \dots a_n \rangle$ where $a_i = (x, y, t)$ and x, y are spatial coordinates and t is a timestamp.*

Hereafter, we consider antenna and location as exchangeable. Note that each antenna information $a_i \in H$ corresponds to a call made by the user when the mobile phone was connected to the antenna a_i, and that no information on the duration of the call is present.

DEFINITION 2 (STOP). *Given the historical activity of a user H, a spatial threshold th_s and a temporal threshold th_t, a stop is the maximal subsequence $s = \langle a_m \dots a_k \rangle$ s.t. $0 < m \leq k \leq n \wedge \forall_{m \leq i \leq j \leq k} max(Dist(a_i, a_j)) \leq th_s \wedge Dur(a_m, a_k) \geq th_t$.*

Therefore, a stop is detected if the user is moving within a region defined by a radius equal to the spatial threshold th_s, and within he time described by the temporal threshold th_t. Note also that, in our definition, we take into account the maximum of the distances among the different antennas. This allows us to handle possible localization errors (e.g.: a user stops but the mobile phone changes antenna in a reasonable short time, thus creating fictitious movements). The procedure to extract the stops for a certain user is described in Algorithm 5.1.

The procedure takes in input the historical mobility of a user, $\mathcal{A}$, the spatial threshold th_s and the temporal threshold th_t. The antennas are buffered until the maximum distance between each pair of antennas belonging to the buffer is not greater than the th_s threshold. If this happens, then the temporal distance between the first antenna and current antenna is checked against the temporal threshold th_t. If we fall above it, then the set of antennas is considered as a stop and the antenna representing the mode is returned as representative stop. Otherwise, the buffer is cleared and the maximum distance check reset, since we consider the user in movement. Note that computing the centroid of the buffer to detect the stop is worthless in this case because of the low spatial accuracy of the data under analysis.

Algorithm 1 Stop_detection ($\mathcal{H}$, th_s, th_t)

Require: Historical mobility $\mathcal{A}$ of the user u, spatial threshold th_s, and temporal threshold th_t
Ensure: Set of stops $\mathcal{S}$
$Buffer = \emptyset$
$max_{distance} = 0$
for all a in A **do**
  for all b in $Buffer$ **do**
    if $distance(a, b) > max_{distance}$ **then**
      $max_{distance} \leftarrow distance(a, b)$
    end if
  end for
  if $max_{distance} > th_s$ **then**
    if $a.time - first(Buffer).time > th_t$ **then**
      $\mathcal{S} \leftarrow mode(Buffer)$
    else
      $Buffer = \emptyset$
      $max_{distance} = 0$
    end if
  end if
  $Buffer \leftarrow a$
end for
if $Buffer \neq \emptyset$ **then**
  $\mathcal{S} \leftarrow mode(Buffer)$
end if
return $\mathcal{S}$

The procedure then returns the set of stops $\mathcal{S}$ performed by a user.

In our experiments, we adopted 1 km as spatial threshold, and 1 hour as temporal threshold, values demonstrated to be realistic in [7].

5.2 O/D Matrix estimation

A first insight regarding the mobility within a city can be captured by extracting the O/D matrix, and it represents the first input for the optimization module. In several works, such as [16] and [7], the authors proposed methodologies to estimate the O/D matrix using GPS data as well as CDR data. We apply a similar methodology in order to estimate the flows between each pair of origin and destination antennas within a certain temporal interval using the set of stops extracted during the previous step. All the antennas can be either origin and destination. Time is divided into 24 hourly intervals. In order to define the evaluated flow between the antenna O and the antenna D, we first introduce the concept of trip. A trip $trip(u, O, D, t)$ is the path between two consecutive stops O and D where $O \neq D$. A trip is characterized by the user u, the stop D as starting point of the trip, the stop D as destination of the trip, and the time t corresponding to the time associated to O. Therefore, the estimated flow between a pair of antenna O, D, in a time interval h, is defined as:

$$flow(O, D, h) = \sum_{u \in U} trip(u, O, D, t) \quad (1)$$

where the $t \in h$.

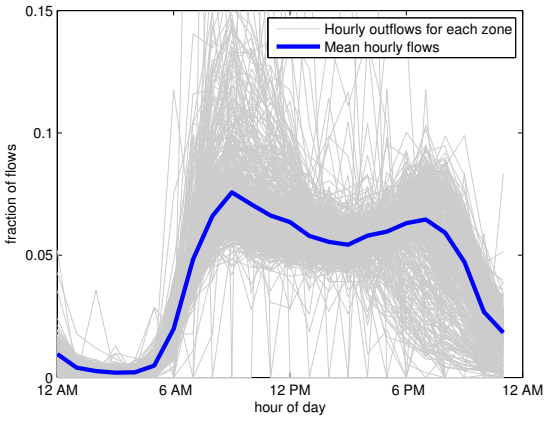


Figure 7: The figure describes, in grey, the hourly out-flows considering each antenna as origin, and, in blue, the mean hourly flow.

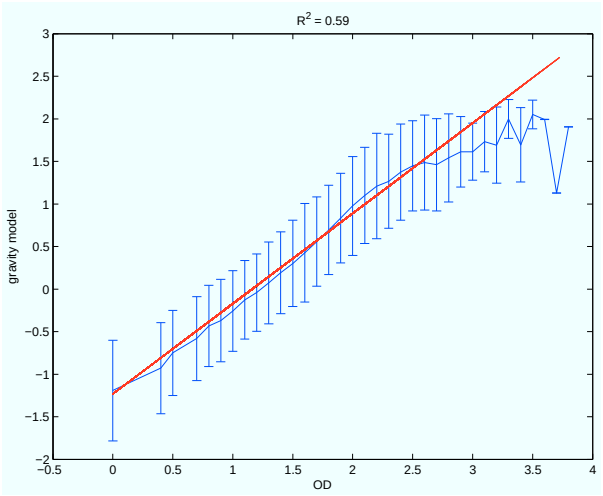


Figure 8: The plot shows the correlation between the flows of each O/D pair and the respective gravity model.

Thus, we sum all the trips of all the users having O as origin and D as destination and they fall in the time interval h.

The hourly distribution of the flows considering each antenna as origin is reported in Figure 7, together with the mean hourly flows. As we can notice, the mean flow follows the typical two peaks trend that describes the overall mobility pattern of cities [28]. An example of the visualization of the outgoing flows of a selected antenna together with its temporal profile is represented in Figure 6.

Furthermore, we validate the process of the estimation of O/D flows through a comparison of our results and the gravity model defined as for each pair of antennas O, D

$$G(O,D) = \frac{O_{out} * D_{in}}{distance(O,D)^2} \quad (2)$$

where with O_{out} represents the sum of all the flows going out from the node O and D_in indicates the sum of all the flows ending in D. We report the results in Figure 8. As we see, there is a linear correlation between the flows detected by our method and the gravity model indicating that the mobility flows in Abidjan follow the similar behavior related to distance as already seen in other cities [7].

5.3 Sequence travel patterns

Transit optimizer usually take into account the flows described by an O/D matrix. The CDR data provide us a finer granularity of the mobility of the users, so that we can take into account also longer sequences of visited cells, in particular the most visited ones among the different users.

To do this, we first defined the daily activity sequence of a user as the concatenation of the different trips performed by a user in a single day.

The goal of this task is to generate additional information to enrich the one that is contained in the O/D matrix, thus we take into account only sequences with at least 3 different locations. Moreover, we discarded sequences containing loops, because we consider more relevant, for the scope of our paper, to analyze only sequences which do not contain the same location twice. More explicitly, we are not interested to extract the patter $Home \rightarrow Work \rightarrow Home$ but patterns that include also a third different location.

Advanced spatio-temporal data mining techniques have been proposed in order to extract spatio-temporal knowledge from mobile objects trajectory datasets as [17], [31] and [10]. Unfortunately, the quality of the data under analysis is too low to apply those techniques. Moreover, our main goal is to present a whole modular system, and the stress is on the overall application of mobile phone data for urban transit optimization, rather than to introduce new analytical methods.

Thus, we focus on which knowledge can be extracted from user-generated mobility data in order to better optimize the transit network of a city.

In this task we adopted a traditional sequential pattern mining algorithm [26] which requires in input a threshold of minimum support min_{sup} in order to extract all the sequences with a frequency higher than min_{sup}. The algorithm has been executed on 199767 sequences longer than 2 stops, and shorter then 12, and with a min_{sup} equal to 0.2%, thus obtaining 30 sequences of length 3 (i.e., there were no frequent sequences longer than 3 stops).

We tuned the min_{sup} parameter to 0.2% due to the fact that the presence of a given user id is not guaranteed across different periods, thus every user has virtually only 14 days to support a given sequence. In other terms, there is only 14 days of mobility data for every single user, and the frequency threshold must be kept low, if we consider that the entire period covers 5 months.

However, these sequences present interesting properties as shown in Figure 9(a) and 9 (b). On the left, we report on the x-axis the frequency of the sequences and on the y-axis the number of users performing that sequence. As we can easily see, the sequences are performed by several users. The same consideration can be done if, instead, we consider the days in which a certain sequence is performed, as reported in 9(b). We can affirm that the detected frequent sequences describe mobility behaviors along different days in this case as well. For these reasons, we consider this set of sequences interesting and usable from the optimization

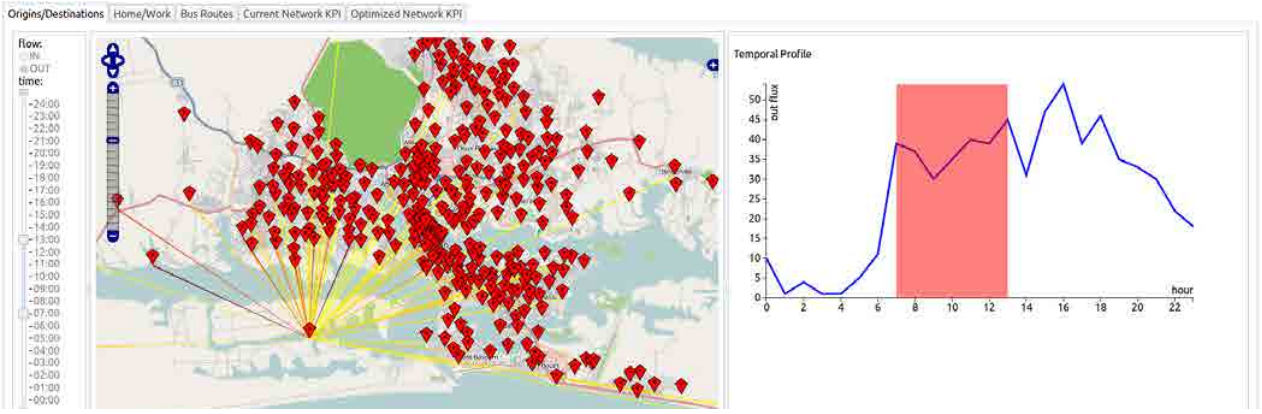


Figure 6: An example of the outgoing flows from a selected antenna (left) and the corresponding temporal profile (right).

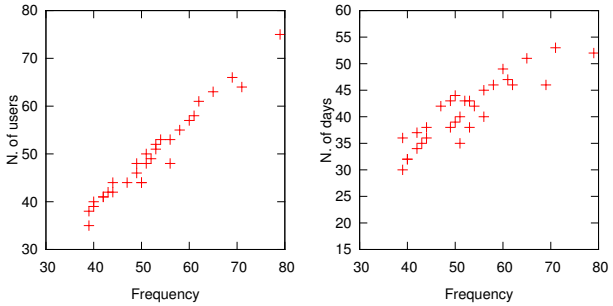


Figure 9: The figures show the relationship between the frequency of a sequence and the user traveling along that sequence (left) and the frequency with the number of days when the sequence was performed.

module since they are able to describe mobility patterns shared by different users and conducted on several days.

5.4 Home and work detection

For each user, we extracted the most likely home and work locations by analyzing the visited locations over the two weeks period.

We followed the method described in [7] to select the home cell, since the criteria proved to be effective. In particular, to detect a home location, for each visited cell, we evaluate the number of nights the user connect to the network in the night-time interval [5] while in that cell, and select as home location the cell with the greatest value. Discarding the home cell, we select as work location the cell with greatest value in the day-time interval (complementary to the night-time). Of course this method to determine home and work locations might result in very inaccurate estimations in case users do not make many calls, or make calls only in a very limited number of days. For this reasons, we defined three accuracy indexes:

- repetitiveness of home cell: number of nights when the home cell is the most visited at night-time, divided by the number of nights where the user connects to the network (ranging from 0 to 1) [8]
- repetitiveness of work cell: number of days when the work cell is the most visited at day-time, divided by the number of days where the user connects to the network (ranging from 0 to 1)
- number of days where the user connects to the network (ranging from 1 to 14)

Figure 11 shows the cumulative distribution of the indexes. To be conservative, we assume that the home and work location estimation is reliable if the three accuracy indexes are respectively greater than 0.5, 0.5 and 7, resulting in a subset of 11020 users. The visualization of the heat maps for the detected home and work locations is depicted in Figure 10.

5.5 Frequent sequence labeling

In this section we try to characterize the 30 extracted frequent sequences of length 3.

For each user associated to a sequence, we label the visited stops as home (H), work (W) or other (O). We then extract the 3-stops sequence and associate them to a string of type 'HWO'. For instance strings of type 'OOO' represent the case when all 3 locations in the sequence do not correspond to either home or work cell. Interestingly, for the 30 extracted frequent sequences, in 80% of cases, both home and work locations are includes for all of the users. This means that the frequent sequences are predominantly associated to commuting. Given this result, we focus our attention on characterizing users' activity sequences. Considering all frequent sequences in an aggregate manner, we can discover that the cases in which the 'Other' location is visited along the commuting route (either the HW or WH direction) is not the most frequent case (see Figure 12). This shows that majority of the discovered frequent sequences are associated to direct commuting followed (or preceded) by visiting another location.

6. OPTIMAL TRANSIT DESIGN

[5] We used 6p.m. to 7.a.m. as the night-time interval based on statistics from `http://news.abidjan.net/h/404484.html`

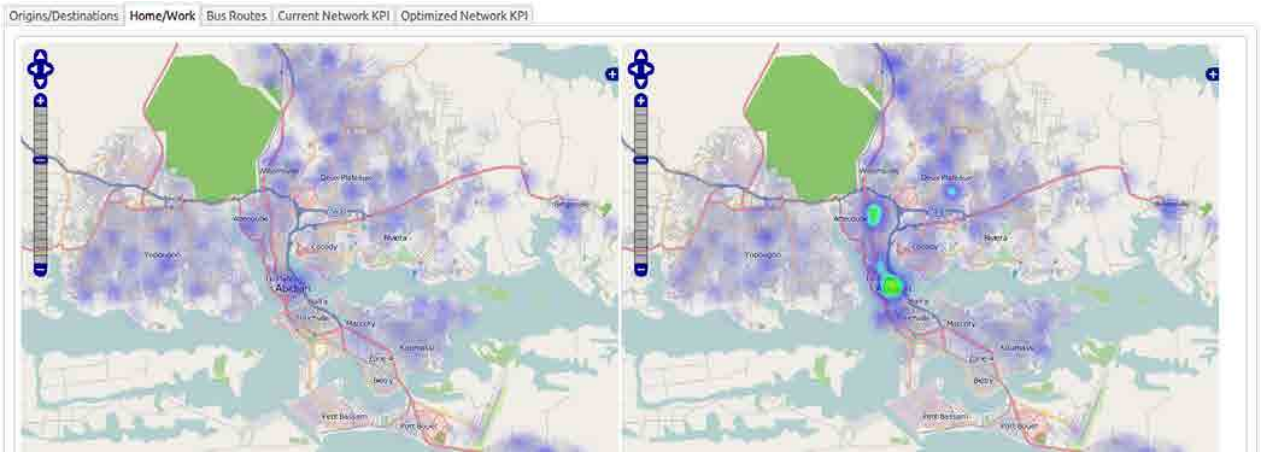


Figure 10: Density of users, mapped by home (left) and work (right) locations.

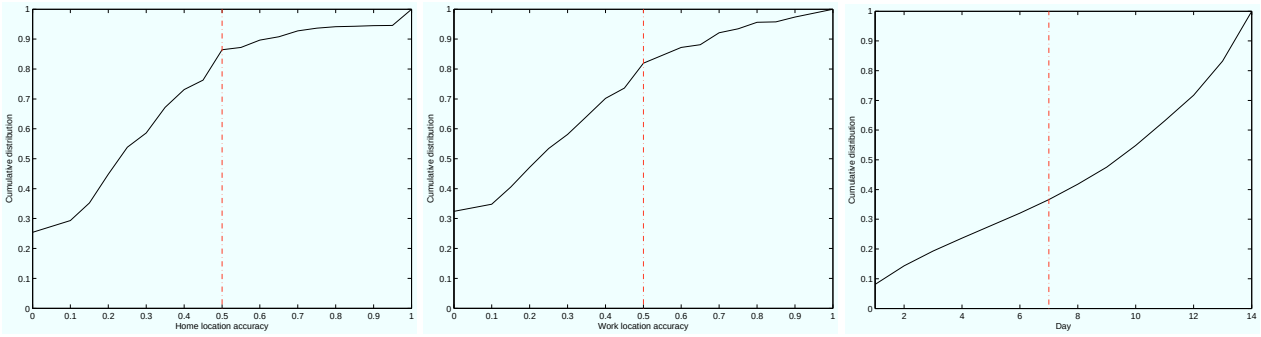


Figure 11: Home and work location accuracy indexes

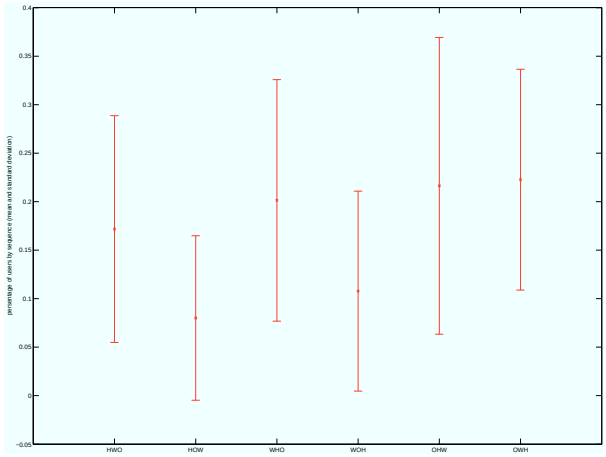


Figure 12: Sequence labeling results.

Given (a) an existing transit network, (b) OD flows derived from mobile data representing travel demand, (c) a set of frequent sequences that serve as candidate new routes, (d) travel time estimates across the network, and (d) a resource budget, in terms of fleet size, we seek to determine an optimal set of new routes and their associated service frequencies, such that passenger journey times city-wide are minimized. A new route is defined by a sequence of stops and has an associated frequency during peak periods.

The problem is strategic in nature as it represents a longer-term decision on the part of a public transport operator. The addition of new routes to the service network are intended to match current supply with revealed demand. From a demand perspective, new routes will attract riders only if they offer competitive service to existing routes. The model therefore considers the user perspective, in terms of optimal strategies, a concept discussed in greater detail below. From a supply side, new routes should fill service gaps and map user activity patterns observed in the data. The optimization routine therefore includes frequent sequences as potential new services. Taken together, a potential new route is only recommended if it directly addresses under-serviced demand, and does so by offering shorter journey times than the ones possible on the existing network.

The proposed model takes the form of a multi-commodity flow problem, with several key differences. As shown in Figure 13, flows over a transit network are complicated due to waiting processes at stops. Transit network links represent both temporal and spatial characteristics. For planning applications, such as this one, an aggregate measure of frequency is typically employed. Services having different frequencies also induce different flows, since users are likely to take services that are more frequent.

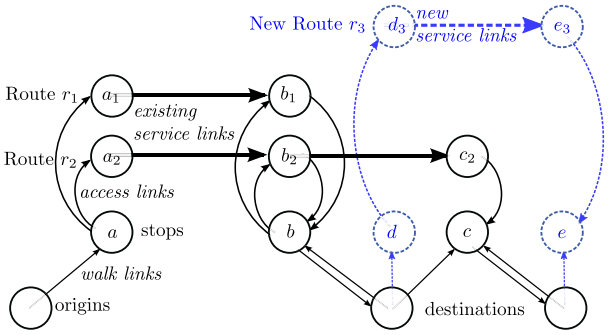


Figure 13: Network model of transit system

User behavior in transit systems involves added complexity of waiting processes and the frequency of routes. Further, users are likely to dynamically change travel paths, depending on which service arrives at a particular stop. Instead of shortest path, users are therefore assumed to follow *optimal strategies* [30]. The process of assigning users to paths along a network, termed transit assignment in the transportation literature, therefore reflects user behavior and allocates flows based on frequency.

Readers are referred to reviews by [11] and [20] on transit network design, that show the prior work, which is omitted for brevity. The proposed model builds on the *optimal strategies* work by [30], that model how transit users are likely to board services. In addition to accounting for the user perspective, the model also seeks to determine frequencies on new routes as shown in Figure 13. Since the network is assumed to be uncongested and links are uncapacitated, there is no travel time penalty for users, should the link flow be very high. The model is therefore both system and user optimal. The candidate set of routes is determined by the frequent sequence patterns determined from the mobile data (see Section 5.3).

The transit network is defined by a set of nodes V (indexed by i) and arcs A (indexed by (i,j)). Passenger demands are denoted by a set of origin-destination flows denoted by K (indexed by k) with the mobile phone data yielding a sampling of the true demand $s(d_k)$. Each od pair k is associated with a node pair (i,j). A subset of nodes, denoted by V', represent nodes at which waiting occurs. Each edge emanating from nodes in the set V' to the service expanded network has an associated frequency parameter q_{ij} and are called *access* links. All links have a travel time of c_{ij}. For access and egress links from the service expanded network $c_{ij} = 0$, and users incur a wait time at the head node i.

The transit network also includes a set of candidate routes R (indexed by r). Let the set NA denote all access links that serve candidate routes. For each route r we define an indicator variable δ_{ijr} that denotes if the route uses link (i,j) and the quantity t_r denotes the round trip time needed to service the route in hours. $\bar{C}$ indicates the maximum vehicles available to service the network. Let g_{ik} denote the following

$$g_{ik} = \begin{cases} s(d_k) & \text{if } i \text{ is origin of } k \\ -s(d_k) & \text{if } i \text{ is destination of } k \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

The decision variables are f_r, $r \in R$ that denotes the design frequency on route r, v_{ijk} denotes the flow on (i,j) for od-pair k, and w_{ik}, $i \in V'$ represent the waiting time at node i associated with od pair k. The optimal transit network design problem can be stated as

$$\min \sum_{k \in K} \left[\sum_{(i,j) \in A} c_{ij} v_{ijk} + \sum_{i \in V} w_{ik} \right] \tag{4}$$

subject to,

$$\sum_{j:(i,j) \in A} v_{ijk} - \sum_{j:(j,i) \in A} v_{ijk} = g_{ik} \forall i \in V, k \in K \tag{5}$$

$$v_{ijk} - q_{ij} w_{ik} \leq 0 \qquad \forall i \in V', \; k \in K \tag{6}$$

$$v_{ijk} - \sum_{r \in R} \delta_{ijr} f_r w_{ik} \leq 0 \qquad \forall (i,j) \in NA, \; k \in K \tag{7}$$

$$\sum_{r \in R} \frac{t_r}{60} f_r \leq \bar{C} \tag{8}$$

$$0 \leq f_r \leq u_r \qquad \forall r \in R \tag{9}$$

$$v_{ijk},\, w_{ik},\, f_r. \in \mathbb{R}^+ \tag{10}$$

The objective (4) minimizes system wide journey times for users which includes travel and waiting time. Along with the flow conservation constraints in (5), constraints (6) allocate flows on existing services based on frequency of service. Constraints (7) are non-linear and relate flows for new services and their respective frequencies. The amount of services introduced are bounded by the fleet size constraint (8) and by upper bounds on frequency (9).

6.1 Separable Approximation

For small and medium sized networks, the non-linear set of constraints in (6) which involve a product of two decision variables can be transformed to a separable function which can then be approximated by a piecewise linear function. The result program introduces new variables but remains linear since the functions are convex.

Define

$$y_{ik} = \frac{1}{2} \left(f_r + w_{ik} \right), \forall i | (i,j) \in NA, \tag{11}$$

and

$$z_{ik} = \frac{1}{2} \left(f_r - w_{ik} \right), \forall i | (i,j) \in NA, \tag{12}$$

and r is the frequency associated with the new service access link in the set NA. Using this transformation, (7) yields

$$v_{ijk} \leq y_{ik}^2 - z_{ik}^2 \qquad \forall i | (i,j) \in NA, k \in K. \tag{13}$$

Replacing the decision variables f with $(y+z)$ and w with $(y-z)$ and approximating the square terms in (13) yields a large scale linear program with $(2|NA| + |A| + |V'|)|K| + |R|$ variables.

For large scale networks, this approach may not be feasible, even though the model is linear. Column generation approaches, similar to the ones proposed in the literature [5] or heuristic approaches can be pursued. For this paper, a heuristic procedure is developed that first picks a set of routes from major origin-destination flows by using the separable programming approximation. Then, the optimal

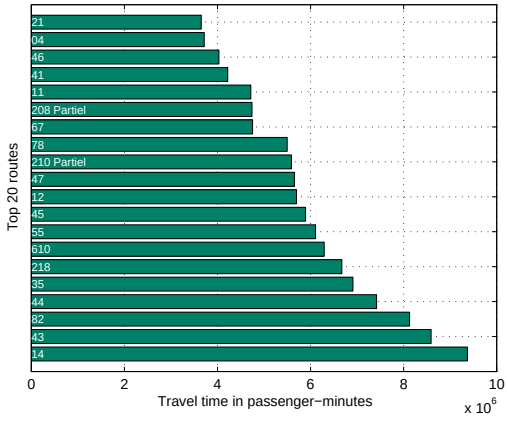


Figure 14: Top 20 routes in Abidjan computed by aggregating for all segments of a route, a product of passenger flows and travel time for each segment

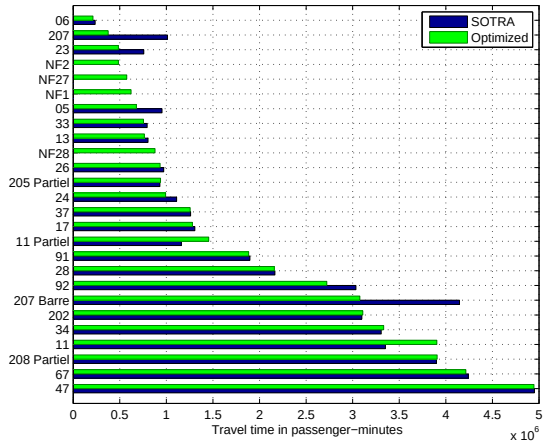


Figure 15: Comparison of route loads before and after the optimization for routes with significant changes

strategies flows for all OD pairs are determined using the new set of routes. The model is implemented using OPL and executed using IBM ILOG CPLEX v12.4 for the city of Abidjan using the existing SOTRA network described in Section 3 and the candidate routes described in Section 5.3.

6.2 Results for Abidjan

Results from two cases representing the before and after of system improvements are presented. In the first case, the only the existing SOTRA network as described in 3 is considered. The second case, the SOTRA network is improved with a fleet increase of 65 routes. To selected the candidate new routes, we have chosen 35 randomly selected shortest paths among already used stops (as it's usually done in these optimization models). Moreover, we have taken the 30 frequent sequences from Section 5.3 and converted them in candidate routes among the visited antennas.

The optimal alignment of new routes and their associated frequencies are determined using the model presented above. The model recommends four additional routes with service headways of 12 minutes.

The resulting flows along the SOTRA network is shown in Figure 16(top).

Expected user flows in passenger minutes can be aggregated by route for each service. The top 20 routes for Abidjan are shown in Figure 14. The wait times at stops are shown in Figure 16(middle).

The new proposed routes and optimized flows along the extended network (SOTRA + new routes) are shown in Figure 16(bottom). The new routes (in red) represent Feeder services that allow improving the overall ridership. Indeed, in the optimized network, system-wide travel times were found to reduce by 10.2%, while waiting time systemwide increased by 2%, on account of new stops that were introduced. While there was no change in travel patterns in areas not impacted by the new services, they did have an impact on passenger flows on 22 existing routes. The difference between existing and optimized services in passenger-minutes is shown in Figure 15. Many services see a decrease in travel times, as the newly introduced services (NF) take up some of the flows. Interestingly, all four new selected routes belong to the frequent sequences extracted in Section 5.3. This is also explained by the fact that the extracted sequences are relatively short and cover high demand ODs.

7. USER CLASSIFICATION

Within our vision of a system where the end user perspective is really taken into account, there are a few questions that arise. Are there specific classes of users served better by the transit system? How do we characterize those users? Can we exploit the information contained in the mobile phone data to address questions of market segmentation?

User classification is then important, from one side, to ensure that the user perspective is not biased by the intensity of mobile phone activity for each specific users, and, from the other side, to assess the performances of the transit network w.r.t the different defined classes of users.

We developed a module of the system that performs the following tasks:

- it calculates, for each users, and for each day of mobile phone activity, the number of calls performed and the total distance traveled;
- it classifies the users on the basis of two attributes, namely the intensity of the mobile phone activity, and the intensity of the mobility activity;
- it takes the KPIs of the transit network from the module described in Section 6, and it joins them with the classes of users;
- it displays the distributions of the KPIs for each different class of users.

We now describe in details how the classes are generated, how the KPIs are merged with the classes, and the results of these two steps on the data analyzed.

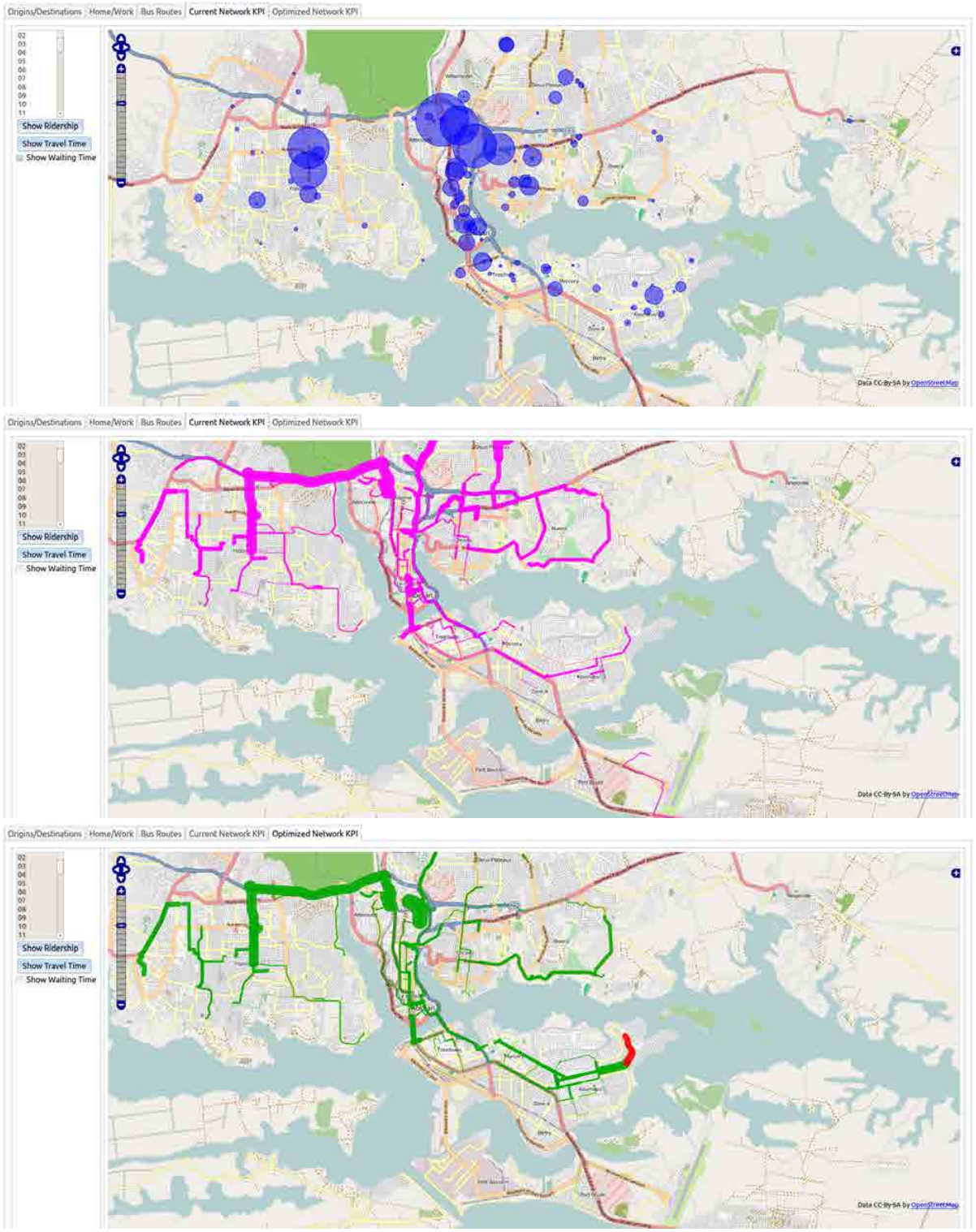


Figure 16: Current SOTRA network. Top: radius of circles is proportional to waiting times at stops. Middle: line width is proportional to the ridership of the line. Bottom: SOTRA network and additional routes (line width is proportional to the ridership of the line).

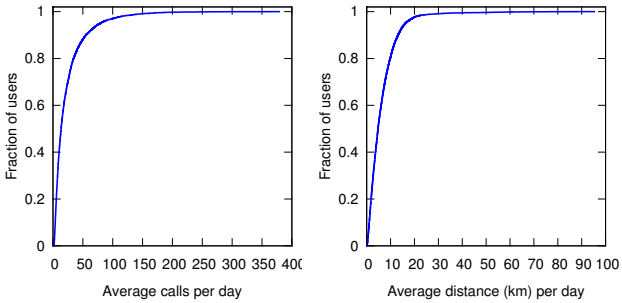


Figure 17: Cumulative distributions of the average calls and distance (in km) per day.

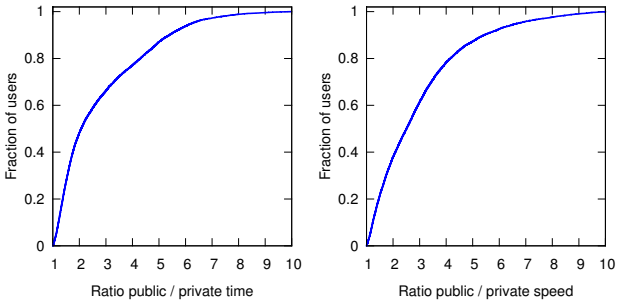


Figure 18: Cumulative distributions of the ratios.

	Mobility	Mobile
Low	5510	5311
High	5510	5709
Total	11020	11020

Table 1: Number of users per class

7.1 User classes

We took the mobile phone call dataset, consisting of several lines of the following form:

```
UserID  Day  Time  Antenna
```

and, for each user, we computed, day by day, the number of calls performed, and the total distance traveled. While the first was taken straightforwardly from the above data, we relied on the stops calculated from the module described in Section 5 to calculate the distance. In particular, after aggregating by day, we computed, for each user, the total distance (in meters) traveled per each day of mobile phone activity[6].

Then, for each user, we computed two numbers: average number of calls placed per day, and average distance traveled (per day of mobile phone activity). We then aggregated by users, and looked at the cumulative distributions of those two functions, reported in Figure 17

For each of these functions, we then divided the population in two classes of intensity, namely low and high, by finding the median of the distributions, i.e., said $F(x)$ the cumulative distribution function of x, we computed m such that:

$$\int_{-\infty}^{m} \mathrm{d}F(x) \geq \frac{1}{2} \text{ and } \int_{m}^{\infty} \mathrm{d}F(x) \geq \frac{1}{2}$$

and assigned each user to the 'low' class whether her/his averages were below or equal than m, 'high' otherwise. Table 1 reports the division of users into classes[7].

The module is then able to label each user with two flags: low/high intensity for mobile phone activity, and low/high intensity for mobility activity. We later refer to them as mobile or mobility profiles or classes.

[6]Due to our strategy of relying on mobile phone data, we cannot assess the mobility of a user in a day without placed calls

[7]The imbalance of the Mobile classes is due to users having values corresponding to the median

7.2 KPIs on different classes

The module described in Section 6 produces, among its output, the KPIs for each O/D pair. In particular, it produces several lines of the form:

```
O D dist_priv dist_pub time_priv time_pub
```

where: 'dist_priv' is the distance calculated on the road network[8] if a user avails of private transportation means; 'dist_pub' is the distance (including walking and public transport) needed by a user of the public transport analyzed to go from O to D; 'time_priv' is the travel time calculated on the road network as above; 'time_pub' is the total travel time (including waiting, walking, and public transport) needed by a user of the public transport to go from O to D.

When designing or assessing a public transport system, in order to ensure the behavioral switch of the travelers from private to shared transport, it is important to understand whether it will be competitive over the alternatives provided by the private or on-demand alternatives. To this extent, we can use the KPIs above to measure the ratios between the total travel time needed to go from O to D by public transport, and the total travel time needed to move by private means. A similar ratio may be computed also on the average speed, but for sake of simplicity here we present the results on the total travel time ratio only, which seems to be more important from the end user perspective.

Thus, from the input above, the module computes, for each O/D pair:

$$time\_ratio(O, D) = \frac{time\_pub}{time\_priv}$$

which can take values in $]0, \infty[$. As we are only interested in seeing how much the users are penalized when using public transport, and in order to discard outliers due to exceptional configurations of the road network and of particular O/D pairs, we restrict our focus on the values in the interval $[1, 10]$, with 1 meaning that there is no difference in the total travel time between traveling by public or private transport, and 10 meaning a user pays 10 times the travel time of the private means when traveling on a bus. Figure 18 reports the cumulative distribution of the time ratio (left) and the speed ratio (right).

By joining the above results on the *time_ratio* and the information about home and work locations computed by the module presented in Section 6, we can assign a computed

[8]Via querying the APIs of a popular online service for travel directions

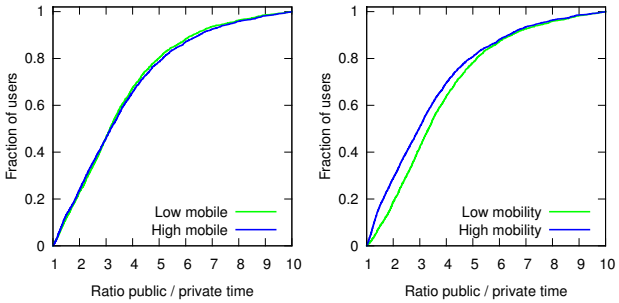


Figure 19: Cumulative distributions of the ratios, by user class.

time_ratio to each user. Then, we can aggregate the results by class of users, and compute the distribution of the ratio by class. Figure 19 reports the cumulative distribution of the ratios for the mobile classes (left) or mobility classes (right).

Two important things may be seen in these figures. First, the ratios are not significantly biased by the amount of mobile phone calls we had in the data for each user. This seems to suggest the fact that the mobility and transport analyses are not affected by the amount of mobile phone calls produced by a user. Second, we see differences on the ratios by mobility classes. In particular, the cumulative distribution of the ratios for the users with a high mobility profile shows that, for any given ratio of the population, users with low mobility profile are more penalized by the public transport. Note that low and high mobility profiles are based on the average distance covered per day, not on the average distance per trip, i.e. we do not distinguish between users performing few long trips, from users performing many short trips. However, these plots seem to follow the intuition reported for instance in [8], where the authors showed by empirical analysis that there is a correlation between users living in areas well served by public transport and their high daily mobility.

7.3 Applications

The methodology provided by this module is powerful and extensible, by, for example, taking into account other KPIs, or classifying users by different categorical attributes. Being the nature of this process rather general purpose, there are a number of applications of the obtained results.

First, the results of the optimization step may be assessed against different classes of users, to evaluate the potential market for a line. This can also be coupled by analyzing, or actuating, on the alternative: dynamic road congestion charges, as well as appropriate bridge or tunnel tolls, may be evaluated on the basis of the potential of the alternative provided by the public transport.

Second, in the line of this, it is possible also to assess the potential user discrimination of a line. Although we have not found specific evidence of this in the available data, it may be possible to discover that specific classes of users are served better by a given line, or in a specific temporal interval during the day, or in a given area of the city. These indicators may be also joined with other socio-economic variables that can help in this direction.

Third, different classes of users may be used to identify different key indexes of the so called "willingness to pay" for the performances of the transport service. For example, it is possible to correlate classes of users with their natural predisposition to wait longer for a connecting bus, or to wait longer in specific areas, and so on.

These are just a few example applications of the results obtained in this analysis, and while we described the working methodology behind this module, we leave for further research and case studies the full exploitation of the results obtained in this section.

8. DISCUSSION

In this paper we have presented a system to optimize the transit network based on mobile phone network data, starting from an existing system. Thus, results of the optimization will depend on the quality of the mobile phone network data in terms of representing real mobility demand, and on having an accurate representation of the available transit network.

Regarding the mobile phone data, we decided to use the Dataset 2 from the Orange D4D Challenge. This allowed extracting mobility information for 500,000 users in a 5 month period, with cell tower location accuracy. For the mobile phone data, a crucial factor to take into account is the localization error, which could limit the minimum size of the spatial units that can be considered and lead to errors in statistical analysis. In this study, the mobile phone data from Orange have been localized associating users to the coordinate of the antenna they are connecting to during a call. This thus introduces localization errors which can be even of several kilometers for rural areas where antennas' density is low.

Other elements that can affect the statistical results include: (1) the market share of the mobile phone operator from which the dataset is obtained; (2) the potential non-randomness of the mobile phone users; (3) calling plans which can limit the number of samples acquired at each hour or day; and (4) number of devices that each person carries. A recent study however shows how mobility estimates based on mobile phone network data was robust to the substantial biases in phone ownership [32].

Moreover, due to the fact that the considered dataset is event-driven (location measurements available only when the device makes network connections) the connection patterns of users are affecting the possibility to capture more or less trips. This could be a particularly important bias, as for each user we can only monitor his/her location during a 2 week period. A larger monitoring period would have allowed improving the accuracy of the extracted mobility patterns (ODs, and travels) as they would have been less biased by the event-based nature of the location information (users making calls). The system we built is able to handle larger datasets that could allow limiting these biases.

Nonetheless, the analyses performed on the inter-event time and the spatial distribution of mobile phone users seem to confirm results shown in related studies [7] that: mobile phone users have fairly frequent calling activities so that their location changes can be tracked through mobile phone traces. Therefore, mobile phone traces represent a reasonable proxy for individual mobility, whose quality and representativeness could be further improved in the future as the penetration rate of smart phones keeps increasing.

Regarding the transit network information, we recognize

that accurate SOTRA transit route and schedule information was not available. We then decided to leverage all available Web information to extract reasonable bus stop location as well as route shape information. Unfortunately we were not able to fully validate the extracted transit network information. We hope this could be achieved in the near future with the help of the local authorities, and potentially with citizen engagement. From the point of view of the designed system, transit network information is considered as an input, and obtained results in terms of KPIs and optimized recommendations for improvements could be re-evaluated in case more accurate inputs become available.

9. FUTURE WORK

AllAboard is a platform with many practical implications. First, by leveraging crowd sourced data, it allows for assessing the market opportunities of the current system, and new lines suggested by the optimization. Second, being able to classify users according to different criteria, it is possible to use the results to assess market segments for each new line. When seen in conjunction, these two features open the opportunity for supporting decisions on several aspects of mobility policy, including better fares, dynamic road or congestion pricing, multi-modal data collection, and user incentives to encourage behavioral switch.

References

[1] R. Ahas, A. Aasa, S. Silm, and M. Tiru. Daily rhythms of suburban commuters' movements in the Tallinn metropolitan area: Case study with mobile positioning data. *Transportation Research Part C*, 18(1):45–54, 2010.

[2] R. Ahas, S. Silm, O. Järv, E. Saluveer, and M. Tiru. Using Mobile Positioning Data to Model Locations Meaningful to Users of Mobile Phones. *Journal of Urban Technology*, 17(1):3–27, 2010.

[3] K. Axhausen and T. Gärling. Activity-based approaches to travel analysis: conceptual frameworks, models, and research problems. *Transport Reviews*, 12(4):323–341, 1992.

[4] A. L. Barabási. *Bursts: The Hidden Pattern Behind Everything We Do*. Dutton, 2010.

[5] R. Borndörfer, M. Grötschel, and M. E. Pfetsch. A column-generation approach to line planning in public transport. *Transportation Science*, 41(1):123–132, 2007.

[6] J. L. Bowman and M. E. Ben-Akiva. Activity-based disaggregate travel demand model system with activity schedules. *Transportation Research Part A: Policy and Practice*, 35(1):1 – 28, 2001.

[7] F. Calabrese, G. Di Lorenzo, L. Liu, and C. Ratti. Estimating origin-destination flows using mobile phone location data. *Pervasive Computing, IEEE*, 10(4):36 –44, april 2011.

[8] F. Calabrese, M. Diao, G. D. Lorenzo, J. F. Jr., and C. Ratti. Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. *Transportation Research Part C: Emerging Technologies*, 26(0):301 – 313, 2013.

[9] F. Calabrese, C. Ratti, M. Colonna, P. Lovisolo, and D. Parata. A system for real-time monitoring of urban mobility: a case study in rome. *IEEE Transactions on Intelligent Transportation Systems*, 12(1):141–151, 2011.

[10] H. Cao, N. Mamoulis, and D. W. Cheung. Mining frequent spatio-temporal sequential patterns. In *ICDM*, pages 82–89, 2005.

[11] G. Desaulniers and M. Hickman. Public transit. *Handbooks in operations research and management science*, 14:69–128, 2007.

[12] G. Di Lorenzo and F. Calabrese. Identifying human spatio-temporal activity patterns from mobile-phone traces. In *Intelligent Transportation Systems (ITSC), 2011 14th International IEEE Conference on*, pages 1069 –1074, oct. 2011.

[13] G. DiLorenzo, J. Reades, and F. Calabrese. Predicting personal mobility with individual and group travel histories. *Environment and Planning B: Planning and Design*, 39(5):838–857, 2012.

[14] N. Eagle and A. (Sandy) Pentland. Reality mining: sensing complex social systems. *Personal Ubiquitous Comput.*, 10(4):255–268, 2006.

[15] M. Gendreau and P. Marcotte. *Transportation and Network Analysis: Current Trends*. Springer, 2002.

[16] F. Giannotti, M. Nanni, D. Pedreschi, F. Pinelli, C. Renso, S. Rinzivillo, and R. Trasarti. Unveiling the complexity of human mobility by querying and mining massive trajectory data. *The VLDB Journal*, 20(5):695–719, Oct. 2011.

[17] F. Giannotti, M. Nanni, F. Pinelli, and D. Pedreschi. Trajectory pattern mining. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '07, pages 330–339, New York, NY, USA, 2007. ACM.

[18] Y. Gong and R. Mackett. Visualizing Children's Walking Behaviour Using Portable Global Positioning (GPS) Units and Activity Monitors. In L. H. and M. Batty, editors, *Virtual Geographic Environments*, pages 295–310. Science Press, 2009.

[19] M. Gonzalez, C. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[20] V. Guihaire and J. Hao. Transit network design and scheduling: A global review. *Transportation Research Part A: Policy and Practice*, 42(10):1251–1273, 2008.

[21] K. Gwilliam. *Africa's Transport Infrastructure*. World Bank Publications, 2011.

[22] E. H. John Krumm. Locadio: Inferring motion and location from wi-fi signal strengths. In *MobiQuitous*, pages 4–13, 2004.

[23] E. Kouakou. *Public Transport in Sub-Saharan Africa: Major Trends and Case Studies*. UITP Ű International Association of Public Transport, 2010.

[24] J. Krumm. Real time destination prediction based on efficient routes. In *Society of Automotive Engineers (SAE) 2006 World Congress*, 2006.

[25] A. Kumar and F. Barrett. Stuck in traffic: Urban transport in africa. *AICD, Background Paper, World Bank, Washington, DC*, 2008.

[26] J. Pei, J. Han, B. Mortazavi-asl, H. Pinto, Q. Chen, U. Dayal, and M. chun Hsu. Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth. In *Proceedings of the 17th International Conference on Data Engineering*, pages 215–224, 2001.

[27] A. Pentland. Eigenbehaviors: identifying structure in routine. *Behavioral Ecology and Sociobiology*, 63:1057–1066(10), May 2009.

[28] D. L. Schrank and T. J. Lomax. *The 2013 urban mobility report.* Texas Transportation Institute, Texas A & M University, 2013.

[29] C. Song, Z. Qu, N. Blumm, and A.-L. Barabasi. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, February 2010.

[30] H. Spiess and M. Florian. Optimal strategies: A new assignment model for transit networks. *Transportation Research Part B: Methodological*, 23(2):83–102, 1989.

[31] R. Trasarti, F. Pinelli, M. Nanni, and F. Giannotti. Mining mobility user profiles for car pooling. In *KDD*, pages 1190–1198, 2011.

[32] A. Wesolowski, N. Eagle, A. M. Noor, R. W. Snow, and C. O. Buckee. The impact of biases in mobile phone ownership on estimates of human mobility. *J. R. Soc. Interface*, 10(81), 2013.

Mobility Modeling for Transport Efficiency

- Analysis of Travel Characteristics Based on Mobile Phone Data

Vangelis Angelakis
vangelis.angelakis@liu.se
David Gundlegård
david.gundlegard@liu.se*
Botond Rajna
botra718@student.liu.se
Clas Rydergren
clas.rydergren@liu.se
Katerina Vrotsou
katerina.vrotsou@liu.se

Linköping University,
Department of Science and Technology

Richard Carlsson
richard.carlsson@ericsson.com
Julien Forgeat
julien.forgeat@ericsson.com
Tracy H. Hu
tracy.h.hu@ericsson.com
Evan L. Liu
evan.l.liu@ericsson.com
Simon Moritz
simon.moritz@ericsson.com
Sky Zhao
sky.zhao@ericsson.com
Yaotian Zheng
yaotian.zheng@ericsson.com

Ericsson Research,
Services & Software

***Abstract* — Signaling data from cellular networks can provide a means of analyzing the efficiency of a deployed transportation system and assisting in the formulation of transport models to predict its future use. An approach based on this type of data can be especially appealing for transportation systems that need massive expansions, since it has the added benefit that no specialized equipment or installations are required, hence it can be very cost efficient.**

Within this context in this paper we describe how such obtained data can be processed and used in order to act as enabler for traditional transportation analysis models. We outline a layered, modular architectural framework that encompasses the entire process and present results from initial analysis of mobile phone call data in the context of mobility, transport and transport infrastructure. We finally introduce the Mobility Analytics Platform, developed by Ericsson Research, tailored for mobility analysis, and discuss techniques for analyzing transport supply and demand, and give indication on how cell phone data can be used directly to analyze the status and use of the current transport infrastructure.

The project presented in this paper a part of the D4D challenge.

Keywords — mobility, mobile phone call data, transportation, travel demand

I. INTRODUCTION

Transport infrastructure has globally been identified to have a significant correlation with the economic growth in a country. Investments in transport infrastructure have been identified to have a positive effect on the economic growth, and with a positive economic growth, investments in public infrastructure are often made. Expansions of the transport infrastructure can be made in different areas of a country and can include several different types of transports (roads, airports, railways, etc.). Since investments of this type are very costly, it is therefore of crucial importance to make careful analysis of the cost-benefit-ratio for the potential investments. Models for cost-benefit-analysis require background information on land use as well as current and future demand for mobility.

The use of cellular network signaling data has the potential to fundamentally change how we can estimate transport models, analyze the efficiency of a current transportation system, and predict its future use. By mapping the cell phone data to the transport infrastructure it will be possible to estimate the current use of the transport system. From the results of such estimations, suggestions for improvements to the existing transport system can be generated. The outcome would be more efficient mobility and, in the long run, increased economic growth.

Furthermore, in developing countries where the cellular networks can provide a much better coverage than traditional sensor infrastructure for traffic and transport, this type of data will be very important to generate decision support information for key infrastructure investments. Decisions taken today on infrastructure development and urban planning can lock cities into mobility behavior patterns for the next 30 to 50 years.

A. Aim and key paper outcome

The aim of this paper is to investigate the potential of mobile phone call data in the context of mobility, transport and transport infrastructure. In the paper, we describe how mobile phone data can be processed in order to be applied in traditional transportation analysis models. We present the Mobility Analytics Platform developed by Ericsson Research, tailored for mobility analysis, and present techniques for analyzing transport demand, analyzing the transport supply, and how cell phone data can be used directly to analyze the status and use of the current transport infrastructure. The analysis will exemplify how mobile phone call data can be used to generate better decision support for infrastructure

* Corresponding author: David Gundlegård, david.gundlegard@liu.se, Linköping University, 601 74 Norrköping, Sweden.

investments, identify bottlenecks of existing infrastructure as well as estimate and predict the state of the transportation network.

Analyses presented in this paper are based on the mobile phone data from the country of Côte d'Ivoire, presented in detail in Blondel et al. (2012).

Unlike fixed infrastructure systems for data collection, cell phone signaling data is not bounded by any transport mode or any specific spatial region. This makes it possible to analyze the travel demand and travel times. The key outcomes of the paper are a set of transport indicators that can be measured using the present type of mobile phone usage data. Based on these indicators, we present use cases for the case of Côte d'Ivoire, applicable also to other regions where the same type of data is available.

The project presented in this paper a part of the D4D challenge.

B. Methodology

In our work, we have developed a generic, flexible, and customizable architecture to extract human mobility knowledge from geo-tagged network data. By implementing this architecture many of the different types of location data can be analyzed while always maintaining the necessary privacy and integrity of the users. We have also defined interfaces that make it easy to integrate different types of applications and services to our system.

Our architecture is based on 3 separated layers which are each well isolated from each other. The separation of the 3 layers ensures modularity enabling the system to be customized according to the type of analysis that needs to be done or the type of data that needs to be collected. It also ensures that privacy sensitive information can be separated from non-privacy sensitive information. We call the layers in our architecture the Data Collection, Analytics, and Knowledge Exposure Layers, see Figure 1.

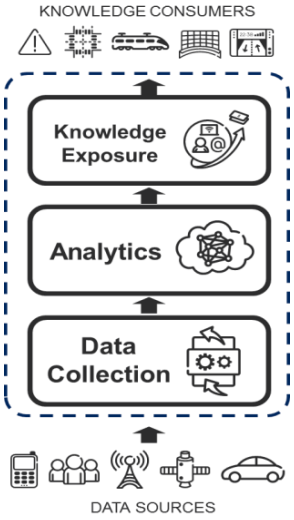


Figure 1: A description of the 3 layer generic architecture for collecting, analyzing, and exposing valuable mobility information from mobile networks.

The main responsibility of the Data Collection layer is to receive different types of geographically-tagged data and transform this raw data to a uniform data format before it is analyzed by the analytics engine of the system. The main functions of the Data Collection Layer are:

1. Data retrieval: The system must be able to retrieve the raw data from the data providing node. From some data sources access to the data will be implemented through a continuous stream of data.
2. Data preparation: every data source has its proprietary data format and the Data Collection Layer needs to be able to convert it to a uniform format used by the analytics engine.

The output of the Data Collection Layer is a set of records which each denotes the recorded location of a user in the network.

Depending on the knowledge that needs to be derived from the data the Analytics Layer needs to implement one or more analytical methods. We call these the Analytics Engines of the system. The Analytics Engines have three major functional requirements: analytical accuracy, processing efficiency and privacy preservation. To make it possible for a wide range of applications and services to make use of the knowledge about human travel patterns derived from the mobile network data a generic method to make the information available is needed. This is handled by the Knowledge Exposure Layer.

C. Outline

The rest of the paper is organized as follows. In Section II the background to the studied case, Côte d'Ivoire, is presented both in term of the current transport infrastructure and the mobile phone data used. Three ways of describing the travel demand is discussed. In Section III, results from numerical analyses of key indicators for the transport infrastructure, based on mobile phone data, is presented. Section IV describes the main functionalities and features of the Mobility Analytics Platform developed by Ericsson Research and in Section V use cases and conclusions are presented.

II. BACKGROUND

A. Côte d'Ivoire

Côte d'Ivoire is located in the west of Africa and has about 19 million inhabitants and has a size which is about 80% of the size of Sweden. The capital of the country is Yamoussoukro, relatively close to the middle of the country, see Figure 2. The city with the largest number of inhabitants is the city of Abidjan. Abidjan is located in the south part of the country.

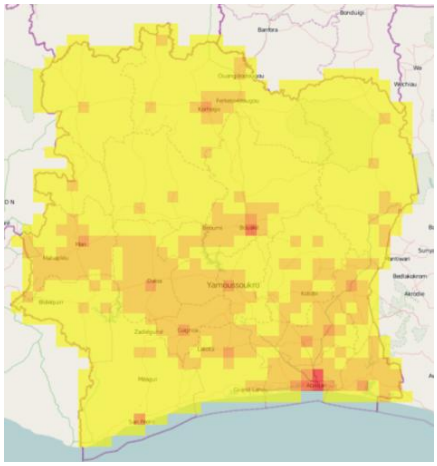


Figure 2: The population density of Côte d'Ivoire (Afripop, 2010).

An overview of the road infrastructure, as presented in the Open Street Map, is presented in Figure 3. The country has about 80.000 km of roads where only 8% of the length of the network is paved roads (CIA World Factbook, 2006). For comparison, Sweden has about 100.000 km of public roads and about 80% of the length of the public road network is paved (Trafikverket, 2012).

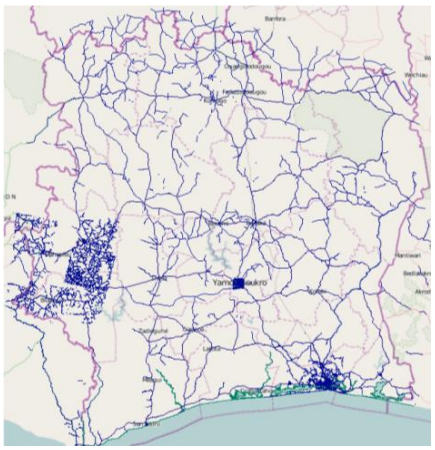


Figure 3: The road and rail infrastructure as presented in the Open Street Map data (www.openstreetmap.org).

B. Mobile phone data for mobility analysis

The mobile phone data for Côte d'Ivoire originate from the mobile operator Orange (Blondel et al. 2012) and consist of 150 days of cell phone data for research purposes.

The data was collected during a 150 days (3600 hours) period between 1st of December, 2011 until 28th of April, 2012. Out of the 3600 hours of data there are 100 hours missing due to technical failure. This time period covers 2.5 billion calls and SMS exchanges. The data includes timestamps, caller IDs, call durations, and antenna codes, for each data exchange. Positions of the calls are identified according to the connected antenna. The position of each antenna is given as the longitude and latitude, slightly blurred due to being sensitive information. The antennas are depicted in Figure 4.

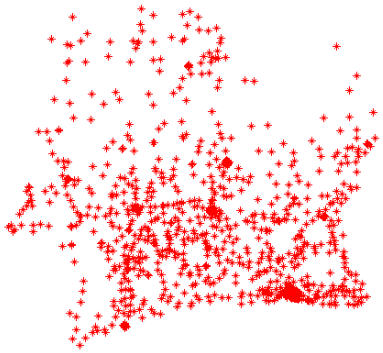


Figure 4: Antenna positions in Côte d'Ivoire.

The data consist of three sets of phone use data. In this paper we focus on one of the sets, the user trajectories with high spatial resolution data (referred to as Set 2 in Blondel et al. 2012). This set has several subsets where each subset is a user trajectory table, where the positions of the connected antennas are described for 50.000 users during a two week period. These two week periods are consecutive for the whole period, however from period to period, the anonymous IDs change, therefore a single user cannot be tracked for a period longer than 2 weeks. There are ten two-weeks periods altogether. Time stamps are rounded to the minute.

Based on this set, it is possible to visualize the movement of specific users. In Figure 5, users which have connected to more than 100 different antennas during a two week period are shown.

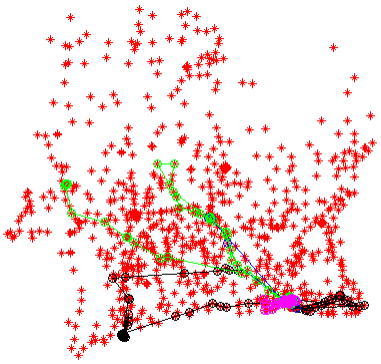


Figure 5: Trajectories for users with more than 100 different antennas registered during two weeks.

It should also be mentioned that this dataset only contain data from call data records, which is a subset of the mobility data that is available in cellular networks. Other types of data that can be collected from the cellular network, e.g., location updates, handover events or measurement reports, would affect the results of the analysis. A detailed description of the data available in cellular networks for the purpose of traffic

management and planning can be found in Gundlegård and Karlsson, 2006.

C. Previous work

During recent years groups from both academia, including MIT Sensible City Lab and Pisa KDD Laboratory and commercial institutes, including IBM Smarter Planet Initiative and Microsoft Research, have started developing techniques for using different types of data from communication networks to understand the interactions among people and their behavior. Example of methods developed include knowledge such as the distance a user travels, important places in people's lives (Isaacman et al. 2011), prediction of user location based on trajectories (Monreale et al. 2009), estimating and predicting urban traffic (Zheng et al. 2011), and methods for understanding group behavior (Zheng et al. 2009).

There has been a number of publicly funded large research initiatives for mobility analysis, such as GeoPKDD (GeoPKDD, 2005) and DATA SIM (DATA SIM, 2011).

Travel time estimation based on mobile phone data has been studied before, see for example Karhumäki, 2002 and Gundlegård and Karlsson, 2009; however, most of the work has been focused on using more detailed location data from the cellular network, e.g. handover events. In Caceres et al., 2012 the correlation between number of calls and the actual traffic flow is investigated. Several techniques were tested to find a best fitting model with help of mobile phone data and traffic counts.

The majority of the previous research work has been focusing on the data mining aspects and not the creation of practical tools for making use of the knowledge derived from the networks. Although many of the research initiatives have made significant findings in the areas of data mining and performing advanced analytics on mobility data there is a lack of any kind of open service platform or other tools for enabling developers to create innovative applications based on the analysis.

III. Analytics Layer and Results

The main focus of the paper is to investigate different options for the analytics layer and how these can be applied for use cases in the context of decision support related to the transportation network and infrastructure.

For our initial analysis and experiments we have used the V-Analytics framework developed by (Andrienko and Andrienko, 2011) and for producing travel flow representations. The bulk of the remaining results presented have been produced using the Mobility Analytics Platform, developed by Ericsson Research, which we present in Section IV.

A. Travel demand

The travel demand is one essential input to models for transportation analysis. The travel demand is normally described in an origin-destination matrix. Given a division of a geographical area and a division of the area into zones, the origin-destination matrix describe the number of trips from each pair of zones, e.g. from zone A to zone B for each pair (A, B). The origin-destination matrix describe the demand in a given time interval. Normally, the origin-destination matrix describes the number of trips that starts at zone A during the specified time interval, going to zone B. Such a origin-destination matrix is called static, since the matrix includes one number valid for the whole time interval, see example in Figure 6.

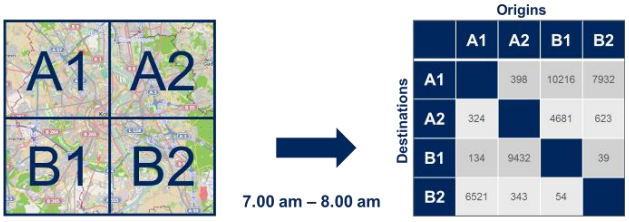


Figure 6: An example of an OD matrix that has been created based on the division of a geographical area into zones. The number in each cell in the matrix represents the number of trips that started between 7 am and 8 am and originated in the zone on the x-axis and ended in the zone on the y-axis.

A key difference with cellular network data and many other data sources when it comes to estimating traffic demand, is that it is possible observe the state directly. In many other sources, such as traffic flow measurements, we are relying on models to convert from the observation to the state of interest. In case of traffic flow measurements, we need to make a traffic assignment to go from travel demand to observed flows, which makes the estimation process much more complicated and time consuming.

In this subsection, we define three types of representations of the mobility and the travel demand.

Travel demand description type A – Static

For the static description of OD travel demand, both the time and the zones are set for the analysis. The originating or terminating zone of the user is calculated by its most common position during the predefined time period.

The predefined periods of time are, for example, set to two time periods, where one can be associated with the "home" zone, and may, for example, last from 10pm to 7am. The other can be associated with the "work" zone and last from 9am to 4pm.

Clustering methods were used to determine origin and destination zones, which can improve the visibility of the travel patterns. In Figure 7, an example is shown for k-means clustering based on antenna locations in Abidjan. However, by including more information in the clustering, like antenna usage, population data and land use it might be possible to reveal more information about which different zone types there are in a city. In the example, five zones were created with k-mean clustering method, based on antenna positions only.

Movements between zones are calculated, and movements within a zone are not taken into account, as this type of analysis does not include a technique to differentiate short trips from non-movements.

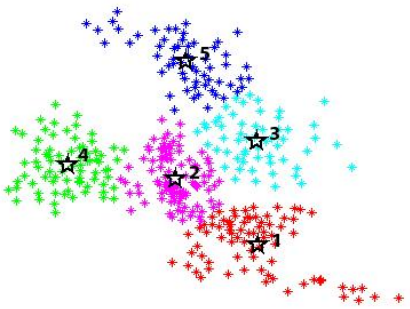


Figure 7: Example of k-means clustering using antenna locations to determine origin and destination zones.

Table 1 shows the OD-matrix for the case where the "home" period was chosen to be between the 7th of December 10pm, and 8th of December 7am and the "work" period from 9am till 4pm on the 8th of December. Table 1 gives the identified movements from "home" to "work", between the zones. It can be observed that most of the traffic goes into zone number 2. In comparison to other zones, 56% of all registered users move to zone 2. The highest mobility is between zone 4 and zone 2, which covers 23% of the traffic between the five zones.

Table 1: OD matrix for a single day between time periods 10pm-7am and 9am-4pm

		Destination				
		1	2	3	4	5
Origin	1	0	63	23	7	8
	2	13	0	39	59	35
	3	17	135	0	18	54
	4	11	259	31	0	25
	5	14	155	92	30	0

The advantage of generating OD matrices based on this definition is that the zones and the timespan of the periods is predetermined and can be set as the user chooses. This is important if the results shall be fused with other travel demand indicators. On the other hand, the time periods have to be decided in advance and applied to the whole data set, even though the daily movement pattern may not be the same for each user.

Travel demand description type B – Dynamic

Another way of defining the OD matrices is to use a dynamic definition of the times when a user is considered as being stationary in order to try to and capture as many trips made by the users as possible.

This definition requires a technique for determining the locations where a user has been stationary. We then count the trips that have been made between these locations and aggregate these into an OD matrix. The zones can be generated in the same way as for type A, described earlier, that is, the geographical area is divided into a number of sub-areas (zones) for which we have data. Each OD matrix consists of an aggregation of all trips made between each zone during, for example, 1 hour.

The method of determining when a user has been stationary is described in detail in (Mellegard, et al. 2011). The algorithm selects the stationary locations based on 4 parameters; two spatial and two temporal. The spatial parameters define the longest distance a point can be from the previous point to be classified as stationary (Trip begin threshold) for a moving user and the maximum distance a point can be from a stationary location in order for the new point to be classified as the user still being stationary (Stationary location size). The temporal parameters defines the minimum time a user has to be stationary in order to record a station (Trip end threshold) and the minimum time between two stations in order to record them as two separate stations (Minimum trip time). To further filter very short trips it also allows definition of the shortest distance of a trip for it to be counted (Minimum Travel Distance). See Table 2 for the threshold values used in the analyses.

The advantage of using this definition for calculating the OD matrices is that we do not need to make any assumptions on the travel times or habits of the users, e.g. such as time of day that they travel or when they are at home or at work. This will allow us to capture trips made by people who travel between many different locations during a day or travel during irregular hours. The disadvantage is that we need to define the zones that we divide the geographical area before that analysis is done which could result in loss of information about a high number of trips between two locations inside one zone. The technique also relies on the correct definition of the four parameters, which typically requires a calibration effort for different types of input data.

This type of definition is suitable when a higher resolution in space and time is required to separate travels, for example when we want to capture travels to other activities than work, say travels to shopping, daycare etc.

This way of calculating traffic demand is implemented in the Mobility Analytics Platform developed by Ericsson Research and can be shown in the Output Explorer of this tool.

Table 2: listing of the values of the different parameters used as input for the algorithms to compute the OD-matrices that have been tested in the tool.

	Result set		
	Abidjan	Yamoussoukro	Set2.0
Analysis parameters: *Geographical zone definition*	10x10 grid of the city of Abidjan	10x10 grid of the city of Yamoussoukro	10x10 grid of the country of Ivory Coast
Trip begin threshold (meters)	500	500	800
Stationary location size (meters)	10	10	10
Minimum travel distance (meters)	500	500	2000
Minimum trip time (minutes)	10	10	10
Trip end threshold (minutes)	50	50	200

The tool allows us to evaluate our methods using different input parameters and we have produced OD matrices using three different zone definitions, one representing the entire country, one representing the area or Abidjan, and one for analyzing the movement in the Yamoussoukro area.

Each result set represents one OD matrix for each hour of the day and the visualization represents the movement towards each zone (terminating), away from each zone (originating), or the net influx of people to a zone (balance). Examples of the results are shown for the Abidjan, Côte d'Ivoire, in Figures 8, 9, 10 and 11.

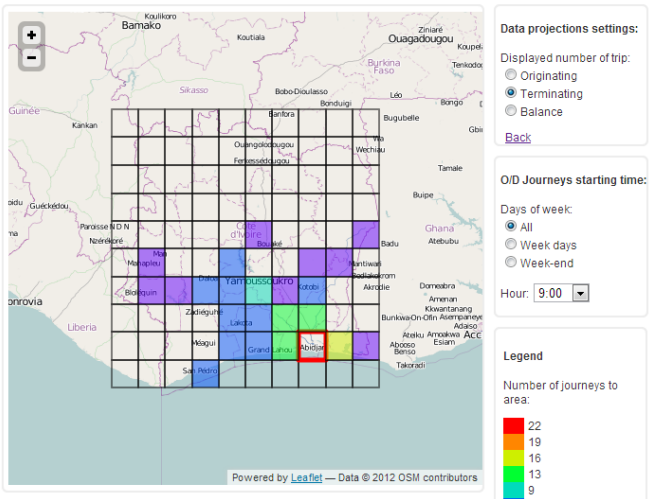


Figure 8: Selecting the Abidjan area from the overview of the entire country shows the number of long distance trips that started between 9 am and 10 am and terminated in Abidjan.

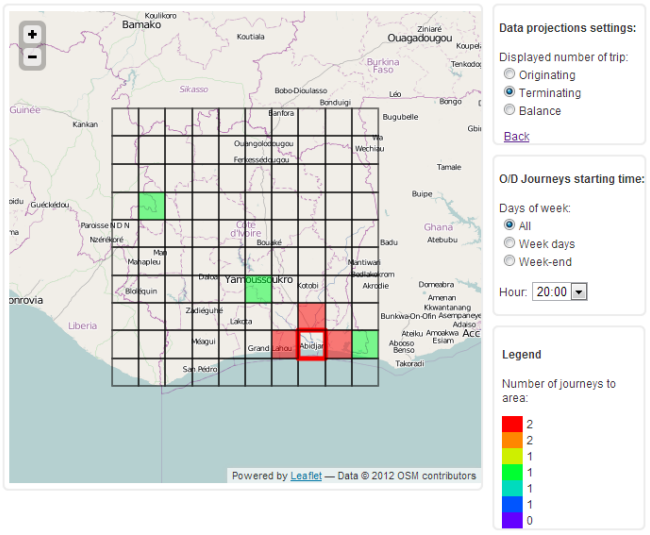


Figure 9: Selecting the Abidjan area shows by adjusting the time selection we can clearly see that the number of long distance trips towards Abidjan peak around 9 am and the decreases during the day as few people start a long distance trip towards the end of the day.

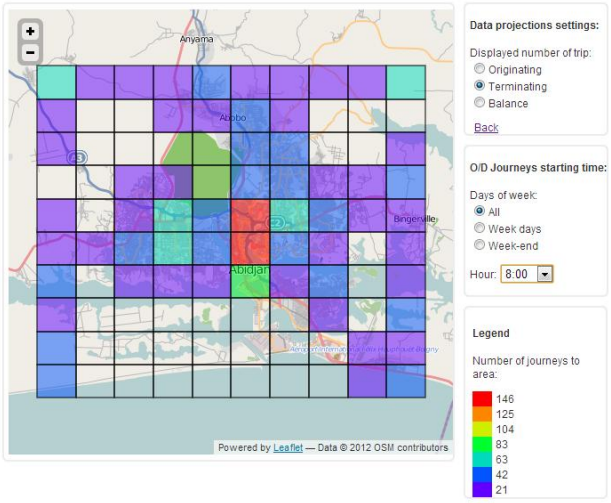


Figure 10: An overview of the movement in the Abidjan area shows that people travel towards the city center early in the morning hours.

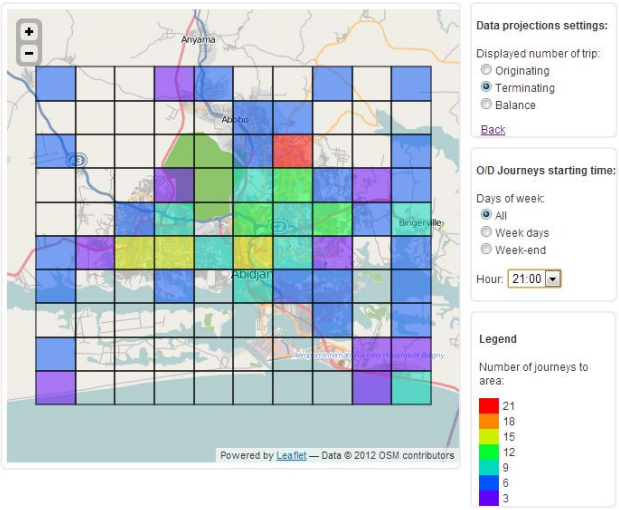


Figure 11: Towards late in the evening the movement is reversed as people travel outwards from the city center with a concentration towards the north eastern part.

Travel demand description type C – Travel flow

Travel demand captures how many users that travels between two zones in a certain time interval. By removing the stationarity requirement of users in the travel demand description of type B definition and reducing the size of the zones, we are moving towards a metric that describes how many users pass a certain location during a specific time interval; hence we call this a *travel flow* metric.

Removing the stationarity requirement, that is, taking all locations samples of a trajectory into account and treating each consecutive pair of samples as a trip, generates as high resolution as possible in time and space given the sampling frequency and location accuracy of the data source.

Since the distribution of the sampling time interval is heavy-tailed we can still get trips that are covering a quite large distance. This can be avoided by reducing the time filtering

window size enough to filter out trips that are too long. However, we need to keep the time window large enough to maintain statistical significance of estimates.

Since the spatial resolution of the dataset is relatively coarse, it is challenging to classify trips by transport mode or specific links in the network. Therefore we use the term *travel* flow to indicate that it is an aggregate of travels that pass by an area, independent of the transport mode and whether a specific link in the road network has been used.

We have applied the generalization and aggregation approach described in (Andrienko and Andrienko, 2012) for aggregating the mobile phone call data into travel flows between generalized places defined around the networks' antennae positions. This is performed in a sequence of steps as follows. First the trajectories are extracted from the mobile phone call data: The calls received/performed by each user are ordered chronologically into a sequence of calls representing the trajectory of this user in space. Second generalized places are extracted by using the antenna positions as seeds around which Voronoi polygons are generated. These polygons define the set of places that the explored area is divided into. The trajectories are then aggregated into moves between pairs of places by defining transitions between them, and counting the number of transitions present. Figure 12 shows a visualization of travel flows for the city of Abidjan.

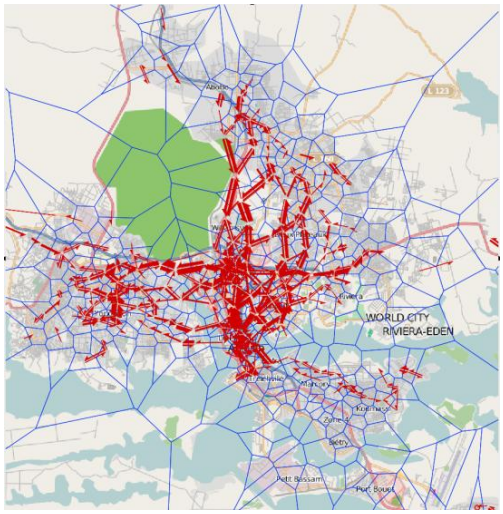


Figure 12: Travel flows between antennas represented with Voronoi polygons. The flow is aggregated for a two hour period in Abidjan.

If we reduce the spatial resolution of the zones, i.e. aggregate antennas into larger zones, while still not requiring any stationarity to separate trips, we get travel flows between zones. The difference compared to the travel demand is that travels that pass through a specific zone without any stop is counted as flow in and out of the zone. Figure 13 shows an example of flows between aggregated areas whereas as Figure 14 shows the same data but filtering for the links with highest flow.

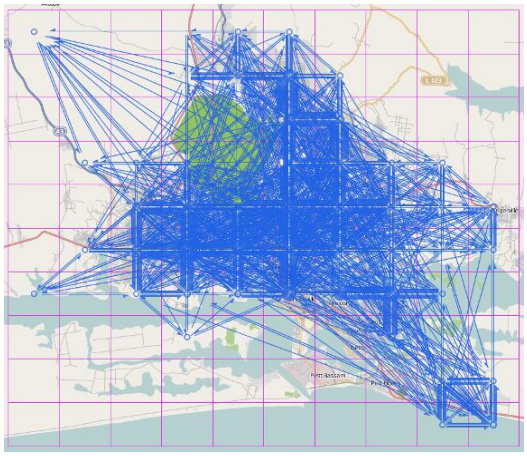


Figure 13: Travel flows between aggregated areas. The flow is aggregated for a two hour period in Abidjan without filtering.

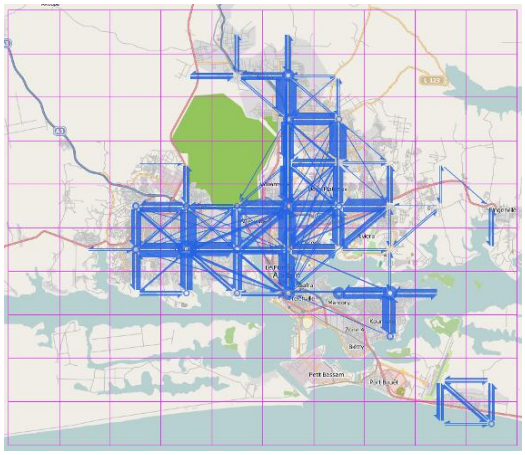


Figure 14: Travel flows between aggregated areas. The flow is aggregated for a two hour period in Abidjan filtering out links with travel flow > 100.

Travel demand for transportation models

Traditional techniques for finding origin-destination matrices include statistical models, entropy-based models and full travel surveys. The techniques require different input data, normally information from road traffic counts and travel time measurements from traffic cameras. Data from cell phone usage can be fused with these traditional measurements, and be used to improve the quality of the output from the techniques. The demand data is used as input to models that predict the transport behavior in more detail, for example, how the demand is split on different travel modes. This is normally done using discrete choice models. These models also require mode choice data in order to be estimated. Choice data may also be possible to infer from cell phone use data. If this is the case, the data can be fused with observed choice data, and therefore contribute to an improved output from the choice model.

In order to benchmark the traffic demand calculated from the cell phone data set we would like to use an independent way of estimating the demand. A classical way of estimating traffic demand is to use a gravity model where the trip attraction between different zones are modeled based on standard parameters such as population density, distance and travel cost. Also explaining factors like socio-economic characteristics and land-use can be integrated in the model. The number of trips between two zones *i* and *j* can be modeled as (Wilson, 1967):

$$T_{ij} = k\frac{O_i D_j}{d_{ij}^2}$$

Where:

T_{ij} = trips originating at *i* with destination at *j*
O_i = total trip origins at *i*
D_j = total trip destinations at *j*
k = adjustment factor
d_{ij} = distance from zone *i* to zone *j*

In our case we have total trip origins and total trip destinations proportional to the origin and destination population density, respectively. The impedance for a pair of zones is a function of the distance d_{ij} between the zones.

Figure 15 shows the gravity model estimation distribution of traffic demand originating from Abidjan to the whole Côte d'Ivorie. The distribution in Figure 15 can be compared with the distribution computed based on the cell phone data, shown in Figure 20 (Bottom). Clear similarities can be seen when comparing the gravity model output with the estimates based on cellular network data shown in Figure 20. Based on this we can conclude that in comparison to a well-established method to estimate traffic demand the cellular network data does at least not seem to contain any larger bias. However, the gravity model output is very rough and static by nature and the cellular network data can most likely improve traffic demand estimates dramatically compared to that.

Except for benchmarking, the gravity model can also be combined with cellular network data. The cellular network data can be used to estimate trip production, attraction and impedance parameters as well as socioeconomic factors. For example the travel times estimated in this paper are typically a better impedance variable than the distance. In order to estimate trip productions with reasonable accuracy, relatively detailed information about cell phone penetration rates and usage is required.

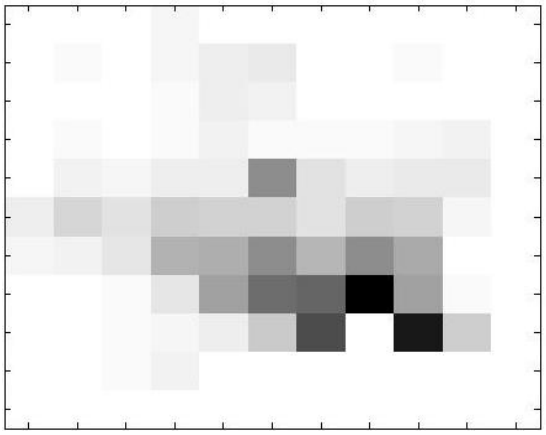


Figure 15: traffic demand proportions originating from Abidjan estimated by a gravity model based on population density and distance between zones.

B. Travel times

Due to the fact that we can only obtain location in terms of antenna positions and that samples are limited to when the user is active, the measurements contain large errors in both space and time domain. The space domain errors limits us to measure travel times for travels that are of a minimum length, and the length requirement is dependent on which relative travel time (average speed) error that can be accepted. The error in time due to sparse sampling limits us to draw conclusions of the minimum travel time instead of the full travel time distribution. However, the minimum experienced travel time is also very useful and a good indicator of travel quality. Figure 16 shows an example of travel times measured in Abidjan, where the heavy tail is due to sparse sampling in time.

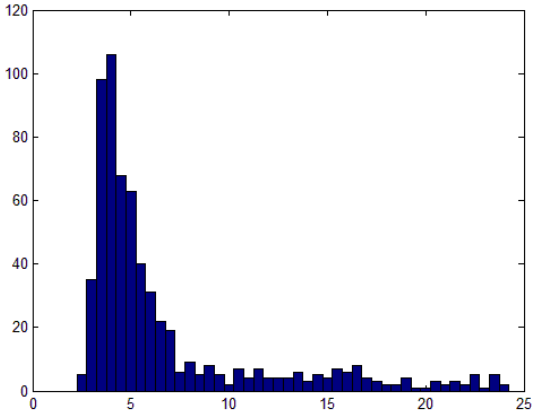


Figure 16: histogram with travel time measurements between Abidjan and Yamoussoukro grouped in intervals of 15 minutes. Travel time measurements larger than 24 hours are not shown.

By dividing the travel time into two parts, one caused by distance and type of transport infrastructure and one caused by queuing delay we are able to identify parts of the network that are congested. We can do this for example by separating the

measurements into peak hour measurements and non-peak hour measurements. By comparing the cumulative distribution function of travel time measurements for the two time periods, where unreasonably high travel times are filtered out, it is possible to identify a travel delay metric.

We have estimated the travel time distribution between Abidjan city and Abidjan airport during off-peak hours (9-16, 18-06, see Figure 17) and during peak hours (7-9, 16-18, see Figure 18) for a time period of six weeks. From the CDFs we see that the minimum travel time is 10 minutes longer for peak hours, indicating that the minimum travel time increases approximately 10 minutes due to congestion in the road network. By combining the two CDFs with travel flow estimates, it is also possible to express aggregated delay metrics like total queuing delay per route and time period.

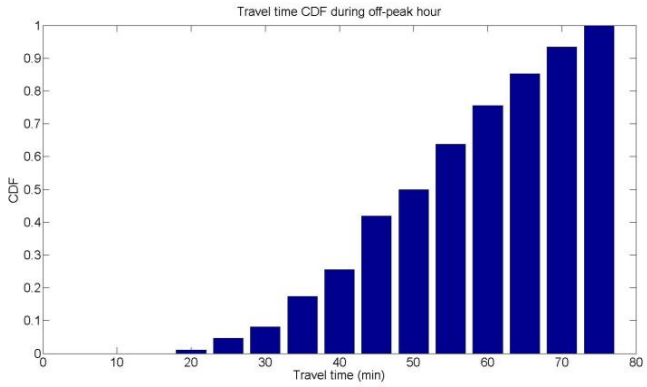


Figure 17: Off-peak hour travel time measurement CDF between Abidjan city and Abidjan airport grouped in intervals of 5 minutes. Travel time measurements larger than 75 minutes are not included.

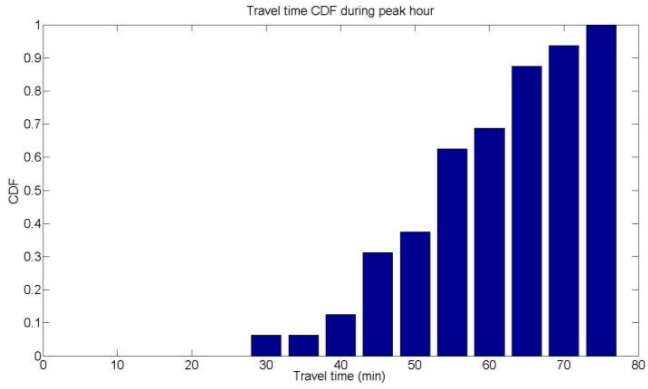


Figure 18: Peak hour travel time measurement CDF between Abidjan city and Abidjan airport grouped in intervals of 5 minutes. Travel time measurements larger than 75 minutes are not included.

IV. The Mobility Analytics Platform

A. Implementation of the 3-layer architecture

In implementing the 3-layer architectural framework we have implemented a modular platform capable of handling a wide range of sources of data, methods for analyzing them via the OD matrix definitions. Our platform implementation allows for high-throughput access to large data sets generated by the Data Collection Layer stored in HDFS™ [ref] and enables distributed processing via an Apache Hadoop server [ref]. At the same time, we cater for user privacy by removing all user-specific information in the early stages of the Analytics Layer. Overall, our aim has been to design the Analytics Engine in such a way that the algorithms can be easily exchanged and modified and that the analytics tasks can be easily configured, scheduled, and controlled through a user friendly interface. Finally, for the Knowledge Exposure Layer we have enabled access to the information for the programmer through an Application Programming Interface (API). For this we enable direct access to the anonymized information in the data model through an HTTP interface designed using the Representational State Transfer (REST) interface architecture [ref]. The API allows any software to easily make use of the OD matrix data to enable new functionality and make decisions based on the estimated travel patterns of the users.

Thus, applications can be designed to make queries for specific data such as "the number of people travelling between two zones on weekday mornings between 9 a.m. and 10 a.m.". This flexibility allows for many different types of applications and services to make use of the API and enables the integration to be done in a way that is optimized for the requirements of the application or service.

Furthermore, to facilitate the configuration, scheduling, and optimization of the analytics process we have created a graphical tool for analyzing the mobility data. Our tool allows the user of the platform to configure input data sources, create new analytics jobs and configure algorithm parameters, as well as view the results of the analysis that is being made available through the Knowledge Exposure API.

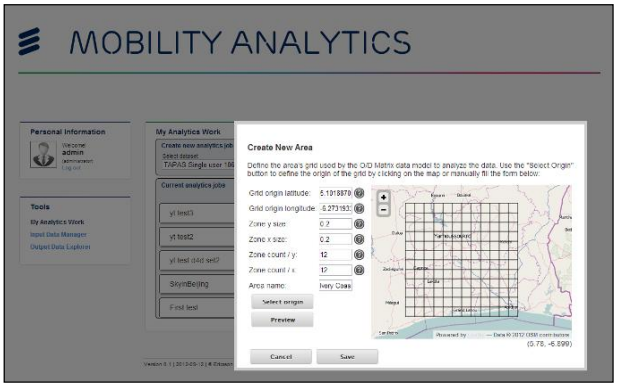


Figure 19: the parameter configuration tool can be used to define the geographical area that is to be analyzed and configure the parameters used in the clustering algorithms.

B. The front-end

Furthermore, to allow a user of the platform we have developed to inspect the information that has been generated from the data we have developed one more tool for interactive visualization that overlays the OD matrix onto a graphical representation of the actual geographic area. We call this tool

"Output Data Explorer". OD matrices that describe the movement of large areas can be very large data structures that are impossible to provide intuition to a human operator. The Output Data Explorer provides an easy way of examining the OD matrices that are the results of the analysis. It can be used to quickly verify the results of the analysis and compare and results of processing jobs with different input parameters before it is made available to applications through the Knowledge Exposure APIs. It can also be used as a tool to directly study the mobility of the users in the area by visualizing and interactive with the OD matrices.

With the Output Data Explorer, by selecting the time of day it is possible to view the total number of trips made to each zone (terminating), the total number of trips made from each zone (originating), or difference between terminating and originating trips from each zone (balance) for the hour starting with a selected time. Furthermore, the Output Data Explorer can display which zones have been the destinations for trips originating in a selected zone.

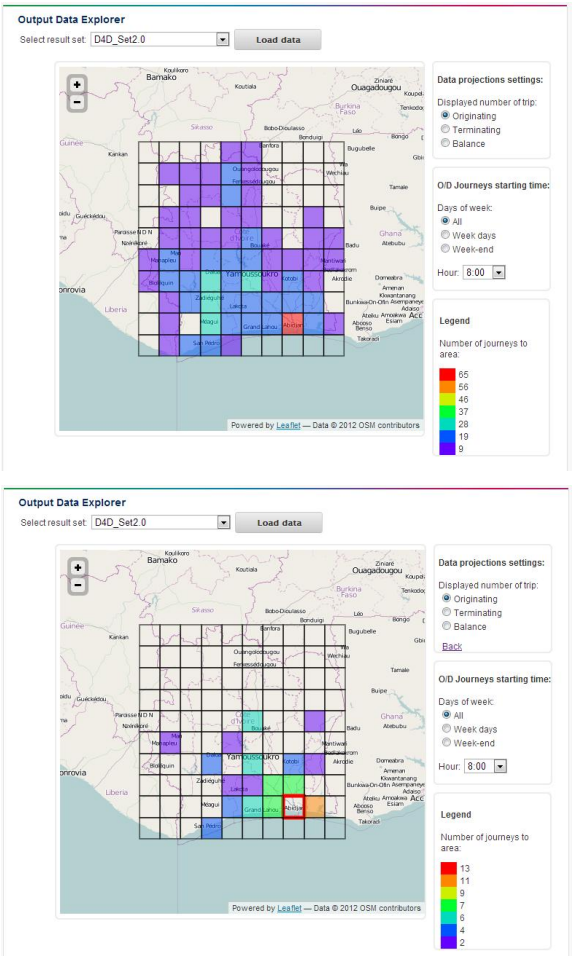


Figure 20: The Output Data Explorer displaying a map of the Ivory Coast color-coded according to the number of trips originating from each zone between 8 am and 9 am each day (Top) and a map of the Ivory Coast color coded according to the trips that started in the Abidjan zone (Bottom).

V. USE CASES AND CONCLUSIONS

Cellular network data has the potential to be used in several types of applications in infrastructure, public transport and road traffic planning. This specific data set comprising a subset of active terminals can definitely be used for strategic and tactical planning; however, it is not clear that this type of low resolution sampling in space and time can allow us to use the data for operational planning or real-time decision support. We are therefore focusing this paper on long term planning applications for transport systems and infrastructure.

In this paper we have demonstrated how to estimate and visualize different types of mobility metrics in both national- and city wide aggregation levels. These mobility metrics can be used to identify different types of bottle necks of the transportation infrastructure, which can be used as input in order to determine where infrastructure investments should be made in order to improve transportation efficiency.

In order to be able to build and adapt the transportation infrastructure efficiently, it is crucial to have reasonable estimates of the travel demand. The travel demand from cellular networks is capturing all types of travel modes, which enables also public transport planning or integrated road and public transport planning, which will be an important area of development in the near future.

The travel flows that are estimated from aggregated movements with higher spatial resolution compared to the travel demand enables an understanding of how the traffic demand is distributed in the transportation network and how it varies over time for different parts of the network. Based on this information it is possible to, for example, make better decisions on where in the network to make sure the infrastructure is maintained properly and where to improve public transport.

The travel time estimates gives a possibility to identify parts of the transportation network which has poor infrastructure, limited public transport or a transportation network that is not well adapted to the traffic demand. The travel time estimates are a very good metric of the quality of service of the transportation network. The travel times are estimated on both national- and city-wide aggregation levels and can incorporate all travel modes.

In case of long travel times we want to distinguish between 1) poor infrastructure or poor public transport and 2) congestion on the road network. To be able to do that we have calculated a travel delay metric based on the difference in travel time distribution for time periods of peak demand and those for non-peak demand. This gives an estimation of a travel delay metric due to congestion aggregated for all people travelling a certain O/D-pair. In these areas other measures than improving infrastructure can also be efficient. This can be for example improving public transport service or measures to spread out the travel demand over a longer period of time during the day.

The mobility metric described above might be possible to estimate in real-time as well, but in order to produce reasonable estimates well calibrated traffic models and most likely better penetration rates or cellular data with higher resolution, e.g.

handover data, is needed. However, the long term estimates will be very useful to determine where it is motivated to estimate real-time traffic information, independent of whether it is based on cellular network data or other traffic sensors.

In developing countries the cellular network is typically much more developed than the traffic and transport sensor infrastructure. However, the traffic situation can be really problematic and the need for well-informed traffic planning decisions is large. Together, this makes cellular network signaling data for traffic planning especially interesting in these countries.

By defining an architecture for a generic platform for analyzing mobility data collected from mobile networks we have shown how knowledge about human travel patterns could be made available for a wide range of purposes. The platform makes it possible to provide APIs to enable entrepreneurs to create any kind of services and applications that make use of this knowledge. We believe that making this knowledge widely available will lead to many new innovations that will benefit new businesses and the development of societies at the same time. By implementing a prototype of a platform utilizing this architecture we have shown that it is technically feasible to implement such a multi-purpose platform and provided an example of how information about human travel patterns can be made available without risking the privacy of the mobile network users.

REFERENCES

N. Andrienko and G. Andrienko, "Spatial Generalization and Aggregation of Massive Movement Data," IEEE Transactions on Visualization and Computer Graphics, vol. 17, no. 2, pp. 205–219, 2011.

N. Andrienko and G. Andrienko, "Visual analytics of movement: an overview of methods, tools, and procedures," Information Visualization, vol. 12, no. 1, pp. 3–24, 2012.

Afripop, www.afripop.org, 2010.

D. Blondel, M. Esch, C. Chan, F. Clerot, P. Deville, E. Huens, F. Morlot, Z. Smoreda and C. Ziemlicki, "Data for Development: The D4D Challenge on Mobile Phone Data", 2012, http://arxiv.org/pdf/1210.0137v1.

T. Karhumäki, "The Utilisation of GSM-Network in Travel Time Monitoring," ITS Workshop on Road Monitoring, Imperia, 2002

D. Gundlegård, and J.M. Karlsson, Handover Location Accuracy for Travel Time Estimation in GSM and UMTS. IET Intelligent Transport Systems, Vol. 3, Issue 1, pp. 87-94, 2009.

D. Gundlegård and J.M Karlsson, Generating Road Traffic Information from Cellular Networks - New Possibilities in UMTS. In: ITS Telecommunications, 2006. p. 1128-1133.

N. Caceres, L.M. Romero, F.G. Benitez and J.M. del Castillo, "Traffic Flow Estimation Models Using Cellular Phone Data", IEEE Transactions on Intelligent Transportation Systems, vol. 13, no. 3, pp. 1430-1441, 2012.

CIA World Factbook, https://www.cia.gov/library/publications/the-world-factbook/geos/iv.html, 2006.

DATA SIM, "DATA science for SIMulating the era of electric vehicles", 2011-, http://www.uhasselt.be/datasim

GeoPKDD, "Geographic Privacy-aware Knowledge Discovery and Delivery", 2005-2008, http://www.geopkdd.eu

S. Isaacman, R. Becker, R. Caceres, S. Kobourov, M. Martonosi, J. Rowland, and A. Varshavsky, "Identifying Important Places in People's Lives from Cellular Network Data". Proceedings of the 9th international conference on Pervasive computing, pp 133-151, 2011.

E. Mellegard, S. Moritz, and M. Zahoor, "Origin/Destination-estimation Using Cellular Netwrok Data". IEEE 11th International Conference on Data Mininig Workshops, pp 891-896, 2011

A. Monreale, F. Pinelli, R. Trasarti, and F. Giannotti, "WhereNext: a Location Predictor on Trajectory Pattern Mining". Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and Data mining, pp 637-645, 2009.

Trafikverket, http://www.trafikverket.se/Privat/Vagar-och-jarnvagar/Sveriges-vagnat/, 2012.

A.G. Wilson, A statistical theory of spatial distribution models, Transportation Research, Volume 1, Issue 3, November 1967, Pages 253-269.

Y. Zheng, Y. Chen, X. Xie and M. Wei-Ying, "GeoLife2.0: A Location-Based Social Networking Service". Proceedings of the 10 th Tenth International Conference on Mobile Data Management., 2009.

Y. Zheng, Y. Liu, Y. J. Yuan, and X Xie, "Urban Computing with Taxicabs". Proceedings of ACM UbiComp, 2011.

MP4-A Project: Mobility Planning For Africa

Mirco Nanni[1], Roberto Trasarti[1], Barbara Furletti[1], Lorenzo Gabrielli[1]
Peter Van Der Mede[2], Joost De Bruijn[2], Erik De Romph[2], Gerard Bruil[2]

[1]KDD Lab,Isti - CNR
Pisa, Italy
email: name.surname@isti.cnr.it

[2]Goudappel Groep
Deventer, Netherlands
emails: pvdmede@, jdbruijn@, edromph@, gbruil@ [goudappel.nl]

Project Abstract

This project aims to create a tool that uses mobile phone transaction (trajectory) data that will be able to address transportation related challenges, thus allowing promotion and facilitation of sustainable urban mobility planning in Third World countries. The proposed tool is a transport demand model for Ivory Coast, with emphasis on its major urbanization Abidjan. The consortium will bring together available data from the internet, and integrate these with the mobility data obtained from the mobile phones in order to build the best possible transport model. A transport model allows an understanding of current and future infrastructure requirements in Ivory Coast. As such, this project will provide the first proof of concept. In this context, long-term analysis of individual call traces will be performed to reconstruct systematic movements, and to infer an origin-destination matrix. A similar process will be performed using the locations of caller and recipient of phone calls, enabling the comparison of socio-economic ties vs. mobility. The emerging links between different areas will be used to build an effective map to optimize regional border definitions and road infrastructure from a mobility perspective. Finally, we will try to build specialized origin-destination matrices for specific categories of population. Such categories will be inferred from data through analysis of calling behaviours, and will also be used to characterize the population of different cities. The project also includes a study of data compliance with distributions of standard measures observed in literature, including distribution of calls, call durations and call network features.

1. Introduction

Population growth, massive urbanization and, particularly, the extensive increase of car use in the last century have led to serious spatial, transport, infrastructural and environmental problems in almost all urbanized areas. As a consequence, since the 1960's urban and transport planning methodologies were developed to forecast future traffic volumes and the expected use of infrastructure and facilities. The purpose of such forecasts is evident: infrastructure and urban planning provide keys to the mitigation and preclusion of transport and environmental problems. Today, urban and transport planning are major tasks of all public authorities.

Although transport planning tools and methodologies are widely used in the western world for more than half a century, in many Third World countries transport planning is scarse due to limited expertise, and lack of resources and, in particular, the unavailability of data that are necessary to utilize these tools and methodologies. Because population growth and economical development has shifted from the western to the non-western world, transport and urban problems are increasing particularly in non-western countries. Still, the urban and transport planning practices of those countries are quite often in its infancy. It is therefore promising, that in the last decades new technologies have emerged, which have great potential to overcome the hurdles that obstruct the much wanted development of new, easy to implement transport planning methods in non-western countries. The current paper presents a proof of concept that by combining mobile phone data traces and road network data from open web sources a basic transport demand model can be build.

Traditionally, to create a transport demand model a number of different data sources are needed: First, an accurate description of the transport network, relevant (multi-modal) traffic counts on key parts of the infrastructure, spatial, socio-economic, mobility and demographic data (e.g. on working and dwelling areas) for the present situation, and scenarios or forecasts of the future. Although most countries do have statistical agencies that provide some of the necessary data, in rapidly developing Third World countries many of these data are missing or of older date.

In many cases traditional 4-step demand modeling is used to build a transport demand model. Four-step modeling comprises (1) trip-generation, (2) trip-distribution, (3) modal choice and (4) traffic assignment (Hensher & Kenneth, 2000). The 4-step model is an aggregated model. It models the travel behavior in a area using averages. Another class of transport models are disaggregated models which model travel behavior at individual level. Both models are aimed at long term (decades) development. More recently, in many western countries long term demand modeling is being supplemented by dynamic, micro- and macroscopic models that are more concerned with actual and in the short-term expected traffic flows and with congestion management on the road network. This is due to the fact that in many western countries infrastructure development is nearing its end and (dynamic) traffic management is the instrument to deal with current problems.

The study area in a transport model is usually divided into small areas called zones. A typical area could be a small neighborhood in a city. A zone represents the origin and destination of traffic within the study area. The trip generation and trip distribution step in the four-step model typically produce an Origin Destination matrix. This matrix is a description in time of traffic flows from any place in the network (origin) to any other place in the network (destination). An O-D-matrix may be produced for a certain period of time (e.g. a peak-period matrix) or for a whole day. By using transport modeling methodologies the flows between origins and destinations can be used to assess expected peak traffic flows on all parts of the network, thus assessing network capacity problems.

In rapidly changing environments, as many urban areas in non-western countries are, infrastructure and transport planning can be enhanced greatly if a transport (demand) model can be used. And not only planning can be enhanced by such a model; also the improved insight in the current situation by an accurate description of mobility behavior and traffic flows on the network provide a valuable instrument to design measures to alleviate problems, such as traffic and transportation management or a better public transport system.

The availability of spatial data on demographics, labor and land use has until now been a prerequisite for establishing an OD-matrix for a transport model. It is a time consuming activity to obtain the necessary data in many developed countries. In most developing countries however the overall availability of data is very limited. As a result the use of transport models has never been a promising option for developing countries.

Therefore, if we could show that a proper O-D matrix can be derived from mobile phone data, and that by using publicly available (free) transport network data and standard transportation modeling software a relevant transportation a basic transport demand model can be build, we provide evidence that also for countries or cities where many data seem to be lacking, now transport demand models can be created. This will on the one hand allow national and local authorities to have a far better understanding of transportation needs and challenges, and will help funding agencies and investors to better assess their potential risks and benefits.

In the next section we will describe step by step the process we have gone through to derive a (as far as we know) a first transport demand model for Ivory Coast and its major urbanization Abidjan.

2. Background

This work tries to combine data mining of GSM traces with transportation modeling methodologies to gain insights into mobility in a monitored area, to allow what-if analysis through simulation or modeling.

GSM data have already been used to describe mobility in several studies, essentially based on the fact that a sequence of geo-referenced calls of users constitutes approximate trajectories of their movements. The key limitations of GSM data are that locations are only approximations and that sampling rate may be low and erratic. Works like [1] try to overcome these issues by working at a large geographical scale and/or under specific conditions (in that case, users where tourists in a large area). In the present work we follow a different approach, and try to exploit the relatively long temporal extension of a dataset to infer more reliable movement information. In particular, an approach similar to [2] (translated from GPS to GSM data) and [13] is adopted, where we try to extract regular movements that repeat consistently in time, which therefore are less likely to be artifacts of the data sampling procedure, and use them to measure systematic mobility in the area (details are provided in next sections). Also, concepts like most favored location, which are exploited in this work, have already been applied in the scientific studies, e.g. [3], but mainly for simple distributions of a population or the recognition of specific activities, such as working, being at home, or leisure.

Macroscopic transport modeling methodology is well established [11,12]. This methodology is mainly implemented through commercial and academic software tools (e.g. OmniTRANS, Visum, Cube, Emme/2, TransCAD), and readily available. These tools can be used only by professionals with accurate knowledge of traffic theory and transport modeling experience. Macroscopic modeling has been used widely by governments and engineering firms to predict future transport network problems and for infrastructure planning. The current paper does not address network or transport planning as such. Its main purpose is to use data mining of GSM traces as an input for transport models, thus integrating these traces as widely and readily available sources of information into the transport planning realm.

3. Data exploration and validation

In this section we explore the data in order to assess their statistical validity and providing insight of possible biases or hidden strategies in the user selection. Moreover we complete the data using a methodology to estimate the coverage of the antennas.

a. Data Exploration

The ten datasets provided are based on anonymized Call Detail Records (CDR) of mobile phone calls and SMS exchanges between five million of Orange's customers in Ivory Coast between December 1, 2011 and April 28, 2012. Each dataset is a set of observation of 50,000 users for two consecutive weeks. This means that each dataset covers 1% of the actual users in the relative period. This leads to a statistical weakness in representation due to possible noise introduced by user selection. Figure 1 shows the distributions of the ten datasets by day of the week. Some datasets have very different shapes, e.g. dataset 9 shows a big decrease in activity on Thursday (day 4) while dataset 5 shows a peak during the same day of the week.

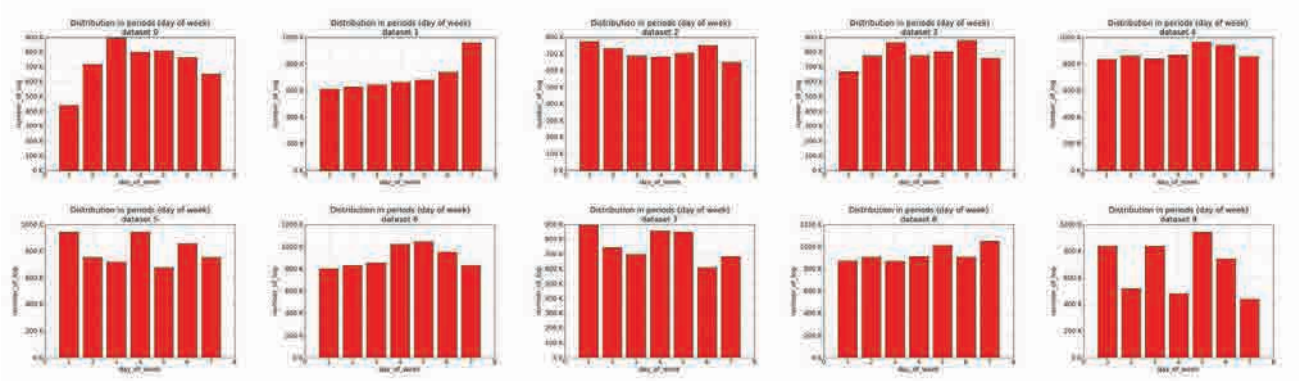


Figure 1: Distribution of the observations in the ten datasets by day of the week

These differences can reflect that some event happened on specific days, but the number of users per dataset is low, and these differences may also be due to unbalanced user selection. Moreover, analysis of the distribution of the observations by hour of the day as shown in Figure 2, shows distributions that are very similar and that follow the same trend: a small decrease in the night, then a fast increase during the early morning followed by a first peak around 8-10 a.m. then a decrease during the afternoon and a second peak around 6-8 p.m. to finish with a fast decrease in the night.

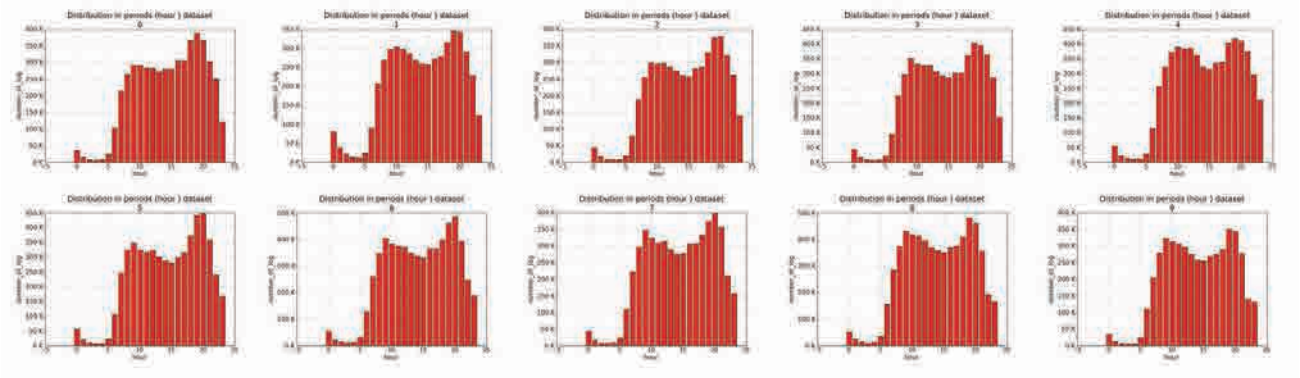


Figure 2: Distribution of the observations in the ten datasets by hours of the day

This kind of distribution has been found in several analyses [4, 5] using different kind of data describing human behavior. The fact that we found this pattern in all datasets, even if anomalies are present, gives strength to our hypothesis that all datasets represent a "standard week" of Ivoirians citizens. For these reasons, we decided to merge the dataset into a single one representing a hypothetical standard week where the number of users is greater and therefore more stable and meaningful. Therefore in the following analyses, when we will refer to the dataset we mean the one obtained by this process.

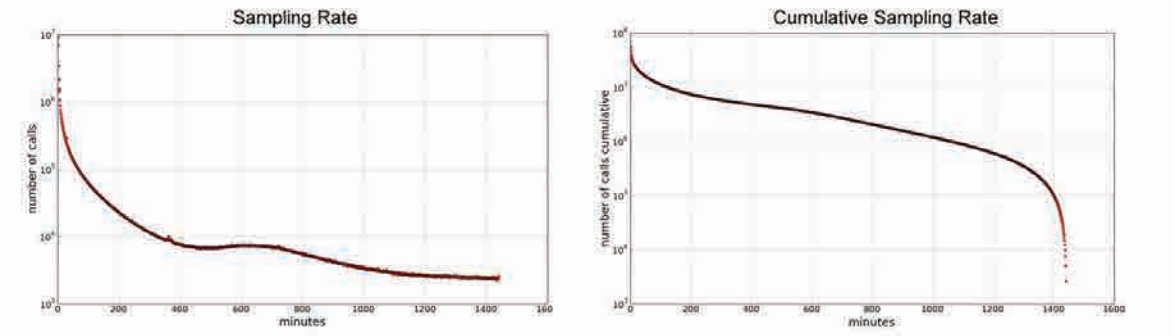


Figure 3: Distribution of the observations in the ten datasets in time w.r.t. the hours of the day

Another aspect to take into consideration is the sampling rate of the data, the temporal distance between two observations of the same user. Figure 3(left) shows the distribution of the sampling rate and Figure 3(right) the cumulative distribution of it. The distribution is 'cut' at 1,440 minutes, which is exactly 24h. We suppose users with at least one call every 24h were selected. This shouldn't affect our analyses, but we must be aware that the population we are focusing on in the literature is called *frequent caller users*.

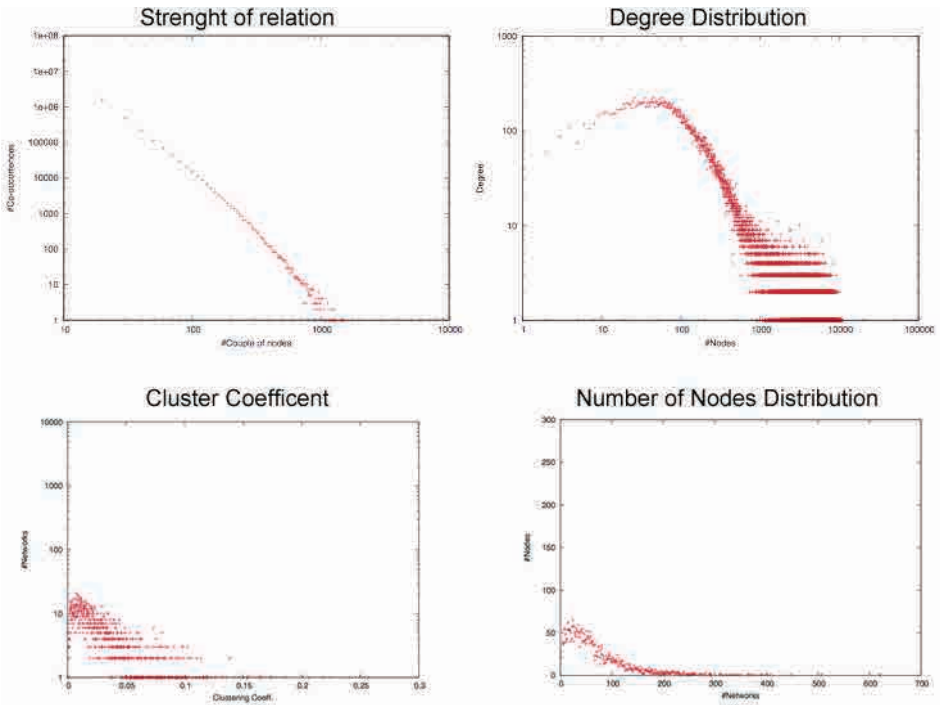


Figure 4: Standard measures for social network analysis

Though we did not use them in our analyses, the communication graph dataset (ego networks) included in this data challenge provided a means to further check the normality of the population described by our data, in comparison with other populations observed in several works in literature. Basically, we extracted the standard measures in graph-theory in order to check if they respect the well-known parameters and if there are not strange effects and biases to be considered. The distribution of the following measures was evaluated (taken from [6]): strength of relation, node degree, cluster coefficient, number of nodes.

The results are shown in Figure 4 and it is possible to notice how they follow the classical distribution as in [6]. This means that the structure of relations in the population under analysis appears not to contain clear biases.

b. Spatial coverage

The data provided contains for each observation the coordinates of the antenna serving the user during a communication, in other words the device is operating in an area covered by that antenna.

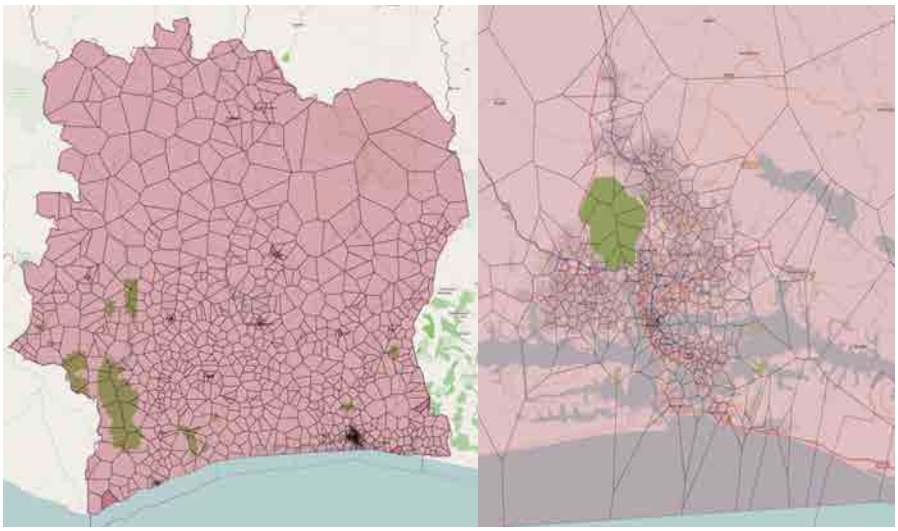


Figure 5: Ivor Coast Voronoi tessellation using the spatial position of antennas (left) and a focus on Abidjan city (right)

In general the coverage of an antenna is influenced by several factors: strength of the signal, the height of the pole, the orientation, the weather, the nearby buildings, etc. Since the provided data does not contain these information and it is not easy to retrieve them from external sources, we apply a well-known methodology in order to estimate the coverage of the antennas using only their spatial location. The method is called as *centroid Voronoi tessellation* (Qu *et al.* 1999) and assumes that the space is partitioned into separate areas, each defined as the set of locations that are closer to our antenna than any other one. The partitioning of the space obtained is shown in Figure 5 and will be used in all the following analyses. The partitioning of the space obtained will be used in all the following analyses.

4. Systematic traffic analysis and transport modeling

The basic events we are interested to spot in the data are systematic trips. Following the approach in [2], we define systematic trips as routine movements that users perform (almost) everyday at (approximately) the same hours. By combining together the systematic movements of each individual in our population, we can

obtain an estimated origin-destination matrix (OD-matrix) that describes the expected flow of people between pairs of spatial locations.

Since the current GSM data are not detailed enough to detect whether a user stopped at a location or initiated a trip within an input sequence, we tackled the problem through a two-step procedure. First, we identified locations that are significant for the mobility of the individual, also called attractors. Second, we identified movements between significant locations that occur with a high frequency, which are later aggregated across the whole population to fill in an OD-matrix. The first step is performed according to the standard approach, also illustrated in [3]: the location where the largest number of calls took place is identified and labeled as L1 (most frequent location). Then, the second most frequent location is identified and labeled as L2. It seems likely that in most cases L1 corresponds to the home location and L2 to work or any other main activity of the individual, or vice versa. The second step is performed over the sequences of L1 and L2 that appear in the traces of each single user. We checked the frequency of movements L1 → L2 and L2 → L1 within specified time slots. Each movement identifies a trip, and if its frequency is high enough, we assume it to be a systematic trip that the user performs during a typical day.

a. Detecting user's attractors

The available data provide the information of the zone from which a phone call is started. Thanks to the large amount of data provided by the telephone operator it is possible to use the spatio-temporal footprint left by the users for the purpose of monitoring their movements in the territory.

Several studies [7] assert that most people spend most of their time at a few locations, and the most important ones may be labelled as home and work. In this section we will explain the methodology used for the extraction of such important locations, which we will call L1 (most important one) and L2 (second most important one) using the frequency of calls made by users.

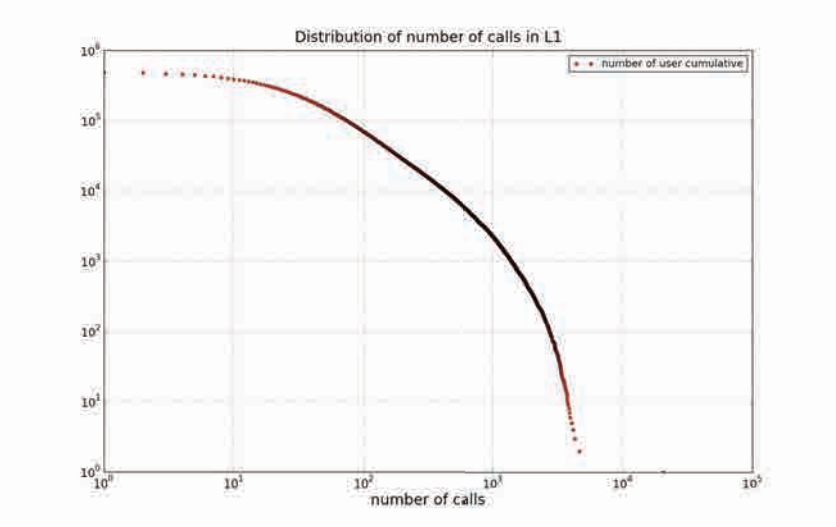


Figure 6 –Example of L1 detection (left), Distribution of number of calls in the most frequent location (right)

The location L1 relates to the antenna from which the user made more phone calls ever. For technical reasons, due to the load distribution of the antenna, it is possible that the reference antenna for different calls made at the same place may be different, even though such antennas are usually close to each other. To eliminate this

effect, we have redefined L1 in order to contain not only the area with most calls, but also all adjacent ones. A visual example is shown in Figure 6, together with a distribution of call frequencies for all L1 areas detected.

It is necessary to point out that, for most users, the call frequency associated to L1 is quite low (though larger than any other one of that user, by definition), thus rising issues of statistical significance and reliability of such a location (see Figure 6). If we consider as a minimum frequency threshold for L1 of one call per day (therefore 15 calls in 15 days) from the area, only for 20% of users (100K) the associated L1 would result meaningful. The rest of this work assumes to work on such subset of locations.

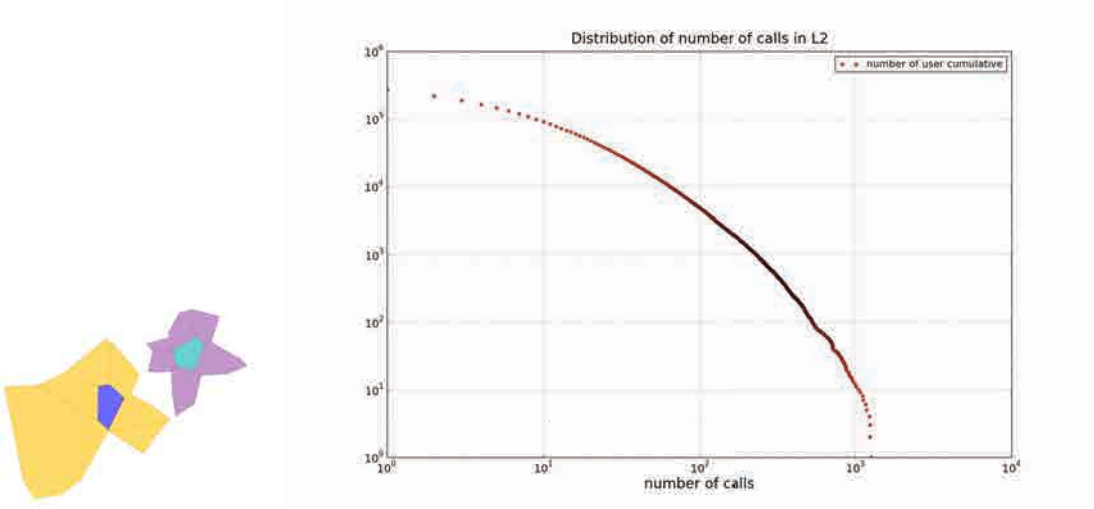


Figure 7 –Example of L1 and L2 detection (left), Distribution of number of calls in the second most frequent location (right)

The L2 is defined as the area that ranked second in terms of call frequency, excluding all areas already absorbed into L1. Figure 7 Shows a visual example of L1 and L2. While in the example L1 and L2 are quite distant (for instance, these individuals might come to the city to work), in some cases they might be adjacent, which happens quite often in city centers where antennas are more dense. Also in this case it is necessary to consider a minimum support of phone calls to identify the significance of the places identified based on the distribution shown in Figure 7(right). The result of this analysis will be used in the next step for the study of the systematic movements between preferred locations.

b. Detecting Systematic Movements

The focus of this analysis is the detection of systematic movements considering two separated time frames: a morning time frame, and an afternoon time frame, in which the users usually move, respectively, from home to work and from work to home.

The first step is to identify the movements performed by individuals from L1 to L2 (L1 → L2) and from L2 to L1 (L2 → L1). It is important to notice that we are looking for movement between these two special areas even if they are not contiguous, i.e. other areas were traversed between them, as shown in Figure 8 (A is distinct from L1 and L2).

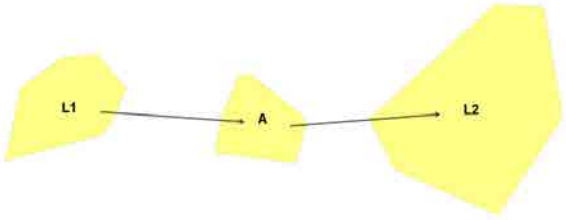


Figure 8 – Example of flow from L1 to L2

The second step consists in selecting only the systematic movements, which is done by applying two different constraints: (i) request a minimum number of movements between the pair; and (ii) request a minimum value for the lift measure of the pattern L1→ L2, which we define as:

$$LIFT(L1 \rightarrow L2) = {P(L1 \cap L2)}/{(P(L1) * P(L2))}$$

Lift measures the correlation between L1 and L2, resulting high if they appear together often w.r.t. the frequency of L1 and L2 taken separately. The main purpose is to normalize the frequency of L1 → L2 w.r.t. the frequency of calls of the user, since otherwise the candidate movements of frequent callers would be excessively favoured in the selection. The constraint on the number of movements is usually adopted in literature to exclude extreme cases where the lift (or other correlation or relative frequency measures) is not significant. More important is the LIFT measure, in fact to select. In our case, after a preliminary exploration we chose to select only pairs that appeared at least 3 times. The threshold for the lift measure was chosen based on the study of its distribution, as shown in Figure 9, selecting the value in which point the slope of the cumulative distribution begins a sudden drop: value 0.7.

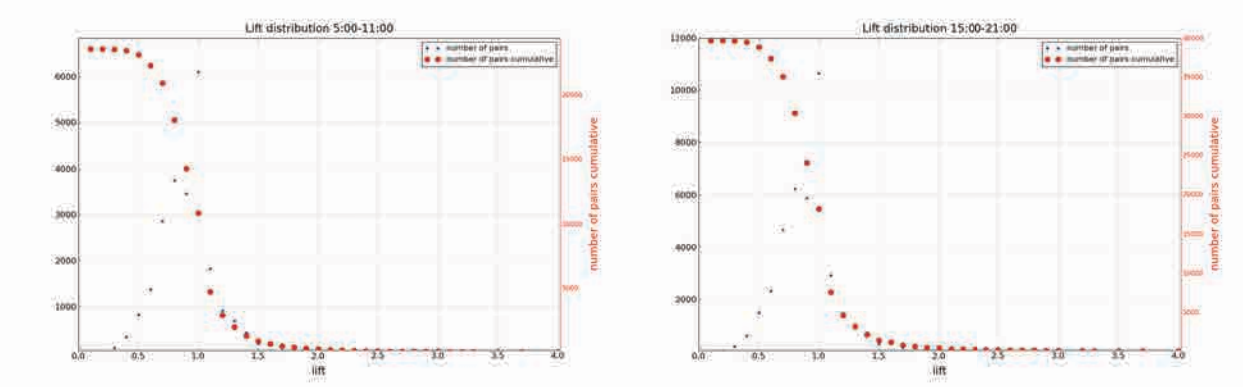


Figure 9 - Frequency distribution of the value of LIFT measure for pairs of frequent movements

In order to define the most appropriate time intervals to represent the morning and the afternoon periods, three key factors should be considered. First, such time intervals should include the peaks in the hourly distribution of road traffic, which we assume to be close to the peaks of the distribution of calls shown in Figure 2 (past experiences, although related to different regions of the world, showed a fairly good correspondence between the overall distribution of calls and that of mobility). The second factor to consider actually conflicts with what the previous one is related to : in order to cope with the sparsity of information for

a single user, our extraction process requires to have a time window as large as possible. Indeed, CDR data enable us to spot a movement only if there have been some calls before and after it. The smaller the time intervals we observe, the lower is the probability of including the calls needed to detect the movement. Empirical evaluations showed that time intervals smaller than 4-6 hours lead to spotting only a small number of movements, thus suggesting to fix a minimum threshold to 6 hours. On the other hand, too large intervals might lead to inclusion of movements outside the intended morning and afternoon periods. Finally, the third factor that can guide the choice of the time intervals is the observation that the overall traffic we detect should balance in terms of direction: the registered flow of movement from any location A to any location B should be reasonably close to the flow we observe from B to A. That holds especially when systematic mobility is considered, which is our case.

To summarize, the optimal time intervals to monitor in our process should cover the traffic peaks, yield reasonably balanced traffic flows and have a duration that is essentially a trade-off between two opposite requirements. In our experiments we explored a small number of alternative settings compatible with the traffic peaks of Figure 2, which are located between 8 a.m. and 11 a.m. (morning peak) and between 7 p.m. and 9 p.m. (afternoon peak). For each setting, we evaluated the traffic balance over all pairs of origin-destination we obtain from the analysis, by means of an index that compares the flows between each pair of locations (A,B) in the two directions as follows:

$$balance(A,B) = \left| flow(A \to B) - flow(B \to A) \right| \Big/ max\{flow(A \to B), flow(B \to A)\}$$

where flow(A ⟶ B) represents the total number of trips from A to B detected on the two time intervals chosen, merged together.

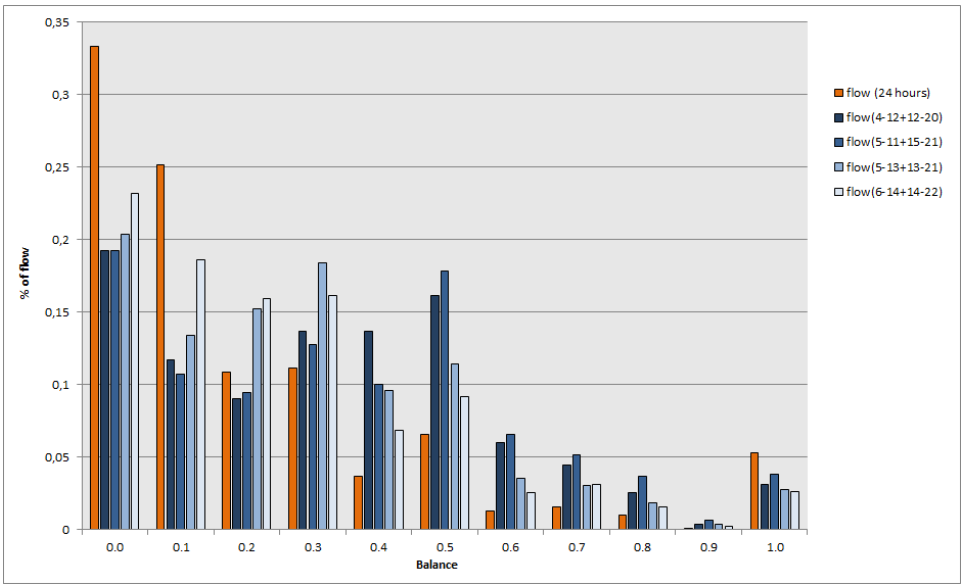


Figure 10 - Distribution of the traffic flows w.r.t. the balance index

Figure 10 shows the overall distribution of the balance index for four different settings (blue bars) together with the same distribution computed over the whole day (orange). In particular, for each bin of values of the index, which ranges from 0 to 1, we plot the percentage of overall traffic flow associated with the bin, in order

to better assess the impact of the bin of the final OD-matrices we are going to build with these data. All the distributions show a peak on the [0-0.1) bin, followed by significant values only up to the [0.5,0.6) bin. That proves that a reasonable level of balancing is provided by the different settings, though they show some differences. In the further steps of our experimentation we adopted the time windows 5 a.m.–11 a.m. and 3 p.m.–9 p.m., corresponding to the central bar in the plot of Figure 10. The reason is that it represents a better trade-off in terms of duration / separation of the two time frames (6 hours, separated by a four-hour gap), and its balance distribution is comparable with the others, though slightly inferior.

c. Systematic O/D Matrix

As previously mentioned, the final goal of our analysis is the synthesis of O/D matrices that summarize the expected traffic flows between spatial regions. In particular, the regions adopted in this study correspond to phone cells of the data provider, which are approximately mapped to the geography through the Voronoi tessellation described in previous sections. Also, our O/D matrices will focus systematic mobility, which represents the core (though not the only) part of traffic. The following figures provide some visual examples of the areas considered as well as the flows obtained at the end of our process.

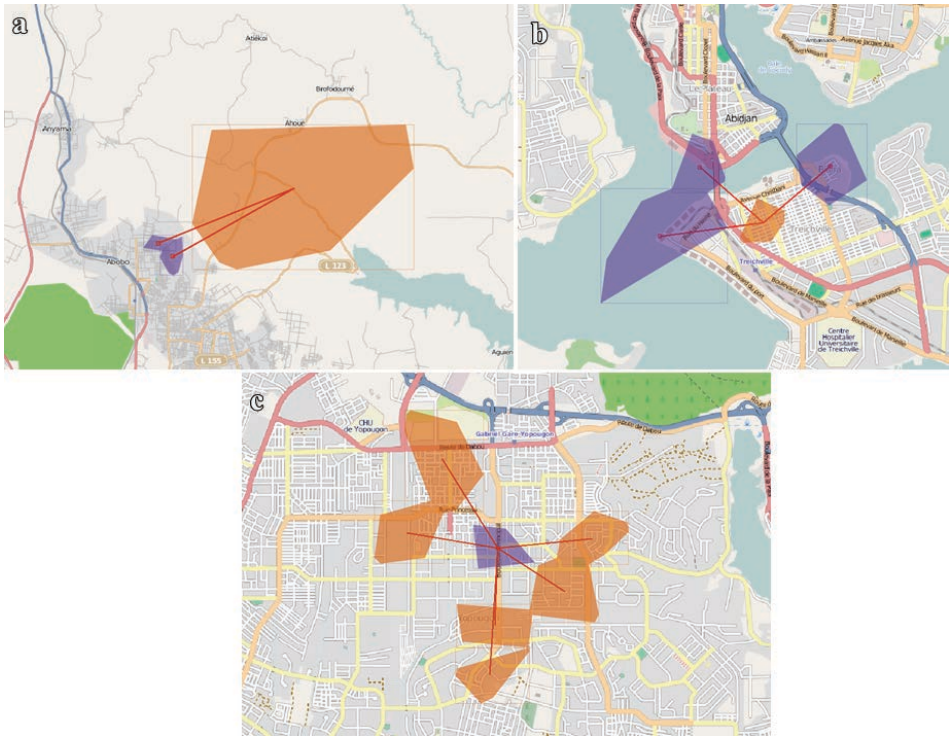


Figure 11: Examples of traffic flows taken from one of the O/D matrices produced: (a) flows from the outskirts of Abidjian to the city; (b) from a single area to other districts; (c) confluence of several districts to a central one.

In Figure 11 some example of intra-city traffic, the first and the second one from one origin to several different destinations, the second one from several origins to a single destination.

d. Building the transport model

Network

We used data of the OpenStreetmap (OSM) road networks for Ivory Coast and Abidjan as a base for modeling. Although of great detail, a drawback of these network data is that many links are not, or not properly,

connected. For route calculation this evidently is problematic. We therefore used an algorithm in a Geographic Information System (GIS) to find unconnected or badly connected roads and connected them properly or at least logically. Furthermore we programmed an algorithm to identify small island-networks. They were either connected to the main network, typically we added a few ferries to connect real islands, or erased because they seemed illogical or unimportant.

The resulting digital road network was imported in OmniTRANS, the software for transport modeling. Network link attributes, such as speed, which are necessary to calculate the shortest route between an origin and destination, were derived from link type information which is available from the OSM-network. The next step was to connect the antennas, e.g. the data carriers, to the network. In anticipation of this step we did not remove all the small roads from the network, which is often done in transport models. By keeping them we could simply connect each antenna to the nearest road. Of course this gives an overload on that particular minor road, but this quickly flattens out as all trips divert in different directions and converge on the major roads.

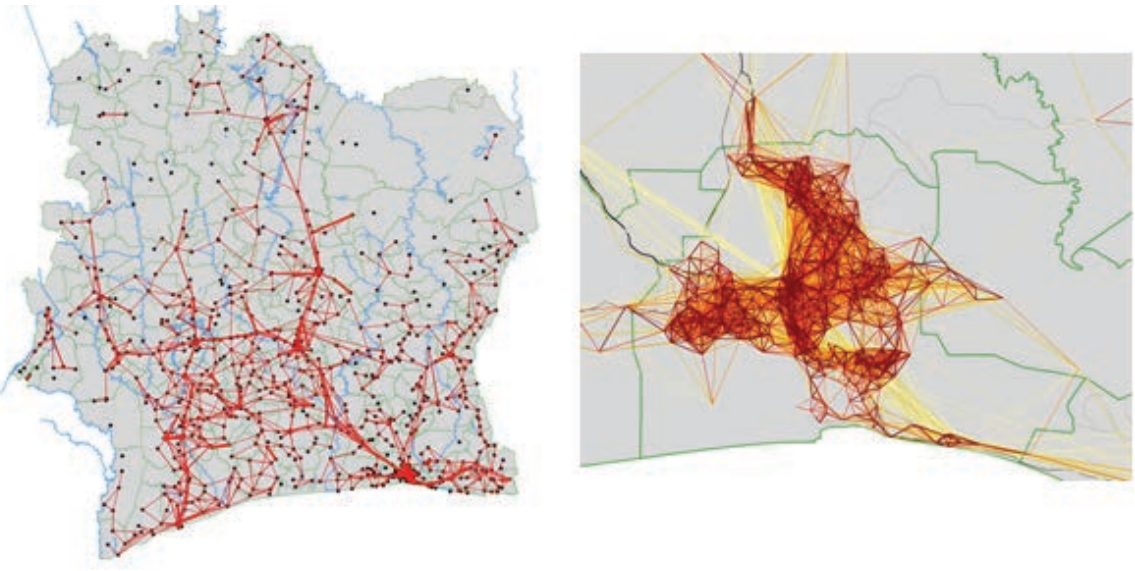


Figure 12: Mobile phone movements in Ivory Coast and Abidjan.

OD matrix

The previously described OD-matrices derived from mobile phone data were imported in the transport model using a simple format (origin number, destination number and number of trips between origin and destination). Different OD-matrices were established for different time periods: an AM peak period (5-11), a PM peak period (15-21) and a 24h-period. Without assignment to a network these data already provide interesting images of movements. Figure 12 shows tower locations in Ivory Coast and movements of mobile phones between tower locations for Ivory Coast and the Abidjan area. These figures already provide a rough idea of the road network and major flows. However, a transport model is needed to gain insight into flows on the road network.

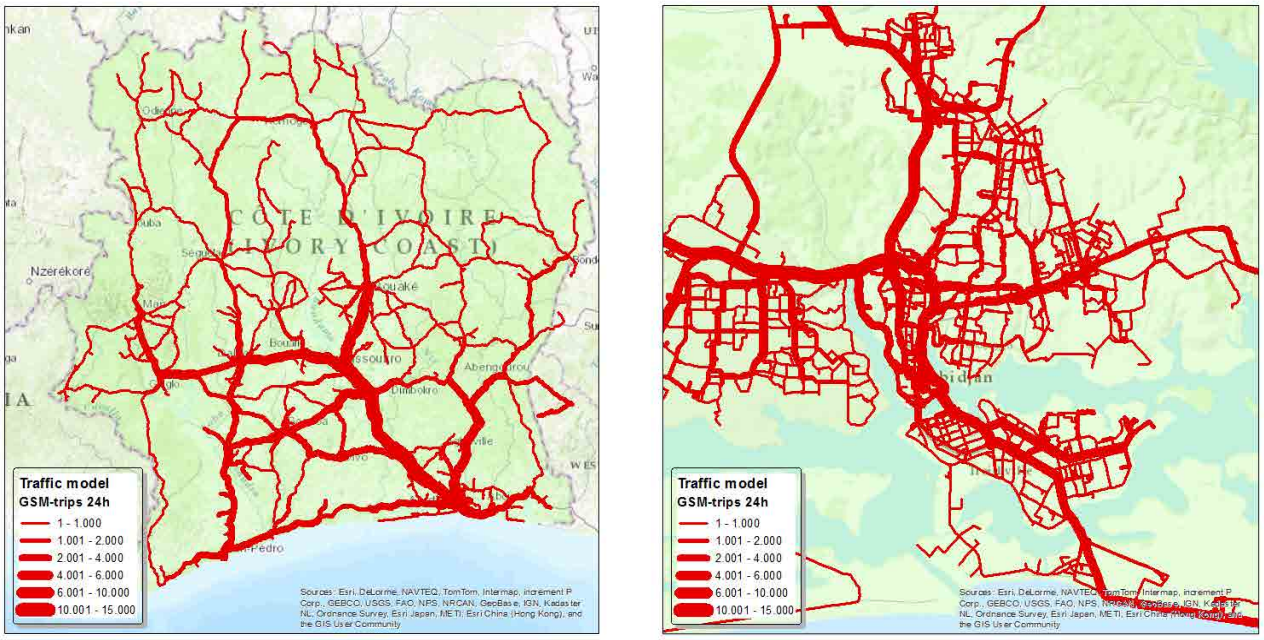


Figure 13: Traffic model for 24 hour period for Ivory Coast (left) and Abidjan Area (right)

Assignment

We used OmniTRANS V6 software to assign OD-matrices to the road network. A simple all-or-nothing assignment technique was used by which all trips on an OD-relation are assigned to the calculated shortest route in time between the origin and the destination. More sophisticated assignment techniques are available, which take into account that different routes may be chosen because of congestion, but since the use of mobile phone data for transport modeling is the central theme of this paper and more data would be needed like the road capacity, we felt that the use of more sophisticated assignment technique was beyond the scope of this research.

Results and interpretation of the transport model

Figure 12 shows assignments of the OD-matrices based on the GSM data for a typical 24 hour period for Ivory Coast and the Abidjan area. All major, national and urban transportation corridors are immediately visible from these plots. Also, the comparison between the linear movements between cell towers in figure 12 and the assigned movements in Figure 13 provides a clear impression of the added value of assigning the movements to the road network. For purposes of readability we have decreased the level of detail in the figures in this paper. The original model plots allow much more detailed analysis of volumes on road links, in both directions of roads.

Figure 14 shows traffic assignment on the network for the morning peak period in the greater (left) and central Abidjan area (right). As can be seen from all assignments the absolute flows, or traffic volumes on the network are very low and do not represent real traffic volumes on the network, since they are based on a selection of the sample provided for this study. Still, as a relative measure all figures show roads with more and less dense traffic. To make the transport model suitable for identifying and exploring current or future transport problems in Ivory Coast, an accurate assessment of absolute traffic flows on relevant parts of the network is necessary. In the following discussion we will deal with what must be done to overcome the limitations of the current model.

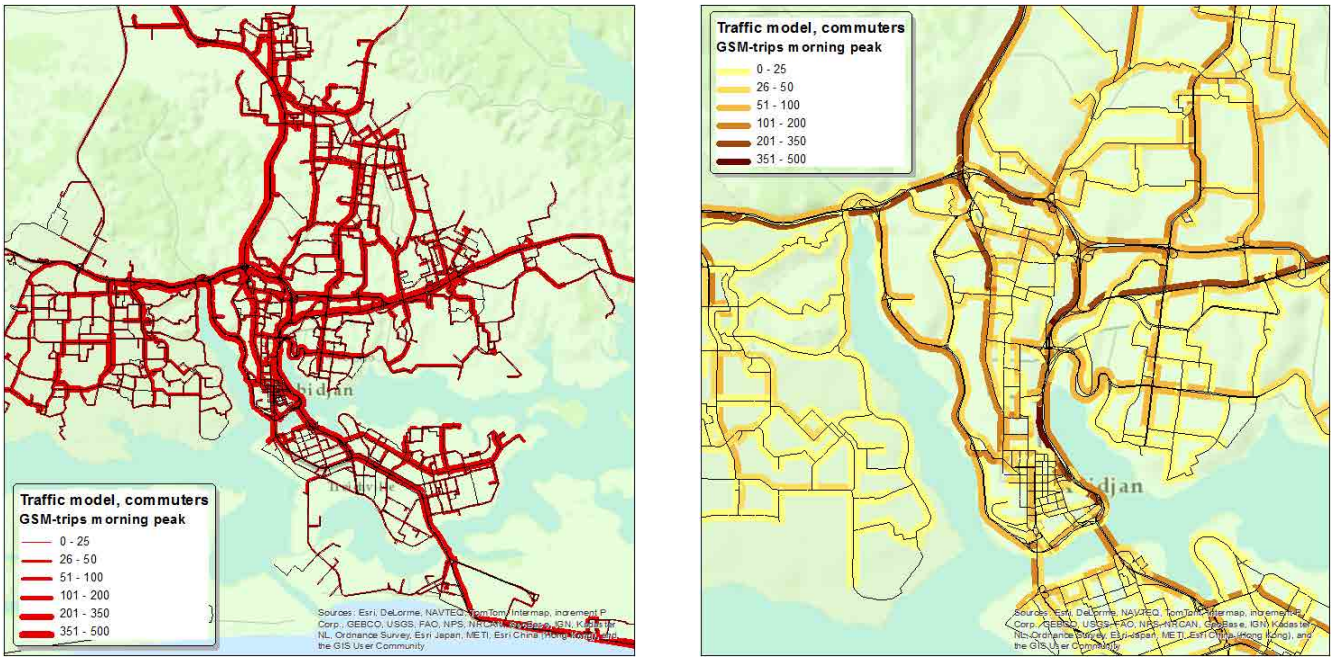


Figure 14: Traffic model for morning peak period for greater Abidjan (left) and Abidjan Centre Area (right).

Discussion

Now, are we where we wanted to be? Do we have a definite traffic model for Ivory Coast and Abidjan, which can be used for planning purposes? The answer to these questions is 'yes' and 'no'. As a proof of concept the current exercise is definitely successful. It is highly rewarding that we were able to create a transport model for an area where we, as researchers, have never been, and for which data were solely obtained from the internet and from a completely new source (GSM call traces). However, the transport model is not yet finished, and this is mainly due to a number of remaining limitations in the data which were available. The good news is, that all these limitations can be solved, but the effort to do this varies.

First, to make an estimate of actual and validated traffic flows on the network, the current model values should be augmented or weighed by a factor. This factor will depend for instance on the market penetration of Orange operated cellphones, cellphone ownership and cellphone usage in Ivory Coast etc. As these levels may differ throughout the country, the weighing-method should take into account local differences. Techniques to do this are already available since also in transport modeling using conventional data, OD-matrices are calibrated from traffic counts. A set of reliable traffic counts in the network should therefore suffice to establish augmented OD-matrices to describe not just the traffic based on mobile phones traces, but the total traffic flows.

A second, more serious limitation in the present model is that it does not make a distinction between the different transport modalities. To achieve such a distinction filters and algorithms must be developed to detect and estimate different modes of travel from the data. The currently available data would hardly allow a distinction of modes that is based on travel speeds. There are two reasons for this. Particularly in congested urban areas different modes tend to travel more or less at the same speed. Therefore, in dense urban areas travel speed is hardly a distinguishing factor for different modes of travel, maybe with the exception of walking and rail. More importantly, the current set of data does probably not allow for a reasonably accurate estimate of travel speeds anyway. This is primarily due to the absence of detailed cell-tower information. More precise information on the radius and directions of the cell tower allows a much more precise estimate of the location

of a cellphone at a certain time and consequently of travel speeds of cellphones. These data are available or can be made available, but were not provided by Orange for the current competition. Furthermore, as car ownership in Ivory Coast is still limited (16 vehicles per 1,000 inhabitants) and a large part of the population (48%) is younger than 14 years, individual car use in the total population will be very limited compared to developed countries. We did not find any usable statistics on the use of cars, busses, trains, taxi's, vans, bicycles, motorcycles, walking etc. However, for modeling purposes, some basic statistics on modal split can be obtained from the same traffic counts as are needed to augment the traffic present in the model and could highly increase the quality of the model.

If we could overcome the above mentioned problems, and it seems absolutely feasible that this can be done, then we will have a model that can be used both for an accurate assessment of the current traffic situation in CI and Abidjan and for forecasting purposes. For forecasting purposes one needs to implement future planned projects in the model and specify the expected general growth in mobility for all modes. Once we have an adequately augmented mode specific OD-matrix, the model immediately allows other calculations and forecasts for environmental impact analysis, such as greenhouse gasses, NOx and PM10 emissions. Of course, for these purposes additional statistics on the Ivory Coast's vehicle park must be incorporated.

5. Discovering social borders

The second part of this project deals with actual social borders within Ivory Coast. The objective of this analysis is to determine the influence of social behavior in a territory [8], and to compare the administrative borders against emerging borders that can be derived from the people movements.

In general, we determine groups of regions such that the inner movements within those groups are more frequent than the movements towards other groups. We thus propose a general framework based on the following steps: (1) Using the same method used in the O/D matrix we create a spatial tessellation whose regions will serve as spatial references; (2) the handovers are used to generalize the movements between the spatial tessellation; (3) they are then coded by means of a directed weighted graph; and (4) the graph is then analyzed to extract the communities within it. In Figure 15 we show the results using different parameters to obtain high and low level communities, we compare them to the political map (top) and the administrative map (bottom).

Comparing the political map with the high level community map we see both similarities and differences. E.g., *region Sud-Comoe* is very similar in both maps except for a small area in the west which is added to this area, reducing *Region de Lagunes* and moving the border towards Abidjan city. Another example, the region in the middle (light green) is completely new extending the influence of the capital *Yomoussoukro* to the south. This suggests that the city of *Bouafle* is socially connected to the capital and less to its own political region. This is highlighted by the choice to invest in a main connection between the two cities[1]. Such a connection is also suggested for *Divo* city and the capital. The present analysis can be helpful in deciding where to invest considering the social needs and optimizing ROI.

[1] http://www.fratmat.info/component/content/article/69-slide/19727-yamoussoukro-bouafle-laxe-routier-reliant-les-deux-localites-est-en-rehabilitation.html. Article of 26 July 2012

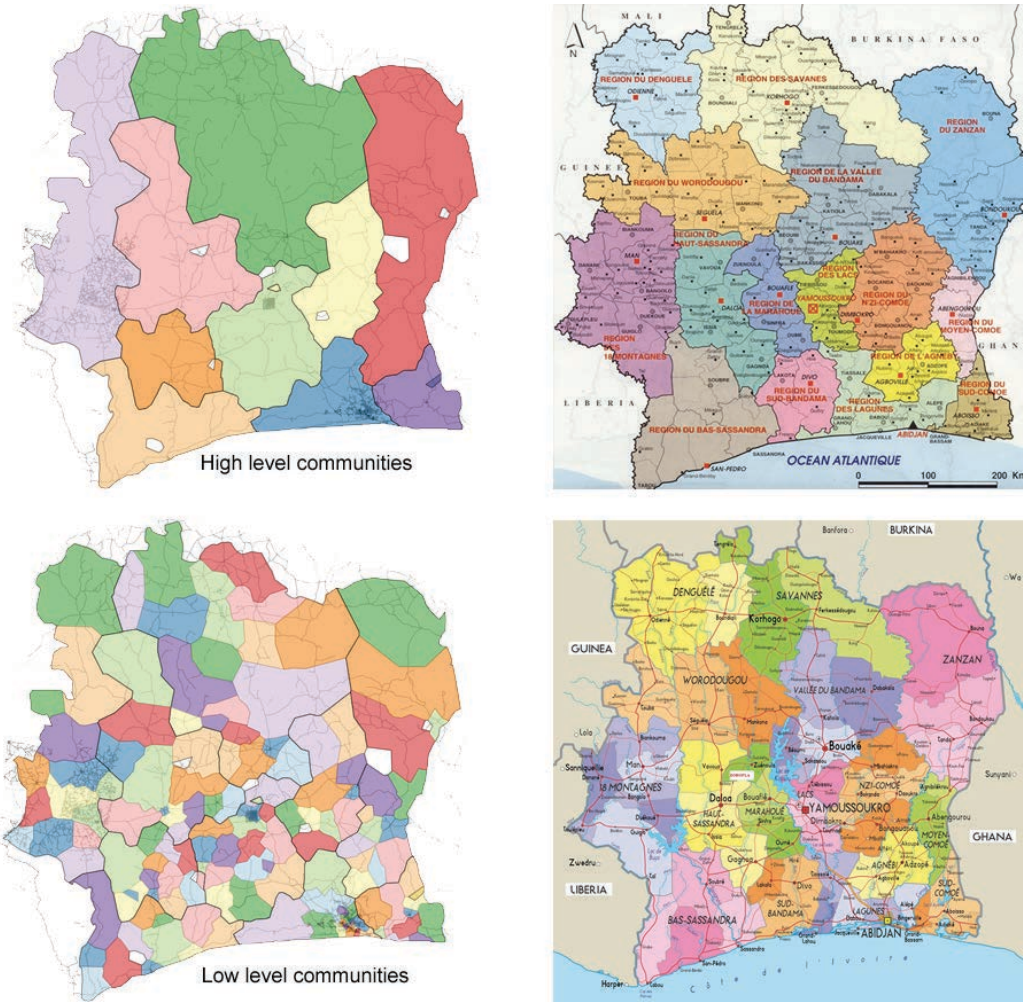


Figure 15: Results of social Borders with at high level (top) and low level (bottom) compared to political and administrative maps.

The comparison between the low level communities and administrative regions shows very similar borders. Only a few regions are divided into more detailed areas. This suggests, as expected, that the administrative zone take the social behavior of the population more into consideration. When we focus our analysis on *Abdjian* city, which is separated into small areas, Figure 16 shows a fine distinction of quartiers of the city.

Figure 17 shows clearly how the economically wealthy district (light green) in the middle is socially closed, and separated from the north quartiers. The same happens for the embassy district and *Cocody* university. As was the case for the high level community, this analysis is a way to consider social borders in urban areas. Such an analysis may be used for planning or redevelopment of the city.

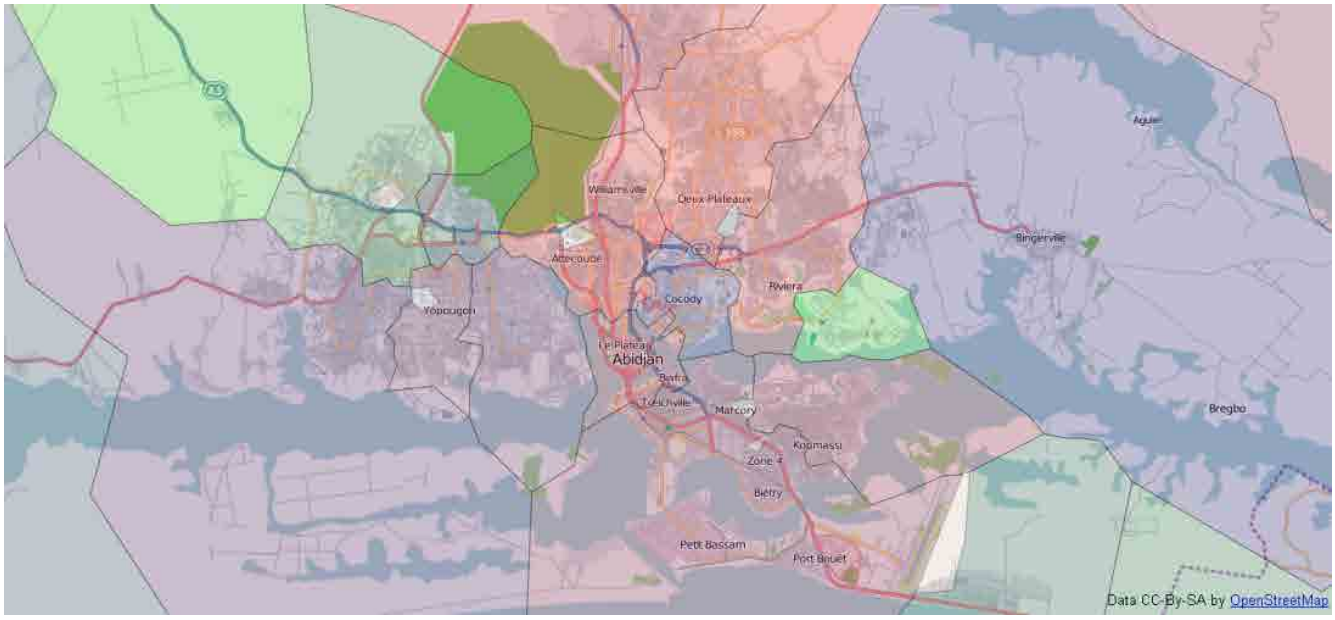


Figure 16: Focus on Abidjian city w.r.t. the low level communities found.

Due to a lack of local expertise we highlight only some clear examples which are immediately evident from the map and visible facilities. We are sure that this tool in the hands of local experts can help to understand the social behaviors at both levels.

6. Analyzing spatio-temporal users behaviors

The availability of huge quantity of mobile phone data stimulates more and more challenging questions that goes beyond the observation of people movements through the territories. What can be done is to deduct some "personal behaviors" starting from the footprints lived by the users. Where they make a call and when, how many time they make a call from a certain location, can be a valid indicator to draw some different user profiles?

In this section we apply an inductive technique on the CDR data based on the Self Organizing Maps to bring out different spatio-temporal user behaviors. In particular, by defining some temporal constraints to filter out the data, we want to identify typical *global profiles* like Resident, Commuters and Visitors of the town of Abidjan. According to common sense and the definition given by The World Tourism Organization[2], we can formalize these categories as follows:

Stated *A* and *B* are two general areas,

- A person is *Resident* in an area *A* when his/her home is inside the *A*. Therefore the mobility tends to be from and towards his/her home.
- A person is a *Commuter* between an area *B* and an area *A* if his/her home is in *B* while the work/school place is in *A*. Therefore the daily mobility of this person is mainly between *B* and *A*.

[2] Wikipedia - Tourism definition. http://en.wikipedia.org/wiki/Tourism

- A person is a *Visitor* in an area *A* if his/her home and work/school places are outside *A*, and the presence inside the area is limited to a certain period of time that can allow him/her to perform some activities in *A*.

To limit the analysis on Abidjan city area, we selected the maximum set of Voronoi cells included in the administrative borders, as shown in Figure 17. This area represents the spatial dimension used to filter out the call events of interest.

Starting from the set of calls made by each user during the 2 weeks of observation, a *Space Constrained Temporal Profile* of each user is reconstructed. According to the definition given in [9], a *Temporal Profile* is a vector of call statistics according to a given temporal discretization, and the *Space constrained Temporal Profile* is a Temporal Profile where only the calls performed in the cells contained within the a certain area, are considered.

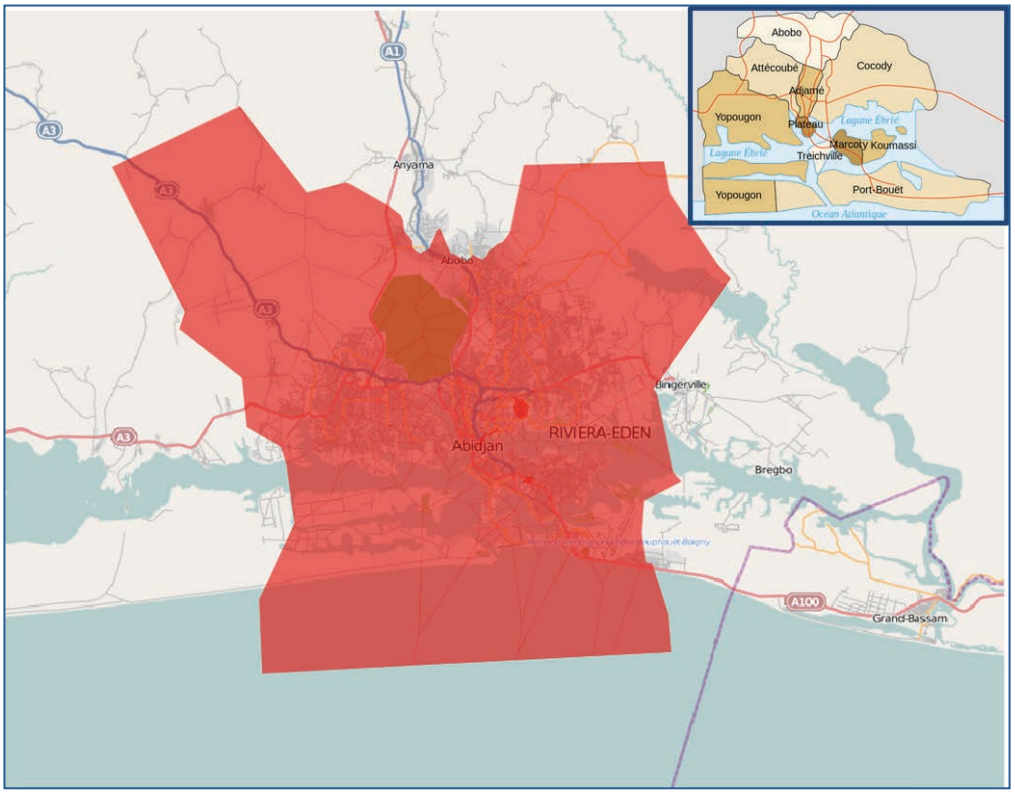


Figure 17: Spatial selection of Abidjan area: Voronoi of the GSM coverage.

Figure 18 shows an example of how the vectors of aggregated Weekly Temporal Profiles (*c* and *d*) can be extracted from the entire set of calls of User 1 (a) and User 2 (b). The values correspond to the number of calls performed by the users during a day and represent the statistics mentioned in the definition. The temporal profiles by week are obtained by flatten the calls over the weekdays in a "representative week". The calls over the same day through the weeks are summed up.

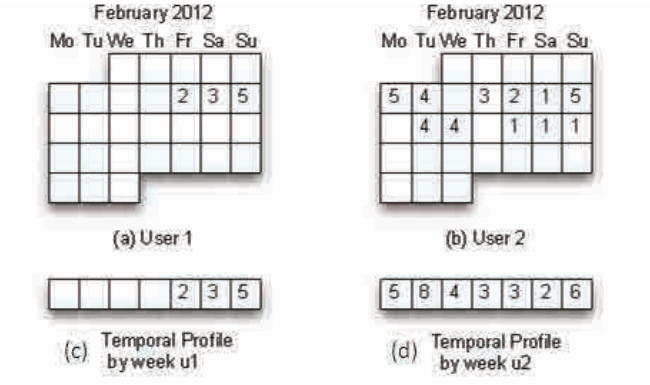


Figure 18: Example of temporal profiles computed starting from the original calls.

In the case study, the spatial component consists of a selection of users and corresponding calls made in the area of Abidjan described by the GSM coverage in Figure 17. The time projection is built by performing 2 temporal operations:

1) the aggregation of the days in weekday and weekend slots, and
2) the splitting of each slot in 3 time bands suggested by the temporal distributions of the calls depicted in Figure 19 representing 3 interesting time windows during the day:
   t1 = [00:00:00 - 07:59:59], Early in the morning when people are usually still at home;
   t2 = [08:00:00 - 18:59:59], Mead day when people are out for work/school or other activities;
   t3 = [19:00:00 - 23:59:59], Late in the evening and night when people are back to home

We call this instantiation *Multi-Dimensional Temporal Profile.*

Figure 19 exemplifies the construction process of the Multi-Dimensional Temporal Profile. Starting from the calls along the days, the presences over t1, …,t3 are computed, then these presences are aggregated over the weekdays and the weekend summing up all of them. The result is a sort of compact representation of the user's behavior measured by his calls. The dataset is then processed by using the SOM algorithm in order to extract the typical global profiles. A SOM is a type of neural network based on unsupervised learning that produces a one/two-dimensional representation of the input space using a neighbourhood function to preserve the topological properties of the input space [10].

In our case, the SOM output is a set of nodes representing groups of users with similar temporal profiles. Figure 20 (b) shows an example of global profile which includes 359 users. Each small square is a temporal slot, the columns are the sequence of weekdays and weekends (for 2 weeks), and the rows identify t1, …,t3. The presence of the individuals is indicated by the colour intensity of each slot (Figure 20 (a)): the black slot indicates no presence, more the colour intensity, the higher the presence in term of number of days. We can interpret this result as a typical profile of people who live outside a city (no presence early in the morning and late at night during the weekdays) but that go daily in the city (maybe for work or school) during typical working hours and that never go during the weekend except sometimes in the night (maybe for leisure or social events). The result of the SOM computation on Abidjan is shown in Figure 20(c). Each block represents a global profile with the same semantics of the personal profiles described above.

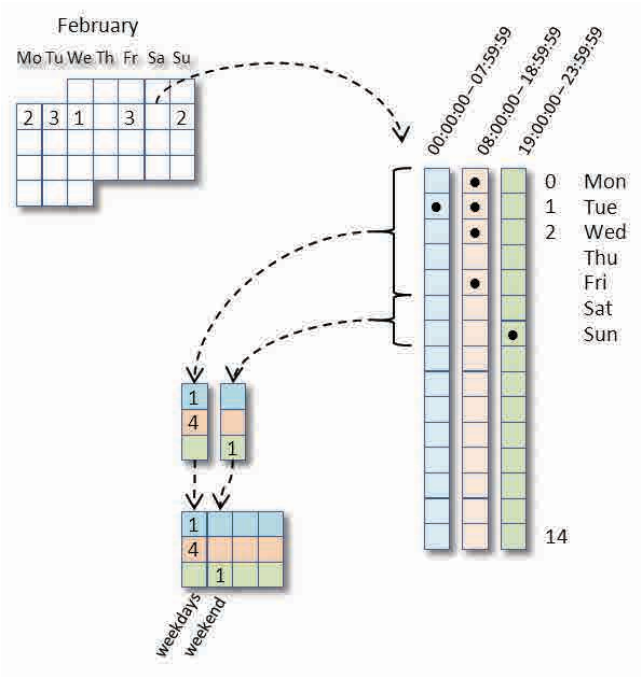


Figure 19: Reconstruction of the *Multi-Dimensional Temporal Profile* used in the Abidjan case study.

We extracted a small set of global profiles to avoid the over specialization of the models: this 3x3 matrix represents the best representation in term of both significance and compactness. Each box synthesizes the global profiles of presence in the city of Abidjan. Each profile is located in the map so that it has the most similar profiles as its neighbours. From the map two kinds of profiles clearly emerge, that take after the Resident and Commuters. In particular the light blue box at the top left corner (made by the contribution of 31610 instances) represents a typical Resident profile: an individual belonging to it is ever present in Abidjan during the two weeks both in the morning, the evening and the night. From here on, we will refer to this profile simply as “Profile R”.

The violet box in the middle line at the right (made by the contribution of 36750 instances) represents instead the typical Commuter profile. There is a strong presence in the weekdays during typical working hours and a slight presence during the weekends, in the morning and night hours. From here on, we will refer to this profile simply as “Profile C”.

The other global profiles can be assimilated to Resident and Commuters with small variability of presence. The fact that there is not an evident and big set of commuters can be explained with the fact that nearby Abidjan there are not closed towns that could generate a daily commuting phenomenon of workers and students. We may believe that the profile of commuters is mainly composed of people coming from just outside the virtual border we drawn.

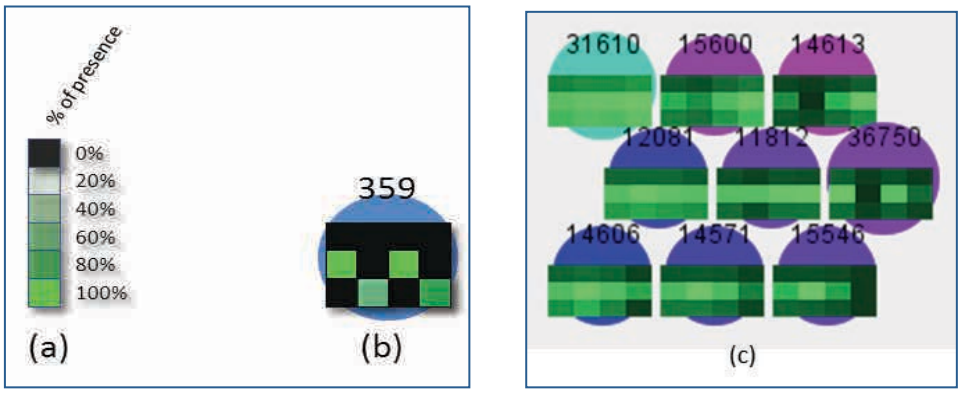


Figure 20: (a) Intensity scale for the presence. (b) Example of global profile which includes 359 users: full presence during the weekdays and half presence during the weekends during the night. (c) Global user profiles in the city of Abidjan.

A profile of Visitor does not come out from the dataset. This can be explained by the fact that the data are only about Ivorian citizens and thus the contribution of foreign tourists is missing, for example.

Figure 21: Preferred Location of the Resident Profile. The home locations (Left) and work locations (Right).

What emerges from the data leads us to believe that the particular geographic location of Abidjan and the nearby environment (the desert isolates the town and there is not an efficient road network), does not encourage the short and mid-term visits of the Ivorians, or maybe there is not a real national tourism phenomenon towards Abidjan. To confirm this hypothesis, however, we would need a wider period of observation (three or more weeks).

Nevertheless, we may validate our profiling by going back to the original call data and studying the spatial distribution of the calls using the formalism of the density map. The idea is to investigate “where” the two preferred locations are sited for Profile R and Profile C.

According to the literature, we can assume that the first preferred location identifies the Home and the second the Work (previously identified as L1 and L2, respectively). From this assumption we built the corresponding density map, where the density is computed as the number of calls that have been made by the users in these locations. The colour scale indicates that the higher the gradation of red, the greater the number of calls that start from that cell.

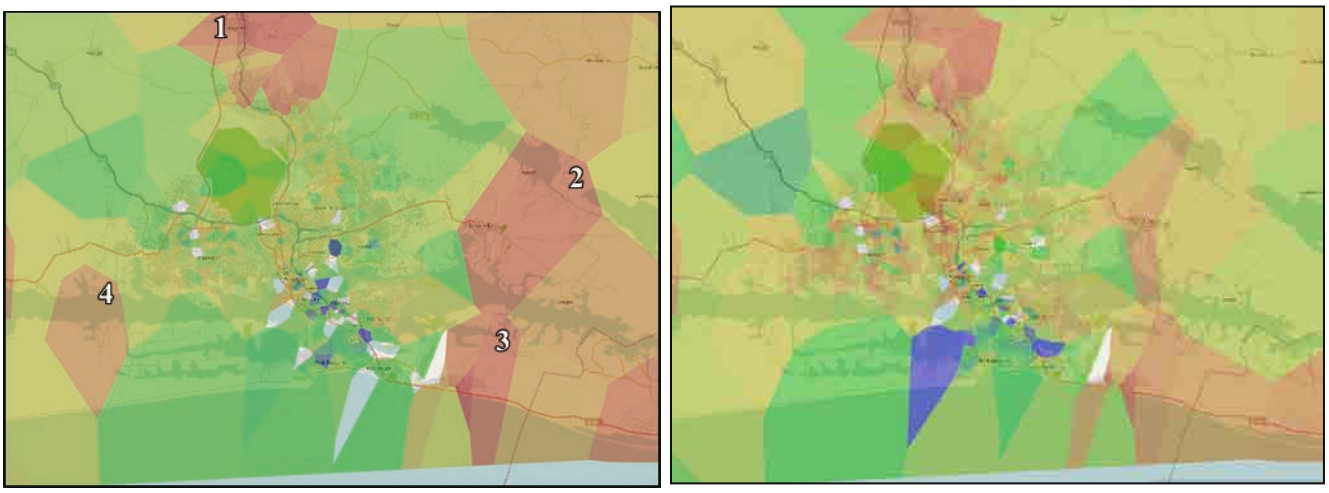


Figure 22: Preferred Location of the Commuter Profile. The home locations (Left) and their work locations (Right).

The map in Figure 21 is obtained from the call behaviours of the Profile R. The map on the left represents the density of calls in the Home locations, while the map on the right represents the density of calls in the Work locations. The behaviour that emerges is consistent with the definition of resident given above because the mobility is mainly inside the borders of the town: both the home and work places are located inside the borders of the town.

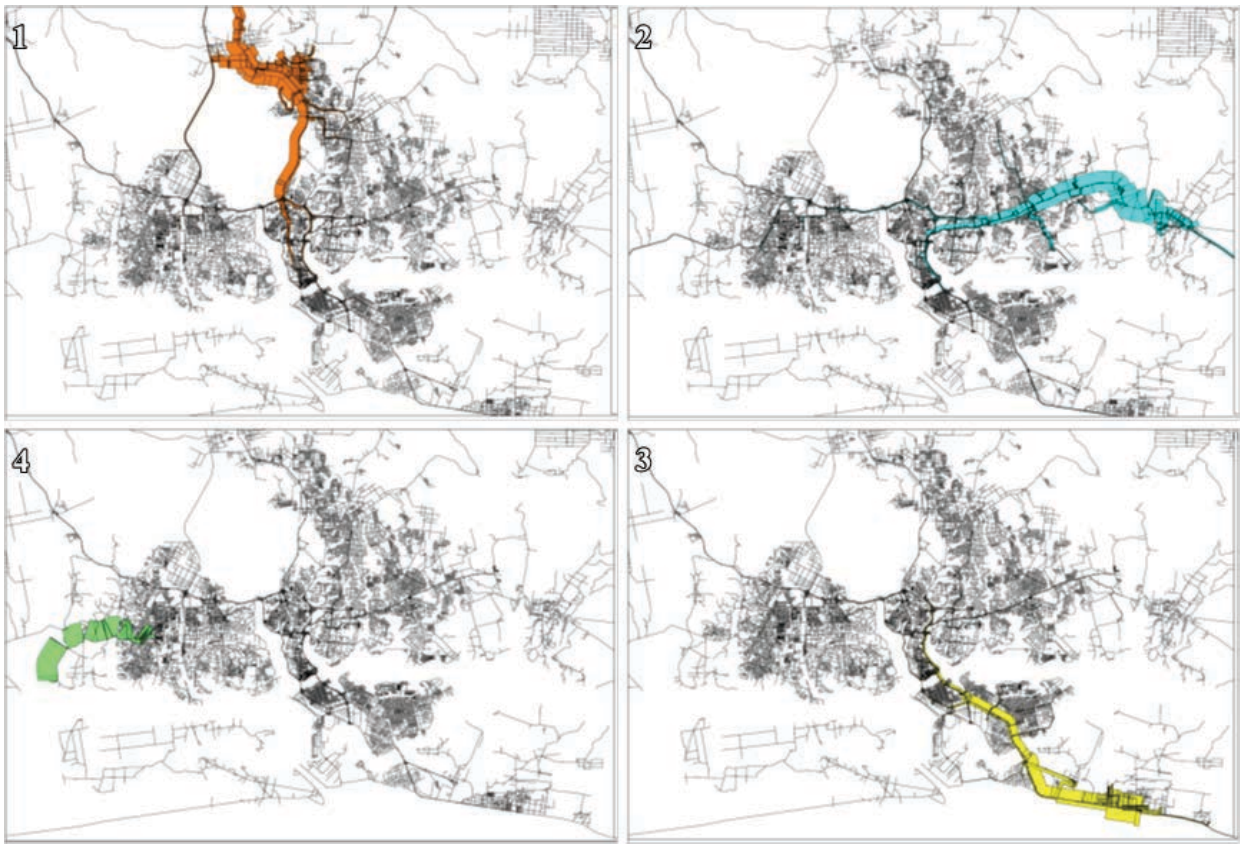


Figure 23: Traffic model for morning peak period in the roads that connect the leaving areas of Figure 22 (left) to the city centre.

The map in Figure 22 is obtained from the call behaviours of the Profile C. As in the previous case, the map on the left represents the density of calls in the Home locations, while the map on the right represents the density of calls in the Work locations. The behaviour that emerges is consistent with the definition of commuter given above, in fact the Home locations are denser outside the borders of the town, while the majority of the working areas are located inside the borders. The transportation model described in Sec. 4d may help to see this flows: considering the four locations in Figure 22 (left) where the density of homes is higher (numbered from 1 to 4), the traffic volumes on the roads in the morning is selected producing the four maps in Figure 23. These maps actually point out that there are notable volumes of traffic on these roads leading into the city confirming the coherence between the results obtained in Systematic movements (done on a subset of users) analysis and the discovery of commuters (done on the whole dataset).

This coherence is particularly promising for the data-mining necessary to create a comprehensive traffic model capable of predicting future situations due to developments of new living areas of business districts, in addition to the discussion in section 4d. As the analysis shows that the structure of the city and the spatio-temporal segmentation of it's population can be revealed from 'just' the call detail records, the practical feasibility of creating a useful traffic model for a developing country looks favourable.

7. Conclusions

The present study shows that even with limited data from GSM traces, as were provided by Orange for this study, it is possible to derive valid information on systematic mobility behavior of people between frequently visited locations for areas that lack information on mobility. From these mobile phone data, origin-destination tables can be created for a chosen geographical area, which can then be used as an input for a transport model of the area. The fact that this can be done, overcomes one of the serious hurdles that until now have impeded the use of transport models in many developmental countries: lack of data. Because also other developments, like the ever-increasing availability of reliable transport network information from sources like Google Earth, Open Street Maps, etc. crucial input for transport modeling is now available for practically any location in the world. It is therefore absolutely worthwhile to take this proof of concept one step further, and create and validate a transport model that can indeed be used by public authorities, engineering firms, investors etc. in developmental countries. In this paper we discussed the most important steps which need to be taken.

At this point it should be stressed, that overcoming the current data limitations requires only limited resources and can lead to a very cost-efficient step in the development of transport models for countries and cities currently lacking such models. Applications like the one presented here also provide opportunities to promote and advance sustainability and address environmental and health concerns. Furthermore, frequently, the data used in existing transport models are quite old, due to the high cost of new data collection and matrix calibration. The methodology presented in this paper allows for a much higher update frequency, thus allowing more up to date models. In rapidly changing environments, as are evident in the current era, the need for 'real-time' data and insights for decision-making is ever increasing.

One of the frequently mentioned obstacles for further or commercial use of GSM data traces, concerns the privacy of mobile phone users. We feel that this issue can and must be tackled and is currently being tackled. The application of data as presented in this study does not concern individual behaviors, but concerns mobility

streams or traffic flows created by *groups* of people. To calculate these flows anonymized individual, temporal data are needed, but the accessibility of these data can be highly restricted. Already in several countries platforms are created within or next to mobile phone operators that make GSM data traces available for analysis and preserve the privacy of mobile phone users.

References

[1] Ana-Maria Olteanu, Roberto Trasarti, Thomas Couronné, Fosca Giannotti, Mirco Nanni, Zbigniew Smoreda and Cezary Ziemlicki. ***GSM data analysis for tourism application****. In 7th International Symposium on Spatial Data Quality (ISSDQ 2011).*

[2] Roberto Trasarti, Fabio Pinelli, Mirco Nanni and Fosca Giannotti. ***Mining Mobility User Profiles for Car Pooling****. In ACM International Conference on Knowledge Discovery and Data Mining (KDD), 2011.*

[3] Balázs Cs. Csáji, Arnaud Browet, V.A. Traag, Jean-Charles Delvenne, Etienne Huens, Paul Van Dooren, Zbigniew Smoreda, Vincent D. Blondel. ***Exploring the Mobility of Mobile Phone Users****. Report on arXiv: 1211.6014 [physics.soc-ph].*

[4] Fosca Giannotti, Mirco Nanni, Dino Pedreschi, Fabio Pinelli, Chiara Renso, Salvatore Rinzivillo, Roberto Trasarti: ***Unveiling the complexity of human mobility by querying and mining massive trajectory data****. VLDB Journal Special issue on Data Management for Mobile Services (2011).*

*[5] Roberto Trasarti, Salvatore Rinzivillo, Fabio Pinelli, Anna Monreale, Mirco Nanni, Chiara Renso, Dino Pedreschi,Fosca Giannotti****: Exploring Real Mobility Data with M-Atlas****. PKDD 2010*

[6] M. Faloutsos, P. Faloutsos, and C. Faloutsos. ***On power-law relationships of the internet topology****. ACM SIGCOMM Computer Communication Review, 29(4), 1999.*

[7] Balázs Cs. Csájia, Arnaud Browetc, V.A. Traagc, Jean-Charles Delvennec, Etienne Huensc, Paul Van Doorenc, Zbigniew Smoredae, Vincent D. Blondelc. ***Exploring the mobility of mobile phone users****. Physica A: Statistical Mechanics and its Applications Journal 2013.*

*[8] Rinzivillo, S, Mainardi S, Pezzoni F, Coscia M, Giannotti F, Pedreschi D. 2012****. Discovering the Geographical Borders of Human Mobility****. KI - Künstliche Intelligenz.*

[9] B. Furletti, L. Gabrielli, C. Renso, S. Rinzivillo. Beijing, China : ***Identifying users profiles from mobile calls habits.*** *UrbComp 2012.*

[10] Kohonen, T. ***Self-Organizing Maps.*** *Information Sciences 2001*

[11] Hensher, D.A. & Kenneth, J. (eds) ***Handbook of Transport Modelling****. Pergamon, 2000.*

[12] Ortúzar., J. and L. G. Willumsen. ***Modelling Transport****, 4rd ed., John Wiley & Sons, 2011.*

[13] Calabrese, F. Di Lorenzo G, Liang Liu & C. Ratti. ***Estimating Origin-Destination flows using mobile phone data****. Pervasive Computing, 4, 36-44. 2011*

Crowdsourcing Physical Package Delivery Using the Existing Routine Mobility of a Local Population

James McInerney, Alex Rogers, Nicholas R. Jennings
University of Southampton, Southampton, SO17 1BJ, UK
{jem1c10,acr,nrj}@ecs.soton.ac.uk

ABSTRACT

In Ivory Coast, as in many developing countries, half the population lives in rural locations, where access to essentials such as school materials, mosquito nets, and medical supplies is restricted. Reaching remote areas is challenging by conventional methods (i.e., using the road network). We propose an alternative method of distribution in which the existing mobility habits of a local population are leveraged to deliver aid. Specifically, we envision that individuals may be asked to bring a package containing aid between two locations that they already usually visit in their daily lives, and that arbitrarily large distances may be covered by chaining together the routine mobility of participants.

Using real cell tower data measuring the mobility patterns of 50,000 randomly sampled individuals living in Ivory Coast (from the Orange dataset), we firstly confirm the feasibility of this method. Secondly, we address the associated technical challenges of this vision in the areas of route optimisation and machine learning. We formulate the routing problem as a Markov decision process (MDP), and use a Bayesian approach to learn mobility habits from very sparse individual cell tower data. In 10,000 simulations, we find that packages take on average 28.2 days to travel an average of 373 km from the source to (rural) destination locations, representing a 83% reduction in total delivery time versus route planning that minimises just the number of participants in the solution path, and replacing 2 days of direct driving time.

1. INTRODUCTION

In Ivory Coast, as in many developing countries (e.g., Ghana, Liberia, Nigeria), half the population lives in rural locations [5], where accessibility to school materials, medical supplies, mosquito nets, and clothing is restricted.

In response to these limitations, alternative methods of aid distribution have emerged in recent years. The Pack For a Purpose is a non-profit organisation that asks tourists who already have a trip planned for one of 47 developing countries to bring small items (e.g., pencils, deflated soccer balls, stethoscopes) in their spare luggage capacity[1]. Local accommodation is coordinated, as part of the scheme, to pass the items on to local schools and hospitals. Another scheme is Pelican Post, which asks donors to send books by post to developing countries[2]. These are promising schemes, but, firstly, they fail precisely when conflict occurs (as do government-run schemes: in Ivory Coast, post-electoral violence in 2011 delayed the distribution of mosquito nets by several months [31]). Secondly, in the case of Pack For a Purpose, they only reach destinations attractive to tourists, so are not applicable everywhere.

[1] http://www.packforapurpose.org
[2] http://www.pelican-post.org

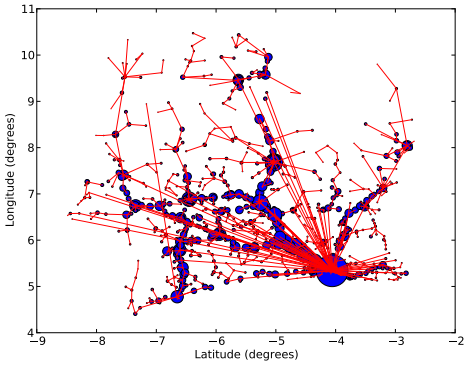


Figure 1: Minimum spanning tree between cell towers in Ivory Coast, where connections are defined by common visitors in the Orange dataset, and the size of node represents *betweenness centrality* (i.e., the number of times the location appears in the shortest path for all possible delivery paths)[19].

They also rely on direct outsider support, when it is arguably preferable to empower local populations wherever possible.

In this work, we propose a new distribution method that uses the natural mobility of a region to move physical packages between two specific locations. In more detail, we wish to take advantage of the pre-existing mobility routines of a set of local participants by asking them to pick up a package from one exchange point (at a location that they normally visit, at a time that they normally visit it) and then drop it off at another exchange point (which is also part of their regular mobility). By chaining together the mobility of several participants, we may cover a large area, possibly a whole country, without having to deploy more expensive and time consuming infrastructure. For example, if we wish to deliver a package of mosquito nets from the capital, Abidjan, to a rural village in the west of Ivory Coast, we may first ask Ibrahim, who lives in Abidjan, but often visits his sister in Gagnoa (a city in the west) on weekends, to pick up the package near his house and drop it off near his sister's house in Gagnoa, when he is there anyway. We may then ask another participant, Phillipe, who lives in Gagnoa, but who works in Taï national park on weekdays (driving past the village each day without realising) to drop the package off at the village on his way to work. In this way, the participants do not have to significantly change their schedules or travel long distances that they would not have otherwise travelled.

While potentially appealing, this vision of crowdsourcing physi-

cal package delivery faces two significant technical barriers in learning and optimisation, arising from uncertainty over human location behaviour (i.e., the delay that each participant introduces between steps in the chain).

In learning, the mobility data of each individual is sparse, limited in duration (i.e., we may only have a few week's worth of data from each participant) and, crucially, cell tower readings are taken only when a call or text message is exchanged from the phone, so there are large periods when no location of an individual is registered at all. Yet, most existing methods for mobility prediction rely on large quantities (covering several weeks) of fairly continuous stream of location readings (either from GPS or constant cell tower monitoring) [26, 27, 2]. Therefore, we seek to develop robust methods of learning individual mobility models from cell tower records spanning only short periods of time with sporadic observability.

In optimising which participants to ask in the chain, even if an accurate model of individual mobility could be learnt for all the participants, the expected delay random variable between stages in the package's journey is unbounded. This makes it infeasible to optimise the selection of participants and the package route (given a specific delivery problem specifying the start time, source location, and destination location). In general, routing under delay uncertainty is a #P-hard problem to solve optimally [20].

Therefore, to address these two key technical challenges, and to assess the feasibility of crowdsourcing aid distribution in general, we make three contributions:

- We use the Orange dataset of the mobility of 50,000 people to show that peer-to-peer package delivery is feasible under three key criteria. Firstly, the size of participant pool only needs to be of the order of several thousand to get at least an 80% coverage of the country (with total area 320,000 km^2). Secondly, each solution path (i.e., chain of participants to deliver on package) is between 2-4 people. Thirdly, these requirements are only mildly worsened when considering only rural destinations for delivery.

- We advance the state of the art in route planning in delay networks by providing an approach that works well with the uncertainties caused by routine human behaviour. Specifically, we show that an exact and tractable solution is possible when using a mobility model belonging to the broad class of *temporal periodic* prediction models. Under this assumption, we show that we can formulate the problem as a Markov decision process (MDP) in which the number of states grows linearly in the number of locations, making the overall algorithm polynomial when using a standard MDP solving method (e.g., linear programming, policy iteration) [23]. Using our approach, simulations indicate that source-to-destination delivery time is reduced by an average of 83% compared to choosing the shortest path (in number of participants).

- To provide realistic transition probabilities to the MDP[3], we present an approach to learning the mobility patterns of 50,000 people from very sparse observations, and formulate that mobility model as a repeated Bernoulli trial that may be used directly in an MDP. Our approach is 125% better than a first-order Markov model, which was previously shown to perform well on sparse location data [17, 1].

[3]N.B., the transition probabilities in the MDP are *not* the same as the transition probabilities of the mobility of any individual participant.

The rest of the paper is structured as follows. First, in Section 2, we consider previous work related to the problem of learning of human mobility patterns and optimising under uncertainty of human behaviour. In Section 3, we present our approach, starting with how we make optimal decisions with respect to the choice of participants and locations for any given delivery problem in Section 3.1. Then, in Section 3.2, we present a learning model that deals with sparse observations. In Section 4, we evaluate the feasibility of the scenario before evaluating our approach to learning and optimisation against several state of the art benchmarks in Sections 4.2 and 4.3. We draw conclusions and outline future work in Section 5.

2. RELATED WORK

The idea of distribution using the natural mobility of a group of people is a reoccurring theme in content distribution using mobile *ad-hoc* networks. Keller et al. (2012) used physical bluetooth proximity data from the mobile phones of a group of people, to initiate exchanges of songs between individuals, but without considering prediction or multi-hop routes (i.e., going via one or more intermediaries) [12]. Vukadinovic et al. (2009) proposed a queuing model of the flow of pedestrian crowds to distribute content among mobile phones [33]. This, and other crowd mobility models (e.g., [16]) attempt to capture short term movements of individuals in crowds, which is a distinct and different problem to extracting *routine* mobility patterns. Specifically, in our work, there is a direct line of assumptions going from the raw historical data to decision-making about distribution (via learning and the formulation of transition probabilities in the MDP) that is not present in such work. Furthermore, these approaches rely on the fact that content may be copied and can exist concurrently on multiple devices, making them less applicable for our routing problem.

Another type of diffusion that attracts research interest is the study of the spread of infectious diseases [14]. Epidemiologists look at the mobility dynamics of a population to identify source regions (from which disease is spread), and likely importation regions (to which disease is spread). For example, Wesolowski et al. (2012) used one year of cell phone data of millions of people to model the human movement between different regions in Nairobi [34]. They considered a graph in which the weight of the edges represents the quantity of people travelling between different locations, which is not sufficient for our problem. Hufnagel et al. (2004) considered a global model of human movement using passenger numbers for flights between the 500 largest airports in the world [11]. Colizza et al. (2007) also used passenger numbers to study the spread of the H1N1 avian influenza virus [6]. Such work is concerned with *aggregate* mobility, in contrast, we are interested in *individual* mobility, because, eventually, we need to ask specific people to contribute. The same argument applies to other large scale mobility studies of populations, which are more analogous to analysing climate than to navigating hot air balloons.

The problem of robust route planning under uncertainty resembles the *Canadian traveller* problem (also known as the *bridge* problem) [21, 4], in which the costs of the edges in a graph are random variables that are observed only as the nodes are visited. The name of the problem originates from the concept of a traveller who has to plan a journey between two locations, where the costs of outgoing edges are random variables that are only observed as a graph is traversed. This differs from our problem because the Canadian traveller assumes that path costs are independent of one another, while we have dependencies between costs as well, i.e., the delay outcome of an earlier stage in the chain affects the delay of later stages. An additional difference is that we observe the ran-

dom variables, indicating delay between locations, only *after* the package has completed each intermediate stage of its journey.

Learning routine mobility models has typically been a separate problem from optimisation. Approaches range from purely temporal ([27, 28]), spatial ([10, 1]), or a combination of both ([26, 9, 8]). Existing datasets made widely available have tended to contain approximately continuously recorded cell towers or GPS (e.g., the Reality Mining dataset in which the cell tower was often recorded every few minutes, even though it contained some missing periods due to data corruption and the user turning the phone off [7]; the Nokia dataset, which used adaptive GPS sampling to keep an updated read on the individual's current location [13]). This has inspired many methods that work well on continuous location updates, which do not perform as well as their headline accuracy (when predicting future location behaviour) on sparse data. We address this issue in our work.

Finally, crowdsourcing teams of participants who function as a chain to achieve a single goal resembles the idea behind the winning entry to the Red Balloon Challenge in 2009 [22]. This work is primarily concerned with the problem of recruiting individuals and verifying their reports, which requires the design of economic mechanisms. In this work, we assume recruitment can be done beforehand by an appropriate method (i.e., we do not address it here) but we do investigate *how many* participants are required to get satisfactory results in delivery. Verification may be a problem in our scenario, however, we note that there is a certain robustness to misreporting (i.e., participants saying that they dropped the package off when they did not) compared with other crowdsourcing domains. This is because, firstly, we can verify whether a participant did indeed visit the exchange point (or final destination), since we have their location, as measured by cell phone proximity, which is harder to fake than reporting information. Secondly, we would quickly be alerted to missing or altered packages by the next participant in the chain, making the untrustworthy participant clearly identifiable. On the other hand, this robustness to misreporting fails when participants collude, but we leave this problem for future work. We note, however, that since we are not asking the participants to conduct commercial activity (we are not rewarding them and we make no profit), such misreporting may be less of a factor in practice.

3. DECISION-MAKING WITH UNCERTAIN HUMAN LOCATION BEHAVIOUR

In this section, we present our approaches towards optimisation and learning with uncertain human behaviour in the package delivery scenario. In Section 3.1, we show how it is possible to tractably find an exact optimal solution to routing under delay uncertainty, given a wide class of mobility model (which we define as *temporal periodic* models). In Section 3.2, we give more detail on our probabilistic model that is designed to function well with sparse mobile phone datasets, and provides the predictions used in optimisation.

3.1 Optimisation Problem

We formulate the optimisation problem sketched in Section 1 as a Markov decision process (MDP), which provides a principled way of making decisions under uncertainty [23]. Decisions in this scenario must specify which participants to ask to pick up the package, from where they should pick it up, and the location they should drop it off once they have it. We assume the delay between pick-up and drop-off is outside the planner's control (so we treat them as random variables here), and completely up to the participant who, when asked, does this according to his/her routine schedule.

In general, an MDP is defined as a tuple (S, A, R, T) where S is a set of states, A is a set of available actions for each state. $R(s, a, s')$ is the function that specifies the cost of doing action $a \in A$ to get from state s to s', and $T(a, s, s')$ is the probability of getting from state s to s' when performing action a. The solution to an MDP consists of an optimal policy, $q(s)$, that specifies the best action to perform for any given state s. Ancillary to this function is the value function, G, which gives the expected value for any state (given that the optimal action is performed). We consider each of A, S, R, and T in turn.

3.1.1 Set of Actions A

We assume that the planner has no direct control over the delay (it is up to the participant's schedule) but we assume that we are guaranteed to eventually reach location w, when performing action a (going to location w), and that the arrival time is revealed only after performing each action, resulting in a transition to state (v, t_v), with unknown arrival time t_v. Given the one-to-one mapping of actions and locations (specifying the destination location) we treat locations as synonymous with actions for the rest of this section.

3.1.2 Set of States S

We define the set of states S in the MDP as the set of tuples describing the possible locations and times (v, t_v) (respectively) of the package. This results in the set $S = \{(v, t_v)|v \in V, t_v = 1, 2, 3, ...\}$. We assume discrete time t to capture the required detail in the scenario without the need for more complex continuous time reasoning. However, even in the discrete time case, we see that there is unbounded number of states in S because the delay in moving between locations is unbounded. This makes the MDP intractable.

To overcome large state spaces, there are a few general approaches including Monte Carlo simulation [29], and value approximation (in which values are computed from features of the states) [35]. One time-specific approach is to truncate the range of values for t to come up with an approximation for the optimal policy [30]. However, the number of states grows as a factor of this truncation limit, so more exact approximations must be traded off with computation time. Instead, we find an exact solution under an additional assumption about the mobility model used to produce the probabilistic delays. Specifically, we show that for a large class of mobility models, namely *periodic temporal* models, the probability of delay, $pr(t_w - t_v|v, t_v, w)$ in going from state (v, t_v) to (w, t_w) is periodic in t_v. This results in an MDP with a linear number of states in the number of locations.

Theorem 1. Let S be the set of states $\{(v, t_v)|v \in V, t_v = 1, 2, 3, ...\}$ in an MDP. If $pr(v|t_v)$ is a periodic function (defining H as the number of possible values it can take) in discrete t_v ($\forall v$), then the number of states is linear in the number of locations, i.e., $|S| = H\,|V|$.

Proof: Let $pr(v|t_v)$ be the probability that a given participant is at location v at time t_v, obtained from a mobility model (which, we emphasise, describes individual behaviour and is distinct from the transition function T of the MDP defined in Section 3.1.3). Since t_v is discrete, we can repeat Bernoulli trials from the distribution $r_{d_v} \sim Bern(pr(v|t_v + d_v))$ for increasing $d_v = 1, 2, 3, ...$ until we get $r = 1$. This is a standard formulation (equivalent to repeated tosses of biased coins), with $pr(d_v|t_v) = pr(v|t_v + d_v) \prod_{d'_v=1}^{d_v-1} (1 - pr(v|t_v + d'_v))$. Since $pr(v|t_v)$ is periodic in t_v, with a maximum of H distinct values, the probability of delay, $pr(w|t_v + d_v)$, *from* any next location w (reachable from v) is also periodic for arbitrary delay d_v. Therefore, $pr((t_v + d_v) \bmod H)$

is a sufficient statistic for $pr(d_w|t_v + d_v)$ (the probability of delay d_w from w), clearly taking at most H values. Using the Markov property of MDPs, only H states are required for each location v (for arbitrary v), resulting in $H\,|V|$ states overall. □

Unlike a truncation parameter, we can easily set H for the specific application of the delay network that needs to be modelled, without bias (i.e., without underestimating the delay). For package delivery, it found it sufficient to set $H = 14$ per week, by considering the probability of a participant dropping off or picking up the package in slots of half a day. Therefore, the state space is now $S = \{(v, t)|v \in V, t \in [1, 14]\}$. We further note that it is precisely the periodic temporal class of mobility model that is most useful in predicting behaviour and planning several days in advance (as we require here), since short term spatial correlations (e.g., a participant tends to go home after visiting the market) do not have much effect beyond several hours.

3.1.3 Cost Function R, Transition Function T

The delay in going from location v to location w is the cost function $R(s, a, s')$, where $s = (v, t_v)$, $s' = (w, t_w)$, and a is the action of routing the package to w. The MDP requires a single cost for each state s and action a pair (marginalising over the destination action), yet we have many participants who can potentially perform that action (i.e., who routinely visit both v and w locations). We consider this problem next.

Theorem 2. There is always a *best person* to do any action at any state in the MDP.

Proof: Let $\{\mathbb{E}(d_{v,w}|t_v, i)|p_i \in P\}$ be the set of expected durations between locations v and w for all the participants in P, when asked at time t_v. Since the expected duration of each participant is fully determined by v, w, and t_v, the state is fully determined by v, t_v, and action w. Therefore, there exists a single participant $p^* = \arg\min_i\{\mathbb{E}(d_{v,w}|t_v, i)|p_i \in P\}$ who always minimises the cost of action/state pair $((v, t_v), w)$ (the best person for that pair). Since we defined states and actions arbitrarily, this holds for the entire MDP. □

We can now assume that the cost function R refers to the delay caused by the best person, as dictated by the state/action context. The cost is then the sum of delays for the best person to pick the package up at location v, and drop the package off at w:

$$\begin{aligned}
&R((v, t_v), w, (w, t_w)) \\
&\quad = \mathbb{E}(d_v|t_v \bmod H) + \mathbb{E}(d_w|t_v + d_v \bmod H) \\
&\quad = \sum_{i=0}^{\infty} W^i(Hi + d_v)pr(d_v) \prod_{d'_v=1}^{d_v-1} \left(1 - pr(d'_v)\right) \\
&\quad + \sum_{i=0}^{\infty} W_w^i(Hi + d_w)pr(d_w) \prod_{d'_w=1}^{d_w-1} \left(1 - pr(d'_w)\right)
\end{aligned}$$

where $W_v = \prod_{d'_v=1}^{H} (1 - pr(d'_v))$ and $W_w = \prod_{d'_w=1}^{H} (1 - pr(d'_w))$, with interpretations of being the interpretation of being the probability of the participant *not* visiting the start and end locations (respectively) for an entire period (e.g., a week for $H = 14$). At this point, we can find the geometric infinite sum using basic algebra to get:

$$\begin{aligned}
&R((v, t_v),w, (w, t_w)) = \\
&\quad \left(\frac{d_v}{1 - W_v} + \frac{W_v H}{(1 - W_v)^2}\right) pr(d_v|t_v) \prod_{d'_v=1}^{d_v-1} \left(1 - pr(d'_v|t_v)\right) \\
&\quad + \left(\frac{d_w}{1 - W_w} + \frac{W_w H}{(1 - W_w)^2}\right) pr(d_w|t_v + d_v)\cdot \\
&\quad \cdot \prod_{d'_w=1}^{d_w-1} \left(1 - pr(d'_w|t_v + d_v)\right) \qquad (1)
\end{aligned}$$

The transition function $T(a, s, s')$ may be found in a similar way, but by considering only whole multiples of the given delay:

$$T(w, (v, t_v), (w, t_w)) = \textstyle\sum_{d_v=1}^{H} pr(d_v|t_v)pr(d_w|t_v + d_v) \qquad (2)$$

where $d = (t_w - t_v) \bmod H$, and we have marginalised out the uncertainty about d_v (the uncertainty in pick-up delay) in Equation 2.

We next address the problem of learning mobility models for individuals, which provides the probability of presence that defined the Bernoulli trial used in Equation 1 and 2.

3.1.4 Heuristic Based on MDP

The MDP we presented in this section can be solved in polynomial time in the number of locations with an appropriate standard method (e.g., linear programming, policy iteration) [23]. Since the number of locations in the Orange dataset was on the scale of 10^3, with a sparse transition matrix, this is certainly a tractable approach for individual solutions. However, in our experiments we want to conduct a large number of simulations using the mobility model, therefore, in practice, we use a heuristic that is based on the MDP formulation.

Specifically, rather than yielding a policy mapping each state to the best next location and best person to ask, we focus only on a policy for best people (as defined in Theorem 2). This was done by using the expected cost of actions as a fixed cost during planning, and finding the minimum cost path using Dijkstra's algorithm [25]. In short, the full MDP chooses next locations and people during run time (i.e., as the scenario unfolds in real time), while, for the purposes of simulation, we chose only the next best person during run time.

3.2 Model for Learning Human Mobility from Sparse Cell Phone Data

We now focus on the problem of getting an accurate predictive probability of presence for any location given the participant and the time $pr(v|i, t_v)$, from which the probability of delay can be derived and used in optimisation (as described in Section 3.1). The Orange dataset consists of a set of tuples for each participant $p_i \in P$ of the form (i, x_i, t_i) indicating that participant i was observed near cell tower x_i (discrete) at date and time t_i (continuous). There are three main features of the dataset that influence the design of the model:

1. **Cell allocation noise**
   The cell tower observations provide discrete measurements on the individual's likely location. These observations are not completely synonymous with location, since there may be a choice of several towers that the phone can connect to (especially in urban environments) at any single location. This allocation is decided by outside factors that we treat

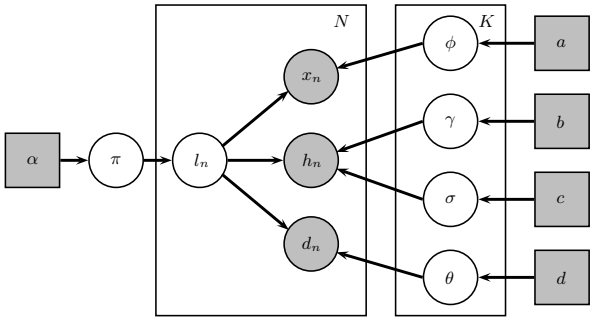


Figure 2: Graphical structure of the Dirichlet process location model, showing conditional independence between the random variables. Shaded nodes are observable and square nodes are fixed values.

as noise (i.e., the network operator's optimal allocation of phones to towers). Our approach needs to isolate the human presence information in the cell tower allocation to phones and ignore other factors. This implies the need to infer the locations, each of which may be statistically associated with several cell towers.

2. **Sporadic observations**
   Since the cell tower is only recorded in this dataset when a phone call or text is made (about 7 times a day, on average) approaches that were designed to be used on continuously collected location data (e.g. eigenvectors [26, 8], variable-order Markov models [2], linear embedding [27]) are not effective. We therefore need a method that can fill in (extrapolate from other observations) large periods of no observability.

3. **Short duration**
   The data on each individual covers a period of only 2 weeks. This, combined with the fact that each day may have only a few (or zero) observations, makes learning challenging. Overfitting is a danger when the training data (i.e., the 2 weeks of observations) contains characteristics that do not generalise to the rest of the individual's behaviour (i.e., beyond 2 weeks). On the other hand, we want our model to make stronger predictions that, say, a purely uniform distribution over locations (which is not going to overfit the data).

These considerations suggest the use of the Bayesian framework, which allows us to assume the existence of latent variables that abstract away from the variability of cell allocation (Factor 1), and make custom assumptions about the smoothness of location (Factor 2). Furthermore, Bayesian non-parametric methods can provide us with powerful guards against overfitting (Factor 3).

In more detail, we assume the existence of latent discrete locations, l_n that are associated with each observation (x_n, t_n), and correspond to places in the individual's routine life (e.g., home, work). To address the problem of filling in large periods of missing data, we assume that behaviour is periodic, as is common in other routine mobility models [26, 27]. Specifically, given that we only see 2 weeks of mobility, we assume both weekly and daily periodicities in behaviour. In the model, we achieve this by decomposing the date/time observation t_n to the discrete day of the week, d_n, and continuous hour of the day h_n. Mixture modelling is a well established method for inferring hidden discrete variables, but the standard approach requires the specification of the number of locations [3]. Therefore, we use a Dirichlet process mixture model (a non-parametric approach) that allows us to also infer the number of locations, K [18]. This is important, as highlighted in Factor 3, because setting K too high (manually) will cause the model to overfit the data[4].

A full generative model for location observations of each individual is therefore the following:

$$\boldsymbol{\pi} \sim DP(\alpha) \tag{3}$$

for each latent location k :

$$\boldsymbol{\phi_k} \sim Dir(a) \tag{4}$$
$$\gamma_k \sim \mathcal{N}(b) \tag{5}$$
$$\omega_k \sim IG(c) \tag{6}$$
$$\boldsymbol{\theta_k} \sim Dir(d) \tag{7}$$

for each observation n :

$$l_n \sim \mathcal{M}(\boldsymbol{\pi}) \tag{8}$$
$$x_n \sim \mathcal{M}(\boldsymbol{\phi_{l_n}}) \tag{9}$$
$$h_n \sim \mathcal{N}(\gamma_{l_n}, \omega_{l_n}) \tag{10}$$
$$d_n \sim \mathcal{M}(\boldsymbol{\theta_{l_n}}) \tag{11}$$

where, first, distribution $\boldsymbol{\pi}$ over latent locations is drawn from a Dirichlet process (Equation 3) that defines the prior probability of each location in the dataset. Second, the four parameters to the model $\phi, \gamma, \omega, \theta$ are drawn from their prior distributions (Dirichlet, normal, inverse-gamma, and Dirichlet, respectively) in Equations 4-7 [3]. These priors were chosen for their conjugacy to the parameter distributions, making the model simpler to infer. Thirdly, for each observation, a latent location is drawn (Equation 8), and the selection of this location defines all the observable information in the dataset (x_n, the cell tower, h_n the continuous hour observation, and d_n, the day of the week). Since x_n and d_n are discrete observations, they can be drawn from multinomials, while h_n (the continuous hour of the day) is drawn from a normal distribution with mean θ_{l_n} and variance ω_{l_n} (Equations 9-11). Defining h_n in this way makes the temporal distribution smooth, allowing us to fill in periods with only a few observations. However, we sacrifice some flexibility with this assumption, i.e., it does not capture multi-modalities in presence for a single location l_n.

The conditional independence assumptions between the random variables are visually represented in Figure 2. Direct inference of all the parameters from the data is not possible in this model, requiring us to either optimise them (via variational approximations), or to perform sampling [3]. We choose the latter, so that we do not have to consider the problem of finding local maxima, which can affect optimisation techniques. We employed a Gibbs sampling scheme, using standard techniques for all the distributions used in the model, adapted from the approach given by [24] to include the multinomial likelihood functions. Using sampling, we can find the predictive distribution for location v given the entire training set $\boldsymbol{X}$ for each individual [3]:

[4]N.B., while it is possible to retrain for different values of K and select the K that gives the highest data likelihood, we do not use this, as to repeat this process for 50,000 individuals would waste a lot of computation time.

$$pr(v|t_v, \boldsymbol{X}) = \sum_{r=1}^{R} pr(v|t_v, \boldsymbol{M^r}) pr(\boldsymbol{M^r}|\boldsymbol{X}) \quad (12)$$

where r is the index of each sample (taken after convergence), t_v is the query time, $\boldsymbol{M^r}$ is the entire set of model parameters found in sample r, and R is the total number of samples.

4. EXPERIMENTAL RESULTS USING THE ORANGE DATASET

In this section we use the real world cell tower mobility data of 50,000 people living in Ivory Coast, measured over 2 weeks, to assess the feasibility of crowdsourcing package delivery in Section 4.1. Then, using the same data, we evaluate our approach to prediction in Section 4.2, and optimisation under uncertainty in Section 4.3.

4.1 Feasibility Study

To assess the feasibility of the idea of crowdsourcing package delivery, we consider three key criteria:

1. **Number of participants required for acceptable coverage** If the number of participants required to cover large geographical areas is too large, then the recruitment and administrative burden would make this alternative delivery method infeasible in practice.

2. **Number of participants required in any specific delivery** Problems would arise if the number of participants in the chain from the source to destination location is too long. Firstly, if each person contributes a non-zero probability of theft or disappearance of the package, then there is greater risk with long chains in practice. Secondly, delay is introduced for each transfer.

3. **Feasibility of delivering to rural locations** If Criteria 1 and 2 are met only for urban locations, then the usefulness of crowdsourced package delivery is undermined, since urban locations have greater infrastructure resources (such as private courier, fast roads, international visitors) that make them easier to reach. Therefore, we want to make sure that we can also get acceptable delivery coverage of rural areas.

To assess these criteria, it was sufficient to consider a simpler instantiation (in this section only) of the problem we defined in Section 3.1 that takes into account the locations that each person in the participant set, P, visited, but does not include the temporal structure in the mobility. This presence graph, $\mathcal{G} = (V, C)$, contains locations V and for all entries $e_{a,b} \in E$. $e_{a,b}$ represents whether there exists a person $p \in P$ who visited both location a and b during their recorded data (with value 1, if true, and 0 otherwise). Since $\mathcal{G}$ contains no uncertainty, it amenable to standard graph analysis algorithms which we detail below. Throughout, we use cell tower as synonymous with locations, though in practice, of course, the source, exchange, and destination locations would be somewhere (less than a few km) near the cell tower in most cases.

4.1.1 Criterion 1: Number of Participants Required

To assess the number of participants required for wide geographical coverage (Criterion 1), we uniformly randomly subsampled participant sets, P', from the global participant set P (containing 50,000 people), for a wide range of different sizes $|P'| = \{10^{0.5i} | i = 1, 2, ..., 9\}$. For each participant set, we then uniformly sampled 1,000 pairs of locations (source and destination) from V representing 1,000 possible delivery problems. Figure 3 shows the distribution over geographical distance between the source and destination locations that were sampled uniformly. The plot is notable for its bi-modality which we interpret as being caused by the fact that the concentration of cell towers in urban locations (especially the capital, Abidjan) favours short distances in the test set. We consider a different (urban to rural) distribution of test locations in Section 4.1.3.

For each test location pair, we used Dijkstra's algorithm to find the shortest path (the standard algorithm can be applied to graph $\mathcal{G}$ because these is no uncertainty about the edge costs) [25]. Figure 4 shows the percentage of location pairs that were feasible (i.e., that had any path between the source and destination locations). The blue line with circular points shows the feasibility for uniform random source and destination locations. We see that the geographical coverage is very poor when there are fewer than $10^{2.5}$ participants. The critical range seems to be around 10^3, when feasibility

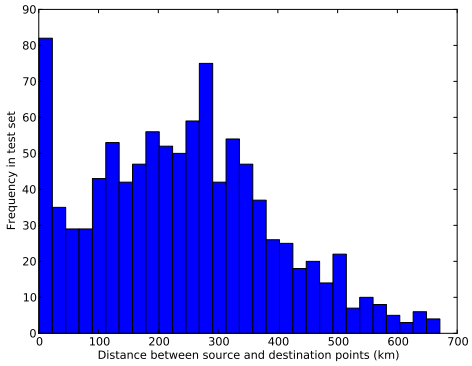


Figure 3: A frequency histogram of the distances of the randomly sampled (source,destination) delivery problems using uniform sampling.

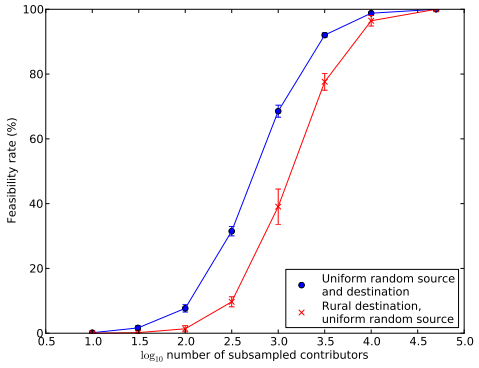


Figure 4: A plot of the percentage of randomly sampled (source,destination) delivery problems that had a solution path of any size, against the $\log_{10}$ size of the number of potential contributors.

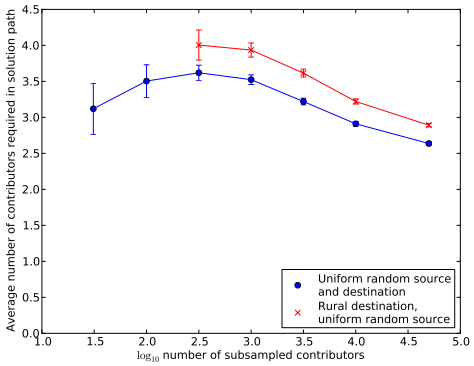


Figure 5: A plot of the average number of contributors required to each specific delivery problem (drawn from the much larger pool of potential contributors) against the $\log_{10}$ size of the potential contributors pool. N.B., a majority of rural destinations are infeasible for pool sizes of less than $10^{2.5}$, therefore we are unable to plot the line below this range.

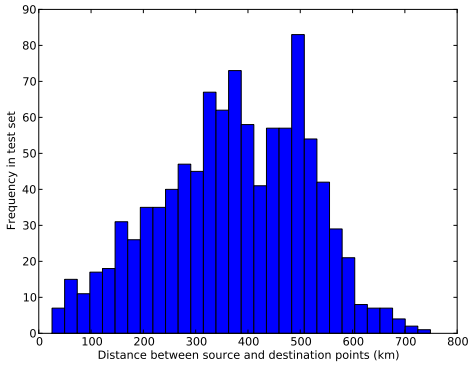


Figure 6: A frequency histogram of the distances of the (source,destination) delivery problems when the destination is sampled from a set of rural locations (while the source is uniformly sampled from all locations).

surges with each new participant. An acceptable geographic coverage, trading off against recruitment/administration costs, appears to be around $10^{3.5}$ participants. N.B., however, this result applies to the approximately 1,200 cell towers that the 50,000 participants visited, though we can see in Figure 1 that these are distributed (geographically) widely in the country. Criterion 1 is therefore satisfied because we can get acceptable geographical coverage without requiring very large numbers of participants.

4.1.2 *Criterion 2: Number of Participants Required for Any Given Delivery Problem*

To assess the number of participants required in any given solution path (Criterion 2), we used the same subsampled participant sets as in Section 4.1.1 and plotted the length of the shortest path against the size of each subsampled participant set in Figure 5. The length of the shortest path indicates how many people are required for any specific delivery problem. The blue line with circular points is the line of interest for Criterion 2, where we see that the number of participants required for any solution path stays within a small range 2-4. Since we can only plot the length of feasible paths in Figure 5, the number of contributors required for specific paths actually increases with the size of the participant subset, initially. However, once path feasibility (indicated in Figure 4) reaches a significant level, the trend is as expected: having a wider pool of participants allows more efficient (i.e., shorter length) paths to be discovered. Note that, since we not considering duration in Figure 5, the lowest cost paths in the full model may require more people. In any case, since the cost for losing the package can be fully specified by the planner, the optimal tradeoff between path length and duration can be found.

4.1.3 *Criterion 3: Rural Distribution*

So far, we have only considered uniformly sampled source and destination test points, which favours urban locations (since there are greater numbers of cell towers in urban areas). We now consider Criterion 3 for rural feasibility, by sampling a set of delivery problems where the destinations are only rural (keeping source locations uniformly sampled, as before). To sample rural destinations, we use auxiliary data from the government-run postal service website[5], which lists the locations of 93 postal hubs in Ivory Coast. The postal service has an office in every major town and city in the country. We converted the names of the locations to latitude/longitude pairs using Google[6] and Bing[7] geocoding services. A set of candidate rural locations was produced by removing from consideration all locations in V that were less than 50 km in Euclidean distance from any postal office. This yielded 97 destination locations. The average distance between source and destination points increases to 373 km (from 241 km previously) under this scenario, making it a much more challenging problem. If one wanted to deliver the aid by conventional means, it would take about 2 days of driving (assuming an average speed of 50 km/h, and that the driver would have to go in both directions, driving about 8 hours a day). If trucks were to deliver all the items in the simulation (of 10,000 journeys), it would take them 20,000 days of driving under these assumptions. However, we have not taken into account the network flow limit of the human participants, which would certainly be a factor at this scale.

The resulting distance distribution over all delivery problems under this scenario is shown in Figure 6. We ran the same analyses for Criteria 1 and 2 with rural destinations, yielding the red lines with crosses in Figures 4, and 5. The feasibility plot in Figure 4 shows that for rural destinations, the participant pool needs to be approximately $10^{0.5} \approx 3$ times as large as for uniformly sampled destinations. Furthermore, the number of contributors required to get an adequate coverage has increased to approximately $10^{3.5}$, demonstrating that the method is still feasible, even for rural destinations. In Figure 5, we see a fairly constant increase in the optimal path length (for any given participant pool), reaching a maximum of 4 participants required per delivery problem, which is still an acceptable level. To summarise, restricting the destinations to be rural certainly makes the delivery problem more challenging, but it is still feasible.

[5] http://www.laposte.ci/bureau.php
[6] https://developers.google.com/maps/documentation/geocoding/
[7] http://www.microsoft.com/maps/developers/web.aspx

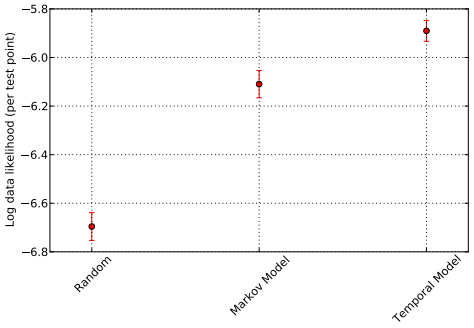


Figure 7: Average $\log_e$ data likelihood (higher is better) of held out test data (size 50,0000) of the temporal model against two benchmarks.

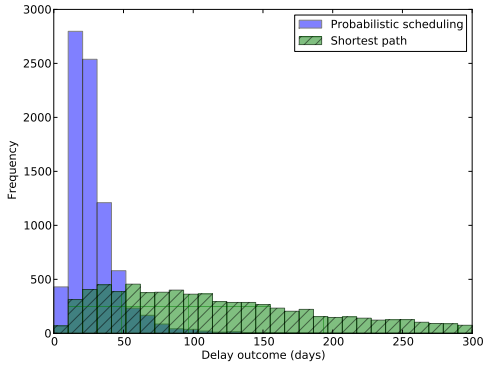


Figure 8: The durations of 10,000 simulated journeys in the rural test set using both the shortest path decision (benchmark) and our probabilistic scheduling/routing approach.

Now that we know that all three feasibility criteria are met, we consider the problem of learning the temporal structure in mobility to enable the minimisation of delay in delivery from source to destination nodes.

4.2 Evaluation of Human Mobility Predictions

In this section, we evaluate our approach to predicting human mobility under considerable data sparsity, caused by the fact that the Orange dataset (for granular locations) contains only 2 weeks of observations for each participant and gives sporadic readings on their locations (on average, 7 readings per day).

We split the cell tower data of 50,000 uniformly randomly sampled people into training and testing sets. The test set contains a single cell tower location reading from each person's data, therefore giving a test set of 50,000 data points. The rest of the data for the same individuals was used in training. To evaluate our approach, we looked at the logarithm of the data likelihood of each test point, which is a standard measure of model accuracy for probabilistic models [3]. We used fairly non-informative hyperparameters $a = 1, d = 1, \alpha = 1$ for the discrete priors (see Figure 2). We used $b = (0.01, 12)$ and $c = (0.01, 3)$ for the continuous temporal priors, referring to the relative mean of precision w.r.t. the data, the mean of the prior, the degree of freedom in the precision, and the inverse mean of precision, respectively.

The result can be seen in Figure 7. To place this result in context, we also evaluated two other models under identical conditions. The first was a purely random model, with a temporal Gaussian distribution (with mean 12pm and standard deviation 6 hours) across each day, and with equal probability mass across all locations and all days. The second benchmark is a first-order Markov model, which has previously been shown to perform well under data sparsity [17].

In Figure 7, we see that our model has an average likelihood that is 125% better than the Markov model (since we are using a $\log_e$ y-axis scale) which, in turn, is 180% better than the random method. We therefore conclude that ours is a good model for predicting human mobility under sparse conditions. For all subsequent evaluation into choosing the optimal participants for a delivery problem, our model as the ground truth in our simulations, which enables use to evaluate a suitable range of outcomes even with only a small amount of training data (N.B., the Bayesian non-parametric method ensures that we do not overfit the data).

We now proceed to evaluate the optimisation element of our work. To do this, we make a few additional assumptions in light of the results we have presented so far. Firstly, since a participant pool of approximately 3,500 people is enough to get satisfactory coverage of Ivory Coast (see Section 4.1), we used participant sets of this size in our optimisation evaluation. Secondly, in order to give a thorough evaluation of optimisation, involving 10,000 simulations, we use our mobility model (given in Section 3.2) as the ground truth, since it performs best under the extreme data sparsity that of the Orange dataset.

4.3 Evaluation of Optimisation Approach

Finally, we evaluated the performance of our approach to decision-making (i.e., which participants to ask, and which intermediate locations to use). To get an adequate idea of performance, we ran simulations 10,000 times, using the heuristic approach in Section 3.1.4 (though the optimal approach in Section 3.1 is also tractable for smaller runs). In Figure 8, we see the end-to-end duration performance of our algorithm against the naïve approach that considers only the shortest path (i.e., the minimum number of contributors), which does not consider the temporal mobility habits of the participants. The average total duration for our optimal approach is 28.2 days, versus 163 days for the benchmark. For this, we used the rural test set, defined in Section 4.1.3, with an average of 373 km between the source and destination locations.

5. CONCLUSIONS AND FUTURE WORK

In this work we studied an alternative method to distribution that uses the existing mobility of local people to send packages large distances. Using data describing the real world movement patterns of 50,000 people, we found some evidence to support the feasibility of this vision. Furthermore, we addressed the technical problems associated with this method, formulating an MDP for optimisation and presenting a Bayesian non-parametric model that performs well under data sparsity.

There are several avenues for future work. Firstly, it is unclear how this approach to delivery would work when large numbers of items are sent simultaneously. Network flow analysis might provide some insight, and possible solutions to this issue [15]. Secondly, we have not addressed the trust issues associated with participants [32]. It is trivial to incorporate a measure of trust in the cost function of the MDP (by adding it to the time delay), but learning this feature is challenging.

References

[1] D. Ashbrook and T. Starner. Using GPS to learn significant locations and predict movement across multiple users. *Personal and Ubiquitous Computing*, 7(5):275–286, 2003.

[2] H. Bapierre, G. Groh, and S. Theiner. A variable order markov model approach for mobility prediction. In *STAMI Workshop at the 22nd International Joint Conference on Artificial Intelligence (IJCAI 2011)*, Barcelona, Spain, 2011.

[3] C. M. Bishop. *Pattern recognition and machine learning*, volume 4. Springer New York, 2006.

[4] D. M. Blei and L. P. Kaelbling. Shortest paths in a dynamic uncertain domain. In *IJCAI Workshop on Adaptive Spatial Representations of Dynamic Environments*, volume 4, page 2, 1999.

[5] CIA Directorate of Intelligence. The world factbook. July 2008.

[6] V. Colizza, A. Barrat, M. Barthelemy, A.-J. Valleron, and A. Vespignani. Modeling the worldwide spread of pandemic influenza: baseline case and containment interventions. *PLoS Medicine*, 4(1):e13, 2007.

[7] N. Eagle and A. Pentland. Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing*, 10(4):255–268, 2006.

[8] N. Eagle and A. S. Pentland. Eigenbehaviors: identifying structure in routine. *Behavioral Ecology and Sociobiology*, 63(7):1057–1066, 2009.

[9] H. Gao, J. Tang, and L. Huan. Mobile location prediction in spatio-temporal context. In *Mobile Data Challenge by Nokia Workshop, in conjunction with International Conference on Pervasive Computing*, 2012.

[10] H. Gao, J. Tang, and H. Liu. Exploring social-historical ties on location-based social networks. In *6th International AAAI Conference on Weblogs and Social Media*, 2012.

[11] L. Hufnagel, D. Brockmann, and T. Geisel. Forecast and control of epidemics in a globalized world. *Proceedings of the National Academy of Sciences of the United States of America*, 101(42):15124–15129, 2004.

[12] B. Keller, P. von Bergen, R. Wattenhofer, and S. Welten. On the feasibility of opportunistic ad hoc music sharing. 2012.

[13] J. Laurila, D. Gatica-Perez, I. Aad, J. Blom, O. Bornet, T. Do, O. Dousse, J. Eberle, and M. Miettinen. The mobile data challenge: Big data for mobile computing research. In *Mobile Data Challenge by Nokia Workshop, in conjunction with International Conference on Pervasive Computing*, Newcastle, UK, 2012.

[14] D. Lazer, A. S. Pentland, L. Adamic, S. Aral, A. L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, and M. Van Alstyne. Life in the network: the coming age of computational social science. *Science*, 323(5915):721–723, Feb. 2009.

[15] J.-Y. Le Boudec and P. Thiran. *Network calculus: a theory of deterministic queuing systems for the internet*, volume 2050. Springer, 2001.

[16] J.-Y. Le Boudec and M. Vojnovic. Perfect simulation and stationarity of a class of mobility models. In *INFOCOM 2005. 24th Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings IEEE*, volume 4, pages 2743–2754. IEEE, 2005.

[17] J. McInerney, A. Rogers, and N. Jennings. Improving location prediction services for new users with probabilistic latent semantic analysis. *In Mobile Data Challenge by Nokia Workshop, in conjunction with International Conference on Pervasive Computing, 2012.*, 2012.

[18] R. M. Neal. Markov chain sampling methods for dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265, 2000.

[19] M. E. Newman. A measure of betweenness centrality based on random walks. *Social networks*, 27(1):39–54, 2005.

[20] E. Nikolova, M. Brand, and D. R. Karger. Optimal route planning under uncertainty. In *Proceedings of International Conference on Automated Planning and Scheduling*, 2006.

[21] E. Nikolova and D. R. Karger. Route planning under uncertainty: The canadian traveller problem. In *Proc. AAAI*, pages 969–974, 2008.

[22] G. Pickard, I. Rahwan, W. Pan, M. Cebrian, R. Crane, A. Madan, and A. Pentland. Time critical social mobilization: The darpa network challenge winning strategy. 2010.

[23] M. L. Puterman. *Markov decision processes: Discrete stochastic dynamic programming*. John Wiley & Sons, Inc., 1994.

[24] C. E. Rasmussen. The infinite gaussian mixture model. *Advances in neural information processing systems*, 12(5.2):2, 2000.

[25] S. J. Russell, P. Norvig, J. F. Canny, J. M. Malik, and D. D. Edwards. *Artificial intelligence: a modern approach*, volume 2. Prentice hall Englewood Cliffs, NJ, 1995.

[26] A. Sadilek and J. Krumm. Far out: Predicting long-term human mobility. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[27] S. Scellato, M. Musolesi, C. Mascolo, V. Latora, and A. Campbell. Nextplace: a spatio-temporal prediction framework for pervasive systems. In *Pervasive Computing*, pages 152–169, San Francisco, CA, USA, 2011. Springer.

[28] J. Scott, A. J. Brush, J. Krumm, B. Meyers, M. Hazas, S. Hodges, and N. Villar. PreHeat: controlling home heating using occupancy prediction. In *Proceedings of the 13th international conference on Ubiquitous computing (UbiComp 2011)*, pages 281–290, Beijing, China, 2011.

[29] Q. M. Shao and J. G. Ibrahim. *Monte Carlo methods in Bayesian computation*. Springer Series in Statistics, New York, 2000.

[30] E. Stevens-Navarro, Y. Lin, and V. W. Wong. An mdp-based vertical handoff decision algorithm for heterogeneous wireless networks. *Vehicular Technology, IEEE Transactions on*, 57(2):1243–1254, 2008.

[31] UN Office for the Coordination of Humanitarian Affairs. Ivory coast: Unrest disrupts malaria prevention bid. June 2011.

[32] M. Venanzi, A. Rogers, and N. Jennings. Trust-based fusion of untrustworthy information in crowdsourcing applications. 2013.

[33] V. Vukadinović, Ó. R. Helgason, and G. Karlsson. A mobility model for pedestrian content distribution. In *Proceedings of the 2nd International Conference on Simulation Tools and Techniques*, page 93. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2009.

[34] A. Wesolowski, N. Eagle, A. J. Tatem, D. L. Smith, A. M. Noor, R. W. Snow, and C. O. Buckee. Quantifying the impact of human mobility on malaria. *Science*, 338(6104):267–270, 2012.

[35] J. H. Wu and R. Givan. Feature-discovering approximate value iteration methods. *Abstraction, Reformulation and Approximation*, pages 901–901, 2005.

Towards a recommender system for bush taxis

Sébastien Gambs
Université de Rennes 1 - INRIA / IRISA
Campus Universitaire de Beaulieu
35042 Rennes, France
Email: sgambs@irisa.fr

Marc-Olivier Killijian,
Miguel Núñez del Prado Cortez
and Moussa Traoré
CNRS ; LAAS ,
7 avenue du Colonel Roche
F-31077 Toulouse, France
Email: {killijian, mnunezde, mtraore}@laas.fr

***Abstract*—To improve the transport efficiency and to reduce the traveller stress, we introduce a recommender system for bush taxis in Ivory Coast whose main objective is to propose to pedestrians potential means of transportation in their neighborhood whose destination match their own destination. The prediction of the next location relies on a mobility model called Mobility Markov Chain. One of the strength of the proposed recommender system is that it is fully automatic as a user does not need to explicitly express his next destination but rather the system tries to infer it based on his past mobility behavior. Moreover, the recommendation algorithm is biased towards suggesting the means of transportation that are the cheapest (if one is available). The preliminary evaluation of the recommender system conducted on one of the D4D dataset shows that approximately 99% of the time in less than 30 minutes, the system is able to suggest a mean of transportation that is at most 1 kilometer away from the current position of the user for an accuracy of the prediction of the next location that is between 30% and 50% depending on the complexity of the mobility behavior of the user considered.**

***Keywords*-Location-based service, Recommender system, Next place prediction, Mobility Markov chain.**

I. Introduction

Transport is a vital economic component for developing countries. Indeed, when a performant transport infrastructure exists, people are able to take advantage of the wide variety of business opportunities, thus increasing their income level and improving their standard of living. However, in many countries, the access to transport remains an unsolved challenge, in particular in fast growing urban areas. Therefore, governments have incentive to develop innovative transportation services that are able to cope with the wide variety of the mobility needs of the citizens of these countries.

In this paper, we tackle this issue by introducing a recommender system for bush taxis whose main objective is to propose to pedestrians potential means of transportation in their neighborhood whose destination match their own destination. More precisely for each user, we first build a mobility model summarizing their behavior called a Mobility Markov Chain (MMC) [4]. The MMC is learnt from the mobility traces of the user, which in the context of the D4D challenge correspond to calls made from cell towers. Afterwards, once a MMC has been learnt, it can be used to predict the next location visited by a user based on the knowledge of his current position. In order to suggest a potential mean of transportation, the recommendation system first starts to guess the next location of the user and then scan the neighborhood for potential vehicles (*e.g.*, bush taxis, colored woro-woros, gbakas and traditional taxis) whose destination matches the next location of the user. The destination of a mean of transportation is also predicted by learning a MMC representing his mobility. One of the main advantage of our recommender system is that it is fully automatic, in the sense that a user does not need to explicitly express his next destination but rather the system tries to infer it based on his past mobility behavior. Moreover, the recommendation algorithm is biased towards suggesting the means of transportation that are the cheapest (if one is available). The preliminary evaluation of the recommender system shows that approximately 99% of the time in less than 30 minutes, the system is able to suggest a mean of transportation that is at most 1 kilometer away from the current position of the user.

The outline of the paper is the following. First, in Section II we describe the related work on recommender systems for taxis. Then, respectively in Sections III and IV, we review briefly the different transportation modes available in Ivory Coast as well as the D4D datasets. Afterwards, we introduce in Section V the Mobility Markov Chain (MMC), which is the mobility model that we used to predict the next location visited by a user before conducting a mobility analysis of the D4D dataset in Section VI. The recommender system for bush taxis is presented in Section VII and then we conduct a preliminary evaluation of its performance in Section VIII, before concluding in Section IX.

II. Related work

In recent years, mobility traces of GPS-enabled vehicle, including taxicabs, have been collected on a massive scale [7]. This collection of mobility traces of taxicabs has fostered the research in urban vehicle transportation, in particular in taxi recommender systems.

For instance, Lee, Shin and Park [8] have proposed a recommender system for taxis. These researchers have analyzed

80,000 traces stored in the taxi telematics system of Jeju island (Republic of Korea). Basically, their positions as well as their speed are reported by the taxis to a central server along with their current status (*i.e.*, free or occupied). From these mobility traces, the researchers first extracted potential pick-up points for taxis by observing the changes of status of taxis from free to occupied. Afterwards, the classical k-means algorithm [9] was used to cluster together pick-up points that are close to each other and the center of each cluster represents a potential recommendation for a pick-up point (in practice the value of k was set to 100 in their experiments). The analysis was also performed by taking into account the time dimension in order to discover time-dependent pick-up patterns. The resulting recommender system is able to propose the nearest beneficial pick-up points to vacant taxis.

Ge and co-authors [6], [5] have developed a recommender system for taxi drivers suggesting rides (*e.g.*, sequence of pick-up points or parking places) so as to maximize the probability of picking-up passengers. The pick-up points are learned from the trail of mobility traces of taxi drivers that are the most successful. Then, the pick-up points are clustered using the k-means algorithm (here also the chosen value of k was 100 for the different experiments). A probability is also associated to each generated centroid measuring the frequency of pick-up events when taxi cabs pass across the corresponding cluster. To avoid the overload of the road due to taxicabs in the same area following the same recommendation, a load balancing approach is applied to distribute empty taxis through multiple paths. Moreover, to reduce the risk of fraud from taxi drivers (*e.g.*, greedy taxi drivers who overcharge passengers by taking unnecessary detours), the trajectory of each taxicab is analyzed in order to identify the ones that are unusually long, combining evidences of frauds through the Dempster-Shafer theory.

Zheng, Liand and Xu [17] have designed an application guiding users to locations in which they can wait for a vacant taxi with the main objective of reducing their waiting time. The mobility behaviors of vacant taxis in the streets of Beijing are modeled as non-homogeneous Poisson processes. The application predicts to users the nearest road segment in which they will find a vacant taxi as well as estimate of the average waiting time (30% of simulation error).

Jing and co-authors [15], [16] have introduced a recommender system for both taxi drivers and passengers. Their model describe the probability for taxi drivers to pick-up passengers while going to adjacent parking places (*i.e.*, places in which drivers wait for passengers), the average waiting time depending on the time of the day as well as the average distance for the next trip of the driver. The passenger recommendation algorithm provides two services. The first service returns a list of nearby parking places with the average corresponding waiting time while the second service outputs a nearby road segment located at walking distance for the current position of the passenger as well as the average associated waiting time before finding an empty taxi.

While most of related works have used very precise mobility traces collected by GPS devices installed directly on taxis, our approach relies on coarser mobility traces from phones providing granularity at the level of GSM towers, which is a more challenging task. However, contrary to previous works, we did not limit ourselves to recommend only standard taxicabs but we also push suggestions for other transportation modes in Ivory Coast. These other transportation modes are detailed in the next section.

III. Modes of transportation in Ivory Coast

Due to the fast growth of the population in recent years in Ivory Coast the population of urban cites have also experienced an important increase, mainly due to the galloping urbanization and rural exodus. These fast-growing cities face enormous challenges in terms of the development of their infrastructure as well as the need to cope with the increasing demand for transport. For instance, Ivory Coast has witnessed the rising of informal modes of transport complementary to the traditional buses and taxi services. Thereafter, we briefly review these existing informal modes of transportation.

1) A *gbaka* is a minibus of 14 to 22 seats covering relatively long commuting distance between the outskirts of a city and its urban center. The gbaka has a predetermined trajectory but can deviate from it due to the occurrence of a traffic jam and some suggestion done by the apprentice-gbaka (*i.e.*, the assistant of the driver of the gbaka). Indeed the apprentice is responsible for spotting potential passengers on the way and makes the driver stop if they find one. Gbakas represent around 27% of the public transport offer in Abidjan [10].
2) The (colored) *woro-woro* is an informal taxi riding only within the limits of a particular city whose color represents the municipality to which it belongs. A colored woro-woro has a maximum capacity of 5 seating places and is continuously patrolling for clients unless he is already full. White woro-woro also exist that are not directly attached to a particular city but rather act as a shuttle between neighboring cities and have predefined and well-known parking places. However, as we believe that their mobility behavior is quite similar to the one of gbakas, for the rest of the paper we will make no distinction between gbaka and white woro-woro. This transportation mean counts for 32% of the public transport offer in Abidjan [10].
3) *Bush taxi* is the cheapest mode of transportation for long distance as well as the most common mean of transportation for inter-urban travel. A bush taxi usually pick-ups passengers at a fixed station but without

	Intra-communal	Inter-communal				Inter-rural	
Trajectory	Woro-woro (colored)	Bus	Taxi	Gbaka	Woro-Woro (white)	Bush-taxi (Badjan)	Bush-taxi (504)
Fixed trajectory	✗	✓	✗	✓	✓	✓	✓
Number of seats	5	32	5	12-32	5-8	12-32	8-10
Vehicle type	Sedan	Bus	Sedan	Bus	Bus	Bus	Van

Figure 1. Summary of the transportation modes available in Ivory Coast.

following any predefined schedule. More precisely, a bush taxi generally start its journey when all its seats are filled. Nevertheless, a bush taxi may stop anywhere on the way to pick up or drop off passengers if required. Two types of bush taxis exist: the Badjan, which correspond to Toyota vans, can accommodate up to 32 passengers while the other type of bush taxis drive Peugeot 504, thus having a more limited capacity of 10 passengers at the maximum.

Figure 2. Toyota bush taxi.

4) Traditional (*i.e.*, metered) *taxi* (similar to European taxis) is a transportation mean that is widely available as this taxi is always patrolling. These last years, the business of traditional taxis has suffered from the development of informal modes of transport such as gbakas, woro-woros and bush taxis. While traditional taxis are the most comfortable mean of transportation, they are also the less preferred one due to their expensive price. A traditional taxi can contain up to 5 passengers at a time and is not restricted to operate within the limits of a particular city. However, these taxis are notorious for overcharging clients. This category counts for 17% of the public transport supply in Abidjan [11].
5) Finally, *buses* is the only organized and formal public transportation system serving the city of Abidjan in Ivory Coast. In particular with the exception of Abidjan, other cities do not have any public transportation system and therefore have to rely on the other informal modes of transport detailed before. Buses are operated by SOTRA (*Société des TRansports Abidjanais*). The existing bus lines cover all the districts of Abidjan. However, buses do not have fixed schedules, are quite disorganized and are poorly maintained. Until the beginning of 2010, buses had the monopole on public transportation in Abidjan, but due to the emergence of informal modes of transport, today SOTRA does not meet more than 12% of the public transport demand [10].

Figure 1 summarizes the characteristics of these different modes of transportation.

IV. Description of the D4D datasets

Thereafter, we briefly review the four datasets available for the Data for Development Orange challenge (D4D challenge) [12], [2].

1) *Aggregated communication between cell towers.* For each period of one hour, this dataset summarizes the number of calls as well as the total communication time that has occurred between each possible pair of cell towers. This dataset also contains the identity of the cell tower that has initiated the call. Note that the calls that have started in a particular one-hour time period are associated with this time period irrespective of the time at which they stop.
2) *Mobility traces (high-resolution dataset).* This dataset contains a random sample of 50 000 active users whose traces have been recorded over a period of 2 weeks. These mobility traces are composed of the identifiers of the cell towers from which users have made a call or send a SMS as well as the corresponding timestamp. This process is repeated for 10 two-weeks period but with other samples of 50 000 users randomly selected, which makes a total of 500 000 users overall.
3) *Mobility traces (low-resolution dataset).* This dataset is composed of the mobility traces extracted from a random sample of users (50 000) over a period of 5 months. Each trace has associated to it a timestamp and some location information corresponding to the position at which the call was made or the SMS

was sent. Contrary to the previous dataset, the only information revealed about the location is at the level of the sub-prefecture instead of the cell tower, which is a much coarser information. A sub-prefecture is an administrative unit and Ivory Coastis actually composed of 255 sub-prefectures. A table containing the location of the centers of these sub-prefectures is also given as part of this dataset.

4) *Communication sub-graphs.* This dataset is composed of the communication graphs of 5 000 users. More precisely, the communication graph of each user is generated by observing all the communications that he had with his contacts as well as the communications that have occurred between his contacts and the contacts of his contacts (which means that the graph considered up to two degrees of separation). A communication graph is obtained by aggregating all the communications that have happened over a two-week period. Thus, for the total period of 5 months, 10 different communication graphs are generated for each user. The identity of the contacts of a user have been pseudonymized by assigning a random identifier to each contact that remains unchanged over the total observation period.

While all these datasets contained valuable information, we have focus only on the dataset containing mobility traces (*i.e.* user id, timestamp, antenna id, antenna latitude and longitude) with high-resolution in the remaining of the paper. Before building on this dataset, we have first tried to analyze the mobility behaviors of the users contained within this dataset. Note that mobile phone datasets can be considered as *sporadic* location data in the sense that they expose a limited amount of data concerning the mobility of users contrary to other geolocated datasets containing the movements of individuals recorded at a high frequency (*e.g.*, every minute or every 5 minutes).

In order to differentiate between individuals and vehicles, we have segmented the users by taking into account the minimal number of different antennas they have visited as well as the minimal number of mobility traces they have in their history. As a result, Figure 3 illustrates the number of users of the dataset having respectively at least 20, 40, 100 and 500 mobility traces with 5 or 10 distinct antennas visited.

For the remaining of the paper, we have considered only users that have a "rich enough" mobility behaviors, which we define as user that have a least 40 mobility traces in their history and that have visited at least 5 different antennas. There are approximately 150 000 users corresponding to this profile in the high-resolution dataset. We choose to disregard users that do not match these characteristics, because for each user we will use the first 20 mobility traces as the *training set* from which their mobility model is learnt while the rest of the traces forms the *the testing set*, which is used

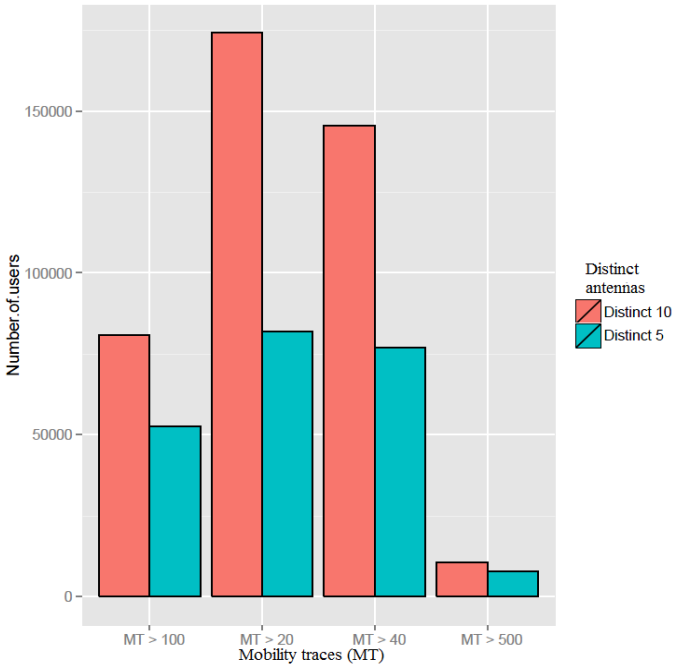


Figure 3. Distribution of the users according to the minimal number of mobility traces and the minimal number of different antennas they have visited available in their history.

for evaluating the performance of the recommender system for bush taxis. In the next section, we briefly review the mobility Markov chain [4], which is the model we used in to capture and represent the mobility of users.

V. Mobility Markov chain

A *Mobility Markov Chain* (MMC) [4] models the mobility behavior of an individual as a discrete stochastic process (more precisely as an ergodic regular Markov chain) in which the probability of moving to a state (*i.e.*, point of interests) depends only on the previously visited state and the probability distribution on the transitions between states. More precisely, a MMC is composed of the followings elements.

- A set of *states* $P = \{p_1, \dots p_n\}$, in which each state is a point of interest (POI). POIs are usually learned by running a clustering algorithm on the mobility traces of an individual, acquired for instance through a GPS-enable device or by taking the list of visited antennas identifiers obtained from his mobile phone records. These states generally have an intrinsic semantic meaning and therefore semantic labels such as "home" or "work" can potentially be inferred and associated to them.
- *Transitions*, such as $t_{i,j}$, represent the probability of moving from state p_i to state p_j. A transition from one state to itself is possible if the individual has a non-null probability from moving from one state to an occasional

location before coming back to this state. For instance, an individual can leave home to go to the pharmacy before coming back to his home. In this example, it is likely that the pharmacy will not be extracted as a POI by the clustering algorithm, unless the individual visits this place on a regular basis.

- The *stationary vector* π (also known as the steady state or the stationary distribution) is a column vector obtained by multiplying an initial column vector ((e.g.), a uniform vector) with the transition matrix repeatedly until convergence. The precise meaning of this vector depends on the type of mobility traces that have been used to build the MMC. For instance, if the POIs have been extracted by running a clustering algorithm on traces acquired through a GPS then p_i represents the percentage of the time spent by an individual in the i^{th} POI. However, in the case of D4D challenge, which consists of mobility traces obtained from call logs, p_i symbolizes rather the probability to send or receive a call or SMS while being location in the antenna corresponding to the i^{th} POI.

In a nutshell, building a MMC is a two-steps process. During the first phase, an algorithm extracts the POIs from the mobility traces. In the context of the D4D challenge extracting POIs amounts simply to read the antennas that have been visited by a particular user while for mobility traces obtained from a GPS-enabled device, this phase is more complex as it involves preprocessing the data in order to remove moving points before applying a clustering algorithm to obtained the POIs. For instance, in a previous work [4], we have used a clustering algorithm called Density-Joinable Cluster (DJ-Cluster) to discover the POIS.

During the second phase, once the POIs (*i.e.*, the states of the Markov chain) are identified, the probabilities of the transitions between states can be computed. To realize this, the trail of mobility traces is examined by chronological order and each mobility trace is tagged with a label that is either the number identifying a particular state of the MMC or the value "unknown". Finally, when all the mobility traces have been labeled, the transitions between states are counted and normalized by the total number of transitions in order to obtain the probabilities of each transition. We refer the interested reader to [4] for the details of the algorithm on how to build a MMC.

A MMC can be either represented as a transition matrix (Table I) or in the form of a directed graph (Figure 4) in which nodes correspond to states and arrows represent the transitions between along with their associated probability. When the MMC is represented as a transition matrix of size $n \times n$, the rows and columns correspond to states of the MMC while the value of each cell is the probability of the associated transition between the corresponding states.

	Home	Work	Gym	Bar	Restaurant
Home	0	1	0	0	0
Work	0,3	0	0,2	0,2	0,3
Gym	1	0	0	0	0
Bar	1	0	0	0	0
Restaurant	0.5	0	0.5	0	0

Table I
REPRESENTATION OF A MOBILITY MARKOV CHAIN AS A TRANSITION MATRIX.

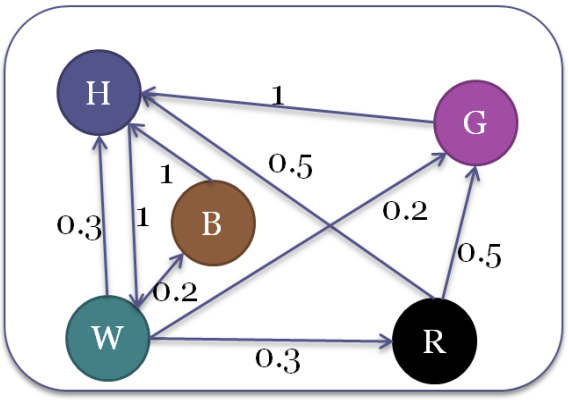


Figure 4. Representation of a Mobility Markov chain as a directed graph.

VI. MOBILITY ANALYSIS OF THE DATA

To be able to recommend different modes of transportation, we first need to understand how to categorize the mode of transport of a user. We choose to define as the possible categories for a user, the ones described previously in Section III (with the exception of buses), namely individual, gbaka, colored woro-woro, bush taxi and traditional taxi. To distinguish between these categories, we propose to rely on the three following mobility characteristics: (Shannon) entropy, predictability and average traveled distance per day. Thereafter, we briefly review the notions of entropy and predictability.

- In general, the *(Shannon) entropy* is a measure of uncertainty regarding the output of a random variable [13]. In the context of mobility, the entropy of a user quantifies the spatial uncertainty about the exact location of a user. For instance, it can be defined as the average number of binary questions that one needs to ask in order to predict the particular POI (*i.e.*, antenna) on which the user is currently located. Considering a particular user u, we can compute his entropy by applying the following formula

$$H(u) = -\sum_{i=1}^{n_u} p_{i,u} \log_2 p_{i,u} \quad (1)$$

in which p_i represents the probability to be located in the i^{th} POI for user u while n_u corresponds to the number of POIs visited by this user. For instance,

consider the situation in which Alice has visited four different POIs forming the following set $\{A, B, C, D\}$. For each of this POI, the number of recorded mobility traces is respectively 40, 20, 10 and 10 mobility traces. Therefore, we have $p_A = 50\%$, $p_B = 25\%$, $p_C = 12.5\%$ and $p_D = 12.5\%$. Applying Equation 1 leads to an average entropy of 1.27 bits for Alice as illustrated in Table III.

POI	Nb of mobility traces	Probability	Entropy
A	40	0.5	-0.51
B	20	0.25	-0.11
C	10	0.125	-0.24
D	10	0.125	-0.41
Total	80	1	1.27

Table II
EXAMPLE OF THE COMPUTATION OF ALICE'S ENTROPY.

- The *predictability* [3] is a theoretical measure quantifying how predictable is an individual based on his MMC model (*cf.* Section V). For instance, consider the scenario in which Bob is currently located on the "Home" POI then based on his MMC, the probability of making a successful guess of his next location is theoretically equal to the maximal outgoing probability transition leaving from this state (*i.e.*, POI), which could be for instance the transition going to the "Work" POI. More formally, the predictability $Pred$ of a particular user u is computed by using the following formula:

$$Pred(u) = \sum_{i=1}^{n_u} (\pi_{i,u} \times p_{max\_out}(i, u)), \qquad (2)$$

which corresponds to the sum of the product between each element i of the stationary vector $\pi_{i,u}$ computed from the MMC of user u, in which $\pi_{i,u}$ is the probability of being in a particular state (for n_u, the total number of states of the MMC of user u) and $p_{max\_out}(i, u)$ represents the maximum outgoing probability leaving from the i^{th} state.

Based on these three mobility characteristics (*i.e.* entropy, predictability and average traveled distance per day), we have analyzed the dataset in order to observe their spatial distributions (Figure 5). First, with respect to the distance, we can observe that most of the individuals do not travel more than 250 kilometers per day on average. Second, we looking at predictability we can see clearly that a threshold of 35% seems to divide the population into two groups. Finally, the distribution of the entropy seems to be spread rather uniformly between 0 and 6 bits.

Based on these observations as well as on the description of the different modes of transportation (Section III), we propose to partition the users contained in the dataset into the following categories whose characteristics are summarized in Table III. We recognize that this proposition of classification is rather heuristically for now and maybe

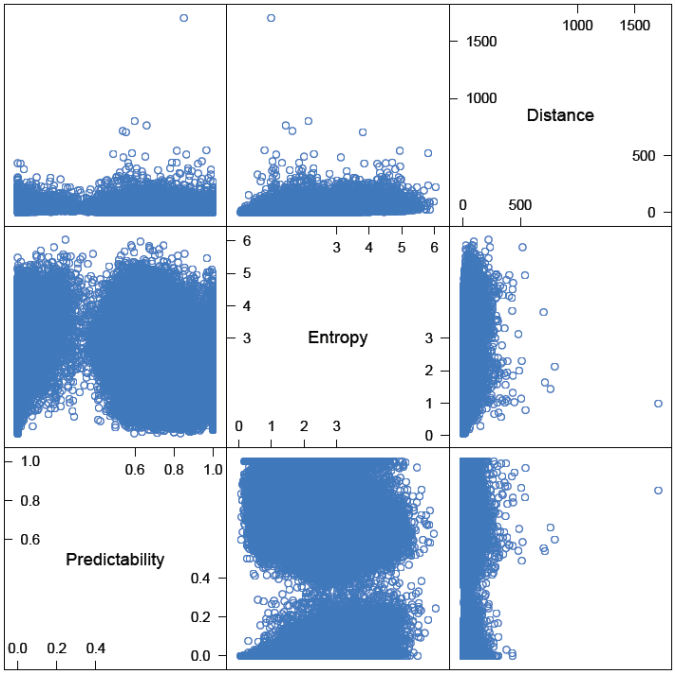


Figure 5. Comparison between the entropy, predictability and average traveled distance per day for the users of the dataset that have at least 40 mobility traces in their history and have visited at least 5 different antennas.

suggest to discussion, thus we consider it only as a first step towards discovering more sophisticated groups of users in the population sharing the same mobility patterns and characteristics. We leave the discovery and characterization of these more refined categories as future work.

Category	Predictability	Entropy	Distance
Individual	High $(0,35 \leq p \leq 1)$	Weak $(e < 3)$	Weak $(0 \leq d < 50)$
Gbaka	High $(0,35 \leq p \leq 1)$	Weak $(e < 3)$	Average $(50 \leq d \leq 150)$
Woro-woro	Weak $(0 \leq p < 0,35)$	Weak $(e < 3)$	Average $(50 \leq d \leq 150)$
Bush taxi	High $(0,35 \leq p \leq 1)$	High $(3 < e)$	High $(150 < d)$
Traditional taxi	Weak $(0 \leq p < 0,35)$	High $(3 \leq e)$	Average $(50 \leq d \leq 150)$

Table III
SUMMARY OF THE CHARACTERISTICS DEFINING EACH CATEGORY OF TRANSPORTATION MODE.

When analyzing the dataset with this characterization, we obtain respectively 77 674 individuals, 1 977 gbakas, 4 861 woro-woros, 8 565 bush taxis and 2 842 traditional taxis.

VII. RECOMMENDER SYSTEM FOR BUSH TAXIS

In this section, we propose a recommender system suggesting to individuals a transportation mode among the following set $TM = \{bush\_taxi, woro\_woro, gbaka, traditional\_taxi\}$, which corresponds to a vehicle corresponding to this mode

of transportation whose position is spatiotemporally close to the one of individual considered and whose destination matches the next location of an individual. Consider for instance, the scenario in which Alice is currently at "Home" at 8AM and she wants to go to the POI corresponding to "Work". She starts the application corresponding to the recommender system on her smartphone in order to retrieve a transport passing in her neighborhood (*e.g.*, at most 1 kilometer for her current position) in the next 30 minutes and whose destination matches her own destination.

More precisely, the recommendation algorithm works in two phases. During the first phase, the recommender system takes as input the location l of the individual and looks for other users belonging to the one of the categories of the set TM that are spatiotemporally close to l. Afterwards, during the second phase, the recommender system predicts $next\_loc\_user$, the next place that will be visited by the user, as well as $next\_loc\_transport$ the next destinations of the neighboring transports in order to find a potential match (*i.e.*, $\mathsf{Dist}(next\_loc\_user, next\_loc\_transport) < \delta$, for δ a predefined distance threshold). In order to minimize the cost of the travel for the user, the system tries to find first a match from the "bush taxi" category. If at least one match is found in this category, then the recommender system returns the match from this category whose destination is the closest to the destination of the user. If no match is found, then the recommender system looks for a match in the following categories until one is found: "woro-woro", then "gbaka" and finally "traditional taxi".

We consider the three following strategies in order to compute the spatial closeness between individuals.

1) The first strategy, which is based on the *current position* of the user, searches for transports located from distance d from the current position l of the individual in the next 30 minutes.
2) The second strategy based on the *trajectory* takes as input the actual position l of the individual and computes the minimal Euclidean distance d from the current position to the path obtained by joining the point of origin and the point of destination of the transport.
3) The third strategy, which is called the *anywhere* strategy, models the situation in which a vehicle is close to the user position (*e.g.*, a taxi picks up a passenger and drives him directly to his next location).

Based on these strategies, we have designed different methods for picking up and dropping off passengers, modeling the difference between the transports that have a fixed or a flexible route. We have applied these strategies to the different transportation means existing in Ivory Coast as summarized in Table IV. For the pick up method, we rely on the current position strategy for all transportation modes. However, for the drop off method, the trajectory strategy is used for bush taxis and gbakas while the current position strategy is applied for the woro-woros and the anywhere strategy is used for traditional taxis.

Transportation	Pick up method	Drop off method
Bush taxi	Current position	Trajectory
Woro-woro	Current position	Current position
Gbaka	Current position	Trajectory
Traditional taxi	Current position	Anywhere

Table IV
PICK UP AND DROP OFF METHODS DEPENDING ON THE TRANSPORT MODE CONSIDERED.

The recommendation algorithm takes as input P, a set of pedestrians (*i.e.*, individuals) and V a set of potential vehicles belonging to the four categories of transportation mode $TM = \{bush\_taxi, woro\_woro, gbaka, traditional\_taxi\}$. For each pedestrian $p \in P$, the algorithm predicts $next\_loc\_user$, the next location visited by the user. Then, the recommendation algorithm looks for possible transports that are spatiotemporally closed to p based on the current position strategy mentioned previously starting from the "bush taxi" category. Once all the potential vehicles have been identified in this category, the next destination $next\_loc\_transport$ is predicted for each vehicle of this category that are spatiotemporally close to the user until a match is found (*i.e.*, $\mathsf{Dist}(next\_loc\_user, next\_loc\_transport) < \delta$). If a match is found, then the recommendation algorithms suggests it to the pedestrian, otherwise this process is then repeated for each transportation mode continuing with the woro-woro, gbaka and then finally the traditional taxi until a match is found. More precisely, the goal is to recommend a transportation mean to a user while at the same time minimizing the travel cost, which explained why we use this particular order. If the recommender system is not able to suggest a transport to a given pedestrian, then it is considered to have failed to provide a recommendation for this particular input.

On order to predict the next location visited by a user (pedestrian or transport), the recommendation algorithm relies on the MMC learnt from the mobility traces of this user. More precisely, the algorithm finds the state of the MMC corresponding to the actual position of the user and then predicts as the next location visited by the user the POI corresponding to maximal outgoing probability leaving from the current state (*i.e.*, ties are broken arbitrarily). In the next section, we will evaluate the practical performance of the recommendation algorithm on the D4D dataset.

VIII. EMPIRICAL EVALUATION

In this section, we detailed the preliminary empirical evaluation that we have conducted to assess the efficiency of the recommender system for bush taxis. Note that contrary

to most of the previous works detailed in Section II, our goal is not to maximize the occupancy of taxis or to minimize the waiting time for passengers. Rather, our main objective is to provide each pedestrian with the recommendation about a transport going in the same direction as his next destination while at the same time selecting the cheapest transportation mode among the following categories: bush taxi, woro-woro, gbaka and traditional taxi. One of the main difficulty compared to previous work is that we do not have detailed GPS traces for the vehicles, instead we only have access to the sporadic exposure of their location data. In order to evaluate, the accuracy of the predictions made by MMCs for next location as well as the recommendation algorithm, we designed two metrics, the *prediction accuracy* as well as the *coverage*, that we described thereafter. These metrics have been assessed on the categories of users introduced in Section IV.

The *prediction accuracy* quantifies the capacity of a mobility model (in our case a MMC) to predict the next location visited by a user based on its current location. In our experiments, we have considered only users that have at least 40 mobility traces and that have visited at least 5 antennas. For each user, we split the trail of mobility traces into two sets of same size: the *training set*, which is used to build the MMC, and the *testing set*, which is used to evaluate its accuracy in predicting the next location visited by a user. Contrary to the predictability (*cf.* Section VI), which is a theoretical measure, the prediction accuracy of a MMC is computed on the testing set, which is completely disjoint from the training set on which the MMC is learnt.

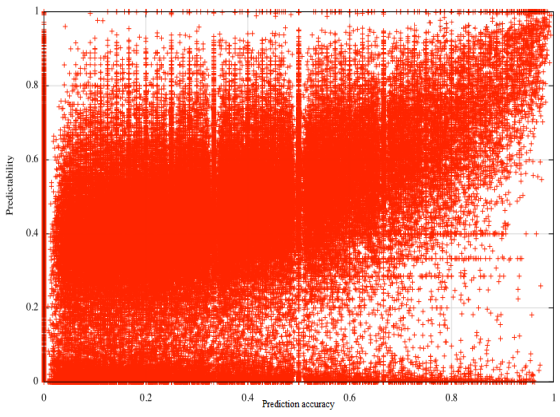


Figure 6. Comparison of the prediction accuracy versus the predictability for the MMCs.

Figure 7 depicts the correlation between the (theoretical) predictability and the (empirical) prediction accuracy. The average predictability is approximately of 50% while the average prediction accuracy is of 30% with an average absolute difference between the two of 28% and a variance of 6.5%. These results are explained by the fact that mobility GSM traces are coarser and less regular than GPS traces. Comparing to previous work done by the authors [3] the prediction accuracy is less than when a MMC is trained on richer traces (*e.g.*, mobility traces obtained from GPS-enabled devices). We observe that some users might display a poor predictability but a high prediction accuracy, we conjecture that this is due to the fact that these users have few antennas in their test set. In contrast, another group of users displays a high predictability but a poor prediction accuracy, due to the presence in the test set of antennas (*i.e.*, states) that the MMC has never seen in the training set when it was learnt.

In a nutshell, the *coverage* measures the percentage of users that have been served by the recommender system. More precisely, considering a set composed of n individuals, we consider for each user the mobility traces of their testing set and we try for each of these traces to suggest a potential transport for the user based on the recommendation algorithm described in Section VII. If a potential match is found then we consider that the system has succeeding in proposing a recommendation to the user, while otherwise we consider that it has failed. The coverage is then averaged over all the pedestrians and all the mobility traces of their test set. To evaluate the coverage, we have used 10% of the users coming from the "individual" category sampled at random (roughly 8000 users), which we believe correspond mainly to pedestrians. The spatiotemporal closeness considered for the recommendation process is respectively of 30 minutes for the time and 1 kilometer for the distance. Recall that the recommendation algorithm looks first for a match in the bush taxi category, before going through the gbaka, then the woro-woro and finally the traditional taxi categories. Overall, we observe a high coverage of 99% for the recommender system, which is decomposed respectively over 72% for bush taxis, 15% for gbakas, 1% for the woro-woros and finally 11% for the traditional taxis. Therefore, we observe that almost 3 times out of 4 the recommender system is able to provide a recommendation that corresponds to the cheapest mode of transportation. Of course, the results we have obtained are only preliminary and we plan to validate them by running the experiments on the complete set of users who have more than 40 mobility traces on their history and that have visited a least 5 different antennas.

IX. Conclusion

In this paper, we have introduced a recommender system for bush taxis in Ivory Coast whose main objective is to propose to pedestrians potential means of transportation in their neighborhood whose destination match their own destination by relying on a mobility model called Mobility Markov Chain. One of the main advantage of our recommender system is that it is fully automatic, in the sense that a user does not need to explicitly express his next destination but

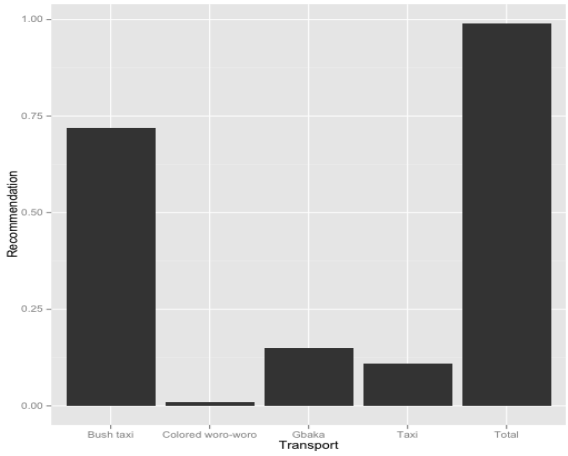


Figure 7. Distribution of the coverage over the different possible modes of transport.

rather the system tries to infer it based on his past mobility behavior. Moreover, the recommendation algorithm is biased towards suggesting the means of transportation that are the cheapest (if one is available). The preliminary evaluation of the recommender system shows that approximately 99% of the time in less than 30 minutes, the system is able to suggest a mean of transportation that is at most 1 kilometer away from the current position of the user for an accuracy of the prediction of the next location that is between 30% and 50% depending on the mobility category to which this user belongs.

As future works, we plan to validate the results of the preliminary analysis on the complete part of the D4D dataset composed of users that have at least 40 mobility traces in their history and that have visited a least 5 different antennas. We also want to evaluate the possible trade-off between the accuracy of the prediction and the coverage of the recommendation algorithm, possibly by combining them into a global metric *à la* F-measure quantifying the success of the recommender system. With respect to the mobility model considered for the prediction of the next location, we also want to apply more sophisticated variants of MMCs that incorporate the notion of time slices (*i.e.*, they can predict the next location depending on the current period of the day) or that can remember the k last visited locations by a user. Moreover, we also want to conduct an evaluation measuring the economical impact on the average cost of travel for a user of using different recommendation strategies. Finally, we acknowledge that the current idea of a recommender system for bush taxis is still very crude and we are currently working on refining it by making it more adaptive and flexible to the current context.

REFERENCES

[1] Daniel Ashbrook and Thad Starner. Learning significant locations and predicting user movement with GPS. In *Proceedings of the International Symposium on Wearable Computers*, volume 6, pages 101–108, Sardina, Italy, February 2002.

[2] Vincent D. Blondel, Markus Esch, Connie Chan, Fabrice Clérot, Pierre Deville, Etienne Huens, Frédéric Morlot, Zbigniew Smoreda, and Cezary Ziemlicki. Data for development: the d4d challenge on mobile phone data. *Computing Research Repository*, 1210(137):1–10, September 2012.

[3] Sébastien Gambs, Marc-Olivier Killijian, and Miguel Núñez del Prado Cortez. Next place prediction using mobility Markov chains. In *Proceedings of the First Workshop on Measurement, Privacy, and Mobility*, volume 3, pages 1–6, Bern, Switzerland, April 2012.

[4] Sébastien Gambs, Marc-Olivier Killijian, and Miguel Núñez del Prado Cortez. Show me how you move and I will tell you who you are. In *Transactions on Data Privacy*, volume 4, pages 103–126, August 2011.

[5] Yong Ge, Chuanren Liu, Hui Xiong, and Jian Chen. A taxi business intelligence system. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 735–738, New York, NY, USA, August 2011.

[6] Yong Ge, Hui Xiong, Alexander Tuzhilin, Keli Xiao, Marco Gruteser, and Michael Pazzani. An energy-efficient mobile recommender system. In *Proceedings of the 16th ACM international conference on Knowledge discovery and data mining*, volume 10, pages 899–908, New York, NY, USA, July 2010.

[7] Junghoon Lee, Gyung-Leen Park, Hanil Kim, Young-Kyu Yang, PanKoo Kim, and Sang-Wook Kim. A telematics service system based on the linux cluster. In *International Conference on Computational Science*, pages 660–667, Krakow, Poland, June 2007.

[8] Junghoon Lee, Inhye Shin, and Gyung-Leen Park. Analysis of the passenger pick-up pattern for taxi location recommendation. In *Proceedings of the Fourth International Conference on Networked Computing and Advanced Information Management*, pages 199–204, Washington, DC, USA, September 2008.

[9] James MacQueen et al. Some methods for classification and analysis of multivariate mbservations. In *Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, Berkeley, CA, USA, July 1967.

[10] UITP International Association of Public Transport. Aperçu du transport public en afrique subsaharienne. http://www.uitp.org/knowledge/pdf/transafrica_fr.pdf, 2009.

[11] UITP International Association of Public Transport. Public transport in sub-saharan africa, major trends and cadse study. http://www.uitp.org/knowledge/pdf/PTinSSAfr-Majortrendsandcasestudies.pdf, 2010.

[12] Orange. D4d challenge dataset. http://www.d4d.orange.com/learn-more.

[13] Claude Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 23(3):379–423, October 1948.

[14] Zhixian Yan, Dipanjan Chakraborty, Christine Parent, Stefano Spaccapietra, and Karl Aberer. SeMiTri: a framework for semantic annotation of heterogeneous trajectories. In *Proceedings of the 14th International Conference on Extending Database Technology*, pages 259–270, Uppsala, Sweden, March 2011.

[15] Jing Yuan, Yu Zheng, Liuhang Zhang, XIng Xie, and Guangzhong Sun. Where to find my next passenger. In *Proceedings of the 13th international conference on Ubiquitous computing*, pages 109–118, Beijing, China, September 2011.

[16] Nicholas Jing Yuan, Yu Zheng, Liuhang Zhang, and Xing Xie. T-finder: A recommender system for finding passengers and vacant taxis. *IEEE Transactions on Knowledge and Data Engineering*, 99:1–14, August 2012.

[17] Xudong Zheng, Xiao Liang, and Ke Xu. Where to wait for a taxi? In *Proceedings of the International Workshop on Urban Computing*, pages 149–156, New York, NY, USA, August 2012.

Real-time streaming mobility analytics*

András Garzó[1,2], István Petrás[1,3], Csaba István Sidló[1,2] and András A. Benczúr[1,3]

[1]Computer and Automation Research Institute, Hungarian Academy of Sciences (MTA SZTAKI),
{garzo, petras, scsi, benczur}@ilab.sztaki.hu
[2]Faculty of Informatics, University of Debrecen
[3]Eötvös University, Budapest

Abstract

In our research we tested distributed streaming algorithms and infrastructures to process large scale mobility data for fast reaction time prediction. We use the D4D Challenge data set as a source to generate, by multiplying with noise, even larger realistic data sets.

Instead of addressing the problem of identifying the exact location and movement of an individual user (that the data set is not sufficiently detailed for), we learn global patterns both on the user level (home, work location, daily routes) and the traffic (typical routes at time of the day).

As a key performance indicator of our applications, we measure the running time and the error of predictions in short range (5 minutes to 1 hour) and long range (daily, weekly) of the location of an individual user and the density in a given area. Over a cluster of a few old dual core servers, we are capable of processing tens of thousands of record in a second. Our results open the possibility for efficient real time mobility predictions of even large metropolitan areas as well.

We demonstrate our solution via a fast reaction visual dashboard application that can form the base of emergency or rescue services as well as provide grounds for ride sharing, traffic planning and optimization, thus saving natural resources.

Key words: Mobility, Big Data, Data Mining, Visualization, Distributed Algorithms, Streaming Data

*The publication was supported by grant OTKA NK 105645 and the TÁMOP-4.2.2.C-11/1/KONV-2012-0001 project. The project is supported by the European Union, co-financed by the European Social Fund.

1 Introduction

About 60 percent of the world's population are using mobile phones, more than 4 billion people, and 12% of them are smartphones, growing more than 20% a year. More than 30 million networked sensor nodes are now present in many important sectors as transportation, automotive, industrial, utilities and retail sectors. The number of sensors is increasing at a rate of 30% a year. All the data produced finds invaluable use beyond their primary domain. Already relative early papers on the use of mobility traces [12] list several potential applications, including "Traffic services: Information concerning the position [...] to deliver news about congestion and suggestions for alternative routes"; "Navigation aids: Information concerning a user's position, direction, and targets interfaced with GIS" and "Urban systems mapping: large amounts of data gathered in anonymous and aggregated form [...] whose dynamics are described on the basis of people's activities and movements in space".

Our goal is to demonstrate the possibility of real time traffic prediction by using the D4D Challenge Data Set[1] as support for real time navigation and traffic optimization. Intelligent Transportation is at the center of worldwide transport policies intending to consolidate efficient and sustainable transport systems and associated infrastructures. The belief is that smart systems can receive, manage and provide valuable information that will allow transport users and operators to deal with a vast variety of traffic situations: congestions, safety, tolling, navigation support, law enforcement, as well as environmental sustainability (e.g. reducing greenhouse gas emissions, fuel consumption, infrastructure maintenance).

Mobility traces are produced at a very high rate and continuously, making every purely static approach hopeless. Data arrive at a breakneck rate, and they should be analyzed, cleansed, stored, categorized immediately and in a highly adaptive manner. The expressions "Large-Scale Data" or "Big Data" refer to datasets whose size is beyond the ability of typical software tools to capture, store, manage and analyze. Telecommunications data is typically various orders of magnitude larger than any similar data resource available before, making any super-linear algorithm essentially useless. This situation pushes towards new algorithms (typically, approximated and/or distributed ones) and new computational frameworks (e.g., MapReduce, NoSQL and streaming data). In all scenarios, however, scalability is crucial factor to enable any future technology to deal with these datasets.

In our research the emphasis is on the algorithmic and software scalability of the prediction method. Although publications with similar goals have appeared, even recent results [1] consider data sets of similar or smaller size compared to D4D. In contrast, we expect that the mobility data of a large metropolis will be two to three orders of magnitude larger, especially since for accurate prediction all communication with the base stations should be considered, not just the ones associated with voice, text or data traffic as provided in the D4D data set. In

[1] `http://www.d4d.orange.com/home`

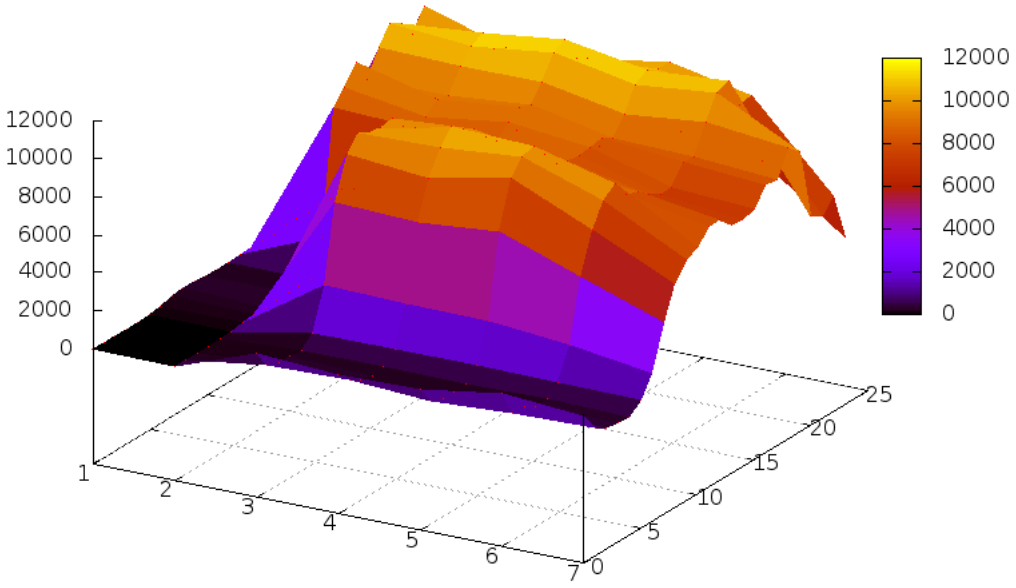


Figure 1: Volume of traffic (vertical axis) as the function of the day of the week (1–7) and hour (0–23) over the horizontal axes. Except for a slight change on Friday and Sunday, we only observe the possible lack of data over Monday and Tuesday night that we consider an artifact.

a metropolitan area of several million people using mobile devices all day, we may expect an event from each device in average in every minute. Hence our target is to be able to process events in the order of 100,000 in a second.

Real time traffic prediction, as opposed to offline city planning, requires processing the incoming data stream without first storing, cleansing and organizing it in any sense. While certain mature frameworks such as Hadoop [15] reach the required level of scalability, they cannot provide mechanisms for streaming input and real time response. Our choice is hence limited to data streaming frameworks. We choose Storm [10], a scalable distributed streaming computational framework developed by Twitter and used by many companies for wide purposes. Storm also supports the MapReduce computational model but allows more flexible data processing operations.

Our data is formed by splitting Fine Resolution Mobility Trace Data Set of the Challenge into two week chunks of the same user mobility and considering

the resulting data as if coming from a single two weeks period. The resulting weekly aggregated traffic volume is shown in Fig. 1, indicating that considering the time of the day only may be a good approximation for user motion.

The fact that only two-week user histories are available in the data set certainly poses limitations for our experiments, as we cannot consider seasonality or specific events, and even weekends are hard to handle due to the lack of a sufficiently long period of observations. Unlike initially planned, we could not perform flock detection (motion of groups) and deviation of real track from expected (map) or permitted (restricted areas) tracks. As another remark, we do not consider cellphone network positioning techniques and their limitations either since, in our opinion, this field is rapidly developing and the availability of technologies will pose no limit on the localization in the near future.

Our prediction method is based on a simple combination of user and cell tower frequent patterns. By also taking the limitations of the data into account, our intent was to produce the simplest yet realistic framework for location prediction. Our system mines typical user patterns such as home and work location at given time of the day as well as typical routes through a given cell, again at the given time of the day such as inbound vs. outbound traffic in the morning and the afternoon. Due to the limitations of the data set, we can mine no unusual patterns such as sports events, accidents, road construction or unusual traffic jam. We believe our framework can very easily be extended to handle and learn such more complex patterns as well.

We evaluate our method both for speed and accuracy. While we consider accuracy as illustration only, we measure both the average distance of the user from her predicted location at her next event and the predicted near future cell density. Since, unlike in a potential real application where service communication logs are also available, we have to deal with varying length holes in the observation: the next voice, text or data traffic of the same user may happen only hours later. For this reason we predict sequences and evaluate always for the next known location after the predefined prediction period.

Speed is evaluated by measuring the number of records that reach the final prediction module in a second. We also increase the rate of records entering the framework for indications on the maximum load without congestion. Both measurements indicate that on an old cluster of dual core machines with only 4GB RAM each, we may reach rates of a few 10,000 records in a second. We may conclude that the anticipated ten times larger traffic of a metropolis can be handled by a reasonable size cluster. We also observe near linear scalability in the number of servers and threads.

The rest of this paper is organized as follows. In Section 2 we describe the elements of the modeling and prediction framework. Section 3 is devoted to describing the streaming data processing software framework and the details of our algorithm. In Section 4 we give our results, both in terms of accuracy and scalability, over the D4D data set. Finally related results are summarized in Section 5.

Figure 2: Sample visualization of user movement in the Abidjan region.

2 Traffic and location prediction modeling

In order to predict user movement and traffic congestion, we give a very simple model that we have implemented in the available short time frame. First of all we note that the data is only approximate as we only know a device cell position when the device is active, i.e. involved in voice or text traffic. For this reason we only learn motion patterns if the two events are sufficiently close. Also, as we wrap the whole year into a two-week period where we use the first for training and the second for testing, we mine no seasonality, not even weekly periodicity. In the description we give pointers to enhancements that can easily be implemented within the system to improve the accuracy of the predictions.

We model the typical motion patterns of each user and of each cell tower as visualized e.g. in Fig. 2. Next we describe a model for each case based on frequent patterns. The final prediction is constructed as the combination of the two models extrapolated for a few hours ahead in time, since information on

user location is partial and we will have to evaluate our prediction first when the user initiates a next event over the network (voice or text).

Throughout our methods, we discretize time into 15 minute intervals of the day and consider patterns related to the given interval. In a more advanced setting, one could easily distinguish weekdays from weekends and slice the day into uneven intervals. By manual slicing shorter intervals can be assigned to morning and afternoon rush hours and longer for the night. By automatic slicing, we could allocate roughly the same number of events into one interval that would make their length uneven similar to the manual case. In this simple experiment we did not apply these ideas.

2.1 User model

Our main assumption is that for most users, their location follows a daily regular schedule, i.e. they leave to work and return home at approximately the same time of the day over roughly the same route. This assumption is confirmed for other data sets e.g. in [7]. We consider one and two-step frequent patterns. In the one step model we predict regardless of where the user is at a given time. The most likely location in this model will for example be home for the night and workplace for the day.

In the two-step model we collect frequent patters of movements with respect to a given time of the day. As a simplifying assumption, we consider p_k^u, the location of user u at time k as a Markovian process that can be factorized:

$$p_k^u\left(c_{i_m}t_{j_m},\ldots,c_{i_2}t_{j_2},c_{i_1}t_{j_1}\right) = \\ p_k^u\left(c_{i_m}t_{j_m}|c_{i_{m-1}}t_{j_{m-1}}\right)\ldots p_k^u\left(c_{i_3}t_{j_3}|c_{i_2}t_{j_2}\right)p_k^u\left(c_{i_2}t_{j_2}|c_{i_1}t_{j_1}\right)p_k^u\left(c_{i_1}t_{j_1}\right) \quad (1)$$

where c_i denotes the ith cell tower and t is the time stamp.

We only store the transition probability of a step $p_k^u\left(c_{i_2}t_{j_2}|c_{i_1}t_{j_1}\right)$, $t_{j_2} > t_{j_1}$ and the probability $p_k^u\left(c_i,t\right)$ of user k being at cell tower c_i at time t.

The output of the prediction is a tree of user paths rooted at the last observed location $\{c_{i_1}t_{j_1}\}$. By pruning very low probability branches, we build a moderate size tree with edges weighted by the transition probabilities.

2.2 Cell tower model

As a cell tower is responsible for a given small region of the country or part of the city, we may learn typical directions of movement depending on the time of the day. For each cell tower we compute the joint probability of frequent paths at a sequence of time $t_{j_n} > \ldots > t_{j_2} > t_{j_1}$ as

$$p\left(c_{i_n}t_{i_n},\ldots,c_{i_2}t_{i_2},c_{i_1}t_{i_1}\right). \quad (2)$$

Histograms of frequent paths ending at the the given cell towers are collected. Paths of length $2\ldots n$ and aggregated and passed to the predictor component periodically. For ease of implementation, we considered $n \leq 3$ in this simple experiment. For general n, the problem can be solved by data stream frequent pattern mining algorithms, e.g. by streaming FP-trees [6].

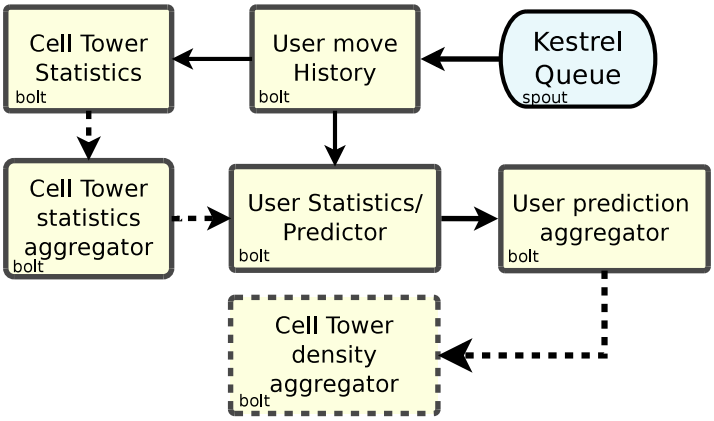


Figure 3: System block diagram of the Storm streaming components. Regular arrows show normal while dashed show the low frequency periodic special "tick" packets.

2.3 Prediction

We generate user step path candidates of length $1, \ldots, n$ and measure the accuracy compared to the real user movement. We fuse the user and cell tower level predictions by a simple linear combination of the probabilities. We aggregate the number of users per cell tower in a time window based on the path predictions. This way we get the cell tower densities in the future that we compare again to the real density at the given time.

3 Streaming framework

In this section we show how we implemented a system responsible for processing the mobility data, store user history data in a persistent store and predict the user path over cell towers.

Our demonstration is based on Storm, a scalable, robust and fault tolerant, real time data stream processing framework. Under Storm a real-time processing job can be defined with the Storm API in a straightforward way. We only have to implement the processing elements using the predefined abstract components: *spouts* responsible for creating input and *bolts* containing the processing step. After that, the components can be connected to each other by streams that transmit data packets called tuples between processing elements as seen in Fig. 3. Data transmission is handled by the framework. We can create an acyclic processing graph with spouts and bolts as nodes and streams as edges. This graph is called the topology that describes how the input data tuples will be processed in the Storm cluster.

After the topology is defined and submitted to a Storm cluster, it will process input until killed. Spouts read input data usually from external sources, e.g.

a message queuing system. Streams from spouts always terminate in a bolt component. Storm can also run multiple instances of spouts and bolts parallel. It can distribute and balance running instances among the nodes in the cluster automatically. Key to a practical application, Storm takes all the burden of scalability and fault tolerance.

In our implementation the raw mobility data was read into the memory by an external application and was put into a lightweight message queuing system called Kestrel [2] with a given frequency. Thus we can simulate how the mobile data arrive in real-time from a service provider.

The spout components in our topology read the input data from the Kestrel queues and send to the user history bolts as in Fig. 3. There can be arbitrary many instances of these components. Each user is assigned to a certain instance. The history tasks preserve history in memory for a certain time window. For persistency, the history components also write these data into a distributed key-value store. Since near real time processing is very important, we tested both Cassandra [9] due to its high throughput writing capabilities and memcached [5] for its efficiency. These distributed key-value stores give persistency to user and cell tower histories and models as well as we write the prediction output for the visualization dashboard.

Next we describe the components of Fig. 3 in detail. We define two types of data flow:

- One regular packet starts from the single spout for each input record that spreads along the thick edges in Fig. 3.
- Periodic aggregation updates move model information along the dashed edges initiated by the special built-in Storm *tick* packets.

We describe our algorithms by specifying the type of data moving along different edges of Fig. 3 and describing the algorithms implemented within each node of the Figure, the bolts that correspond to the steps described in Section 2.

- The spout element emits tuples (u, a, t) of user, cell and time stamp.
- User history elements send fixed length sequences of (a, t) tuples of the past steps both to the last cell statistics bolt for recording the new user location and to the previous cell for counting frequencies through the cell.
- User history elements send trees rooted at the current location (a, t) weighted with the transition probabilities as in equation (1).
- Cell statistics elements periodically submit the frequent patterns to a single cell statistics aggregator bolt.
- The cell statistics aggregator bolt periodically refreshes the cell frequent patterns as in equation (2) to all user statistics predictor bolts.
- User statistics predictors emit the aggregated future history of the user in a form of rooted trees as in equation (1). This element is used in the current experiment to measure the accuracy of the user location prediction.

- User prediction aggregator periodically emits the predicted density of all cells seen in the prediction of the given user for aggregation by the single cell density aggregator element. This element is used in the current experiment to measure the accuracy of the cell density prediction.

4 Experiments

In this section we describe our measurements for speed, scalability and quality. We also describe our demo visualization system. To emphasize scalability in the number of threads and machines, we ran our experiments over a Storm 0.9.0-wip4 cluster of many old dual core Pentium-D 3.0GHz machines with 4GB RAM each.

In order to simulate a real-life deployment, we feed the Storm cluster with data from queuing systems. Each queue can emit approximately 20,000 records in a second and this rate can be multiplied if more than one machines over the cluster serve queues. Note that Storm can, even in these low memory machines, easily cache the entire data set and hence instead of the rate of entering Storm, we measure the rate of processing by the last prediction element, the user prediction aggregator, in Fig. 3.

In the first experiment we test how many servers are needed to process the output of a single queue of rate 20-30,000 in a second. To avoid misleading figures due to caching, we ran the system for 5 minutes (approximately 6M records) before starting to measure the predictor element processing rate. In Fig. 4 we see that the system enters real time processing with roughly 20 servers, after which the rate at the last element is equal to that of the queue output.

Next we start increasing the rate of packets entering the system by deploying more than one Kerstel queues at different servers. After a sufficiently large rate, the server farm will not be able to process the stream real time, however in Fig. 4 we observe linear scaling in the number of servers even for very high input rates.

As an illustration, we measure the quality of the predictions of our simple methods. We compare both user location and cell density predictions to the actual data by splitting it into a training and a testing period. Cell density can simply be measured by averaging, in time, the root relative squared error measure.

User location accuracy is somewhat trickier in that we have only interrupted knowledge of the user location with blank time intervals corresponding to no traffic generated by the user. Also note that the prediction may skip periods of time: in equation (1) we discretize time into 15-minute intervals but the subsequent timestamps t_{j_i} in a frequent pattern may differ by an arbitrary multiple. Hence we use a dynamic time warping algorithm that matches the next real and predicted user locations. In this sense the mean accuracy, i.e. the fraction of correctly prediction user location in a two-day testing period is 87.7%.

We developed a visualization demo application that shows the predicted and

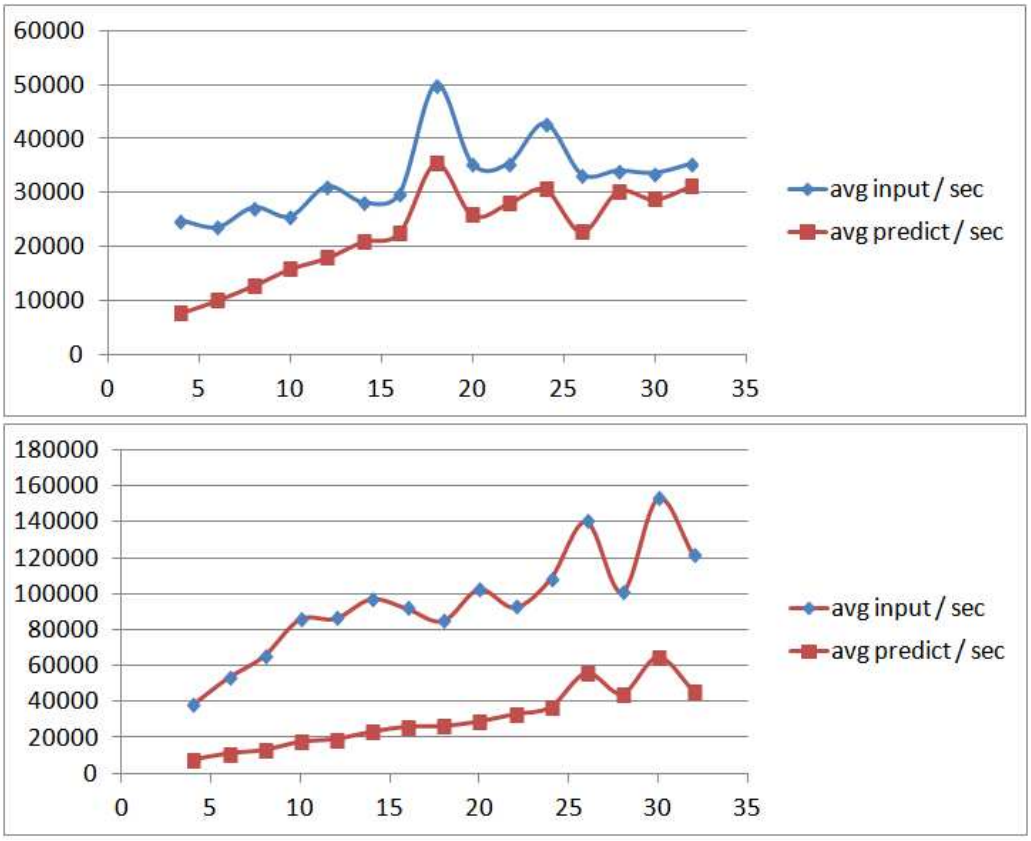


Figure 4: Rate of packets accepted (input per second) and rate of packets used for prediction (predictions per second) as the function of the number of servers in the Storm cluster. **Top:** single input queue with a maximum input rate of 20-30,000 in a second. **Bottom:** the same chart with five input queues residing at five different servers.

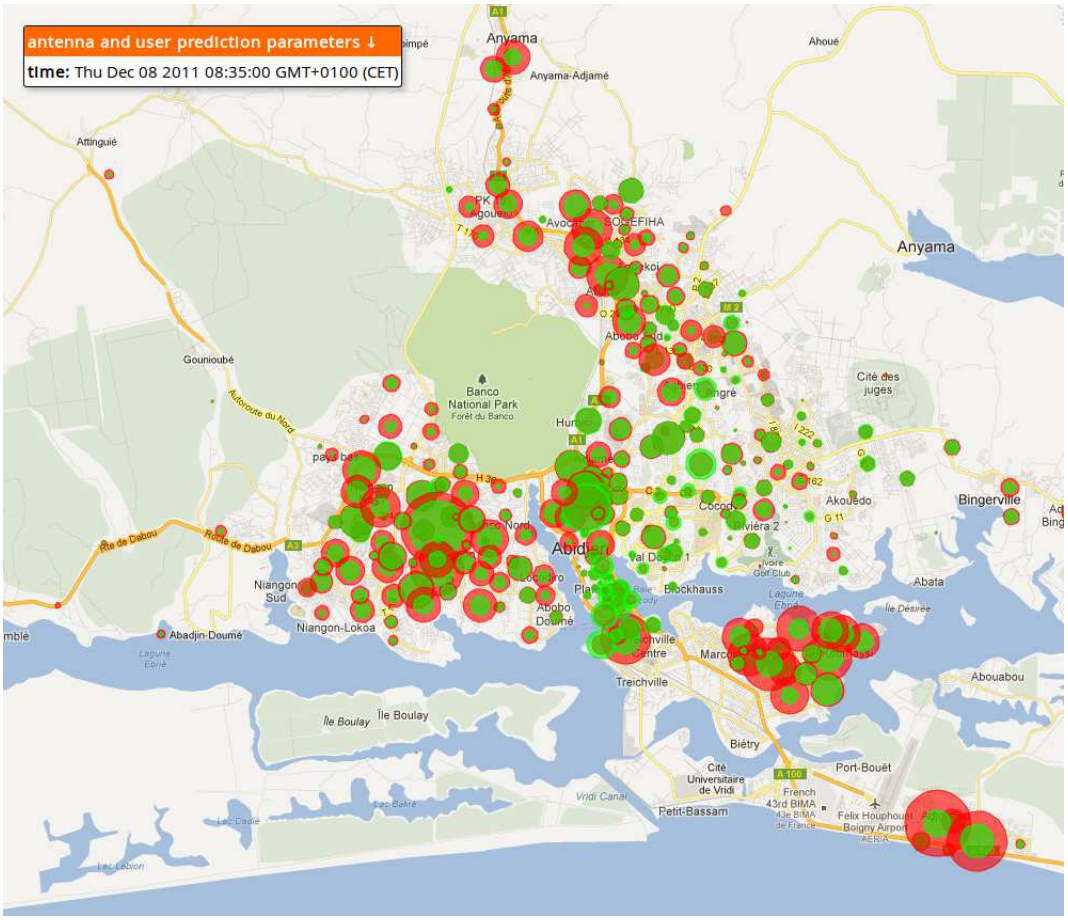


Figure 5: Visualization of the cell traffic prediction. Red circles show actual sizes while green is the prediction.

real cell density (Fig. 5) as well as the predicted and real trajectories of randomly selected users. The demo is accessible by obtaining access from the authors, at `http://bigdatabi.sztaki.hu/d4d/`.

5 Related Results

The idea of using mobility traces for traffic analysis is not new: in [12] a case study of Milan while in [1] of New Your City suburbs is presented. However these results consider static pre-collected data. We aware of no prior results that address algorithmic and software issues of the streaming data source. Mobility and City Planning is considered by major companies. IBM released redguides [8, 13] describing among others their IBM Traffic Prediction Tool [13]. Un-

fortunately little is known about the scalability and the technology of existing proprietary software.

Several recent results [7, 14] analyze and model the mobility patterns of people. In our results we rely on their findings, e.g. the predictability of user traces.

Most important in our research is the use of distributed software frameworks for scalability and fault tolerance. Major social media companies have all developed their software tools [4]. Mobility data naturally requires data streaming frameworks [10, 11, 3]. Apparently Storm is a mature framework that let us fit learning and implementation within the short time frame of the D4D Challenge.

6 Conclusions

In this preliminary experiment we demonstrated the applicability of data streaming frameworks for processing mass mobility streams for real-time traffic and motion prediction. Given more detailed data, our framework is suitable for detecting flock (motion of groups), deviation of real track from expected (map) or permitted (restricted areas) tracks. Our results open the ground for advanced experimentation regarding the quality of large scale mobility prediction suitable for example for navigation applications.

7 Acknowledgment

Among several colleagues who draw our attention to the scalability issues of mobility data and inspired our research, we are especially grateful, in alphabetic order, to Tamas Bodis (Vodafone Hungary), Csaba Fekete (NSN Hungary), Zalan Heszberger (Ericsson Hungary and TU Budapest), Attila Medvig (NNG) and Janos Tapolcai (TU Budapest).

References

[1] R. A. Becker, R. Caceres, K. Hanson, J. M. Loh, S. Urbanek, A. Varshavsky, and C. Volinsky. A tale of one city: Using cellular network data for urban planning. *Pervasive Computing, IEEE*, 10(4):18–26, 2011.

[2] E. D. Dolan, R. Fourer, J.-P. Goux, T. S. Munson, and J. Sarich. Kestrel: An interface from optimization modeling systems to the neos server. *INFORMS Journal on Computing*, 20(4):525–538, 2008.

[3] M. Dusi, N. d'Heureuse, F. Huici, A. di Pietro, N. Bonelli, G. Bianchi, B. Trammell, and S. Niccolini. Blockmon: Flexible and high-performance big data stream analytics platform and its use cases. *NEC TECHNICAL JOURNAL*, 7(2):103, 2012.

[4] D. Eyers, T. Freudenreich, A. Margara, S. Frischbier, P. Pietzuch, and P. Eugster. Living in the present: on-the-fly information processing in scalable web architectures. In *Proceedings of the 2nd International Workshop on Cloud Computing Platforms*, page 6. ACM, 2012.

[5] B. Fitzpatrick. Distributed caching with memcached. *Linux journal*, 2004(124):5, 2004.

[6] C. Giannella, J. Han, J. Pei, X. Yan, and P. S. Yu. Mining frequent patterns in data streams at multiple time granularities. *Next generation data mining*, 212:191–212, 2003.

[7] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[8] M. Kehoe, M. Cosgrove, S. Gennaro, C. Harrison, W. Harthoorn, J. Hogan, J. Meegan, P. Nesbitt, and C. Peters. Smarter cities series: A foundation for understanding ibm smarter cities. *Redguides for Business Leaders, IBM*, 2011.

[9] A. Lakshman and P. Malik. Cassandra: A structured storage system on a p2p network. In *Proceedings of the twenty-first annual symposium on Parallelism in algorithms and architectures*, pages 47–47. ACM, 2009.

[10] J. Leibiusky, G. Eisbruch, and D. Simonassi. *Getting Started With Storm*. Oreilly & Associates Incorporated, 2012.

[11] L. Neumeyer, B. Robbins, A. Nair, and A. Kesari. S4: Distributed stream computing platform. In *Data Mining Workshops (ICDMW), 2010 IEEE International Conference on*, pages 170–177. IEEE, 2010.

[12] C. Ratti, S. Williams, D. Frenchman, and R. Pulselli. Mobile landscapes: using location data from cell phones for urban analysis. *ENVIRONMENT AND PLANNING B PLANNING AND DESIGN*, 33(5):727, 2006.

[13] S. Schaefer, C. Harrison, N. Lamba, and V. Srikanth. Smarter cities series: Understanding the ibm approach to traffic management. *Redguides for Business Leaders, IBM*, 2011.

[14] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

[15] T. White. *Hadoop: The Definitive Guide*. Yahoo Press, 2010.

D4D Challenge – Report

Visualization of traffic

Jernej Bodlaj · Monika Cerinšek · Vladimir Batagelj

February 15, 2013

Abstract The paper deals with the problem of mobile network traffic visualization of Ivory Coast. This mobile network is quite dense and consists of more than a thousand base stations - antennae. To draw it clearly, we have to develop some special approaches. Our aim is to visualize traffic (duration and numbers of calls) between antennae as clear as possible. All traffic diagrams, we use, are two-dimensional and geographically oriented, but ideas behind vary. We use classic node and links diagrams in which the links are represented with lines. We introduce dispersed line diagrams, which are used to make representations more realistic. Further we display data with star diagrams. We generate diagrams mostly programmatically using a specific program, developed for the challenge. Some partial data diagrams are included in the work, e.g. filtered traffic diagrams and some supplementary diagrams are included to make descriptions of methods more transparent. We included also two videos of traffic evolution in time and diagrams of antenna clustering.

Keywords mobile network analysis · mobile network traffic visualization · line-diagram · dispersed-line diagram · star diagram · traffic evolution video

J. Bodlaj
Hruška d.o.o., Kajuhova 90, 1000 Ljubljana
Tel.: +386-15423614
Fax: +386-59022240
E-mail: jernej.bodlaj@hruska.si

M. Cerinšek
Hruška d.o.o., Kajuhova 90, 1000 Ljubljana
Tel.: +386-15423614
Fax: +386-59022240
E-mail: monika@hruska.si

V. Batagelj
University of Ljubljana, FMF, Department of Mathematics, Jadranska 19, 1000 Ljubljana, Slovenia
E-mail: vladimir.batagelj@fmf.uni-lj.si
URL: http://pajek.imfm.si

1 Introduction

In this report we show results of our work in regard to the application to the D4D challenge offered by France Telecom - Orange. We focus mostly on static geographical visual representations of traffic between base stations in Ivory Coast. Traffic is represented by either the duration of calls or by the number of calls between arbitrary antennae. In an attempt to not fully neglect a time dimension of mobile network, we also rendered two videos to visualize the evolution of the network in time. We got some insight also on the correlation between types of traffic, how duration and number of calls are related and further, how traffic is related to the distance between pairs of antennae where it was carried. We present also some results of antenna clustering in a quite unique way and some other diagrams are included, which show only the most dominating traffic in two distinctive manners.
Our work is mostly hypothetical and experimental. We used some novel techniques to show the traffic which will be explained in details later.

2 Data

All still pictures in sections 3.1, 3.2, 3.3 were obtained from data in set 1 described in [1]. The cumulative sum of durations and numbers of calls between antennae from the whole observation period was aggregated and used to generate two cumulated networks. The first network consists from antennae as nodes and from aggregated sums of durations between them as links. The second network has antennae in place of nodes and aggregated sums of numbers of calls between them for links. Both networks were used as a source also for analysis in sections 3.4, 3.5, 3.7, 3.8, where we were looking for correlations between duration and number of calls and distance of calls.
Data for videos were obtained by the same principle like cumulated networks for still images in previous paragraph, but instead from the whole observation period, a period of individual day was taken into consideration to generate each frame of the video.
Diagrams from section 3.9 were generated from data in a second set described in [1]. The procedure of data transformation into useful form is precisely described in section 3.9.

3 Analysis

3.1 Standard nodes and links approach

Distribution diagrams of numbers of calls and durations of calls between antennae show geographically the number and duration of calls passing through every region, i.e. pixel of the diagram of a country. Sizes of diagrams are in order of 5000×5000 pixels as Ivory Coast is approximately square like country. A trivial diagram was constructed by a number of lines, connecting the

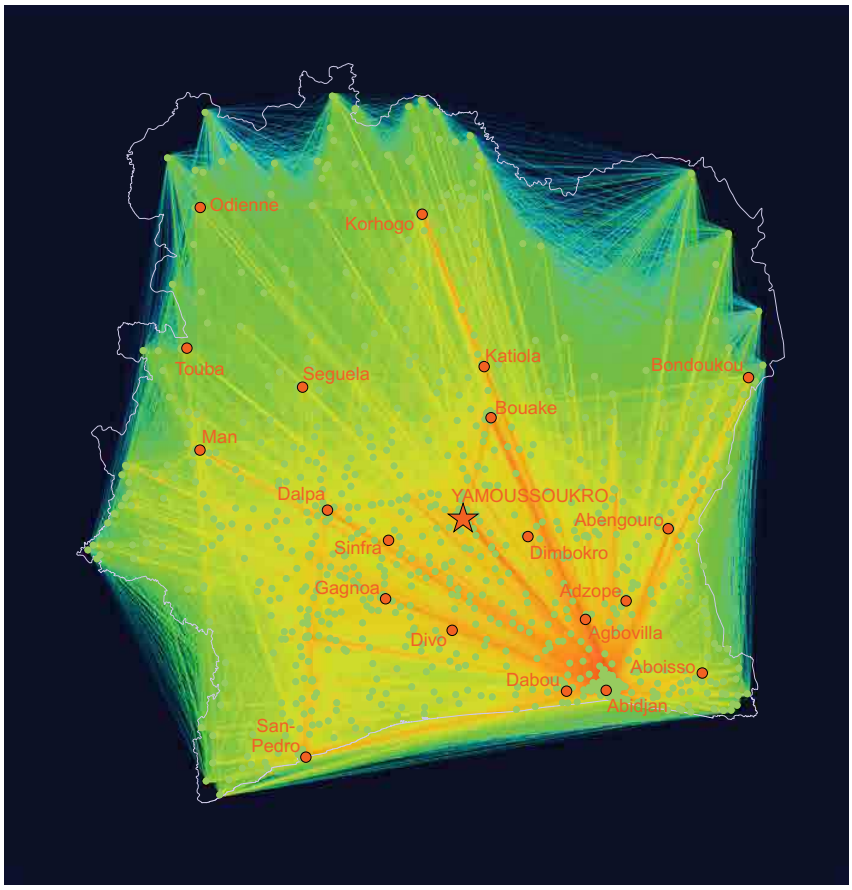


Fig. 1 The diagram of duration of calls between antennae, displayed in logarithmic scale on Fig. 3.

locations of pairs of antennae and with the intensity linearly proportional to the number of calls or duration for the pair of antennae in question for the whole period of observation. Lines were drawn additively, which means line intersections have an intensity of sum of intersecting lines. A weakness of the trivial line diagram is in a high probability of having nearby regions (in the most severe case two neighbor pixels on the diagram) with a high intensity difference. The effect may be clearly observed in a form of sharp edges on both trivial diagrams for duration and numbers of calls on Fig. 1 and Fig. 2 respectively.

We generated video representations[1] of a time evolution of networks of duration of calls and numbers of calls. Each frame in each video takes cumula-

[1] http://zvonka.fmf.uni-lj.si/netbook/lib/exe/fetch.php?id=project%3Ad4d%3Aindex&cache=cache&media=project:d4d:pub:d4dvideos.ra

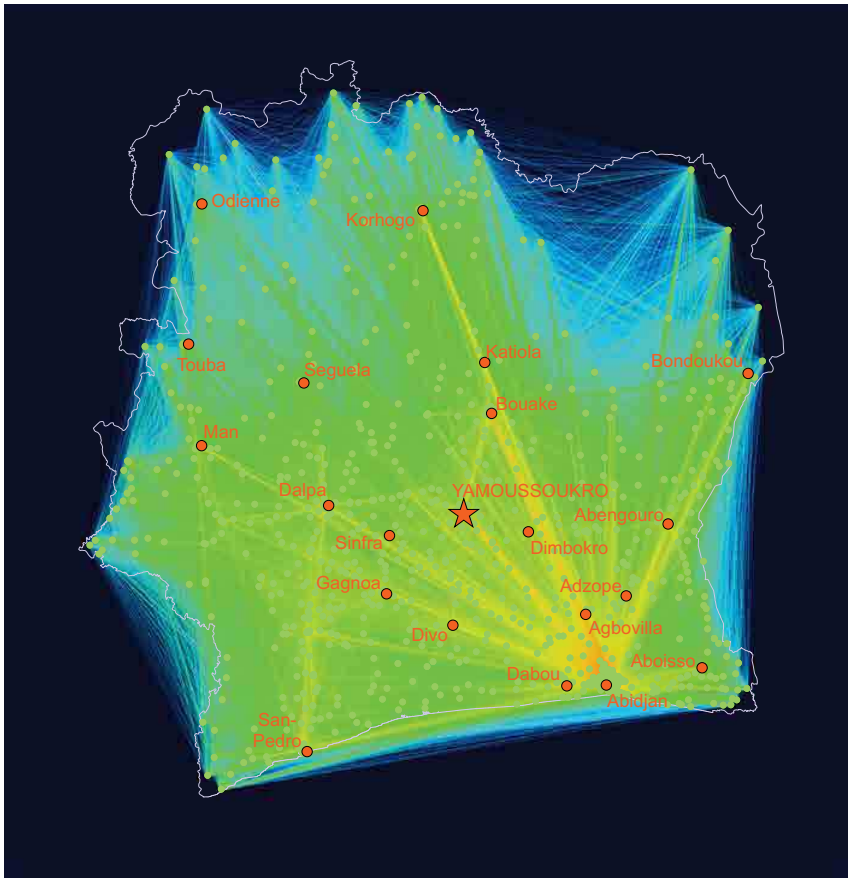


Fig. 2 The diagram of numbers of calls between antennae, displayed in logarithmic scale on Fig. 3.

Fig. 3 A scale used throughout all color-like diagrams. Deep blue represents low values, bright red represents high values.

tive traffic from one day between all pairs of antenna. The datum and a day of weak is written in the left corner of each video, otherwise video frames do not differ in any way from still images on Fig. 1 and Fig. 2 and all comments about those two figures also apply to these videos. A note about normalization; both videos were individually normalized to the highest value of all their frames.

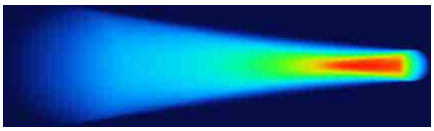


Fig. 4 An example of a dispersed line-like sample with $r_1 = 60, r_2 = 16, d = 348$. All parameters are in pixels.

3.2 Dispersed line representation

To overcome the problem of edginess, we constructed dispersed line-like probability samples and map them additively between antennae pairs instead of lines and obtained dispersed diagrams on Fig. 6 and Fig. 7. The idea behind dispersed line-like probability samples is that a caller is located, for the sake of simplicity, entirely randomly somewhere inside the radius of source antenna and not necessarily at the exact location of the source antenna and the receiver is also randomly located somewhere inside the radius of sink antenna. A hypothetical call or its duration would be represented by the line between the random positions of the caller and the receiver.
Additively drawing a sufficient number of lines between randomly positioned hypothetical callers and receivers inside their antennae radii renders a non-normalized approximation of the sample for the chosen antenna pair, which we then normalize to a sum equal to 1. A sufficient in a previous sentence for us means enough to be more or less unable to see any improvement in sample by adding more lines. A sketch of two antennae, i.e. antenna pair, with their radii, random positions of two callers and receivers, each inside their antenna radii and communications between them, represented by pixels on lines, are depicted on Fig. 5. As lines are drawn additively, the intersection of both communication lines is, by addition, as twice as intensive as each line by itself. Please note, that samples could be generated analytically, but due to our simple presumptions, high accuracy is not required. We intuitively determined the radius of each antenna as a 70% of distance to the nearest other antenna. Ideally samples would be prepared for every antenna pair, but computationally the approach would be unacceptable. As we would require more than a million different samples, we generated a data bank of approximately 5500 samples arranged by three parameters: radius of first antenna, (5 – 140 pixels), radius of second antenna, which is always smaller than radius of first antenna (2 – 107 pixels), and distance between both antennae (5 – 1203 pixels). We generated more samples with smaller distances and less with larger distances, as the effect of for instance increasing a long sample for n pixels causes less error than increasing a short sample for the same n pixels. An example of one sample ($r_1 = 60, r_2 = 16, d = 348$) is displayed on a Fig. 4. When mapping samples between real pairs of antennae, we always picked the sample, which fitted best to the real parameters and in cases when second antenna radius was larger than the first, antennae radii were swapped and the chosen sample drawn in reverse.

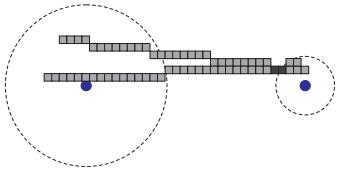


Fig. 5 A visual representation, how samples are generated. A pair of antennae with their radii is displayed and two random lines from inside source and to inside sink antenna radius are shown and their intersection is, by addition, as twice as intensive as each line by itself.

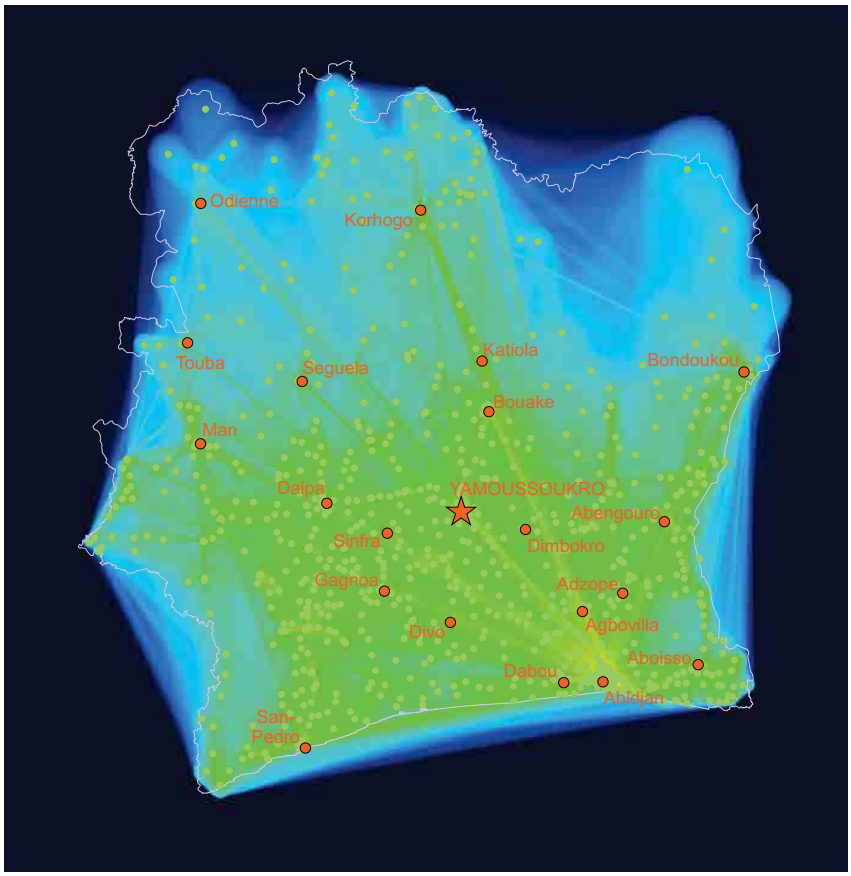


Fig. 6 The diagram of duration of calls between antennae, generated with line-like dispersed samples, displayed in logarithmic scale on Fig. 3.

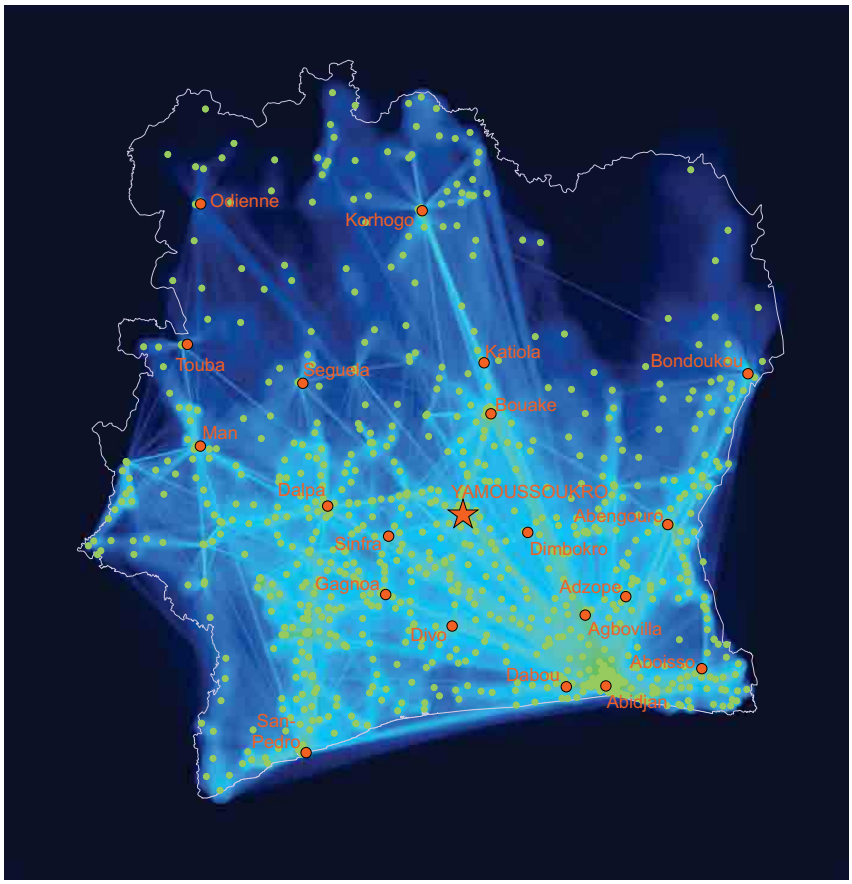


Fig. 7 The diagram of numbers of calls between antennae, generated with line-like dispersed samples, shown in logarithmic scale on Fig. 3.

3.3 Activity star representation

We generated pictures on Fig. 8 and Fig. 9 the same way as the trivial line diagrams, except instead of every line, we drew a pair of mutually pointing star rays as on Fig. 10. The intensity of rays is proportional to the duration of calls or number of calls from owner antenna, i.e. is the same as the intensity of replaced line. The size of stars is proportional to the overall strength of the antenna, which is the sum of all durations or numbers of calls respectively from this antenna.

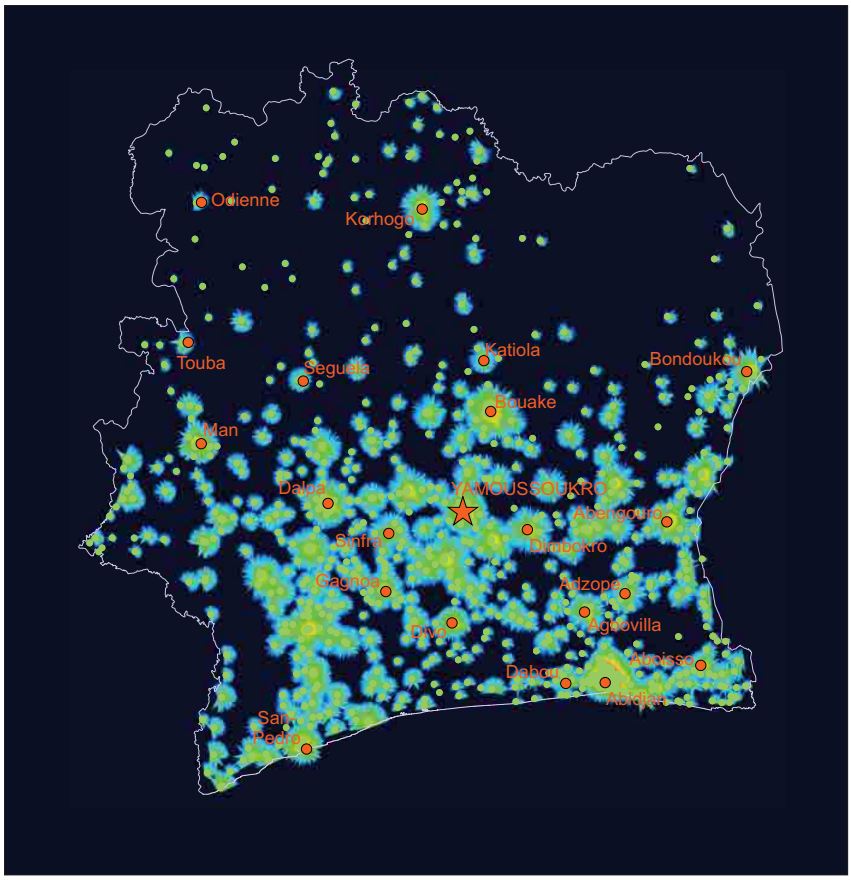


Fig. 8 A star diagram of duration of calls between antennae, shown in the logarithm scale on Fig. 3. A direction of each star ray points to the antenna where the communication is going. Size of the star is proportional to the overall duration of calls from this antenna.

3.4 Correlation between number of calls and duration of calls

On a scatter plot diagram on Fig. 11 the correlation between normalized number of calls and normalized duration of calls is displayed. A whole cumulative traffic (calls and duration) in a complete period of observation was determined for each antenna source and normalized to the number of receiving antennae. The diagram clearly shows the entities are in a linear correlation.

3.5 Duration and number of calls in relation to their distance

We expected that the distance of calls should somehow be related to their duration and analyzed it. We grouped pairs of antennae into 500 classes of different

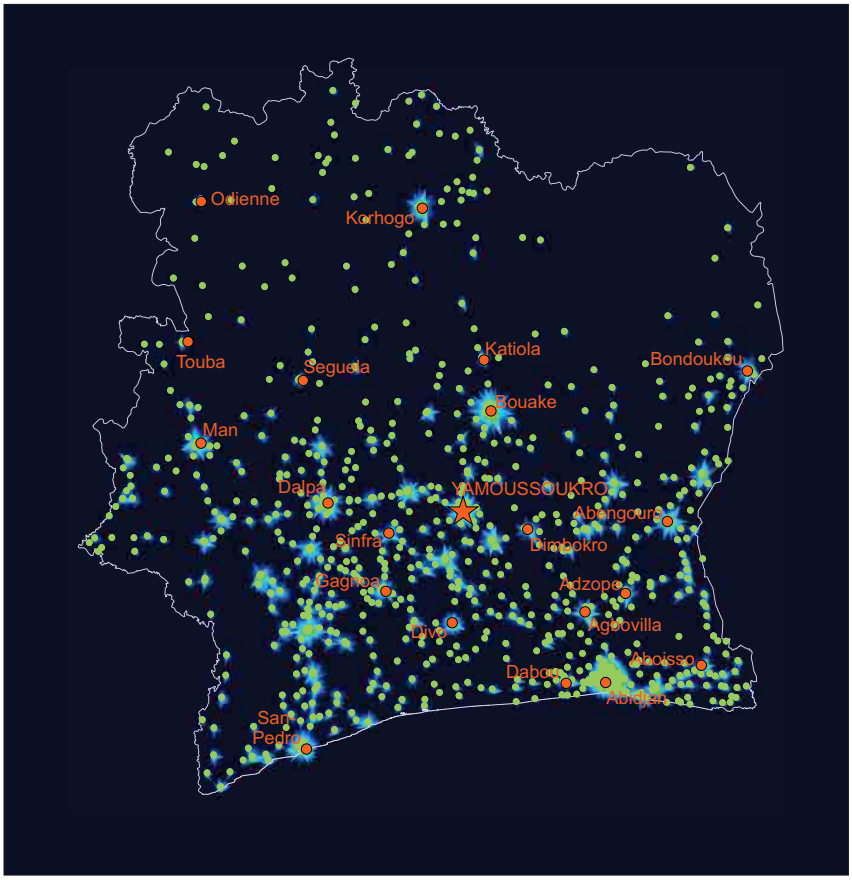


Fig. 9 A star diagram of numbers of calls, displayed in the logarithm scale on Fig. 3. A direction of each star ray points to the antenna where the communication is aimed. Size of each star is proportional to the overall number of calls from this antenna.

Fig. 10 A star ray, used in star diagrams to represent the selected direction.

distances, from $0m$ to the distance of two most remote antennae (approximately $764km$), summed up durations of calls in each class and normalized them with the number of pairs and further with the hypothetical probability of appearance of this distance class. Probabilities of distance classes were approximated by the probability of randomly picking lines from a square of the size of Ivory Coast, which meet the distance class in question. The distribu-

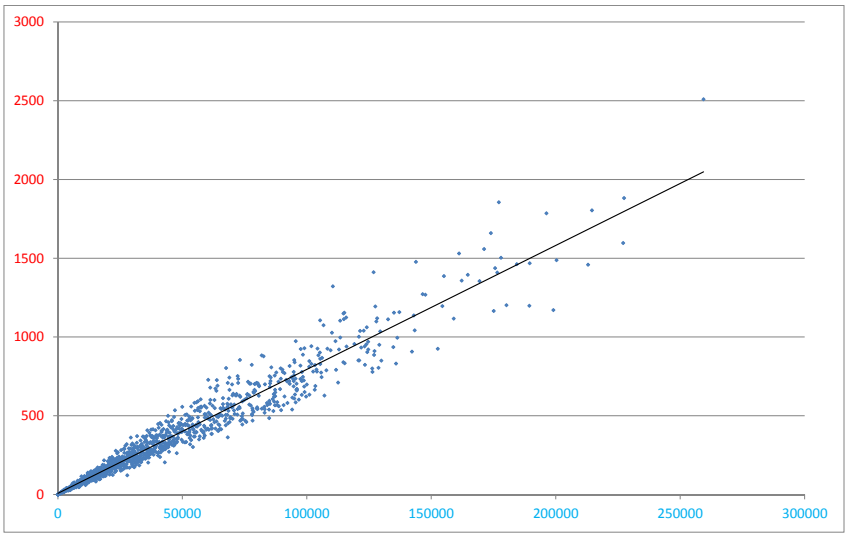


Fig. 11 Scatter plot diagram of numbers of calls versus durations of calls for each antenna. Red are numbers of calls, blue are durations of calls.

tion of probability of picking a random line, i.e. a pair of random points or positions of two hypothetical antennae, in a square is explained in section 3.6.

Distinctive peaks formed on the diagram of normalized durations in distance classes on Fig. 12. We suspected that peaks correspond to pairs of large cities, where many antennae are close together and consequentially many pairs of antennae of similar distance are formed. Pairs of large cities are displayed with red dots on the diagram. We expect, that the cumulative duration of calls between two large cities is proportional to population of both cities and might be inversely proportional to their distance. Many other aspects should be considered in this regard, but we chose a product of city populations, divided by their distance for a measure of city pair importance. This measure takes upper expectations into account and reflects in height of red dots in the diagram on Fig. 12. We expected at least one evidently important city pair to be aligned with each of the most prominent peaks. Some obvious pairs of cities clearly stand out: Abidjan:Korhogo, Abidjan:Man and Abidjan:Bouake, etc.

The same logic, as for the correlation diagram of duration of calls versus distance of calls, applies to the correlation diagram of numbers of calls versus the distance of calls. Peaks are slightly more pronounced in this case. See Fig. 13.

3.6 Distribution of line distance in a square

The distribution of probability of randomly picking a line from a chosen distance class in a square is displayed on Fig. 14. We originated from the idea

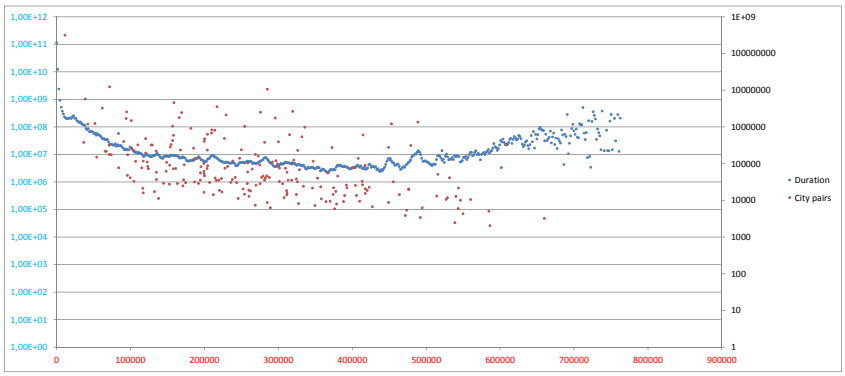


Fig. 12 Correlation diagram of duration of calls related to the distance of calls. Red - distance class and distance between cities in meters, blue - normalized duration per distance class, corrected with a probability of this class, black - measure of city pair importance.

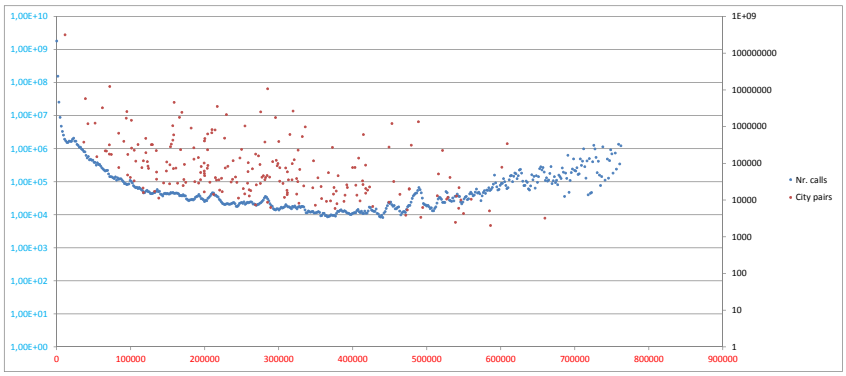


Fig. 13 Correlation diagram of numbers of calls related to the distance of calls. Red - distance class and distance between cities in meters, blue - normalized number of calls per distance class, corrected with a probability of this class, black - measure of city pair importance.

that the Ivory Coast is approximately square country and therefore picked a unit square as a shape to describe the area, where all antennae are positioned. While antennae are not positions sharply on the border of the country, we enlarged the unit square for 5% in width and height. We then generated millions of lines with randomly positioned endings in the enlarged square and grouped them into 500 equidistant classes inside the interval of $[0, \sqrt{2}]$. All lines, longer than $\sqrt{2}$ were ignored. (Remember, our new square has a side of length $a_1 = 1.05$.) To get a normalized distribution of class probability, we divided counts of lines in each distance class with the number of lines in all distance classes. These probabilities were later used as normalization factors in analysis in section 3.5.

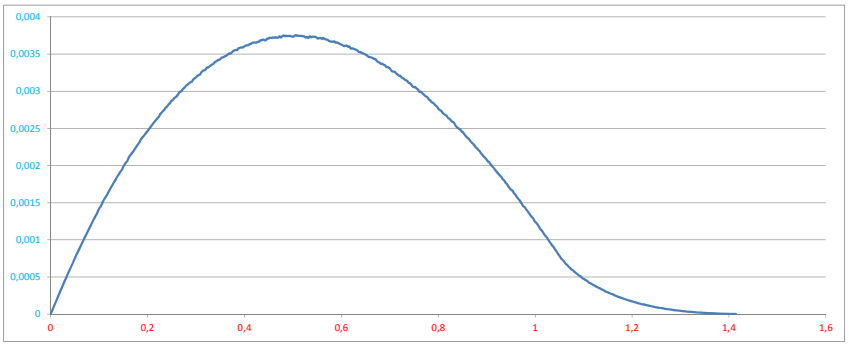


Fig. 14 Probability of randomly picking a line with a length suiting one of 500 uniform classes of lengths in the interval of $[0, \sqrt{2}]$ from a square with sides of length equal to 1.05. If a picked line is longer than $\sqrt{2}$, the picking is not considered as valid and it is repeated. Probability is displayed in blue and class length in red.

3.7 Number of local calls and local duration of calls versus the area antennae cover

We have not found any obvious correlation between the number and duration of local calls, i.e. calls made from antenna to the same antenna, versus the relative area the host antenna covers. The relative area of each antenna was determined as a square of a distance from antenna to the nearest other antenna.

3.8 Filtered traffic diagrams

An observation was made to the most prominent durations and numbers of calls for every distance class, defined in section 3.5. As we grouped pairs of antennae into distance classes, we can easily show some of the most prominent ones. On Fig. 15 and Fig. 16 traffic of the three most prominent pairs of antennae is displayed. In the other aspect, a thousand of the most prominent durations and numbers of calls, regardless of distance class, are displayed on Fig. 17 and Fig. 18.

An interesting view to a network of antennae may be obtained by keeping only a part of the most prominent lines between each antenna and its neighbors in relation to the maximal line, i.e. traffic, from that antenna. On Fig. 19 a network of antennae with lines stronger than one forth of the most powerful line from each antenna are kept. As we can see, even though one would expect 1/4 of the maximal output traffic is a low boundary, almost all lines have values below that limit and are therefore removed. The network of antennae interconnections falls apart to form many connected components which are depicted by various colors.

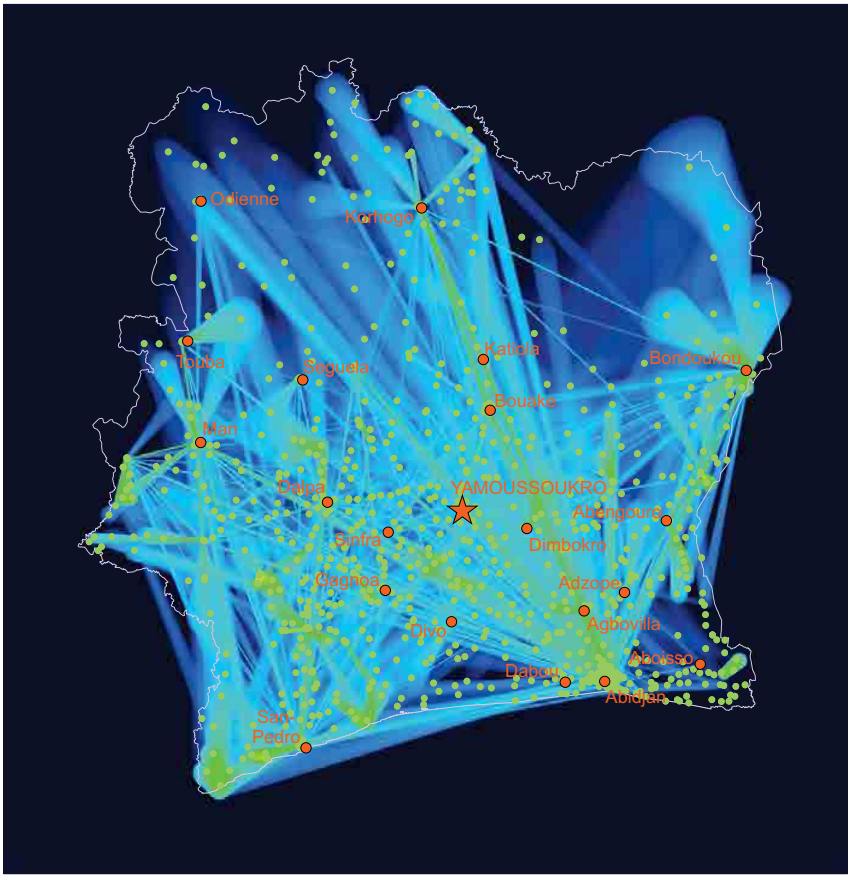


Fig. 15 Three of the most prominent durations of calls between antennae in pairs for every one of 500 uniform distance classes, displayed in logarithmic scale Fig. 3.

3.9 Dispersed cluster representations of daily and weekly behaviour of antennae

We generated dispersed images (6000×6000 pixels) of clusters of antennae by representing each antenna with a round sample (1000 pixels in diameter) of specific color, added at the position of each antenna and possibly overlapping with many samples of other antennae of the same or different cluster. The color of sample was determined by the cluster to which the chosen antenna belongs. The idea is to visually and geographically show where each cluster dominates. The shape, i.e. the intensity, of the sample was determined by the function (1), where x is a distance from center of sample to the edge of sample at $x = 1$. No specific background behind the (1) exists. We only tried to make it bell-like and with the function value slowly approaching towards zero as

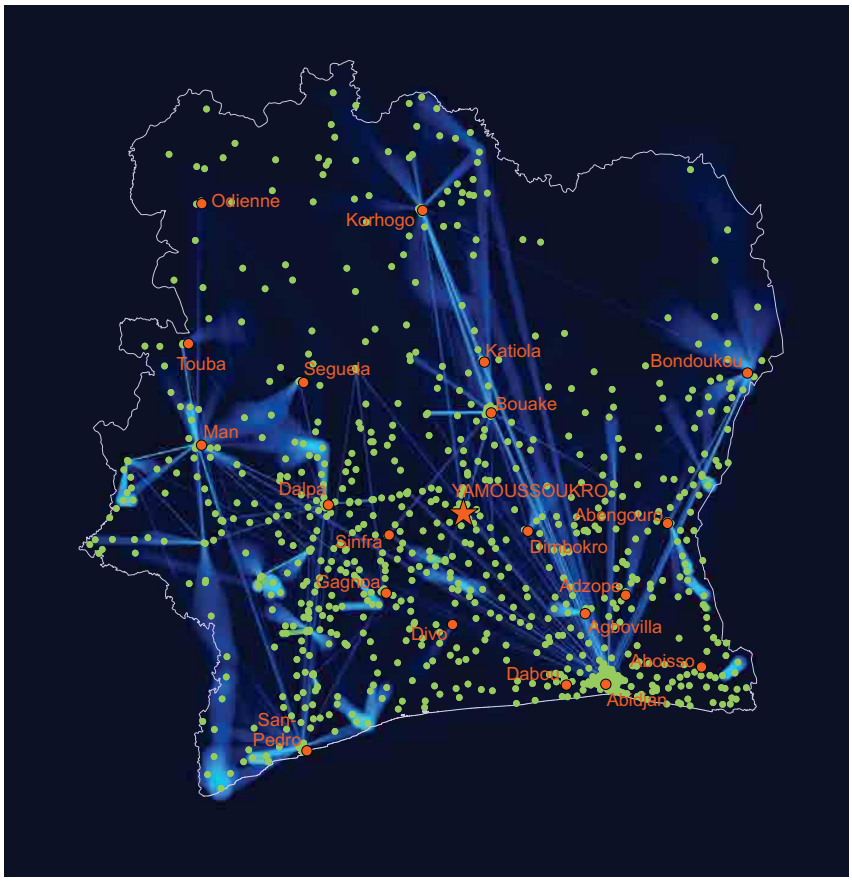


Fig. 16 Three of the most prominent numbers of calls between antennae in pairs for every one of 500 uniform distance classes, displayed in logarithmic scale Fig. 3.

x approaches 1, both with a goal to produce the most eye appealing final results. A visual representation of the function is given on Fig. 20. We show each picture normalized to the highest intensity; either red, green or blue and in the logarithm scale.

$$f(x) = (1 - x^2)/(40 \cdot x^2 + 1), x \in [0, 1] \quad (1)$$

Given a data about calls made in 10 14-days periods we cannot only analyse the trajectories of users. The other possible way of the analysis is the analysis of an activity of antennae. Given exact timestamp of each call enables us to see the activity of each antenna at selected hour in a day or at selected day in the week. Antennae with no activity are not included in this set.

We transform data from the set 2 into two-mode networks and one of them is also a network of antennae as one set of vertices and days in a week as

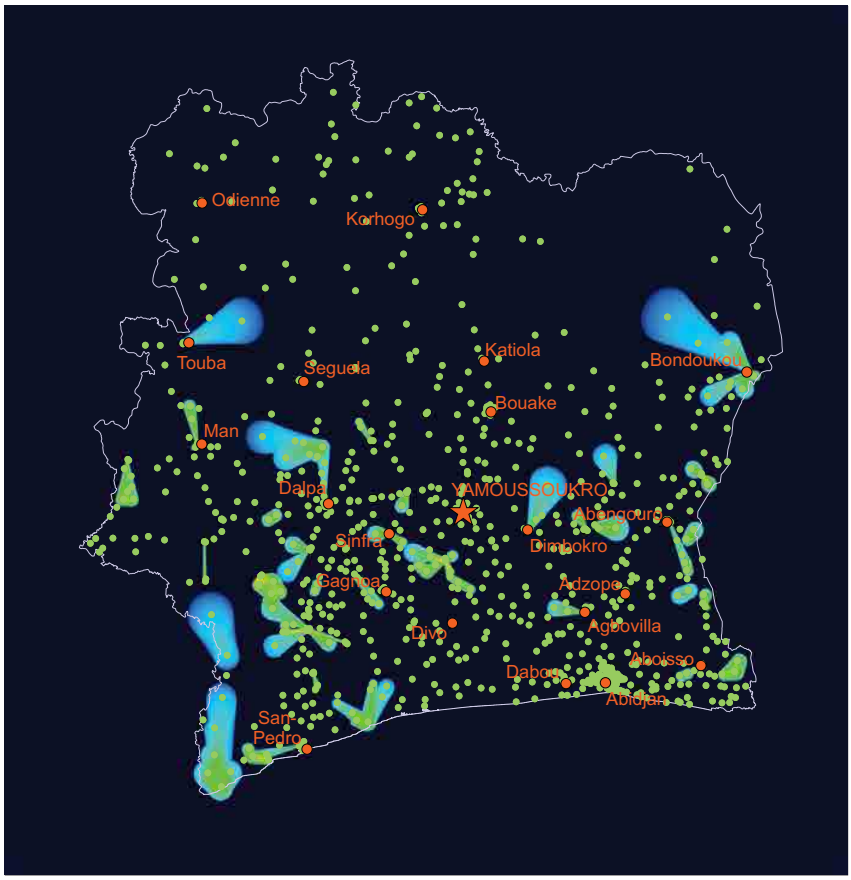


Fig. 17 A thousand of the most prominent durations of calls between antennae pairs, displayed in logarithmic scale Fig. 3.

another set of vertices. Each antenna is connected to the day in a week if there was made at least one call from this antenna at the timestamp that determines selected day in a week. The value of the connection from the antenna to the day in a week is equal to the number of calls that were made in that day from that antenna. Because there are seven days in a week, every antenna has at most seven connections.

A degree of a vertex that represents an antenna tells us the minimum number of days in which the antenna was active – some calls were made from that antenna. Let us look at the Fig. 21. Antennae are represented with dots of different colors. The area around each antennae is also colored in the same color and this kind of a visualization helps the viewer to easily recognize the patterns of antennae of different colors. The area around Abidjan is almost white, because there is settled a lot of antennae, so their surroundings cover

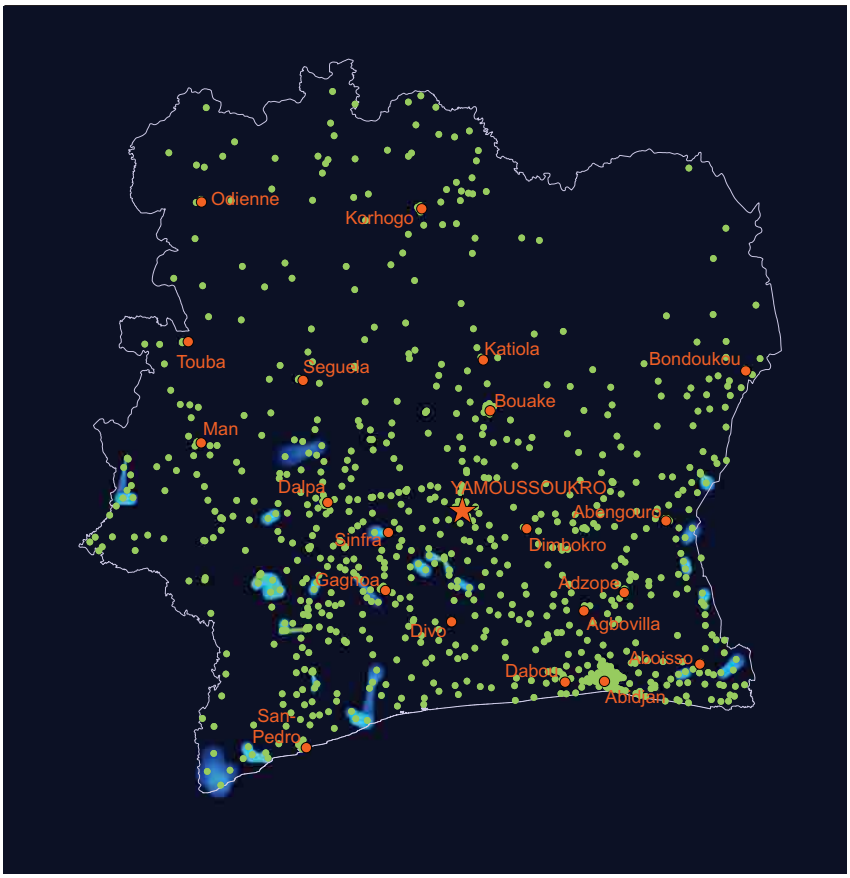


Fig. 18 A thousand of the most prominent numbers of calls between antennae pairs, displayed in logarithmic scale Fig. 3.

each other and the colors sums to the almost white color. Red dots dominate and those dots represent antennae with at most full week activity. Still, there are 3.54% of those that are not active through the whole week. Most of those antennae are in the south-east part of the country, more specifically in Abidjan and its surrounding region. Green dots represents antennae that are active at most 6 days in a week, the blue ones represent antennae with at most 5 days of an activity per week, yellow dots are for 4 days of an activity, magenta dots for 3 days of an activity, cyan dots for 2 days of an activity and violet dots for at most one day of an activity per week.

Another two-mode network that we produced is a network of antennae as one set of vertices and hours in a day as another set of vertices. Each antenna has at most 24 connections and we wanted to see the distribution of a number of hours of an activity of antennae.

Daily Commuting in Ivory Coast: Development Opportunities

Marco Mamei, Laura Ferrari
Dipartimento di Scienze e Metodi dell'Ingegneria
University of Modena and Reggio Emilia, Italy
Email: {marco.mamei, laura.ferrari}@unimore.it

***Abstract*—The creation of effective policies to improve life conditions in developing countries pass through accurate estimates of the distribution and dynamics of people and resources in the environment. The widespread adoption of cellular networks opens new possibilities to obtain such information directly from data on network usage.**

Origin-Destination (OD) matrices describing the overall mobility patterns across a region are important and versatile tools on which to ground policies and interventions in a region. OD matrices are important to organize and prioritize the creation of infrastructures (e.g., the road network), to support innovative services, and also to manage public health policies and epidemiological studies.

In this paper, we present an approach to automatically extract OD matrices describing daily home-work commute from data collected by cellular networks.

As home-work commute represents a large fraction of the overall people mobility, identifying such mobility patterns at a national scale is important to realize effective policies fostering the development of the region.

In particular, we present two applications that take advantage of the extracted information and could have an impact in the development of the country. The first application aims at providing a novel low-cost service for the delivery of physical and digital goods across the country. The second application tries to infer epidemiological risk on the basis of the identified mobility patterns, thus supporting public health policies.

Results have been tested on a real dataset comprising cellular network's traces of 500,000 users in Ivory Coast for a period spanning 5 months.

I. Introduction

The unprecedented volume of data currently being generated in the developing world by the use of telecom networks allows to uncover people behavioral patterns across a number of domains, in particular, with regard to mobility and communication patterns. All this information can provide the basis for setting up effective policies fostering the development of a given region. For example, the analysis of people whereabouts patterns (as detected by mobile phone use) has been used to quantify the impact of human mobility on malaria epidemics [18] and to describe the dynamics of residency in slums neighboring developing cities [17]. Such kind of information can be extremely valuable in the definition of public health and urban planning campaigns.

One of the most important and *versatile* information that can be extracted from telecom networks' data is Origin-Destination (OD) matrices describing the overall mobility patterns across the region. An OD-matrix assumes that the region under study is partitioned into a finite set of zones $S_i : i = 1, ..., n$ and records the number of trips from any origin zone S_i to any destination zone S_j. OD-matrices can be compiled on the basis of different time-extents (e.g., considering trips happening in a given hour/day), and by also considering the purpose of the trip (e.g., home-work commute).

Traditionally OD matrices are costly and difficult to obtain, especially in developing countries, because they are based on travel diaries made every few years, which quickly become obsolete and strongly rely on provided reports.

In this paper we propose a mechanism to extract the *daily home-work OD-matrix* from a dataset of mobile phone call detail records. In particular, the approach has been applied to the "Mobility traces: coarse resolution dataset" of the D4D Challenge (*www.d4d.orange.com*) comprising timestamps and location information for calls and SMS made by a sample of 500,000 users during a 5-months period.

Daily home-work OD-matrix can serve many purposes. As home-work commute describes a large fraction of the overall mobility in the region, the OD matrix represents a primary source of information to plan and organize the transportation/road network and the region's infrastructures in general. In addition, the process of identifying the daily home-work OD-matrix immediately brings the estimate of population distributions during the most important part of the day (home and work places). Such kind of estimates are a fundamental prerequisite for the accurate measurement of the impacts of population growth, for monitoring changes and for planning interventions.

From a complementary perspective, information about people mobility can support the development of innovative services and start-ups (e.g., in the logistic sector) that could support the economy of the country.

Finally, human mobility and people distribution is of primary concern for epidemiological studies and for the identification of effective public health policies. Also in

this case, home-work commute is at the basis of a large fraction of people encounters and whereabouts and it is thus a fundamental information to understand the surge and spreading of epidemics.

In the following of this paper, we first describe the approach we used to automatically extract the daily home-work OD-matrix from telecom data (Section 2). Then in Section 3 we highlight possible applications of the extracted mobility patterns. The first application aims at providing a novel low-cost service for the delivery of physical and digital goods across the country. The second application tries to infer epidemiological risk on the basis of the identified mobility patterns, thus supporting public health policies. Section 4 shows the limitations of our work, while in Section 5 we present some related work in the area. Eventually, Section 6 concludes and discusses avenues for future work and application in the context of developing countries.

II. Origin Destination Matrices

A. Dataset

The dataset we adopted (D4D Challenge SET3 [2]) is based on anonymized Call Detail Records (CDR) of phone calls and SMS exchanges between 500,000 anonymized Orange's customers in Ivory Coast between December 1, 2011 and April 28, 2012. CDR location information is provided on a regional resolution. The whole geographic area of Ivory Coast has been divided into 255 zones (subprefecture) and each CDR is located in one of the subprefectures. Overall, provided data consists of a set of TSV files of about 14.3 GB.

Each element (CDR) of the dataset comprises the following fields:

- **user-id** is an integer representing the user
- **connection-timestamp** is the UTC time when the event took place
- **subpref-id**. is the id of the subprefecture where the event took place.

In addition, another file associates each *subpref-id* with the geographic coordinates of the barycenter of the corresponding subprefecture region. Moreover, we retrieved from the D4D Challenge Web site the shape-files of the subprefecture regions used to show the results of our analysis on a map.

B. Home and Work Places

The main step in creating the daily home-work OD matrix consists in identifying home and work places of each user. This task is simplified by the natural clustering of CDR events induced by the CDR localization at the subprefectures level.

Following an approach similar to [3], [10], we considered all the network events of each user. Then we clustered all the events on the basis of the corresponding subprefecture.

Relying on commonsense assumptions, we defined a time window associated to "home-based events" comprising all the events happening from 9pm to 6am, and a time window associated to "work-based events" comprising all the events happening from 11am to 4pm.

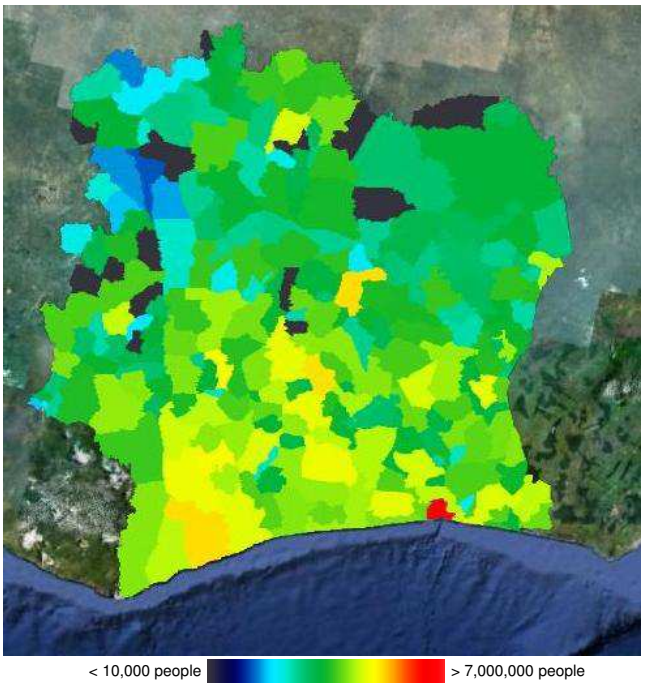


Figure 1. Home-based population density. Subprefectures have been colored according to the log of their population, so as to account for outliers (i.e., the Abidjan subprefecture). Numbers are linearly scaled to the actual Ivory Coast population estimated to be about 20,153,000 people by the World Bank in 2011. For better visualization, find kml file from *www.agentgroup.unimo.it/d4d*

For home discovery we weighted all the clusters (i.e., subprefectures) by the number of events happening in that cluster within the "home" time window. The subprefecture with the highest weight is classified as the home of the user. For work discovery we applied the same approach using the "work" time window.

Aggregating all the results we obtain the distribution of the population sample during home and work time windows across the country. Fig 1 represents the home-based population density. We also set up a Web page to download kml files of our results: *www.agentgroup.unimo.it/d4d*.

To validate the resulting distribution, we try to cross correlate our results with two other sources of data. On the one hand, we obtained an estimate of the resident population by subprefecture from the Geonames dataset (*www.geonames.org*) that aggregates data from a number sources on the Web. On the other hand, we obtained another estimate from Afripop (*www.afripop.org*). Afripop data [15] were collected in 2010 at $100m^2$ resolution. We aggregate this data at the subprefecture level and compared the results. Figure 3 shows correlation results among ours results and these comparison datasets. The table in Fig. 2 summarizes

Comparison	R^2	Mean abs. error
Afripop VS. Ours	0.77	50183
Geonames VS. Ours	0.75	75670
Afripop VS. Geonames	0.99	23622

Figure 2. R^2 and Mean absolute error among our results and the two comparison datasets, namely Afripop and Geonames.

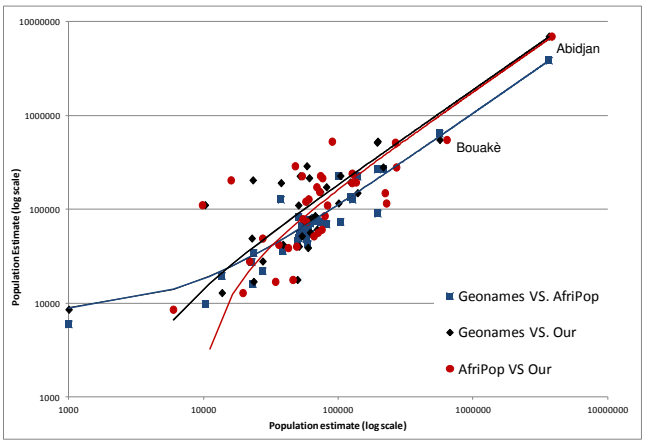


Figure 3. Correlation results among ours results and the Geonames and Afripop datasets.

results: it shows R^2 and Mean absolute error among our results and the two comparison datasets. In general it is possible see a good correlation among the data $R^2 \simeq 0.75$. The mean absolute error around 50,000 persons is reasonable considering linear scaling approach to scale the population in our dataset to the whole population of Ivory Coast.

In a second experiment, we applied the algorithm to extract home and work places to individual months of the dataset. Figure 4 shows that the number of users living in a given subprefectures across months is rather stable. Although this is not appreciable from our data, this kind of approach could reveal large-scale trends like the demographic grow of large cities like Abidjan and Bouakè.

C. Daily Home Work OD Matrix

On the basis of the extracted places it is possible to estimate the OD matrix for working days, by counting the number of users who live in a given subprefecture S_i and commute to work to a subprefecture S_j. The resulting count is the value OD_{ij} of the matrix. Given N the number of subprefectures ($N = 255$ in our case), we obtain a N^2 matrix describing the overall mobility demand in the country. The highest values in this matrix lay on the main diagonal OD_{ii} as most people live and work in the same subprefecture, but movements among different subprefectures can be appreciated.

This type of data can be used to compute the number of people that commute through the country, thus allowing to understand the mobility demand for a given road segment. In order to compute this information, for convenience of matrix

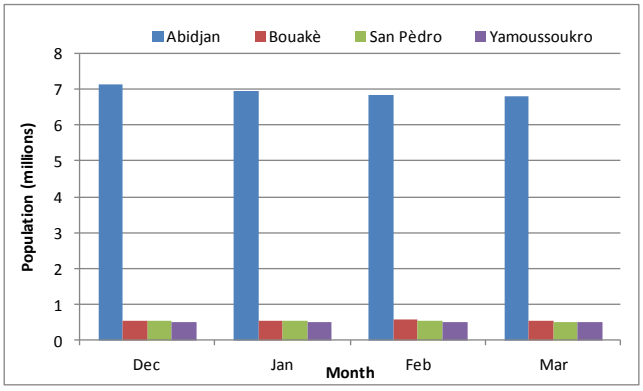


Figure 4. Number of users living in a given subprefectures across months.

operations, we also organized the OD matrix in a *mobility demand column vector* D comprising N^2 rows.

We then tried to extract a simplified model of the road network in the country. In particular, for any two *confining* subprefectures S_i and S_j we extracted the shortest road connecting the two via the Google Transit API. The results is a set of M roads overall defining a road network for the country.

We then applied a simple greedy geographic routing algorithm [14] to compute the set of road segments involved in the commuting between any two subprefectures.

On the basis of these results, we compute a road network matrix $Rnet$ comprising N^2 columns each associated with a possible path between any two subprefectures, and M rows each associated to a road segment. The matrix $Rnet_{ij}$ has value 1 if the geographic routing algorithm includes the road segment i in the path j, it has value 0 otherwise.

Given the above notations, we can compute the vector representing the mobility demand for a given road segment as:

$$F = Rnet \cdot D$$

F is a row vector with M columns. Each value F_i is the number of people commuting through the i^{th} road.

Figure 5 illustrates the result of this process: the size of each road segment is proportional to the associated mobility demand. Not surprisingly, the roads around Abidjan are those experiencing the heaviest traffic.

III. Applications

In this section we present two applications that take advantage of the extracted OD matrix. We tried to focus on applications that could have an impact in the development of Ivory Coast. The first application aims at providing a novel low-cost service for the delivery of physical and digital goods across the country. This kind of applications could support the developing economy of the country. The second application tries to infer epidemiological risk on the basis of

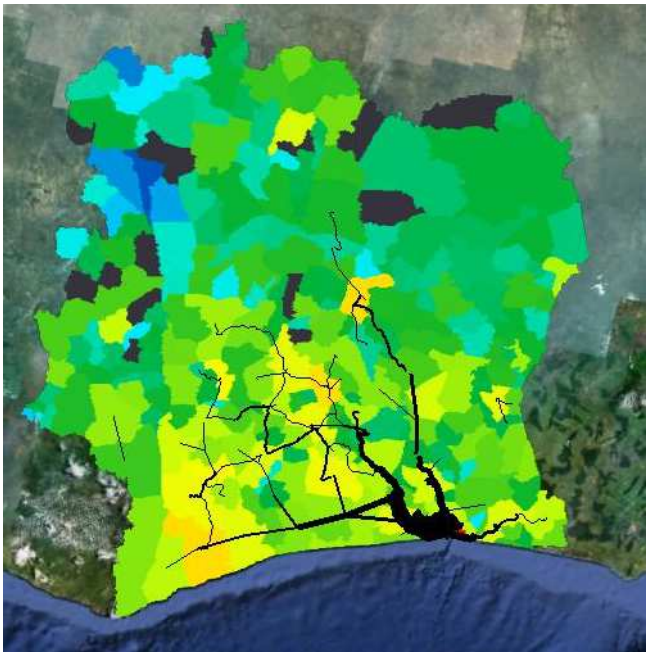


Figure 5. Model of the Ivory Coast road network weighted by the estimated daily OD home work commute traffic (only major roads are visualized). For better visualization, find kml file from *www.agentgroup.unimo.it/d4d*

the identified mobility patterns. This kind of studies could have an impact in the development of public health policies across the region.

A. Social Delivery

A number of start-up companies in the developing world have shown the potential of crowdsourcing for the growth of the economy. TxtEagle, for example, offers crowdsourcing and market research in developing countries by recruiting a large number of users on the fly [5].

An innovative application in this direction could enable a novel low-cost service for the delivery of physical and digital goods across the country. As shipping items in the developing world might be problematic/not accessible, the application could rely on a network of crowdsourced "postmen" delivering an item to the intended destination.

Using data on daily commute patterns (OD matrix) it is possible to identify a set of users (the postmen) who can pass an item to each other, while undergoing their daily commute, so that the item is eventually delivered to the destination. For example, to send an item from a peripheral subprefecture like Maffere to Abidjan, the application could dynamically hire a person commuting from Maffere to Aboisso, who will rely the item to another person commuting from Aboisso to Bonoua, who will finally rely the item to a person commuting from Bonoua to Abidjan, thus reaching the destination (see Figure 6).

The approach could be used to transport physical objects, but also digital ones. For example, we could think of

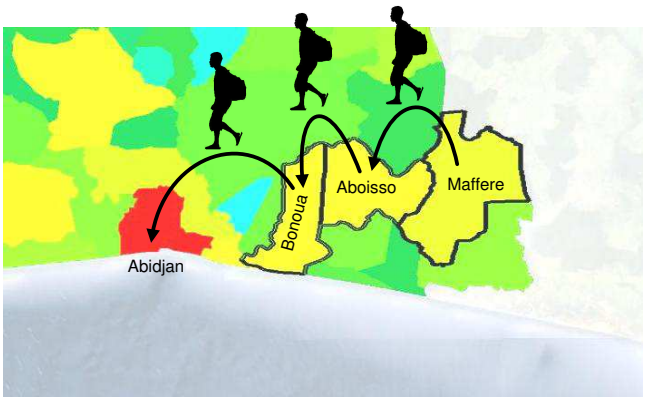


Figure 6. Social delivery application.

an application collecting all the network traffic (emails, files, Google requests, social network's posts, etc.) produced by a rural village that it is not covered by an adequate network infrastructure. The application would encrypt and store such data in the phone of the postmen until they get a suitable network connection. Postmen could then collect the responses and deliver them back to the village as sort of "human routers". This would create a low-cost Delay Tolerant Network (DTN) with large latency and using social routing [19].

Naturally, this kind of approaches would integrate suitable incentivization, security and privacy protecting mechanisms similar to those experimented in the context of the DARPA Red Balloon challenge [12] and of the Tag Challenge [13].

Although we did not really implement such an application, we tried to estimate some expected performances using the extracted OD matrices. In particular, we run an experiment (with a number of simplifying assumptions) to estimate the number of days to ship an item from one subprefecture to another one.

In the experiment, we assumed that for every 100 persons commuting between two subprefectures, 1 person participates to the social delivery application and will collaborate to the shipment. We considered two different approaches with respect to how routing takes place. In the *road-based* approach, we consider the road network defined in the previous section and assume that any postman carries the item only for a single road. Then the item is relayed to the next postman. In the *commute-based* approach, postman carries the item until their destination (i.e., work or home places) possibly spanning multiple roads. In both the scenarios postmen route the item following a greedy geographic routing mechanism. We then simulated the delivery process using the extracted commuting patterns. Figure 7 illustrates the number of days required to ship an item from the subprefecture of Yamoussoukro to some other subprefectures (sorted by the geographical distance from Yamoussoukro). Resulting negative values represent

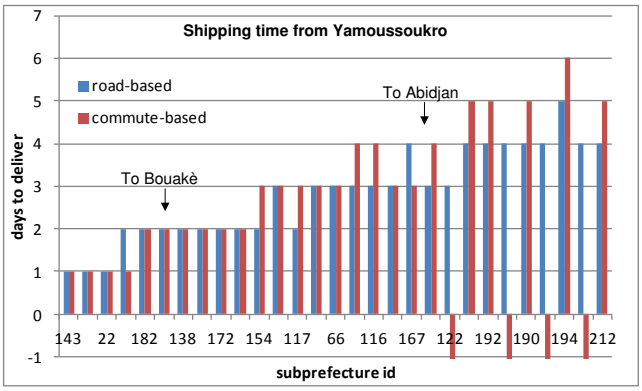


Figure 7. Days required to ship an item from the subprefecture of Yamoussoukro to some other subprefectures (sorted by the geographical distance from Yamoussoukro). Negative values represent the situation in which the greedy geographic routing fails and the item cannot be delivered to the intended destination.

the situation in which the greedy geographic routing fails and the item cannot be delivered to the intended destination (of course, a real application could implement better routing strategies). Looking at the figure it is possible to see that the number of days linearly increases with the distance from the source (as expected). Rather unexpectedly instead, the number of days for the *commute-based* approach is often larger than for the *road-based* approach and it exhibits a larger number of routing failures. This is because the *commute-based* approach tend to create more swirling routes (to follow the postmen home-work commute) that degrade performances.

B. Epidemiology

Epidemiology is another important area in which accurate information on people distribution and mobility patterns plays a fundamental role [15], [11], [18]. In fact, movements of infected humans and contacts among individuals can increase the dispersal of the disease and enable its propagation.

The approach to extract mobility patterns presented in this paper can provide effective information on which to ground public health policies to deal with the epidemiological risk. In particular, the extracted daily commuting patterns are particular important. For example, in the case of malaria [18] notices: "*The vast majority of travelers that will affect malaria parasite dispersal are those moving within a country between regions of variable malaria receptivity on a daily or weekly basis*".

On the basis of the extracted data, we conducted two kind of studies: first, following the approach presented in [11], [18], we tried to understand malaria diffusion across the various subprefectures of the Ivory Coast. Second, we tried to combine a general SIS epidemic model [16] to our mobility data to identify the epidemiological risk of the various subprefectures.

Malaria Diffusion in Ivory Coast

In this first application, we try to quantify the impact of the extracted daily home-work commute to the diffusion of malaria in a given subprefecture. Movements of infected humans can in fact increase the dispersal of malaria parasite *Plasmodium falciparum* beyond what would be possible for the mosquitoes hosts alone. Following [11], [18], we separate malaria diffusion between: *importation by visitors* and *importation by returning residents*. *Importation by visitors* comprises individuals who got infected at their home location and bring malaria to their work location. *Importation by returning* residents comprises individuals who got infected at their work location and bring malaria back to their home location.

We downloaded, from the Malaria Atlas Project, the spatial distribution of Plasmodium falciparum malaria endemicity map in 2010 in Ivory Coast. Data consists in a grid of $1km^2$ resolution mapping the age-standardized P. falciparum Parasite Rate ($PfPR_{2-10}$). We then aggregate all the grid cells into the corresponding subprefecture and compute the average value of $PfPR_{2-10}$ for all the subprefectures. The result is a a measure of the proportion of people infected for each subprefecture.

On the basis of such an information, we compute daily *importation by visitors* for commuting between home place i to work place j as $PfPR_{2-10}(i) * OD_{ij}/200$, assuming an average duration of infection of 200 days [11], [18].

We then tried to compute the probability of a person being infected while visiting another subprefecture. For each cell of the grid, we converted $PfPR_{2-10}$ values into the Entomological Inoculation Rate (EIR). using the mathematical model described in [18], [9]:

$$EIR = max(0, -1.573 + 7.74 \cdot PfPR_{2-10})$$

EIR is a measure of transmission intensity based on the number of bites by infectious mosquitoes per person per year. Following [18] we computed the probability P of each individual acquiring an infection in a daily visit to a subprefecture as:

$$P = 1 - (1 + \alpha \cdot b \cdot EIR/365)^{-1/\alpha}$$

α and b are parameters of the model that we set to $\alpha = 4$ and $b = 0.55$ according to [9].

We then aggregate all the grid cells into the corresponding subprefecture and compute the average value of the probability P of an individual acquiring an infection for a daily visit to the subprefecture.

On the basis of such an information, we compute daily *importation by returning residents* for commuting between the home place i to work place j as $P(j) * OD_{ij}$.

Figure 8(top) shows a map of Ivory Coast around Abidjan where subprefectures are color-coded according their malaria endemicity ($PfPR_{2-10}$). The map shows the main paths (90th percentile) for malaria diffusion in terms of *importation by visitors*. Figure 8(bottom) shows the same map where subprefectures are color-coded according the probability P of an individual acquiring an infection for a daily visit to the subprefecture. The map shows the main paths (90th percentile) for malaria diffusion in terms of *importation by returning residents*. For better visualization, find kml file from *www.agentgroup.unimo.it/d4d*.

In general results, show that importation by returning residents for the Abidjan subprefecture is much higher (1 order of magnitude) than importation by visitors. This can be explained considering the movements patterns toward the south-east subprefectures that are close to Abidjan and presents higher endemicity values. Such subprefectures have a large impact in importation by returning residents (because of the people from Abidjan working there), but a small impact in importation by residents because they are much less inhabited than Abidjan.

Epidemiological Risk

In a second application we tried to apply the extracted daily commuting patterns to the evaluation of the epidemiological risk in the country. In particular, we simulated the fictional situation of an epidemic outbreak in a given subprefecture. Our idea is to estimate how the contagion would diffuse across the country as a consequence of people movement behavior. The goal is to identify those subprefectures that are more critical to a potential epidemic outbreak. This kind of analysis could be very important for the set up of public health policies and for the prioritization of screening activities.

We created a *Susceptible-Infected-Susceptible* (SIS) model to simulate the epidemic outbreak [16]. The SIS model is defined as follows:

1) Each user can be in one of two states: Susceptible (not currently infected) or Infectious (infected).
2) We defined an endemicity map associating to each subprefecture the daily probability $P_{endemic}$ of contracting the disease. Each day, a fraction of susceptible users become infected according to the endemicity distribution.
3) When a user is infectious, he can infect the other users in the same subprefecture. A susceptible user in a subprefecture with a fraction $F_{infected}$ of infected persons becomes infected with probability $P_{infect} \cdot F_{infected}$.
4) When a user is infectious, he recovers after $DAYS_{infect}$ days. Once recovered, the user becomes susceptible again.

To simulate the mobility patterns of individuals we adopted the extracted daily commuting behavior. We simulated the epidemic outbreak starting at each subprefecture for 30 days. In particular, we setup the simulation with the following parameters: $P_{endemic}$ distribution has value 0.1 in the starting subprefecture, while is has value 0 in all the other subprefectures. $P_{infect} = 0.8$, $DAYS_{infect} = 10$. Figure 9 shows the fraction of infected individuals at the end of the 30 days for different starting subprefecture. We provide some kml files with time animations of the epidemic simulation process from *www.agentgroup.unimo.it/d4d*.

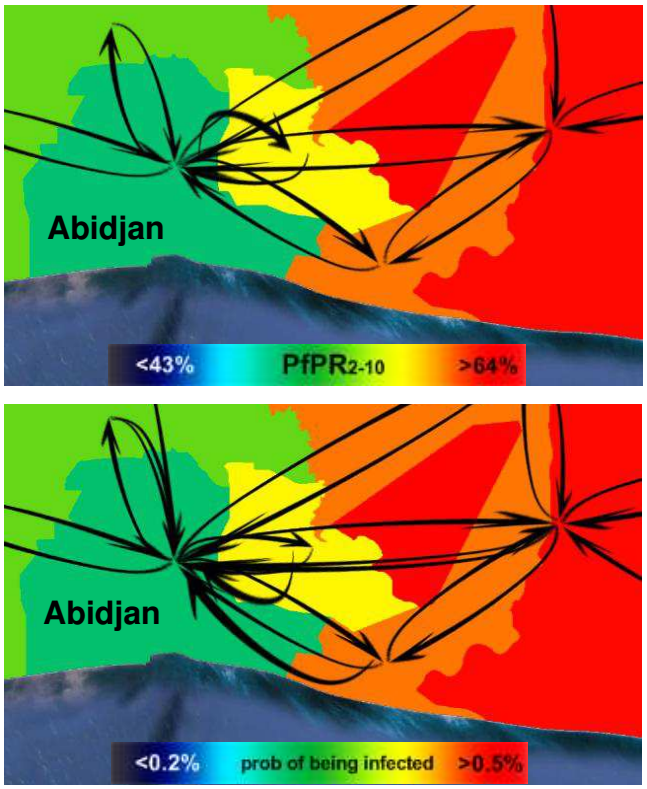


Figure 8. **(top)** Daily parasite importation by visitors, subprefectures are color-coded according their malaria endemicity ($PfPR_{2-10}$). Arrows represent the main paths (90th percentile) for malaria diffusion in terms of *importation by visitors*. **(bottom)** daily importation by returning residents. Subprefectures are color-coded according the probability P of an individual acquiring an infection for a daily visit to the subprefecture. Arrows represent the main paths (90th percentile) for malaria diffusion in terms of *importation by returning residents*. For better visualization, find kml file from *www.agentgroup.unimo.it/d4d*

Results are rather surprising; it is possible to see that epidemic outbreak staring in major cities (e.g., Abidjan and Bouakè) are not the most critical. Because of mobility dynamics, some less populated subprefectures might play a more important role.

As simplified and fictitious this example might be, the adopted approach could be useful to conduct more focused simulation and to possibly help in the set up of public health policies.

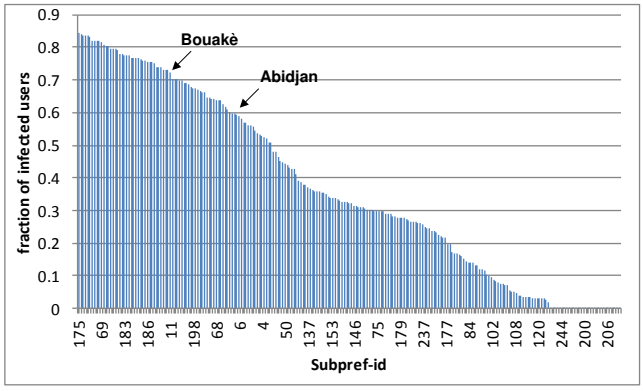


Figure 9. Fraction of infected individuals at the end of the 30 days for epidemic diffusions starting at different subprefecture.

IV. Discussion of limitations

The results presented in the paper might be affected by some bias introduced by the input data and the methodology used to sample the home and work places. We considered all the users in SET3 of the D4D challenge dataset without applying any pre-processing technique to disregard those users having a low number of CDR events. Given the fact that some users have a number of phone logs so low that it does not reflect their mobility behavior, we discovered some rare erroneous home/work estimates. We have experienced a few cases in which the detected home and work places for a given user are too far to be reasonable. However, as we considered 500,000 users, these rare errors had little influence (less than 2% of the home/work pairs are distant more than 100 km). To correct these errors we tried to delete from the OD matrix all the entries with a commuting distance greater than 50 km. The results we obtained after this trimming process are in line with the ones presented in the paper.

Another issue is related to the validation of the obtained work places. While we evaluated resulting home places by means of cross correlation with other data (see Fig. 3), we did not find suitable data on which to validate work places. In any case, the main goal of this paper is to measure the mobility demand and to provide interesting applications of the extracted information that might have an impact in the development of the country. Thus, our home/work estimation, even though not fully validated, provides an interesting mechanism for future elaborations.

Finally, some bias could have been introduced in the epidemiological risk applications. We set parameters and assumptions according to the best of our knowledge, but we are not expert in public health and epidemiology. While better techniques may be applied, we believe that the idea of using OD matrices from mobile phone dataset to infer epidemiological risk, may support public health policies in the developing countries. In our future work we are planning to collaborate with public health experts to discuss about the most appropriate assumptions and parameters to set up the models.

V. Related Work

Fueled by the "recent" availability of data on cellular network use patterns, a number of researchers recognized the application potential in automatically identifying users' important places and mobility patterns. Among many others, the works [3], [1], [10] try to automatically extract the home and work places for the user, and to construct OD matrices on top of that. The work in [3] divides the area under investigation in a grid where the side of every cell is 500 meters long. For each cell of the grid, they count the number of nights the user connects to the network in the nighttime interval (6pm - 8am) while in that cell, and select as a home location the cell with the greatest value. Similarly, the work place is estimated as the most frequent stop area on weekday mornings between 8am and 10am. The approach is evaluated by comparing estimated results with census statistics, result shows a good correlation between the two estimates. The work in [1], [10] proposes an approach to identify important places in people's lives and in particular home and work places. The approach clusters the antennas on the basis of their mutual distances. Then it applies logistic regression on the basis of when CDR events happened to determine whether a cluster represents an important place or not. Finally, home and work places are identified as those important clusters having the majority of events between 7pm and 7am (for home) and 1pm and 5pm (for work places). Similarly to these works our approach estimates home and work locations on the basis of the time when network events take place. In our work, the location of events is given at the subprefecture level thus we already have events clustered at this scale.

An increasing number of researches is also trying to applying such data to address challenges of developing countries, in particular with regard to data on human health, movement, communication, and financial transactions. In this context the Artificial Intelligence for Development community (*http://ai-d.org*) comprises projects including optimizing the allocation of malaria eradication resources in Kenya, detecting behavioral anomalies associated with outbreaks of cholera in Rwanda, quantifying the dynamics of slums in Nairobi, uncovering patterns in regional communication data associated with the spread of HIV and contraception norms in the Dominican Republic, and assessing the social impact of previous policy decisions ranging from road construction to the placement of latrines throughout the developing world. The OD matrices we obtained are a very useful and *versatile* instrument to support most of the above and similar applications.

VI. Conclusion and Future Work

We presented a general approach to extract OD matrices related to daily home-work commute from a dataset of mobility traces derived from cellular network's usage. Resulting matrices are a very useful and versatile instrument to represent overall mobility patterns and people distribution in the region. Accordingly, they can support a wide range of development activities from planning and prioritization of infrastructures, to epidemiological studies and public health campaigns, to innovative applications supporting novel activities and applications in the country.

Our future work in this direction can proceed twofold.

On the one hand, the presented approach could be extended to capture individuals' mobility patterns more accurately. In particular, mobility routines other than home-work commute could still represent an important fraction of the overall mobility and thus have an impact in a number of scenarios (e.g., weekends' mobility patterns are outside of the home-work commute and impact in human encounters and mobility demand). With this regard, approaches like [6] and [7] can be fruitfully applied to extract mobility patterns at a finer scale. In addition, analyzing data on a larger temporal extent could reveal long-range patterns like the gradual urbanization of the country and migratory trends.

On the other hand, it would be interesting to explore the other datasets provided by the D4D challenge [2]. For example, aggregated traffic data (SET1) can be used to identify and possibly predict overcrowded and underpopulated events [8]. From another perspective, communication graphs among individuals (SET4) can be used to better infect co-presence and encounters among people with notable impacts to the applications we described [4].

Most interestingly, it would be important to combine the results of different analysis together to get a multi-faceted representation of the trends happening in the country

Finally, it would be important to actually realize and explore applications on the basis of such data. This would allow to actually appreciate the practical usefulness of the proposed approach and to gain insight in further direction of improvement.

References

[1] R. Becker, R. Cceres, K. Hanson, S. Isaacman, J. Loh, M. Martonosi, J. Rowland, S. Urbanek, A. Varshavsky, and C. Volinsky. Human mobility characterization from cellular network data. *Communications of the ACM*, 56(1):74–82, 2013.

[2] V. Blondel, M. Esch, C. Chan, F. Clerot, P. Deville, E. Huens, F. Morlot, Z. Smoreda, and C. Ziemlicki. Data for development: the d4d challenge on mobile phone data. In *arXiv:1210.0137v2*, 2013.

[3] F. Calabrese, G. D. Lorenzo, L. Liu, and C. Ratti. Estimating origin-destination flows using mobile phone location data. *IEEE Pervasive Computing*, 10(4):36–44, 2011.

[4] F. Calabrese, Z. Smoreda, V. Blondel, and C. Ratti. The interplay between telecommunications and face-to-face interactions: a study using mobile phone data. *PloS ONE*, 6(7):e20814, 2011.

[5] N. Eagle. Txteagle: Mobile crowdsourcing. In *International Conference on Internationalization, Design and Global Development*, San Diego (CA), USA, 2009.

[6] N. Eagle and A. Pentland. Eigenbehaviors: Identifying structure in routine. *Behavioral Ecology and Sociobiology*, 63(7):1057–1066, 2009.

[7] L. Ferrari and M. Mamei. Classification and prediction of whereabouts patterns from reality mining dataset. *Pervasive and Mobile Computing Journal*, 2013.

[8] L. Ferrari, M. Mamei, and M. Colonna. Discovering events in the city via mobile network analysis. *Journal of Ambient Intelligence and Humanized Computing*, 2013.

[9] P. Gething, A. Patil, D. Smith, C. Guerra, I. Elyazar, G. Johnston, A. Tatem, and S. Hay. A new world malaria map: Plasmodium falciparum endemicity in 2010. *Malaria Journal*, 10(378), 2011.

[10] S. Isaacman, R. Becker, R. Cceres, S. Kobourov, M. Martonosi, J. Rowland, and A. Varshavsky. Identifying important places in peoples lives from cellular network data. In *International Conference on Pervasive Computing*, San Francisco (CA), USA, 2011.

[11] A. LeMenach, A. Tatem, J. Cohen, S. Hay, H. Randell, A. Patil, and D. Smith. Travel risk, malaria importation and malaria transmission in zanzibar. *Scientific Reports*, 93(1), 2011.

[12] G. Pickard, I. Rahwan, W. Pan, M. Cebrian, R. Crane, A. Madan, and A. Pentland. Time critical social mobilization: The darpa network challenge winning strategy. In *arXiv:1008.3172*, 2010.

[13] I. Rahwan, S. Dsouza, A. Rutherford, V. Naroditskiy, J. McInerney, M. Venanzi, N. Jennings, and M. Cebrian. Global manhunt pushes the limits of social mobilization. *IEEE Computer*, PP, 2013.

[14] I. Stojmenovic. Position based routing in ad hoc networks. *IEEE Communications Magazine*, 40(7):128–134, 2002.

[15] A. Tatem and C. Linard. Population mapping of poor countries. *Nature*, 474(36), 2011.

[16] E. Vynnycky and R. White. *An Introduction to Infectious Disease Modelling*. Oxford University Press, 2010.

[17] A. Wesolowski and N. Eagle. Parameterizing the dynamics of slums. In *AAAI Spring Symposium on Artificial Intelligence for Development*, Palo Alto (CA), USA, 2010.

[18] A. Wesolowski, N. Eagle, A. Tatem, D. Smith, A. Noor, R. Snow, and C. Buckee. Quantifying the impact of human mobility on malaria. *Science*, (338):267–270, 2012.

[19] E. Yoneki, P. Hui, and J. Crowcroft. Wireless epidemic spread in dynamic human networks. In *Workshop on Bio-Inspired Design of Networks*, Cambridge, UK, 2007.

ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2013) 1–11

Procedia Computer Science

www.elsevier.com/locate/procedia

Building a minimal traffic model from mobile phone data

Michael Zilske, Kai Nagel

Technische Universität Berlin, Transport Systems Planning and Transport Telematics, [zilske|nagel]@vsp.tu-berlin.de

Abstract

We investigate setting up a traffic simulation scenario only from a high spatial resolution set of mobile phone trajectories and an OpenStreetMap road network model. Mixed road traffic is modelled as the result of a choice for each user between a simulated congested car mode and a non-congested alternative mode parameterized by its speed.

Keywords: traffic simulation

1. Introduction

Transport planning is a multi-faceted exercise, necessitating many inputs from many different sources. One such source are simulations of the transport system. With such simulation system, one can insert a considered infrastructure or policy measure into the model, and observe the simulated reactions of the transport system. Downstream modules, such as the calculation of emissions (e.g.[1]), accessibilities (e.g.[2]) or inputs to a cost-benefit-analysis, can be attached (e.g. [3]). An important issue with such models is that it is rather time-consuming and expensive to put them together; for example, the model for the German national assessment exercise takes several years to put together, which may be prohibitive especially for a developing country.

In this situation, it is interesting to consider alternative, and possibly faster and cheaper, approaches. With the availability of OpenStreetMap (OSM) data, one major obstacle has been removed. We have consistently found that it is possible to base traffic simulation models based on that data in spite of some shortcomings [4], and while the data quality of OSM differs heavily between urban and rural areas, it approaches that of commercial network data for large cities.[5] The main issue is that OpenStreetMap data does not contain flow capacities, i.e. the maximum number of vehicles that can leave a link during an hour. Instead, that number is estimated from road category information.

The other missing item in order to bring such a model up and running is the demand. The demand contains, in some way, information about all trips that are made from one location to another during one day. Such demand data is typically obtained from surveys. Two typical sources are the census or similar information (which typically contains, besides the home location, the work or education location), or trip diaries, which are in many countries obtained from asking a sample of the population about how, and at which locations, they spent a certain day. However, sometimes such information is not available, or it is difficult to procure, for example because of privacy issues. In this situation, the generation of synthetic

Highway tag	Lanes	Free speed (km/h)	Capacity (veh/(l*h))
`motorway`	2	120	2000
`motorway_link`	1	80	1500
`trunk`	1	80	1500
`trunk_link`	1	50	1500
`primary`	1	80	1500
`secondary`	1	60	1000
`tertiary`	1	45	600
`minor`	1	45	600
`unclassified`	1	45	600
`residential`	1	30	600
`living_street`	1	15	300

Table 1. Link attributes used for the values of the OpenStreetMap highway tag.

demand from electronic sources has become an increasingly active research field. Some of these investigations have a focus on route choice (e.g. [6]), while others derive origin/destination matrices [7][8]. Most of them employ some sort of intermediate model of behavior or trip generation before assigning traffic to the road. In consequence, the guiding focus for the present paper is the question in how far a meaningful traffic simulation can be constructed directly just out of OpenStreetMap network data, and anonymous cell phone traces as provided by the "Data for Development" (D4D) challenge.[9]

The paper is structured as follows: In section 2, we describe the data sets used for creating the supply (network) and demand (population) data. Section 3 covers the construction of the simulation model. In section 4, we give some results of parametric simulation runs. In section 5, we discuss some of the issues we faced while constructing the model and directions for future research.

2. Data description

2.1. Road network

The road network data for this scenario is based on OpenStreetMap. The OpenStreetMap data was converted to a simulation network by assigning attributes related to traffic flow to road segments. This is done based on the value of the `highway` tag, a road classification scheme particular to OpenStreetMap.[4] A graph representation of the road network is constructed from the OpenStreetMap data, where each intersection becomes a vertex and each road segment becomes a link. The precise geographic embedding of the road segments is discarded, and only its length is stored as a link attribute, along with its capacity, maximum speed in uncongested state, and number of lanes. The values we used are given in figure 1. The resulting network for the entire country has approximately 22,000 nodes and 63,000 links.

According to the CIA world factbook web site, the length of the road network of Côte d'Ivoire is 80,000 km, of which 6,500 km are paved. The OpenStreetMap documentation features an overview of the international equivalence of highway tags, but no country in Africa is currently included in this overview. The combined length of all edges labelled with any value of the highway tag is 29,300 km. The combined length of all edges labelled with `highway=primary`, `highway=motorway` or `highway=trunk` is 9,000 km, suggesting that these categories together already contain some roads classified as unpaved by the other source, and that the rest of the network should definitely be considered unpaved. Still, the data only accounts for less than half of the reported network length, so it is to be expected that siginificant parts of the network, most of it probably in rural areas, is not available in the model. This contrasts with coverage in the more economically developed countries, where the part of the road network not covered by OpenStreetMap can be considered negligible for traffic modelling purposes.

Upon inspection, certain areas of the country seem to have a good coverage, notably the area of the economic capital, Abidjan. For this reason, and because our simulation approach has so far been applied mostly for urban areas, we decided at this point to focus on the city.

2.2. Mobile phone sightings

The mobile phone data under consideration here consist of two sets of individual trajectories collected over a study period of 150 days: One set of high spatial resolution (HSR) data, and one set of long term (LT) data. The HSR set consists of trajectories generated from billing data, and tracks 50,000 individuals. The individuals are drawn from the customer base of one mobile phone operator, which claims five million customers, at an estimated total population of twenty million.

Locations are given by the number of the cell phone tower with which the mobile phone was communicating at the time of the record. Locations are only recorded while the user is in a call. After every two-week period, the population sample is redrawn, so it is not possible to track a single individual over more than two weeks. The LT set tracks another sample of 50,000 individuals, but over the whole study period, at the price of providing much lower spatial resolution, namely on the level of sub-prefectures. In this experiment, the idea was to directly generate a pattern of daily traffic from trajectories, so we use the HSR data set.

3. Traffic simulation model

The simulation is a loop which consists of:

- traffic flow simulation
- scoring
- replanning

This loop operates on an initial population of individiual agents which is generated from the mobile phone data. The following sections describe the phases in detail.

3.1. Initial Demand

Côte d'Ivoire has an estimated population of 20 million, according to the World Bank population data set on `google.com/publicdata`. We generate a synthetic one percent sample population by creating 200,000 synthetic agents. We simulate no particular day, but a typical work day, so we overlay four arbitrary days of mobility traces, each taken from a different sample (size 50,000) of mobile phone customers from the HSR data set. This means that we are combining data from multiple days to build a population for a supposedly average working day. Clearly, this is not the same as having 200,000 samples from one day. It is, however, still better than the alternative, which is expanding the first 50,000 samples to 200,000.

Each agent is equipped with a mobility plan which consists of an alternating list of

- geo-located and timed activities
- leg descriptions, including mode of transport and route

Agents essentially divide their time between conducting an activity, and travelling.

For each agent, an initial mobility plan is devised which is consistent with the data: The agent has to be in cell C_i at time t_i for every reading i. This leaves many degrees of freedom. In particular, it is not known whether the agent is travelling when a reading is taken. Any partitioning of the time into activities and trips which is consistent with the readings is in principle admissible. We start by defining every reading i to be an activity which ends at the time t_i the reading is taken. If several consecutive readings happen at the same cell, only the latest of those is considered. At this time, the agent will start travelling towards the location of the next reading $i + 1$ and, upon arrival, will stay there until the time t_{i+1}. During this time, the agent is considered to be conducting an activity. Activity locations are fixed to a geographical point which is randomly drawn from the Voronoi cell C_i of the tower where the reading was taken. It is assumed that activity locations have direct access to the road network, so they can be positionally identified with links. Each randomly drawn point is therefore snapped to the end of the nearest link, which is considered to be the activity location.

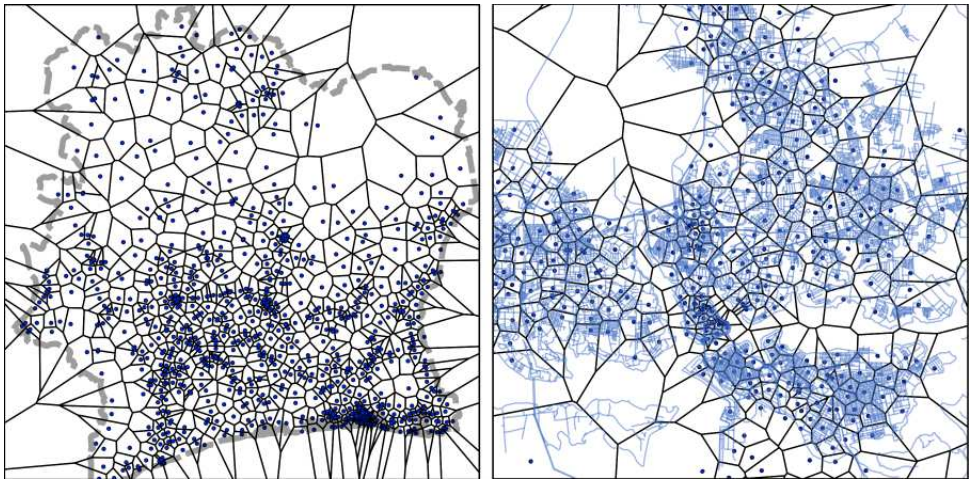


Fig. 1. Cell tower locations with their Voronoi cells. Comparing Voronoi cell sizes on a country-wide scale (left, overlayed with national border) and on the scale of the Abidjan urban area (right, overlayed with road network) show the large variation in cell size.

The plan is then checked for initial feasibility. Each leg is routed through the road network on the fastest route at maximum uncongested vehicle speed, as per the link attributes stored in the road network model. The travel time is summed up, and it is checked if the agent would under these assumptions be able to reach the next activity location in time. If this is not the case, the plan is redrawn. This procedure is iterated 20 times, and if no feasible plan has been found by then, the case is considered pathological and discarded. This does not necessarily mean that the input data does not resemble a real trajectory. Incomplete road network data is the most likely reason for these cases.

Finally, we filter and keep only those plans with at least one sighting in the Abidjan urban area, which is our study area.

The result of this initial demand generation process is a 1% synthetic population sample where every agent uses a car for every trip and tries to take the freespeed-fastest route through the road network. This initial population is then fed into the simulation loop for relaxation, the steps of which are described in the next paragraphs.

3.2. Traffic flow

The mobility simulation concurrently executes the mobility plans of the agents. Agents leave their activity location at the scheduled activity end time and head for their next destination. The road traffic is simulated using the queuing model of traffic flow[10]. In this model, the limited flow capacity of links is honored, so that in every time step, only as many vehicles can exit a link as specified. If more vehicles enter a link than its flow capacity admits, vehicles will accumulate on the link until its storage capacity is hit, which is determined by the length and width of the link. When a link is full, no more vehicles can enter, causing the congestion to propagate back upstream. The simulation records time use for each agent: The points in time where agents depart from and arrive at their activity locations are fed back to the agents as experience to evaluate the utility of the executed plan. Note that simulated vehicles are considered uniform. In this experiment, in contrast to explicitly modelling different vehicle types and their interaction [11], we model mixed traffic as a combination of uniform car traffic and the possibility to switch to an uncongested mode, which is described below.

3.3. Plan scoring

The agents evaluate the outcome of their plan with a simple utility-based approach. Time spent travelling is considered to contribute negative utility. Since, in this study, we do not have an activity model which

```
<person id="10008_1">
 <plan>
  <act type="sighting"
   link="91273255_1059606493_1059606739_R"
   x="-455429.63088982203" y="593487.4664630938"
   end_time="07:42:00" />
  <leg mode="car"
   dep_time="07:42:00"
   trav_time="00:04:15"
   arr_time="07:46:15">
    <route type="links">
     91273255_1059606493_1059606739_R
     91273255_1059606493_1059606739
     91273253_1219665629_1059606739_R
     [...]
     125239948_338881494_338881537</route>
   </leg>
   <act type="sighting"
    link="125239948_338881494_338881537"
    x="-454496.5669239445" y="593286.9239709259"
    end_time="17:36:00" />
   <leg mode="car"
    dep_time="17:36:00"
    trav_time="00:00:31"
    arr_time="17:36:31">
     <route type="links">
      [...]
     </route>
    </leg>
    <act type="sighting"
     link="30630786_338406146_338406157"
     x="-454806.9268885712" y="593254.6327337974"
     end_time="18:35:00" />
 </plan>
</person>
```

Fig. 2. Example for an agent plan. This person had 3 call records on the examined day. From the location of the first call record (("sighting", location randomized within the cell) to the second, the free speed travel time is only about 4 minutes, from which the actually experienced, congested travel time may differ enormously. Upon arrival at the second location, the agent will wait until 17:36:00 and depart for its final destination.

would allow comparing the relative utility of time spent at one location with another, we consider time spent in activities to have no contribution to the utility function. Since we have no prior knowledge about the trip strucure, and our modelling decision to split trips at the locations of the GSM readings is somewhat arbitrary in this respect, we consider the disutility of travelling to be linear in total travel time.

$$U = \beta_{trav} \cdot t_{trav} \tag{1}$$

Since in this experiment the utility function does not have any other terms, the value for β_{trav} is arbitrary, as long as it is negative.

3.4. Replanning

After each iteration, the agent population has the opportunity to change their mobility plans in reaction to the outcome of the mobility simulation. In this study, agents have three replanning options. Two of them are creative. Agents chosing these options produce a new plan and execute it in the next iteration of the mobility simulation. These options are route choice and mode choice. The third is switching plans, in which agents retry a previously executed plan from their plan memory based on its previously experienced utility. In each iteration, 10% of the agent population consider their route choice and mode choice, respectively. The remaining 80% can switch plans.

Route choice. Agents reconsider their route through the road network. Instead of taking the least-cost path based on free speed travel times, the link travel times computed as an outcome of the last iteration of the traffic flow simulation are used. In the first iteration, the traffic will concentrate on main roads, leading to high traffic volumes and high traffic times. Agent reconsidering their route will divert to smaller roads in the next iteration.

Mode choice. In the initial population, all trips are done by car. Our model summarizes all alternatives to driving a car in a second mode. Agents which choose this mode are not routed through the network at all. They experience a travel time calculated from the free-speed car travel time between the origin and destination locations, times a travel time factor which characterizes the mode. These agents do not interact with other agents while travelling. They are not impeded by other travellers and do not contribute to congestion themselves.[12] Depending on the travel time factor, a certain share of the population remove themselves from the road network. Note that the modal split is not part of the input data, but an output of the simulation, dependant in particular on the travel time factor.

Switching plans. Every agent has a fixed-size plan memory, set to size 5 in this experiment. Agents which are assigned the option to switch plans pick one of their previously tried plans uniformly at random, and switch to that plan with a probability depending on the difference between the most recently experienced scores of both plans:

$$p_{ij} = 0.01 e^{\frac{s_j - s_i}{2}} \tag{2}$$

In this equation, p_{ij} is the probability of switching frm plan i to j, s_i is the current score of plan i, and 0.01 is the probability of switching between equally scored plans. The simulation is iterated until the system reaches a relaxed state. We consider this to be the case as soon as the average agent score (i.e. travel times) and the mode share have stabilized.

4. Implementation and Results

The scenario was implemented using the MATSim agent-oriented transport simulation software (`www.matsim.org`). In order to be able to run experiments on a desktop computer, we decided to simulate a 1% sample of the synthetic population described in the previous section, scaling the network capacity accordingly. In our experience with the process, this is sufficient to pick up large-scale characteristics of the system. One simulation run takes about an hour on a 2.2 GHz Intel Core i7 MacBook. A run consists

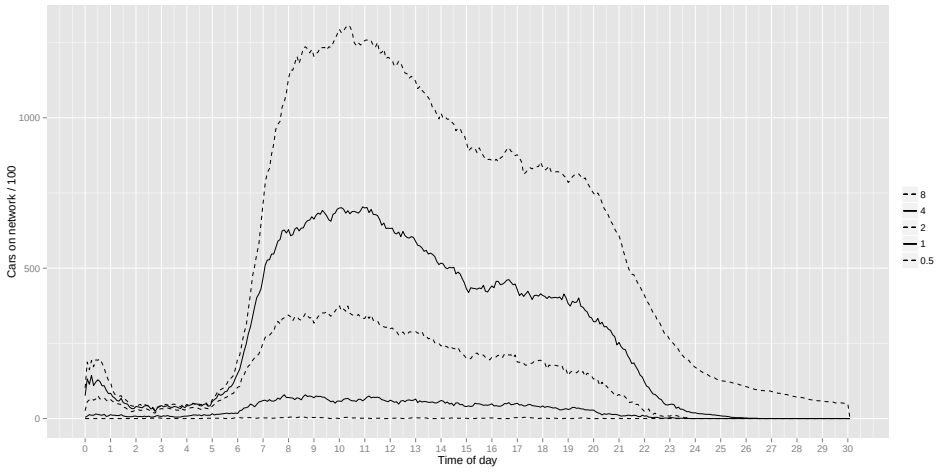


Fig. 3. Number of cars en-route over time of day, plotted for different values of the alternative mode travel time factor. One agent represents 100 travellers. Note that as only sightings from a single day are used to construct the artificial population, agents are expected to be at their final location by midnight at the latest. For factor 4, the network is already too full to permit this. If the alternative mode is faster than driving (factor 0.5), the network is, expectedly, cleared.

of 180 iterations of the simulation loop, which we found to be enough for the quantities presented in this paper to stop drifting. For the last 30 iterations, the creative replanning options, namely route choice and mode choice, are disabled, and agents only switch between existing plans. This is done to eliminate the bias introduced by having a large fraction of the agent population take new routes or the alternative mode without regard for their possibly low utility.

We produced several parametric simulation runs, varying the travel time factor for the non-car alternative. As can be seen in figure 3, a travel time factor of 4 already leaves some agents unable to reach the point of their last sighting before midnight, which means that this state of affairs would be clearly inconsistent with the data. Factor 2 still seems admissible. The travel time factor can be interpreted as how long car travel times along a path in the congested network need to become compared to its free-flow state so that agents travelling along that path will be moved towards using the alternative mode.

In figure 4, we plot the resulting share of car drivers over the travel time factor. Since a factor of 4 is already considered inadmissible, the prediction would be a share of not much more than 0.2.

Even with a factor of 4, i.e. a situation which is quite congested according to Fig. 3, no systematic or directed traffic jam patterns emerge. The network just seems too full overall. This is quite different from other similar studies (e.g.[13, 14]), where we always found quite well-structured congestion patterns, in particular into the city during the morning peak. Further inspection of the results leads to the observation that most of the congestion in our model seems to be away from the freeways, on the secondary road network. That is, under congested conditions traffic is unable to get out of the secondary street network. Once the traffic makes it onto the primary network, the model displays few if any restrictions. Clearly, this statement would need to be verified on the ground before being a possible basis for planning decisions. It could, for example, also be a consequence of the demand generation, which, in particular because of spurious cell handovers, may generate a lot more local traffic than there is in reality. If such verification on the ground would corrobate that the local congestion effects are over-estimated, then methods to remove those spurious cell handovers from the demand generation would need to be inserted into the model.

Figure. 5 displays the probability density of the total travel time per person per day. One notices a peak near 0.2 hours for the car mode, and near 0.6 hours for the non-car mode. While the 0.6 hour value seems plausible for a reasonable walk length of 30 minutes and an average number of daily trips of less than 2 (given in [15]), the 0.2 value seems way too low, suggesting again that we overestimate the number of local (very short) car trips.

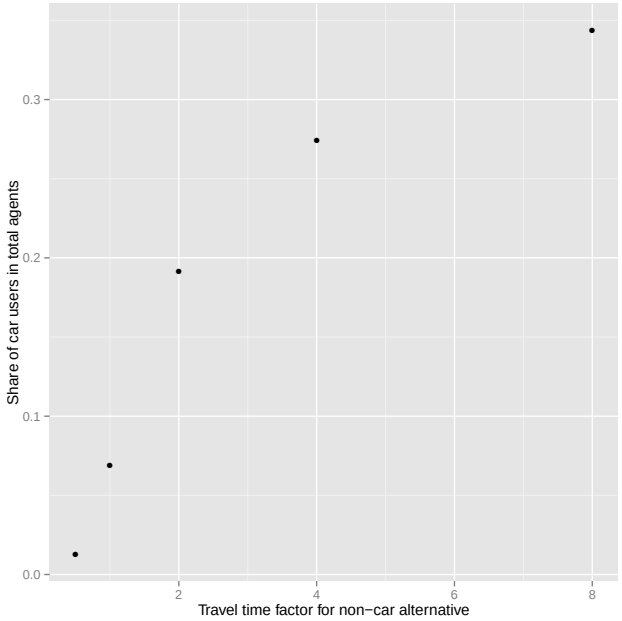


Fig. 4. Number of agents using a car, plotted over the alternative mode travel time factor. Note that an agent either does or does not use a car for all trips of the simulated day. We do not consider mode choice for individual trips.

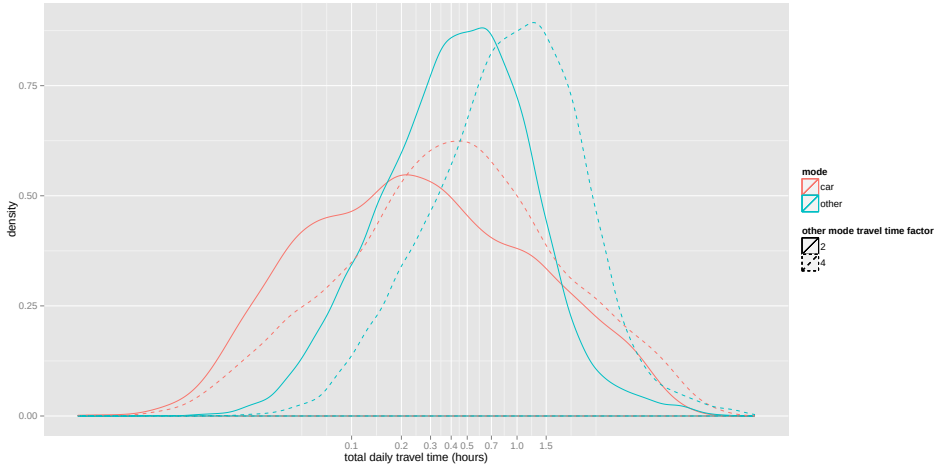


Fig. 5. The daily time spent traveling, for the driving and non-driving sub-populations.

5. Discussion and outlook

Modelling road network access. In the present scenario, as well as in the MATSim software package, the assumption is that every activity location has direct access to the modelled road network. For rural areas in developing countries, this is clearly not met, if only because the OSM-based network model accounts for less than half of the presumed length of the actual road network. In this paper, we focus on the Abidjan urban area, but for a country-wide study, it may be worthwhile to improve on this aspect.

We therefore intend, in future version of MATSim, to explicitly account for the distance between the random location and the network by modelling trips in several stages: network access, network travel and network egress. The travel time for the network stage is determined by the traffic flow simulation, as in the present paper. Travel times for network access and network egress are determined by multiplying Euclidic distance by some factor which represents an unknown mode of travel through unknown terrain with an unknown detour. Conceptually, this would be done by extending the road network with virtual access links which orthogonally connect activity locations to the nearest road network link. This would model a mode of travel composed of several stages, like walking to the next road, being picked up by a motorist, and continuing by car.

Imputing behavioral meaning. Most if not all similar studies go the path that they first attempt to create plausible daily activity plans from the mobile phone records and only then move on to an assignment of the traffic onto the traffic infrastructure. For the present investigation, we have deliberately chosen to immediately assign the mobile phone data to the road network, without an intermediate interpretational layer. There were two reasons to do so:

- We believe that plausible traffic patterns can already be obtained without that intermediate step, and that it saves a lot of time in order to get such simulations up and running. This could, for example, be important for situations with limited budget, or with situations with time pressure such as, say, disaster relief. Clearly, the claim that the results are realistic would need to be checked, for example by traffic counts data on the ground. Such data is, however, fairly straightforward to obtain, for example by, on a particular day, employing someone to stand next to the roadside and count vehicles.
- We believe that it is possible to impute the activity chains also *after* the traffic assignment. In fact, we believe that it may be better to do so, since the interpretational layer always means a loss of information that may still have been in the raw data, such as the deletion of seemingly implausible sightings, or certain variations in the temporal structure from one day to the next. With the approach discussed in this paper, one could always keep the original plan based on the mobile phone data, but generate multiple alternative plans for every synthetic traveler that would be consistent with the mobile phone data. For example, it could be assumed that some phone calls would actually be done en-route, or that some activities would carry on after the last phone call at a certain location. Out of these multiple interpretations of the mobile data, the system could converge to a set of interpretations that is most consistent with other data, such as, for example, time-dependent traffic flow data. This will be the subject of future work.
- An alternative approach might be to use a time use survey, which is available in many countries, as additional data input. Time use surveys are similar to trip diaries in that they follow persons over days, but in contrast to trip diaries they typically do not register locations. Advantages of time use surveys over trip diaries include that they are considerably cheaper to obtain since the geocoding of the locations is expensive, at least with traditional approaches, and they have fewer privacy issues. In consequence, time use surveys are available in many places where trip diaries are not available. In addition, it may even be possible to use a time use survey from a neighboring city or country if the cultures are sufficiently similar. The imputation of activity chains from those time use surveys could be done in ways similar to those pointed out above: For each given sequence of cell phone sightings, one would select all possibly matching activity chains, or a randomly drawn subset. A data assimilation algorithm would then pick those activity chains most consistent with directly and anonymously measured data, such as traffic counts.

Statistical bias. Trajectories sampled from mobile phone users alone are most probably biased. For example, not all members of the population have a phone, persons with a phone have vastly different calling patterns, and trips of persons who make fewer calls will be underreported. There is some indication that, for the purpose of mobility studies, such bias may not be as dramatic as it seems [16]. We believe that the approach discussed above, which is to do the data assimilation with the traffic model already up and running, might also help here. More specifically, one could imagine to give different statistical weights to every synthetic person. Based on other data, like for example time-dependent traffic flow data, one could re-weight the synthetic persons in order to bring the simulation closer to the data. This would presumably increase the weights of those types of persons that were under-weighted in the data, and decrease the weights of those types of persons that were over-weighted. Clearly, the approach will not work if certain types of persons are not included at all. We will investigate these issues in future work.

6. Conclusion

- The investigation demonstrates once more that it is possible to use publicly available OpenStreetMap data as the basis for traffic simulations. In the present situation, the coverage of the rural areas was still insufficient. However, one can either assume that this will improve over the years, or one could dispatch special investigations to insert the missing information if that turns out to be necessary for a specific study.
- The investigation also demonstrates that it is possible to obtain traffic patterns from mobile phone sightings without any layer of interpretation whatsoever. The traffic patterns look plausible; however, some verification would be necessary to decide if they are close enough to reality in order to use the model for policy analysis.
- A parametric study demonstrates that the model is sensitive to the performance of non-car modes.
- Sec. 5 discusses a method how the model could be systematically improved further if additional data is available. As explained, such possible data coould consist of, e.g., time use surveys or traffic counts.

References

[1] B. Kickhöfer, F. Hülsmann, R. Gerike, K. Nagel, Rising car user costs: comparing aggregated and geo-spatial impacts on travel demand and air pollutant emissions, in: T. Vanoutrive, A. Verhetsel (Eds.), Smart Transport Networks: Decision Making, Sustainability and Market structure, NECTAR Series, Edward Elgar, 2012, in press, also VSP WP 11-16, see www.vsp.tu-berlin.de/publications.

[2] T. Nicolai, K. Nagel, High resolution accessibility computations, VSP Working Paper 13-02, TU Berlin, Transport Systems Planning and Transport Telematics, see www.vsp.tu-berlin.de/publications (2013).

[3] D. Pearce, C. Nash, Social appraisal of projects: A text in cost-benefit analysis, Wiley & Sons, London, 1981.

[4] M. Zilske, A. Neumann, K. Nagel, OpenStreetMap for traffic simulation, in: M. Schmidt, G. Gartner (Eds.), Proceedings of the 1st European State of the Map – OpenStreetMap conference –, no. 11-10, Vienna, 2011, pp. 126–134, see sotm-eu.org/userfiles/proceedings_sotmEU2011.pdf.
URL sotm-eu.org/userfiles/proceedings_sotmEU2011.pdf

[5] D. Zielstra, A. Zipf, Quantitative studies on the data quality of OpenStreetMap in Germany.
URL http://www.giscience2010.org/pdfs/paper_187.pdf

[6] N. Rieser-Schüssler, M. Balmer, K. Axhausen, Route choice sets for very high-resolution data, Transportmetrica iFirst (2012) 10021. doi:10.1080/18128602.2012.671383.

[7] J. Ma, H. Li, F. Yuan, T. Bauer, Deriving Operational Origin-Destination Matrices from Large Scale Mobile Phone Data.
URL http://www.mygistics.net/docs/MobileOD.pdf

[8] F. Calabrese, G. Di Lorenzo, L. Liu, C. Ratti. Estimating Origin-Destination flows using opportunistically collected mobile phone location data from one million users in Boston Metropolitan Area [online].

[9] V. D. Blondel, M. Esch, C. Chan, F. Clerot, P. Deville, Data for Development: the D4D Challenge on Mobile Phone Data, arXiv.org.
URL http://arxiv.org/abs/1210.0137

[10] C. Gawron, An iterative algorithm to determine the dynamic user equilibrium in a traffic simulation model, International Journal of Modern Physics C 9 (3) (1998) 393–407.

[11] A. Agarwal, M. Zilske, K. Rao, K. Nagel, Person-based dynamic traffic assignment for mixed traffic conditions, VSP working paper 12-11, TU Berlin, Transport Systems Planning and Transport Telematics, see www.vsp.tu-berlin.de/publications (2012).
URL https://svn.vsp.tu-berlin.de/repos/public-svn/publications/vspwp/2012/12-11/

[12] M. Rieser, D. Grether, K. Nagel, Adding mode choice to a multi-agent transport simulation, Transportation Research Record: Travel Demand Forecasting 2009 2132 (2009) 50–58. doi:10.3141/2132-06.

[13] K. Nagel, Towards simulation-based sketch planning: Some results concerning the Alaskan Way viaduct in Seattle WA, Tech. Rep. 08-22 (2008).
URL https://svn.vsp.tu-berlin.de/repos/public-svn/publications/vspwp/2008/08-22

[14] K. Nagel, Towards simulation-based sketch planning, part II: Some results concerning a freeway extension in Berlin, Tech. Rep. 11-18 (2011).

[15] World Bank. Non-Motorized Transport in African Cities [online].

[16] A. Wesolowski, N. Eagle, A. M. Noor, R. W. Snow, C. O. Buckee, The impact of biases in mobile phone ownership on estimates of human mobility, Journal of The Royal Society Interface 10 (81). arXiv:http://rsif.royalsocietypublishing.org/content/10/81/20120986.full.pdf+html, doi:10.1098/rsif.2012.0986.
URL http://rsif.royalsocietypublishing.org/content/10/81/20120986.abstract

A Tale of Peoples' Movement Patterns in Developing Countries

Kuldeep Yadav, Amit Kumar, Vinayak Naik, and Amarjeet Singh

Indraprastha Institute of Information Technology (IIIT), Delhi, India
Email: kuldeep@iiitd.ac.in

Abstract

Aggregated mobility patterns of mobile users gives deep insights on how people travel across a country which are required for large scale decisions such as transport, city infrastructure etc. With the availability of massive cellular data, studying large scale mobility patterns have become easier and recently, there has been lot of research work on using cellular data to characterize human mobility in different cities. Much of this work have been done for developed countries and to the best of our knowledge, there is no work on characterizing human mobility patterns in a developing county primarily due to un-availability of data.

In this work, we make a first attempt in finding and analyzing aggregated mobility patterns of people in a developing country i.e. Ivory Coast. We found that there is a large difference in daily ranges as well as number of regular places visited by mobile users in Ivory cost as compared to findings of a similar study done in US. We have also found that a give subprefecture's residents tend to travel more often than others.

1 Introduction and Motivation

Location has been an integral part of a person's context because it can be used to infer several key attributes of a person's mobility i.e. places that she visits, frequent traveling routes, interactions with other people etc. The number of mobile phone subscribers are over 6 billion covering nearly 80% people across the world. Today's mobile phones give an ideal platform for collecting location data primarily from two sources, either using a mobile device or from a cellular network end. Most of location studies done till now are based on data collected from individual's mobile phones [11, 10]. While these location studies offer deep insights in to a person-specific mobility [9], they failed to give broader mobility patterns (i.e. at a scale of city or country) due to limited number of users. A city or country scale mobile patterns are important to analyze because they can

offer critical insights in domain such as transportation, city infrastructure etc. For instance, a city level mobility data can be used to answer some of following questions:

1. What is the typical distance travelled by citizens?
2. How many people travel on weekends or holidays as compared to weekdays? How much public transport needed for weekends?
3. Can we categorize residential and industrial areas of a city? How many people are likely to take a specific route based on aggregated people's home and work locations?

In case of cellular network, identifier of a cell tower (popularly known as Cell ID) is collected as part of Call Detail Records (CDRs) when a phone connected to the network make/receive a phone call, send/receive a SMS or MMS or have an active data connection. With over 80% mobile penetration across the world, Cell ID data collected from cellular network provides an opportunity to perform analysis to find large scale mobility patterns which were nearly impossible before. Recently, there has been research work in which CDRs are used to study human mobility. One of first research work with CDRs is done by Gonzalez et al [1] which found that human mobility is highly redundant in spatial as well as temporal dimension. This work focussed on modeling an individual's mobility pattern using a CDR dataset of about 1,00,000 mobile users.

There are some other related work which have used CDR data in diverse application scenarios i.e. forecasting socio-economic trends [18], characterizing urban areas [19], characterizing human mobility patterns [17] and studying disease spread [2]. Here, our focus will be mainly on research work related to characterization of human mobility. Isaacman et al analyze daily travel of people living in Newyork and Los Angeles using a metric *daily range* which represents the maximum distance travelled by a phone user in day [14, 12]. This work reveals several interesting patterns such as people in Los Angeles travel two times more as compared to New York during their regular travel. Also, Isaacman et al report considerable difference in people's movement across different days of week (i.e. weekdays and weekends) as well as according to different months of an year (i.e. summer or winter). In a subsequent work, Issacman et al build algorithms to identify important places in a person's mobility history using CDR data. Using ground truth derived from few volunteers, they have found that estimation of user's important location such as "Home" and "Work" location could be done with an error of 1 mile [16].

Till now, mobility characterization studies are restricted to developed world. The penetration of mobile phones is growing in developing world and studying human mobility for developing world using CDRs data can throw some critical insights too. We believe that human mobility patterns in developing countries may be different from developed world due to many reasons such as quality of transportation, socio-economic status etc. In this paper, we do a detailed characterization of human mobility in a developing country i.e. Ivory Coast

using a publicly available CDR dataset shared by one of biggest mobile operator in that country [1]. We also compare our aggregated statistics with the studies done in developed world. Specifically, our contributions and organization of this paper are as follows:

1. Section 2 describes the dataset which we have used in the analysis for this paper. For all our analysis, we have used fine grained trajectory dataset of $50,000$ people.

2. Section 3 presents aggregated travel patterns of mobile users in the given dataset. We found that nearly half of users do not move from a single location where as a subset of people do long distance travel. We discovered that Friday's and Sunday's are the most preferred days for longer distance travel.

3. Section 4 presents the place visiting patterns for all the mobile users in our dataset. From our analysis, we have found that people in Ivory coast have less number of regular (important) places than people in US.

4. Section 5 discuses the aggregated travel patterns of users living in three different subprefectures in Ivory Coast. One of subprefecture residents travel range is significantly higher than two other subprefectures. A similar result was seen in case of US study too.

5. Finally, we have a discussion in Section 6.

2 Dataset

The dataset used in this paper was acquired as part of Orange D4D challenge. It has phones calls in Ivory Cost during 5 month (from December 2011 to April 2012). The original dataset contains 2.5 billion records, calls and text messages exchanged between 5 million anonymous users. For this paper, we analyze the fine-grained mobility traces of $50,000$ people which has data for subsequent 20 weeks but every two weeks, there is change in user IDs to preserve privacy of mobile users. Unless and until specified, we will be using the dataset of first two weeks throughout the paper, however we have found that our findings are consistent across all the periods of 2 weeks in this dataset.

3 Aggregated Daily Travel Patterns

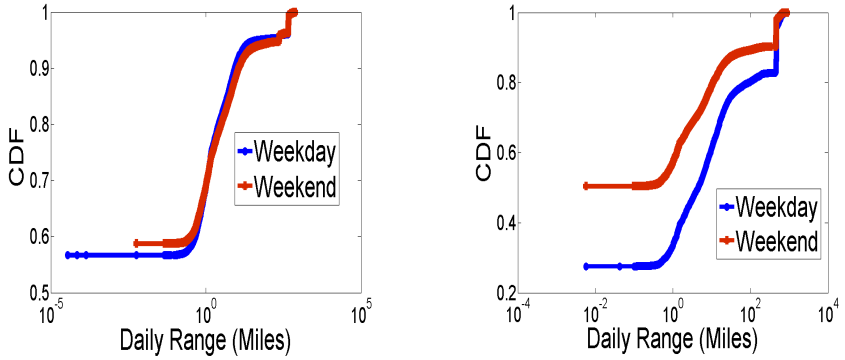
From the CDRs location data, we wanted to analyze aggregated daily mobility ranges of people. Isaacman et al [12] used a metric *daily range* which represents the maximum distance traveled by a person in day. For instance, if a person

[1] http://www.d4d.orange.com/home

visit locations $\{C_1,C_2,C_3,....C_k\}$ in a day then the *daily range* will be the maximum pair wise distance between these locations.

$$dailyrange(d) = maximum(distance(C_i, C_j)) \ \forall i, j \in (1, k)$$

Because, CDR location is recorded only when a person makes call or SMS, it may miss some of the location names which are visited by users but did not got recorded. In this work, we are interested in aggregated mobility patterns and we hypothesize that effect of missed location names will be minimum. It is hard to aggregated this effect due to scale of data collection. However, previous studies [12, 14] have found that *daily range* can give a lower bound of a person's travel and have minimum error when compared with ground truth of few volunteers.



(a) CDF of median daily ranges (b) CDF of maximum daily ranges

Figure 1: CDF plots of median and maximum daily ranges computed from trajectories of 50,000 users.

Percentile	**Weekdays Median Daily Range**	**Weekdays Maximum Daily Range**	**Weekend Median Daily Range**	**Weekend Maximum Daily Range**
Min	0	0	0	0
2	0	0	0	0
25	0	0	0	0
50	0	4.48	0	0
75	1.47	27.43	1.65	6.96
98	463.09	592.97	463.21	478.38
Max	760.60	842.58	760.60	884.66

Table 1: Median and maximum daily ranges computed from trajectories of 50,000 users.

In the given dataset of fine grained trajectories of 50,000 people, we have computed daily range for each day for every user. After that, we have computed median daily range and maximum daily range per user. Median daily range for each user represent the most frequent (regular) travel that she takes most of the days where as maximum daily range represents the maximum distance she travels on some days which are likely to be infrequent. For instance, study done in US [12] found that a user is more likely to do a long distance travel on weekends. Figure 1 presents the CDF plots of median and maximum daily ranges of $50,000$ users for two weeks. There is not much difference in user's median daily ranges (i.e. regular movement) among weekdays and weekends. But, we could see a significant variance in people's mobility on weekends using maximum daily ranges as lot of users prefer to stay at home during weekends where some percentage of users chose to travel large distances. Table 1 present the percentile values of median and maximum daily ranges comparison for the same time period. Some of the main observations are as follows:

1. On weekdays, $50th$ percentile of median daily range is zero mile which represents that more than half of users were staying at the same place (Cell ID) mostly. However, $50th$ percentile of median daily range is 4.48 mile on weekdays, it represents that some of these users made some occasional trips in two weeks duration but nearly 25% of users stayed only in same place.

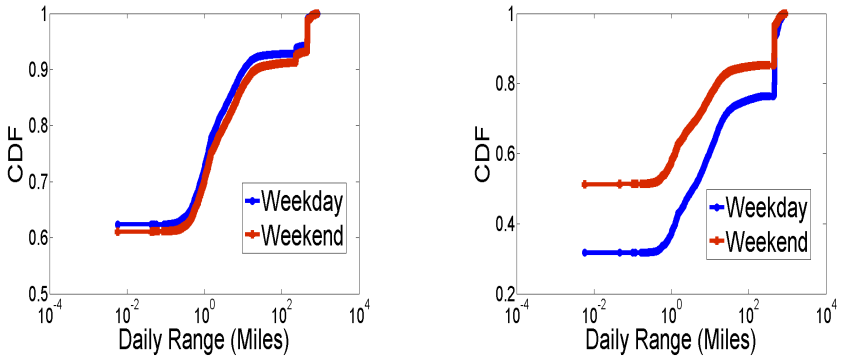
2. On weekends, $50th$ percentile of both median and maximum daily ranges are zero which represents that more than 50% of people choose to stay at home in the whole duration. This phenomenon can also be seen by reduced $75th$ percentile of maximum daily range from 27.43 mile on weekdays to 6.96 on weekends.

3. Table 1 shows that some people prefers to travel large distances on weekdays as well as weekends. Even though, $75th$ percentile decreases, $98th$ percentile remains the same for median as well as maximum daily ranges.

In the given dataset, there were 10 different time periods of two week each. We have found that these observations as well as percentile values are consistent across all time periods. For instance, Table 2 and Figure 8 provides CDF plots and percentile values for the same set of users in a different time period.

3.1 Weekday vs Long Distance Travel

Above analysis showed that users travel farther distances on weekdays as well as weekends and number of users who travel on weekdays are higher compared to weekends. Further, we were interested in finding out on which day of the week, users are more likely to do long distance travel. We have done following analysis to find it.



(a) CDF of median daily ranges (b) CDF of maximum daily ranges

Figure 2: CDF plots of Median and maximum daily ranges computed from trajectories of 50,000 users. (Time Interval: 27/2/2012- 11/3/2012)

Percentile	**Weekdays Median Daily Range**	**Weekdays Maximum Daily Range**	**Weekend Median Daily Range**	**Weekend Maximum Daily Range**
Min	0	0	0	0
2	0	0	0	0
25	0	0	0	0
50	0	3.58	0	0
75	1.25	81.50	1.49	8.48
98	466.34	640.87	467.62	568.71
Max	851.15	876.26	835.44	884.66

Table 2: Median and maximum daily ranges computed from trajectories of 50,000 users.(Time Interval: 27/2/2012- 11/3/2012)

1. For every user, we computed maximum daily range of each week and recorded the day on which it was achieved. It was computed for 10 different time periods (20 weeks).

2. For each day of the week in a time period, we count the number of users who achieved maximum daily range on that particular day.

3. For each day of the week, we have computed the average number of users across 10 different time periods.

Using the above process, we found that most users do their long distance travel either on a Friday or Saturday as shown in Figure 3 while Saturday being the most preferable day. Interestingly, Sunday was one of the least preferred day for long distance travel as very less people did their long distance travel on that day.

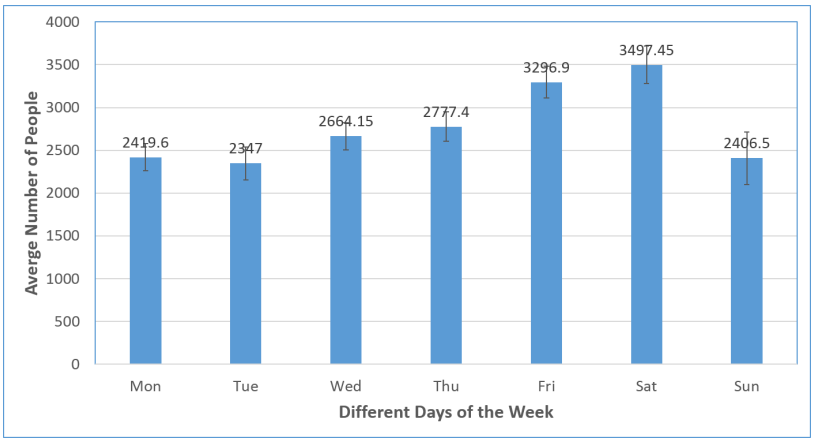


Figure 3: Weekdays vs average number of people who preferred traveling long distance. Saturday was the most preferred day for long distance travel

3.2 Comparison with Existing Studies

From our analysis with Ivory coast data, we have found daily travel range of people in this country is significantly lower than daily travel ranges reported by earlier work [12]. Also, a large number of population do not travel in their regular days (50th percentile is zero for median daily travel range) while in all the US cities, 50th percentile for all users was greater than 2 miles. This effect could be observed due to following reasons:

1. Cell tower density may be sparse in Ivory coast as compared to US. The places to which users travel may be very close to their home and Cell ID does not change.

2. In the given dataset of 50,000 users, most of data is of housewives's mobile phones and they may not be moving much.

In case of Los Angeles and New York, it was noticed that people travel their maximum distance on weekends i.e. Saturday and Sunday. Also, they have considered Friday as a part of weekend because a large number of people do long distance travel on that day too. In our analysis, we have observed that Friday's and Saturday's are indeed two most preferable day for doing long distance travel with a surprising notable exception of Sunday's where less users as compared to other weekend days do long distance travel. In fact, number of people who traveled on Sunday's are even lesser than some of weekdays.

4 Aggregated Place Visiting Patterns

Mobility of a typical user is predicable across different days because there are some frequently occurring (regular) places such as "Home" & "Workplace" [1].

Even though, user may visits several locations in the given duration, she tends to return to one of the regular places. There has been growing interest in logging different places that a user visits in day with the help of sensors such as GPS [11], WiFi [21] as well as using GSM-based information [20]. These fine-grained places information can be used in variety of applications i.e. personal history logging, place based reminders, computing exposure of pollution etc. However, CDR location data is very coarse and places extracted from these location events may not be suitable for some of these applications. Previous work [16] have found that "Home" and "Work" locations can be inferred from CDRs.

We are interested in analyzing aggregated place visiting pattern of all the users in our dataset. Previous studies have shown that a user's phone may connect to different cell towers even it she stays at same places [20]. We used this information to cluster nearby cell towers into one location using following algorithm [16]:

1. For complete duration of data, sort Cell IDs according to number of distinct days they were observed by a user's phone. This step helps us in gaining an understanding of importance of a Cell ID because a user is more likely to make a call from a regular visiting place.

2. After ranking of Cell IDs for every user, we use Hartigans leader algorithm to cluster nearby Cell IDs. It starts with the highest ranked Cell ID and then combine other Cell IDs if they fall with in a threshold distance (t_d). This process is repeated until every Cell ID in a user's movement history assigned to one cluster.

The inherent assumption in above algorithm is that the all the places are at least t_d distance away from each other. Isaacman et al have found that t_d equal to 1 mile works well in their dataset which was collected for Newyork and Los Angeles [16]. To find a good values of t_d, we have performed an experiment where we varied value of t_d from 0.5 mile to 5.5 mile and computed average number of clusters for $50,000$ users. As it is seen from Figure 4, average number of clusters nearly remains same, if value of t_d is equal or bigger than 1.5 mile. We select value of t_d equal to 1.5 miles for further experiments.

For finding places visited by a user, we have used the algorithm described above with value of t_d equal to 1.5 miles. As shown in Figure 5, most of users (about 29%) visits only one place in the whole duration. Large number of users (about 67%) have visited at most 3 different places only in the whole duration. Some users visits unusually high number of places, for instance 6% of users visited more than 10 different places in a duration of 2 weeks. In real-world, there are some users (i.e. taxi driver) who are more mobile other others and probably, visits to high number of places correspond to those set of users.

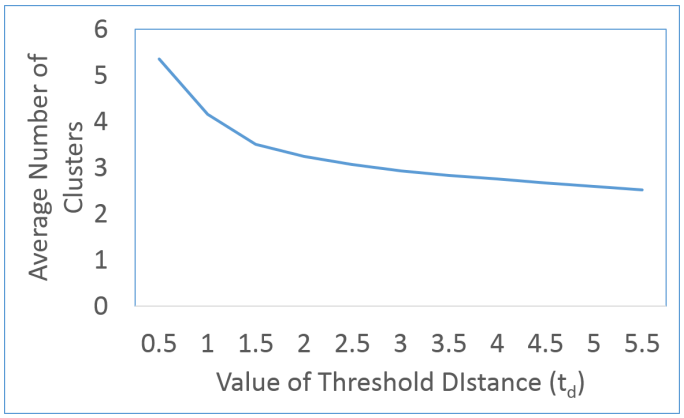


Figure 4: Effect of changing t_d on the average number of places

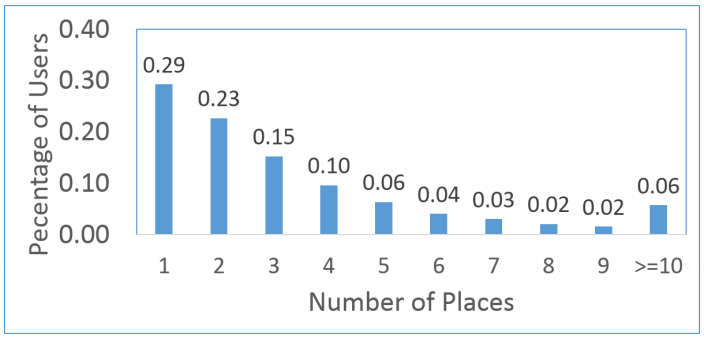


Figure 5: A histogram showing total number of places visited by users. While some users visits large number of places, most of them (67%) visits at most 3 places.

4.1 Regular Places

As described earlier, users are likely to visit many places and not all of these places are regular for a user. The most regular places for a user are likely to be "Home" and "Workplace" and users spent significantly higher amount of time in regular places as compared to places which are occasionally visited. Hereby, we define a metric *place support value* which is representative of regularity of a place in a user's mobility. For instance, *support value* of a place P_i for a given user U_k is computed as follows:

Support value (U_k, P_i) = *Number of days on which* P_i *was visited by* U_k/*Total number of days data availble for* U_k

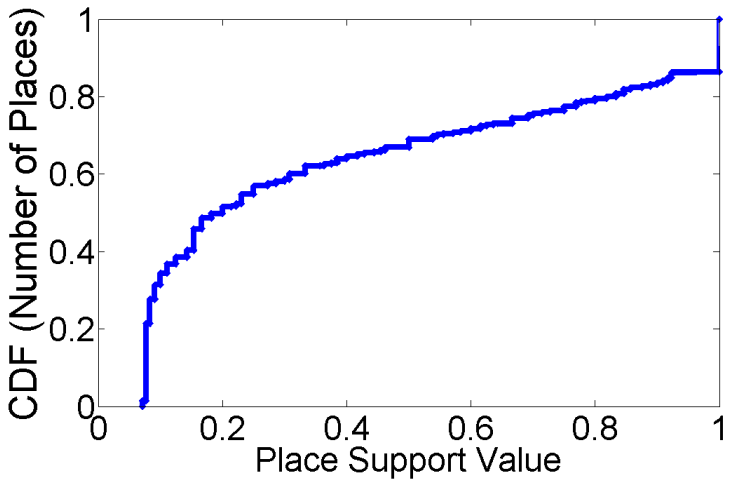


Figure 6: CDF plot showing support values of all the places visited by users. Majority of places have low support value indicating users only visited them occasionally

In the above equation, total number of days represents the days for which user U_k data is available in the given dataset and duration. For this experiment, we have considered the first two week of data only. After extraction of places, we compute a support value for each place. A CDF plot for support values of all places across all user is shown in Figure 6. As it can seen from the Figure 6, nearly 30% of the places visited by users have support value greater than 0.5 which means that users visit these place more than half of total days. Also, 60% of total places have a support value less than 0.3 which means that these places are visited less than one third of whole duration.

Next, we looked at how many regular places that a person visits in a given duration. Figure 7 shows the histogram of number of users w.r.t. number of (regular) places with two different support value threshold. If we use support value threshold greater than equal to 0.3, it means that a place has to be visited

on more than one third of the total days and similarly, it holds for support value threshold equal to 0.5. Majority of users (about 91%) had at most two regular places in their mobility profile when support threshold was equal to or greater than 0.3. For some users, we did not found any regular place which may be caused due to lack of location events (CDRs) or absence of fixed calling pattern. Comparing Figure 7 with Figure 5, we conclude that while users may visit large number of places in a give duration, their regular places remains only few. Also, we were able to extract regular places from location events generated in CDRs.

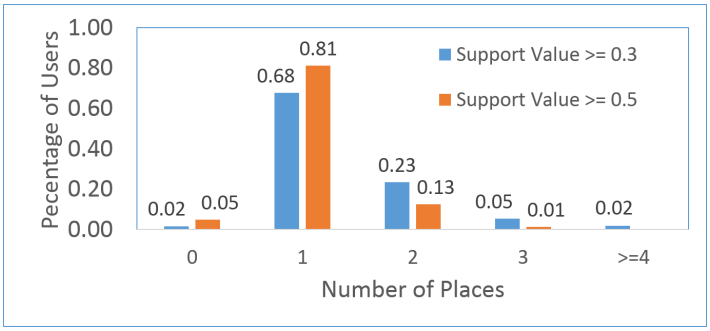


Figure 7: A histogram showing number of regular places visited by users. Majority of users had at most two regular locations

4.2 Comparison with Earlier Studies

For most number of users (about 25%), number of important (regular) places in US study was 5 in Los Angeles and New York. In case of Ivory coast dataset, most of users (68%) were restricted to only one place. One of the bias in case of Ivory coast is that data duration is only 2 weeks compared to more than 11 weeks duration of US study. However, we have analyzed different time periods of data in case of Ivory coast and found nearly same distribution for every week. Also with the larger duration, number of distinct places may rise but for a typical user's mobility profile, number of regular places are unlikely to change much. After comparing distribution of number of places, our conclusion is that people in developed countries tend to visit more regular places than ones living in developing countries such as Ivory coast.

5 Place Specific Daily Travel Patterns

Previous studies have shown that there is a difference in daily travel ranges of people living in different cities/regions. Isaacman et al have shown that people living in Los Angeles's daily commute distance is nearly two times larger than people living in Newyork [12]. We were interested in finding out if this pattern holds in case of a developing country such as Ivory coast too.

In the given dataset, there is no "home" (city) location specified for users. As Ivory coast is divided into 255 different subprefectures, we decided to assign a home subprefecture for all the 50,000 users. We have used the process described in Figure 8a to compute home location (subprefecture) for every user. Step 1 & 2 are already described in earlier sections. Once, we calculate regular places visited by a user. A *home probability value* is calculated for each regular place, which is defined as follows:

Home Probability (U_k, P_i) = *Number of days on which Cell IDs associated with* P_i *were seen in night/Total number of days on which Cell IDs associated with* P_i *occurs in mobility pattern of user* U_k

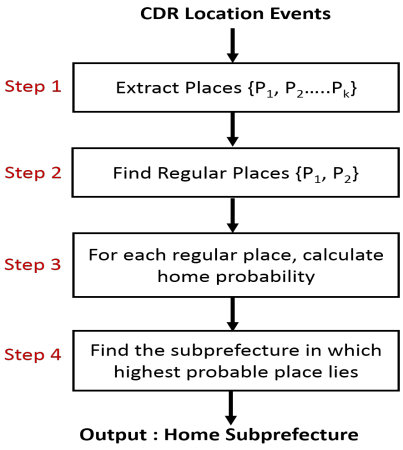
Home probability value works on assumption that user is more likely to spend most of the night time at her home. However, this assumption may not be true for all the users but we expect that it will be true for majority of users. After calculating *home probability value* for all regular places, they are ranked in decreasing order of home probability value and top most regular place is considered to be the home place. Subsequently in step 4, we found the subprefecture in which this place lies and call it as home subprefecture.

We have applied the process given in Figure 8a to all the $50,000$ users and founds home subprefecture for all of them. Figure 8b give a visualization of assigned home subprefectures of all the users, some of subprefectures were home location for high number of users i.e. one subprefecture was assigned to 11826 unique users. There were about six subprefectures which were assigned as home location for more than 1000 users. We have repeated this same experiment for other time periods too and our algorithm was able to successfully assign home subprefectures to all users and distribution of users among subprefectures also remained the same.

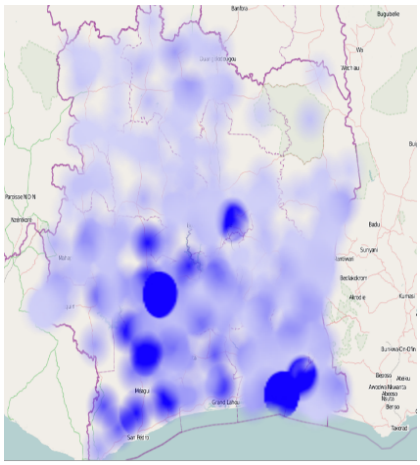
We have picked top three subprefectures having users 11826 (subprefecture ID : 60), 7153 (subprefecture ID : 50), 2257 (subprefecture ID : 59) subsequently and compare their aggregate daily travel ranges. Table 4 compares median daily ranges of users living in three different subprefectures for weekdays. In their regular travel on weekdays, users in subprefecture 59 does not travel much as compared to subprefecture 50 in which large number users do long distance travel regularly. Similarly on weekends too, a significant number users living in subprefecture 50 prefers to travel as compared to other subprefecture where less people do travel on weekends Figure 9b. In all the subprefectures, nearly 25% of users do not travel any distance on regular week days.

5.1 Comparison with Existing Studies

In case of US study [12], researchers have found that people in Los Angeles tend to travel on regular basis and distance is nearly twice as much of New York. In Ivory coast data too, we have found that people who live in subprefectures 50 tend to travel on regular days which is significantly higher than other two subprefectures. Overall, our findings here are consistent with the fact that

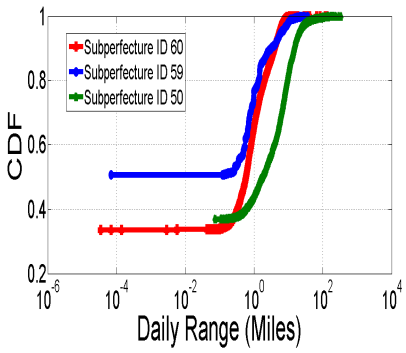


(a) A step by step process to assign home subprefectures for users

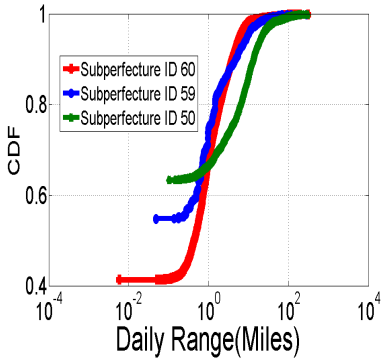


(b) Heatmap of users in different subprefectures

Figure 8



(a) Weekdays median daily ranges for users living in different subprefectures, X-axis is using log scale



(b) Weekdays median daily ranges for users living in different subprefectures, X-axis is using log scale

Figure 9

Percentile	Median Daily Range (Sub ID : 60)	Median Daily Range (Sub ID : 50)	Median Daily Range (Sub ID : 59
Min	0	0	0
2	0	0	0
25	0	0	0
50	0.58	1.79	0.00
75	1.50	7.76	0.96
98	6.80	34.31	9.84
Max	130.09	329.00	35.82

Table 3: Median weekdays daily ranges for users living in different subprefectures. Users living in Subprefecture ID 50 travel long in their daily travel

Percentile	Max Daily Range (Sub ID : 60)	Max Daily Range (Sub ID : 50)	Max Daily Range (Sub ID : 59
Min	0	0	0
2	0	0	0
25	0.81	6.58	0.64
50	2.71	16.06	2.35
75	6.40	36.14	10.36
98	132.51	208.32	69.68
Max	373.94	397.91	339.19

Table 4: Maximum daily ranges on weekdays for users living in different subprefectures. Many users living in Subprefecture ID 50 travel long distance which is several order magnitude higher than other two subprefectures

people living in different areas may have different aggregated mobililty patterns.

6 Discussion

In this paper, we present an extensive analysis of aggregated mobility pattern of mobile users in Ivory coast using CDRs data of $50,000$ users with nearly 20 weeks of data. Our analysis have resulted in several interesting insights which were not seen earlier in the studies done in case of developed countries. We believe that once these insights can be combined with other kind of data (such as economic status of a subprefecture), they can result in useful patterns. As a future work, we want to analyze the people's movement patten across different subprefectures which can be directly used in domains such as transportation.

References

[1] Barabi A., Understanding individual human mobility patterns," Nature 453, 779-782.

[2] Wesolowski, Amy, Nathan Eagle, Andrew J. Tatem, David L. Smith, Abdisalan M. Noor, Robert W. Snow, and Caroline O. Buckee. "Quantifying the impact of human mobility on malaria." Science 338, no. 6104 (2012): 267-270.

[3] Yadav K., Naik V., Singh A., Singh P.,Kumaraguru P., and Chandra U.,Challenges and novelties while using mobile phones as ICT devices for Indian masses: short paper, NSDR'10.

[4] Vu L., Do Q., and Nahrstedt K., Jyotish: Constructive approach for context predictions of people movement from joint Wifi/Bluetooth trace, PerCom 2011.

[5] McNamara L., Mascolo C., and Capra L., Media sharing based on colocation prediction in urban transport, MobiCom'08.

[6] Burbey I.E., Predicting Future Locations and Arrival Times of Individuals. Doctoral Thesis, Blacksburg, Virginia, April 2011.

[7] Scellato, Salvatore, et al., Nextplace: A spatio-temporal prediction framework for pervasive systems. Pervasive Computing (2011): 152-169.

[8] Petzold, Jan, et al. "Comparison of different methods for next location prediction." Euro-Par 2006 Parallel Processing (2006): 909-918.

[9] Y. Chon, H. Shin, E. Talipov, H. Cha, Evaluating Mobility Models for Temporal Prediction with High-Granularity Mobility Data, in Proceeding of 10th IEEE International Conference on Pervasive Computing and Communications (PerCom12), 2012,

[10] Y. Chon, E. Talipov, H. Cha, Autonomous Management of Everyday Places for Personalized Location Provider, IEEE Transactions on Systems, Man, and Cybernetics, Part C: Application and Reviews (SMCC), to be appeared, 2011.

[11] Y. Chon, H. Cha, LifeMap: A Smartphone-based Context Provider for Location-based Services, IEEE Pervasive Computing, vol. 10, no. 2, pp. 58-67, April-June. 2011.

[12] Isaacman, Sibren, Richard Becker, Ramn Cceres, Stephen Kobourov, James Rowland, and Alexander Varshavsky. "A tale of two cities." In Proceedings of the Eleventh Workshop on Mobile Computing Systems & Applications, pp. 19-24. ACM, 2010.

[13] Noulas, Anastasios, Salvatore Scellato, Renaud Lambiotte, Massimiliano Pontil, and Cecilia Mascolo. "A tale of many cities: universal patterns in human urban mobility." PloS one 7, no. 5 (2012).

[14] Isaacman, Sibren, Richard Becker, Ramn Cceres, Stephen Kobourov, Margaret Martonosi, James Rowland, and Alexander Varshavsky. "Ranges of human mobility in los angeles and new york." In Pervasive Computing and Communications Workshops (PERCOM Workshops), 2011 IEEE International Conference on, pp. 88-93. IEEE, 2011.

[15] Isaacman, Sibren, Richard Becker, Ramn Cceres, Margaret Martonosi, James Rowland, Alexander Varshavsky, and Walter Willinger. "Human mobility modeling at metropolitan scales." In Proceedings of the 10th international conference on Mobile systems, applications, and services, pp. 239-252. ACM, 2012.

[16] Isaacman, Sibren, Richard Becker, Ramn Cceres, Stephen Kobourov, Margaret Martonosi, James Rowland, and Alexander Varshavsky. "Identifying important places in peoples lives from cellular network data." Pervasive Computing (2011): 133-151.

[17] Becker, Richard, Ramn Cceres, Karrie Hanson, Sibren Isaacman, Ji Meng Loh, Margaret Martonosi, James Rowland, Simon Urbanek, Alexander Varshavsky, and Chris Volinsky. "Human mobility characterization from cellular network data." Communications of the ACM 56, no. 1 (2013): 74-82.

[18] Frias-Martinez, Vanessa, Cristina Soguero-Ruiz, Malvina Josephidou, and Enrique Frias-Martinez. "Forecasting Socioeconomic Trends With Cell Phone Records." (2013).

[19] Vieira, Marcos R., Vanessa Frias-Martinez, Nuria Oliver, and Enrique Frias-Martinez. "Characterizing dense urban areas from mobile phone-call data: Discovery and social dynamics." In Social Computing (SocialCom), 2010 IEEE Second International Conference on, pp. 241-248. IEEE, 2010.

[20] Bayir, M.A.; Demirbas, M.; Eagle, N.; , "Discovering spatiotemporal mobility profiles of cellphone users," World of Wireless, Mobile and Multimedia Networks & Workshops, 2009. WoWMoM 2009. IEEE International Symposium on a , vol., no., pp.1-9, 15-19 June 2009.

[21] Vu L., Do Q., and Nahrstedt K., Jyotish: Constructive approach for context predictions of people movement from joint Wifi/Bluetooth trace, PerCom 2011.

D4D Challenge
Commuting Dynamics 4 Change

R. Maestre[1], R. Lario[1], M. Muñoz[1], R. Abad[1], J. Gonzalez[1], A. Martín[1], E. Perez[2], and JL. Fdez-Pacheco[3]

[1]Paradigma Labs (PL) , Paradigma Tecnológico (PT) ,
rmaestre@paradigmatecnologico.com
[2]GIS Laboratory , Centre for Humanities and Social Sciences (CCHS) , Spanish National Research Council (CSIC) , esther.perez@cchs.csic.es
[3]Complutense University (UCM) , jlfpsaez@trs.ucm.es

15-Feb-2013

Abstract

Our idea is to use the geolocation data from the antennas processing the mobile phone calls in order to know which sub-prefectures the customers have been getting around. The main goal of our project is developing spatio-temporal models to detect commuting patterns for the different sub-prefectures, including some other factors related to the region and/or time: wealth, development, infrastructure, investment, grants, etc. By means of GIS technology, we will be able to apply our generated models to the gathered data and to analyze their correlations over the Ivory Coast surface, working with geographical layers: landcover, roads map, railway lines, water sources, etc. Consequently, the reached conclusions from our study will be properly visualized, allowing a better explanation of the findings. With a bigger amount of data gathered for a longer period, more interesting and accurate trends could be discovered, allowing us to calculate associated coefficients.

Our analysis models will provide coherent data to support a correct urban design and will mean a monitoring tool for development, specially related to population dynamics. In the near future, some other measures could be included. For instance, hospitals and police stations locations, their calls rate... Thus, we could know its real use, being able to improve their service to the citizens: dangerous areas, crowded hospitals, etc.

Keywords: Human Dynamics, Commuting, GIS, Data Analysis, Mobile Communications.

Contents

List of Figures

Commuted Communities: Commuting Dynamics 4 a Change

For decades, Africa has been receiving cooperation from the so-called countries of the "North". There have been several models of cooperation with this continent, from bi-lateral country agreements passing through religious and non-governmental organizations to new models of financing development. None of them is free from dark sides and controversial issues in this "culture of aid" (Moyo, 2009). The African continent doesn't need that kind of helping agreements which submerge the continent in a dark hole and position it on unequal relations with International Financial institutions which pretending good wishes create difficult conditions that keep on enslaving the continent on new ways of Neo-colonialism. African countries are in need of endogenous conditions for creating their path to democracy and their own government empowering. We understand development as a process linked to the capacity of African people to decide about its own future(SEN et al., 2000). In this sense, our proposed ways of cooperation pass through the vision of an horizontal structure where sharing-knowledge conditions are created in order to allow the empowering in every direction. As the one we introducing on this article, our proposals are based on the new technologies and the way they can help to develop new aspects as far as planning and researching are concerned in any country they are implemented.

The emergence of new technologies and virtual tools in the field of development cooperation had an early stage in the late nineties(De Jong et al., 2005). In that period, the projects developed were aimed at providing hardware structures and networks to disadvantaged communities or with communication difficulties due to the orography within their territory. Nevertheless, our proposals on this article are based on software tools that have been already experienced in countries such as Ivory Coast where the development conditions are allowing to start testing some of these tools.

GIS based technologies have been sufficiently proved in some sustainable development projects carried out in Africa[1] related to international cooperation as well as in local development where the cross-sectorial work is critical and the shortage of funding make it, sometimes, impossible (Mitchell, 1997), (Craig et al., 2002). Recently the GIS has a crucial role in the democratic processes in Africa, as we can see in the creation of the Census 2012 in Rwanda where the GPS data collection will enable them to build a comprehensive GIS database, which shows boundaries, location of schools, hospitals and markets covering 17,700 villages[2].

Also in South Africa, part of our team had the experience of working with GIS as a tool to identify agricultural and rural areas where to start community vegetable gardens as well as managing the cattle in some projects implemented in KwaZulu-Natal region.

As we said, the two tech experiences gathered on this article look deeply their contribution to serve as tools to facilitate the cooperation in two aspects:

a) It will allow to develop better ways to plan and manage the territory since their GIS based structure will permit us to work with geographical layers: landcover, roads map, railways lines, water sources, etc. Consequently, this information will allow to visualize and reach some conclusions in order to design and formulate better cooperation and local development projects based on the municipalities.

b) This tool will become an useful resource in social research(Fielding et al., 2008) on deterritorialized and transnational realities(**?**) since it brings new possibilities of comparative analysis and the creation of resources ad hoc.

[1] http://www.esri.com/library/brochures/pdfs/gis-for-africa.pdf
[2] http://www.humanipo.com/news/1279/GIS-based-enumeration-kicks-off-in-Rwandas-2012-census

About us and why we face this challenge

Paradigma Labs[3] (PL) core values are conducted by one motivation: "To figure out the fuzzy dynamics between Humanity and Technology", providing tools and methods to study, display and understand these dynamics. Therefore, an international challenge whose research subject can be chosen freely as long as it relates to an objective of development and improving quality of life for people, quickly held our attention.

GIS laboratory[4] at Centre for Humanities and Social Sciences (CCHS), Spanish National Research Council[5] (CSIC) is a multidisciplinary group with a huge experience in Remote Sensing and Geoprocessing, providing a quality support for plenty of research projects carried out at CSIC.

Our final aim has been detecting geospatio-temporal patterns in order to obtain useful knowledge to better manage the country resources. For example, if we could predict the traffic intensity segmented by road, week day and hour, then other secondary roads could be suggested or the budget for the most used ones could be increased. Through mobile communications, a specific user can be tracked along the day, not only by means of 'call comunications' but also thanks to applications running on their handsets: IMS, RSS... The dataset provided by Orange is a sample, and it only uses 'call communications', however, with the whole set of data (i.e.: app and call comunications), we strongly believe more accurate and complete models could be discovered helping to identify new kinds of dynamics.

From our own experience studying and modeling several kinds of Human Dynamics such as the ESF project DynCoopNet(Solana and Alonso, 2012) and while developing a Business Intelligence Tracking Tool on Twitter (Marin et al., 2012), we can claim there are two main exploring perspectives: the Geographical one and Temporal one. We believe a mathematical model related to Human Dynamics must be managed with these two viewpoints. The Temporal component is useful by providing a tool to go backward and forward in order to get a more detailed understanding of the dynamics, not only moving across the timeline, but creating temporal windows to group events. The Geographical component provides a more high-level understanding related to the human mobility across the space in different levels and relating it to some other spatial features. Mixing both components in a final and single visualization has led our study during the project.

Consequently, in this paper, we propose a Geospatio-Temporal Model. A Geospatial Model, because user interactions with geolocated antennas are analyzed and treated, and a Temporal Model since several time windows are used to group these user dynamics. The combination of these two variables is used and displayed by a GIS. Initially, several results are showed supporting the project main conclusions. However, what's really important is the whole process for handling the data, that is, the code, tools and methodology, which will be available to the researcher community, allowing to study more deeply the dynamics. For instance, a Standard Kernel Density estimation (KDE) aims to produce a smooth density surface of spatial point events over a 2-D geographic space(Bithell, 1990; Alegria et al., 2011), final dynamics visualization across the several days of the week will be shown by means of KDE, in order to understand and prove which and where the maximum commuting peaks are.

We have focused on the Commuting concept, which could be defined as follows: *Commuting* is regular travel between one's place of residence and place of work or full-time study (Wikipedia, 2012), but sometimes it refers to *any regular or often repeated traveling between locations when not work related*. Our first commuting approach is defined like: "Mobility patterns through inferring dynamic users

[3] http://labs.paradigmatecnologico.com/
[4] http://humanidades.cchs.csic.es/cchs/sig/
[5] http://www.csic.es/

movements grouped by temporal windows".

A *commuter* or *dynamic user* is defined as an user changing their antenna location within the studied temporal window (i.e.: each temporal window groups the whole user communication for a specific hour). Among these temporal windows, *non-commuters* or *static users* have been removed, i.e.: users who do not change their antennas locations within the temporal range. The justification to remove these users comes to focus our study on users that are moving into this temporal windows and perform micro-displacements. It is common that the same user performs these two kind of dynamics within the same temporal window. Note that we are not quantifying the distance, but only the fact of changing from a particular antenna to another one.

State of the art

Nowadays, the world has nearly as many cell phone subscriptions as inhabitants[6]. For the first time, the majority of humanity is linked and has a voice. Consequently, plenty of phone communications are being generated continuously everywhere, and, what is more relevant, they are being tracked: geolocation, start/end times... This is the key, mobile phone companies record data which are very closely associated with behaviour of people.

Analyzing these data in a proper way discloses a great deal of social knowledge (behaviour modeling, people mobility patterns, trends and outliers) which can be applied in countless and different areas[7]: transportation, urban planning, commuting, tourism, traffic congestion, demography, sociology, economy, advertising and commerce, public health... Even without Internet connections (e-mail, IMS and so on), that is, focusing only on speech-calls and text messages, there is a vast amount of information which can be 'read' to reach further conclusions. The ability to understand the patterns of human life by analyzing the digital traces that we leave behind will transform the world, specially poor nations. Reality mining of behaviour data is just the beginning.

Let's describe a really interesting project (Eagle et al., 2009) about behavioural data. By collecting communication traces into an organization and studying the underlying patterns, some key outcomes of interest are revealed: social network structure, inference of friendship and proximity levels, individual satisfaction... This is achieved with temporal data such as call logs, location, phone status, near bluetooth devices, cell antenna ids, application usage(e-mail). Comparing these behavioural data with traditional self-report data show important conclusions.

Regarding D4D datasets (there are only 4 and they contain really simple data), note how they have caused many and varied studies from all teams. As far as we are concerned, we discussed several ideas: antennas network optimization in traffic terms, geospatial-temporal detection of real use for public services (hospitals, schools, police stations...), commuting patterns detection and alike.

Precisely, it has been the human urban mobility approach which we chose as the core of our project. It is so because it is a reality very tied to ordinary people daily lives, so that its study can reveal clues to improve quality of life.

Below, we can see a few current research projects showing how identified commuting patterns are really useful to understand human motion dynamics better and to perform accurate plans and actions:

a) Exploring spatio-temporal commuting patterns in a Moscow university environment allows making more appropriate decisions to decrease the automobile dependence of students, promoting the non-motorized and public transportation. It is a green initiative looking for sustainability: reducing pollution and noise, avoiding congestion, improving public health and urban planning...

b) Classifying different urban areas based on their mobility patterns from mobile phone data. The results can be used to better understand these dynamics allowing more efficient environmental and transportation policies for the time being and for the future (since due to the regularity of the individual trajectories, it can be claimed that human mobility is highly predict).

c) Time patterns and geospatial clustering based on mobile phone network data provide accurate statistics about mobility of people, population density and economic activity with detailed regional

[6] http://www.huffingtonpost.com/2012/10/11/cell-phones-world-subscribers-six-billion_n_1957173.html
[7] http://www.insead.edu/v1/gitr/wef/main/fullreport/files/Chap1/1.6.pdf

and time resolution.

d) Visual analytics system to study people's mobility patterns from mobile phone data. This tool allows to deeply analyze where, when and who from the calls of people, allowing different kinds of aggregations.

As it can be seen, communication data are everywhere (*we are social animals*) and they can be used to obtain really interesting and high-value findings. Imagine, once we know the nature and meaning of these data, it is as if we had access to a lot of complete, reliable and immediate surveys. Honestly, we strongly believe that the future lies in knowing how to process this kind of data to get unique results. MIT's Technology Review has recently identified **reality mining on mobile communications** as one of '10 Emerging Technologies That Will Change the World'.

Several studies have proven the utility of Geographic Information Systems (GIS) for commuting analysis (flows intensities and directions), because of its efficiency dealing with data with a geographic component. This is the case of people displacement patterns on a particular portion of space. That, combined with its ability to represent data over the territory through geovisualization techniques, makes GIS one of the most used tools for this kind of studies.
The conclusion of all of those studies points to the valuable information that can be extracted from commuting, in terms of city growth, decisions about cities and companies placement or people disposition to migrate or dealing with long journeys to their job or leisure places.

There are many applications and use cases in the bibliography, including: the analysis of work commuting in a city, in terms of time and distance to locate spatial variations (Wang, 2000), the analysis of people behaviour patterns related to their leisure activities, the study of georeferenced commuting patterns to elaborate models that predict workplace contacts which result in disease transmission (Chrest and Wheaton, 2009), the study of dispersión trends or concentrations in specific studied areas, the study of average distances spent on displacements as well as spent times relating them with companies and city location patterns, the study of the differences between mobility patterns of national and immigrant employees (Llano Verduras, 2006), the evaluation of optimal routes (Thériault et al., 1999) the analysis of costs and transport problems in intraurban and interuban structures (Zhan et al., 2008), etc.

Problem description & Hypothesis

As we saw in the state-of-the-art section, we can extract knowledge from mobile communication datasets. Thus, the solution proposed in this paper is based upon the hypothesis of mobility patterns to predict common and well-known, geographical and time-based models to manage roads and infrastructures in a correlated way with the results figured out.

The figure below shows a theoretical commuting model proposed as a main pattern.
The model shows two peaks, p_1 in the range 07:00-08:00 and p_2 in 17:00-18:00. This first approach to modelize these dynamics sets up the same height for both peaks p_1, p_2. However, numerical results will show that the height of each peak depends on the day of the week. A central valley is defined between 09:00-16:00, with an uniform displacement distribution.
An ad-hoc mathematical model is defined in the next section in order to confirm this hypothesis, consisting on focusing, filtering and processing the main data to contrast the assumption.

The main idea behind the two main peaks, p_1, p_2, and the central valley in the model, is that people cover larger distances in their displacements early in the morning i.e.: p_1 related with the common business activity. After the first peak p_1, people stay in these target destinations, working, eating, etc ..., but in a more static point of view and always performing displacements. The last point in the hypothesis approach shown in Figure 1 is the second peak p_2, when people return to their destination or the last business activites took place.

The geographical behaviour of the proposed dynamics, always mixed with the temporal component, will by contrasted using GIS tools in order to visualize the expansions and contractions in the main points shown by the hyphotesis: p_1, p_2 and the central valley: an expansion when the maximum displacement is reached on the first peak p_1, and a partial contraction when the central valley is reached. Another displacement expansion when peak p_2 is reached and its corresponding contraction around p_2, when it is declining.

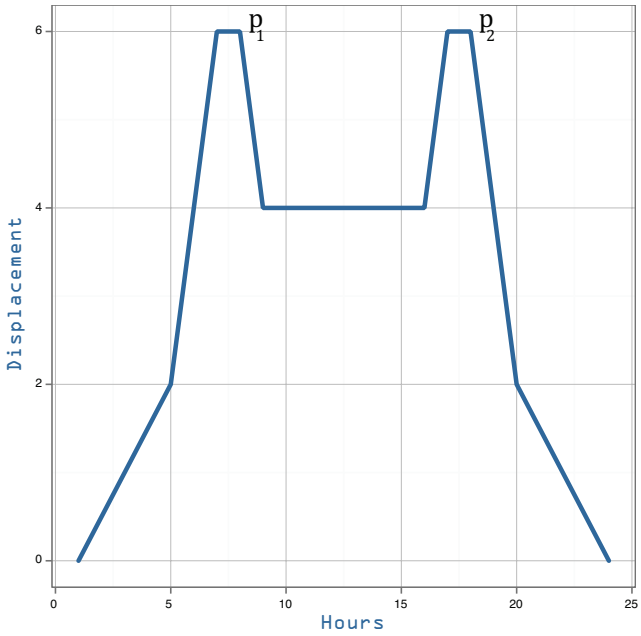


Figure 1: Theoretical Commuting Model

Mathematical model

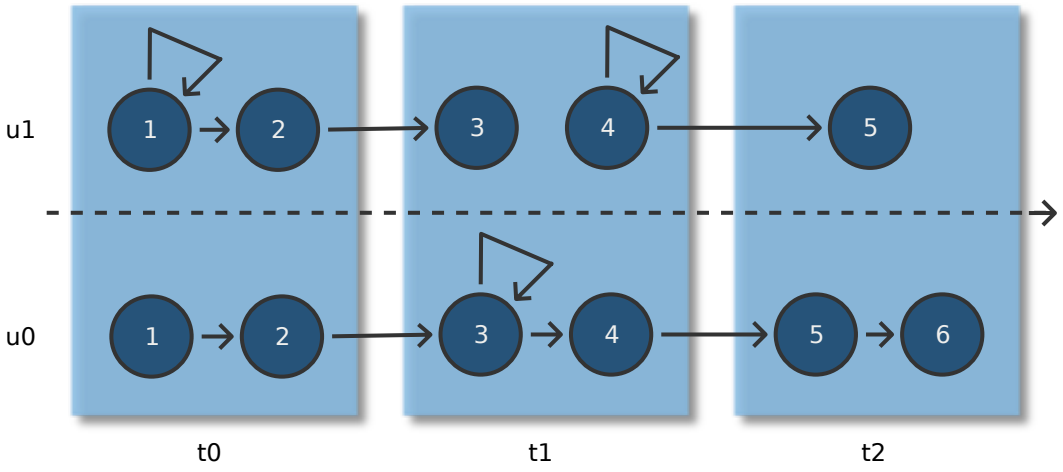


Figure 2: Dynamic and static user patterns

The figure above shows two commuters u_0, u_1 represented on the vertical axis, grouped by three time windows t_o, t_1, t_2. A time window is defined as t_n, where $n \in \{0, .., 24\}$. Each t_n groups the communication traces of the whole set of commuters in a 60-minute lapse.

Let $\vec{P_i} = (p_{ix}, p_{iy})$ be the position vector where p_{ix} is related with the geographic latitud and p_{iy} is related with the geographic longitud of each position.

Formally, a commuter trace during a particular time window and related with a specific user is defined as follows:

$$T = (\vec{p_0}, \vec{p_1}, ..., \vec{p_n})$$

where, $\vec{p_n} \in \mathbb{R}^2$ and $n > 1$.

For instance, regarding with user 0 and time window 0 we have the next user trace $T = (\vec{p_1}, \vec{p_2})$. Regarding with user 1 and time window 0 we have the next user trace $T = (\vec{p_1}, \vec{p_1}, \vec{p_2})$ and so on.

Also, two functions are defined in order to measure the distance (Cook, 2012), given a set of points expressed as spherical coordinates:

$$D(\vec{p_0}, \vec{p_1}) = \text{acos}(\sin(\phi(p_{0x})) * \sin(\phi(p_{1x})) * \cos(\theta(p_{0y}) - \theta(p_{1y})) + \cos(\phi(p_{0x})) * \cos(\phi(p_{1x})))$$

where:

$$\phi(x) = (90 - x) * \frac{\pi}{180}$$

$$\theta(x) = x * \frac{\pi}{180}$$

therefore the function related to the distance and regarding to a specific user into a temporal window is defined as follows [result in Km]:

$$U = 6373 * \sum_{i=0}^{n-1} D(\vec{p}_i, \vec{p}_{i+1})$$

The second function is related to the number of antenna connections into a trace. The key point is to count only the dynamic transitions, i.e.: remove the self edges over a given trace as follows:

$$\mathrm{S}(p_0, p_1) = \begin{cases} 0 & \text{if } D(p_0, p_1) = 0 \\ 1 & \text{if } D(p_0, p_1) > 0 \end{cases}$$

therefore, the function regarding a specific user into a temporal window is defined as follows:

$$N = \sum_{i=0}^{n-1} S(\vec{p}_i, \vec{p}_{i+1})$$

Methodology

Let's describe how we have faced D4D, enumerating the different phases of our project and highlighting the corresponding milestones. We really think that explaining how this work was carried out can be useful both to better illustrate our conclusions and results, and to give ideas for similar projects.

First of all, once we had clearly understood the D4D bases, we studied all provided datasets to be certain of what kind of data were available. Next, we began with the research work: getting information of Ivory Coast, studying some papers about behavioural patterns obtained from mobile phones traces, looking for new related datasets...

With this knowledge, we were prepared to decide which lines of work would be more interesting (without forgetting the cooperative and development goal apart from the scientific one) and, what's more important, being aware of our own time constraints and our team skills –being realistic is crucial.

After some discussion, we agreed to focus on the 2nd dataset 'Individual Trajectories: High Spatial Resolution Data (SET2)' (Blondel et al., 2012) , since it seemed to be the most adequate one for our approach. We conducted our analysis according to the following stages:

1) Processing all traces, grouping them by user and sorting them chronologically, as hourly time series. Paying attention to imprecise or weird traces, which must be filtered.

2) Calculating different magnitudes (absolute, relative and normalized ones) and their mean, median and dispersion to display visual charts, which helped us to discover correlations and to identify 'Temporal Commuting Patterns' for each week day.

3) Handling antenna locations from the previous processed traces, allowed us to identify 'Geospatial Commuting Patterns'. Firstly, we represent networks graphs and some static maps (snapshots of commuters motion). Later, we were able to create animated and detailed maps (Kernel Density, Grids...) which made easier to see crowded areas, related highways... during the days and all across Ivory Coast.

4) Eventually, an online and interactive web-based animation was developed. This geovisualization technique is advantageous in that neither specialized GIS knowledge nor software is required, and it enables change over time visualization that would be difficult to see with static or paper maps. The interface combines raster maps produced in the ArcGIS environment and vector data [PANTALLAZO de la APP final en web]. User interaction is facilitated through the inclusion of buttons on the interface (play controls, modal tab, zooming and panning).

As can be seen, the whole process to obtain the results has been carried out step by step. We had a planning which was useful, but the really important thing was the fact of planning, not the planning itself. There will be unexpected events which required the team to adapt itself to new circumstances.

In the end, we would like to remark how, although assigning particular tasks to different team members looking for productivity, all of us have tried to be involved in all areas.

Results

Here below a brief enumeration of the achieved results during this project.

a) A designed and implemented mathematical model to detect geospatial-temporal commuting patterns.

b) A set of charts and maps which illustrate the previous model, making easier to deduce interesting findings.

c) An on-line application[8] to display all this information in a friendly and customizable way.

Now, let's describe deeply them, as well as other partial findings.

According to the proposed model, a very important feature has been deduced for the Commuting Dynamic. As seen in the picture below, there are two time zones when people perform more phone calls from their handsets than usual. Static and dynamic users have not been distinguished, that is, both self-edges and transition edges are counted together.

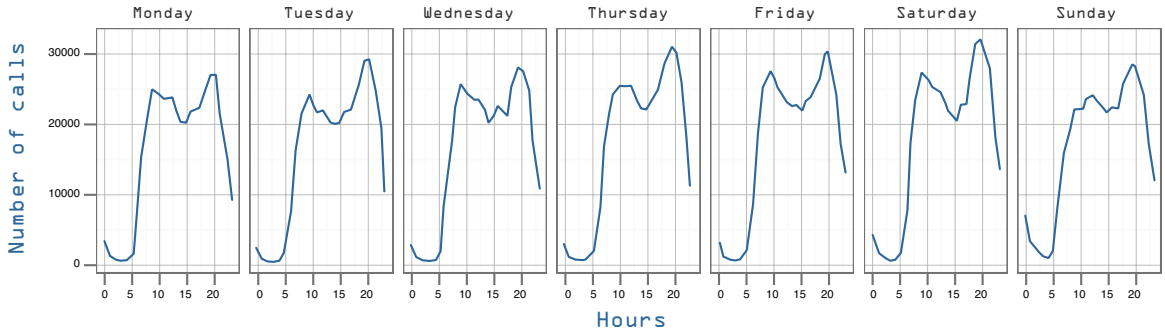


Figure 3: Total amount of calls grouped by Week days

Let's extract some ideas from the previous chart, focusing on what commuters could be probably doing (as it will be seen shortly, most of the phone calls belong to 'dynamic users'): they get up early in the morning and start to perform more and more phone calls from their handsets until $p1$; next, the amount of phone calls decreases slightly, keeping itself more or less balanced (workhours, lunch) until the beginning of $p2$; later, people leave the office and plan the rest of the day (errands, leisure...), what is reflected on a marked rise in the amount of phone calls.

Since it is really complicated to measure the displacements of the people (antenna locations instead of users locations, missing antena identifiers...), what will be analyzed is the amount of callers who are really commuters, that is, dynamic users. This one will be the essential magnitude of our research, leading us to figure out the Commuting Dynamics key-features for different regions and times. Eventually, all conclusions deduced from the G/T Model (charts, formulae...) will be ultimately tested with the final GIS visualization as the main core of the work. In this last phase, geographical displacements are estimated so that they can be plotted in a dynamic and interactive map which makes easier to detect peaks and trends.

With this new chart, it is the ratio between commuters and all users what it is being emphasized.

[8] http://labs.paradigmatecnologico.com/d4d/

During a particular day (24h), there is no need to know if a concrete displacement is or not longer than other. What it is being highlighted here is the movement detection itself, in 1h windows which group all users calling within them. The colour of the chart shows how many users are really commuters.

Datasets have been processed according to the proposed G/T Model, filtering non-commuters when commuters tracks have been calculated. Here below, seven charts display Commuting Dynamics for each week-day. As shown in the results, there is no correlation between the number of dynamic users and the maximum displacement peaks, actually, at the same time the number of dynamic users are growing, the central valley and the two maximum commuting peaks are always presented in the sample.

These seven charts below show a fitter correlation with the theoretical commuting model proposed in this paper (Figure 1), displaying two high peaks and a lower central valley. However, more qualitative data is necessary to figure out the performance of the first high peak, because it is not related to the number of dynamic users. Furthermore, a first approach can be applied to say that few dynamic users (in comparision with the mean) travel longer distances in this first peak, specially in the first uphill to the first peak. This is a common pattern figured out from the proposed model. However, it can be explained because there are people starting their travel from further distances early in the morning, since the most of the people live near their work places and travel nearer distances than the first group.

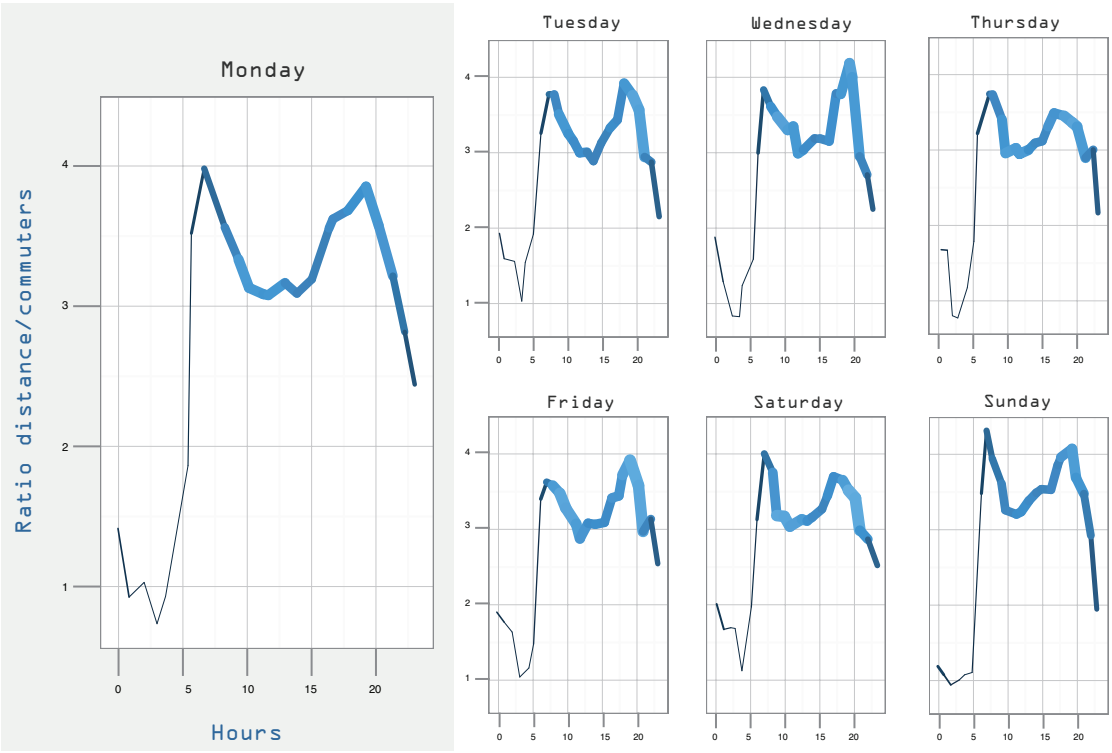


Figure 4: Dynamic users displacements

Density analysis takes known quantities of some phenomena and spreads them across the landscape based on the quantity that is measured at each location and the spatial relationship of the locations of the measured quantities. From the position of the antennas of mobile telephony and the weight estimated for each of them, depending on their traffic intensity, a kernel function has been used to calculate the density of features in a neighbourhood around those features. It calculates a magnitude per unit area from point features (antennas) using a kernel function to fit a smoothly tapered surface to each point. The result is a surface, a raster image. The original information of the antenna positions was expressed in geographical spherical coordinates (longitude, latitude), but the kernel functions needs to work with projected information to calculate the distances, therefore a previous conversion of the same ones was performed projecting them to a projection system UTM using the datum WGS84 and the zone 30N. Kernel Density calculates the density of point features around each output raster cell; the kernel function is based on the quadratic kernel function described in Silverman, 1986. Conceptually, a smoothly curved surface is fitted over each point. The surface value is highest at the location of the point and diminishes with increasing distance from the point, reaching zero at the 'search radius' distance from the point. Only a circular neighborhood is possible. The volume under the surface equals the 'population field' value for the point. The density at each output raster cell is calculated by adding the values of all the kernel surfaces where they overlay the raster cell center. The cells nearer the measured points, the antennas weight, receive higher proportions of the measured quantity than those farther away. (ESRI©, 2012).

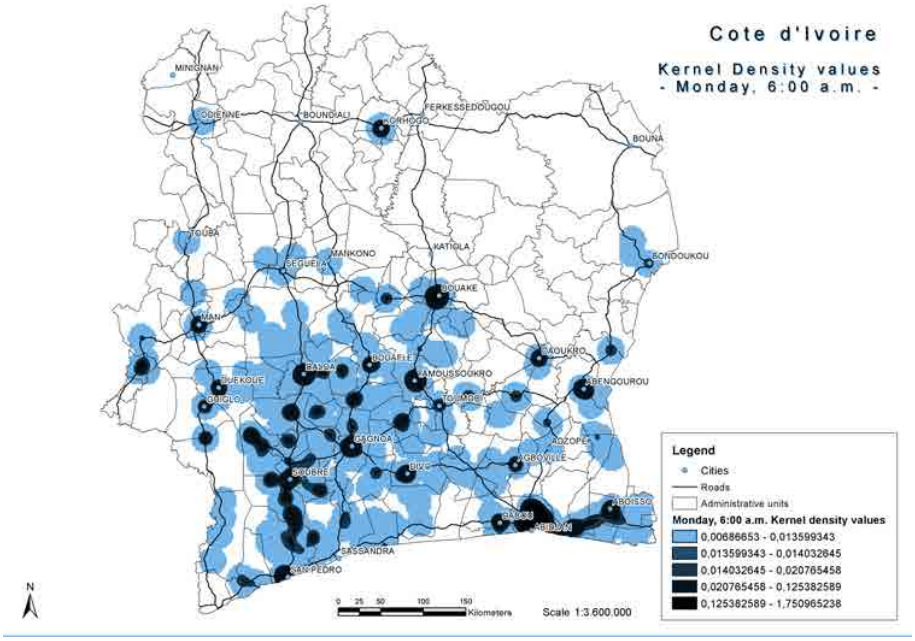
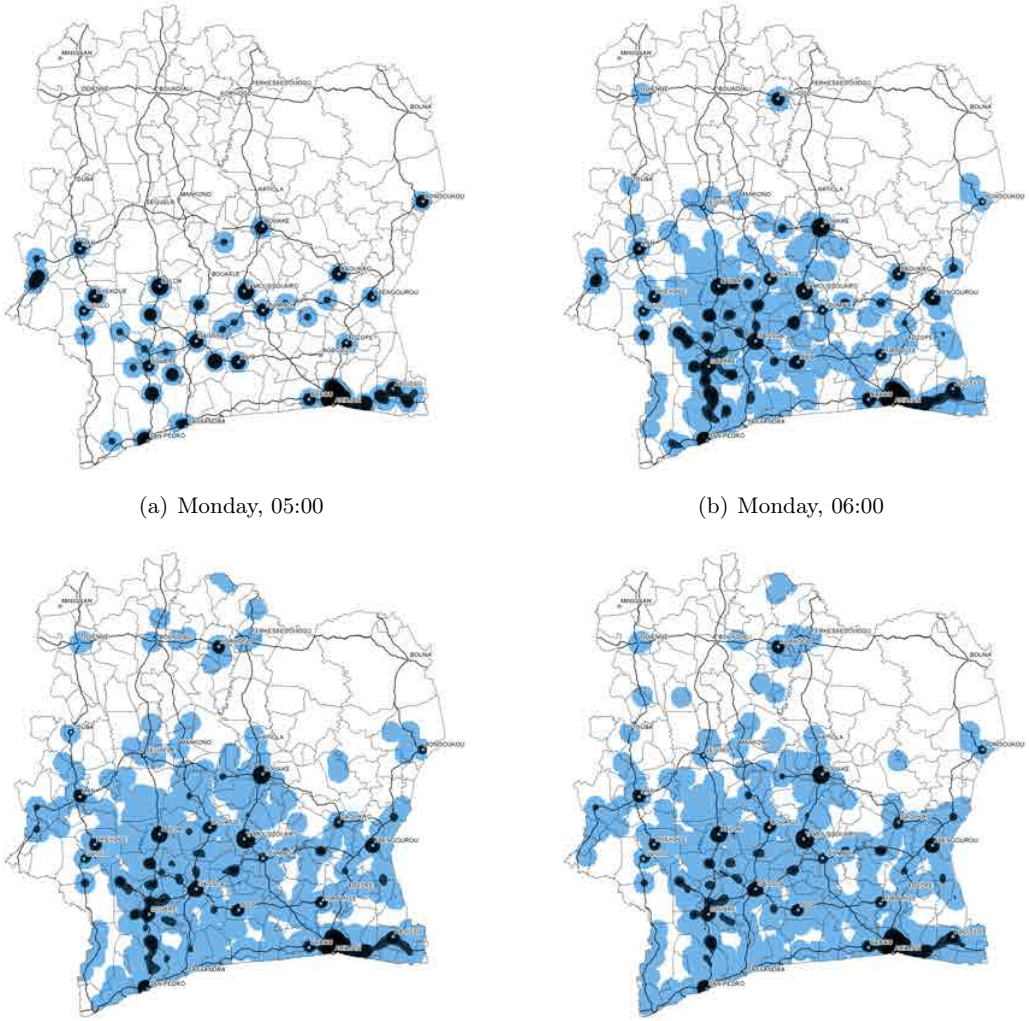


Figure 5: Kernel Density sample

A visualization of the displacement with a Kernel Density estimation using GIS with the antennas positions and users displacements like a directed graph has shown the contraction and the expansion of the commuting dynamics across the 24 hours of the day. The firsts expansion visualization through the kernel density estimation is clearly showed when peak p_1 starts to grown between 05:00 and 07:00 reaching the central valley at 8:00.



(a) Monday, 05:00

(b) Monday, 06:00

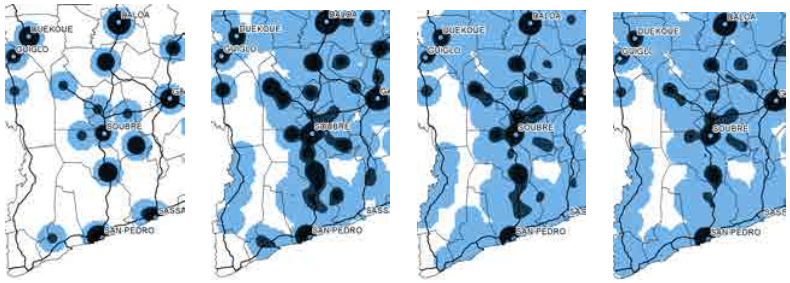
(c) Monday, 07:00

(d) Monday, 08:00

Figure 6: KDE evolution while peak p_1 is reached

Also, while a normalize displacement meassure is distributed across the country in the valley hours, a reduction of the displacement of the dynamics users can be appreciated especially on roads where KDE had a high intensity. Above picture, Monday 06:00, shows cleary a high intensity kernel distributions, coinciding with the maximum displacement moment calculated and showed in the Figure 4 on Monday. This values should be identify independently each week day commuting displacement distribution, however, the ranges of peaks p_1, p_2 and the central valley are leading for a regular temporal pattern.

Next pictures shows in a more detailed way the reduction on the KDE on the roads between main cities. We have selected a sample zone in order to show the correlation between the numeric distance calculated by means of the proposed mathematical model in Section and the KDE calculated by means of GIS technology, to test the final aim proposed in this study.

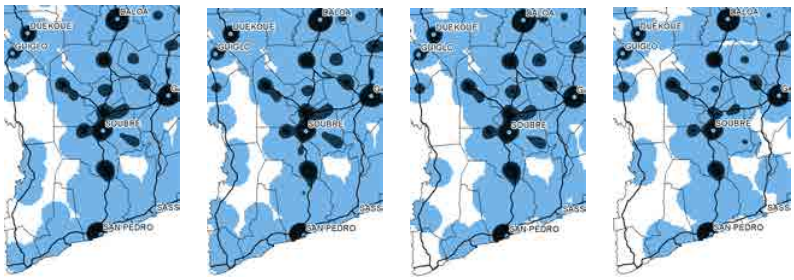


(a) Monday, 05:00 (b) Monday, 06:00 (c) Monday, 07:00 (d) Monday, 08:00

Figure 7: KDE focus on p_1

Therefore, the pictures above show the evolution of KDE in the focus location when maximum displacement p_1 is reached at 06:00, after this hour, KDE becomes softer at 07:00 and 08:00 starting the central valley. Through KDE we can figure out that the median displacement in the central valley 08:00 and go on until p_2 has a greater value as numerical results leads.

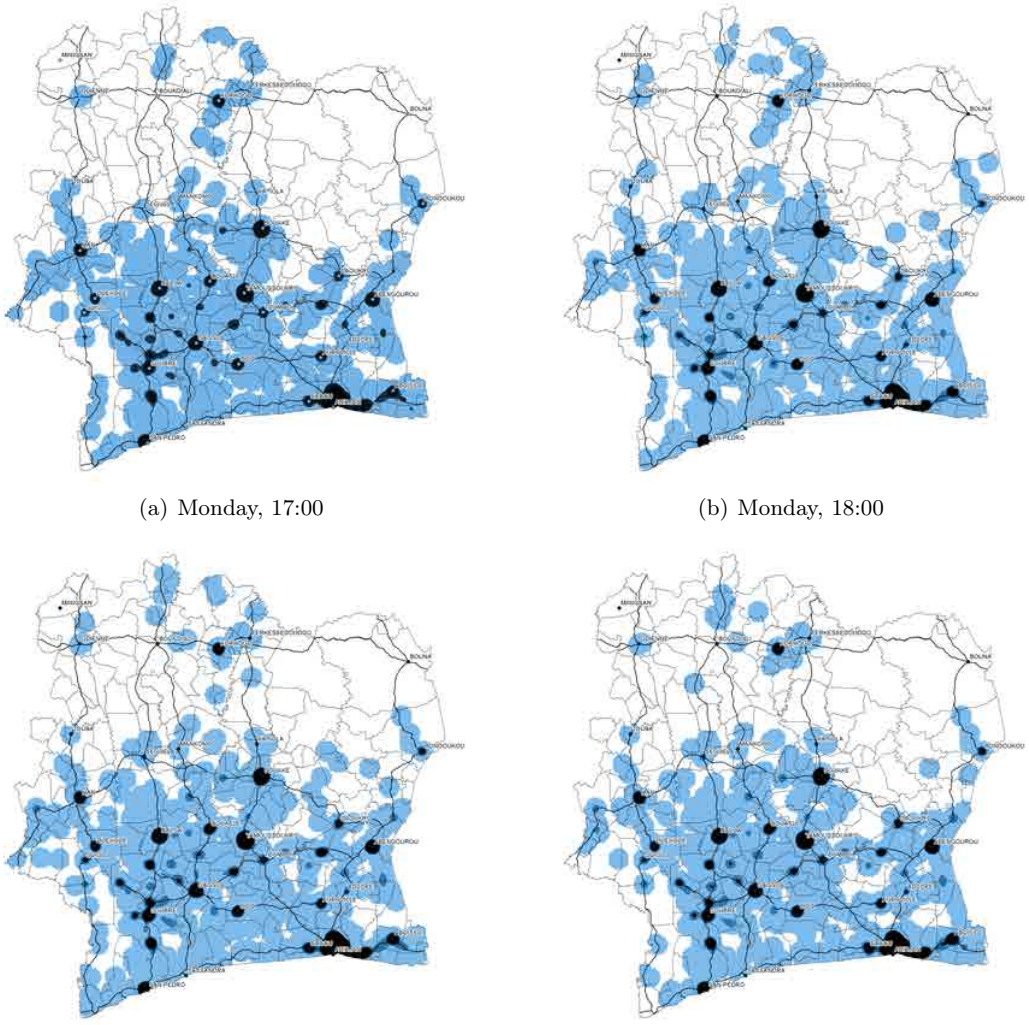
The central valley has a uniform distribution of KDE and also we can observe in the next picture how the intensity increases when peak p_2 is reached at 19:00. We propose the same location to explore and correlate the numerical results of p_2 with the final visualization of the dynamic taking into account the two variables.



(a) Monday, 18:00 (b) Monday, 19:00 (c) Monday, 20:00 (d) Monday, 21:00

Figure 8: KDE focus on p_2

In a more detailed form, the whole extension picture is provided by the nextfour pictures representing the KDE of dynamic users across the country.



(a) Monday, 17:00

(b) Monday, 18:00

(c) Monday, 19:00

(d) Monday, 20:00

Figure 9: KDE evolution while peak p_2 is reached

Conclusions

After having completed the project, we have demonstrated with a practical example how relevant conclusions can be extracted by analyzing this kind of mobile phone networks data. In particular, a couple of general ideas have been confirmed:

a) Mobile phone traces are all over the world and they hold plenty of high value information.

b) Among possible behavioral patterns extracted from a), knowing people dynamics, and specially commuting patterns, constitutes a valuable tool to improve infrastructure and public services both for the time being (detecting crowded/empty areas or periods) and for the future (predictions).

Focusing on D4D, we could remark these partial findings:

a) Distinction of *commutter* and *noncommuter* groups and their evolution during each week day and for different Ivory Coast regions.

b) Guessing of common usage of mobile phones, in amount of calls terms, during a day and in different cities.

c) Identification of a morning peak (09:30) and an evening peak (19:30), in maximum displacement terms, values have been normalized using the median of total amount of commuters.

d) Identification of lower valley between both peaks.

As the main conclusion for this project, it could be claimed that, based on the collected Orange mobile phone traces in Ivory Coast during the observed period of time, **a couple of commuting peaks have been identified for each day of the week, with a more defined pattern for work days and for big cities (Abidjan, Bouaké, Daloa, Yamoussoucro)**. Moreover, **people motion between the outskirts and the city center in the morning, and vice versa in the evening could be detailed for each particular city, highlighting those road segments with more traffic**.

Apart from the 'commuting' conclusions, there is another one which deserves to be exposed. After some discussions, we decided to create an accesible, user-friendly and customizable tool so that this kind of data could be actually profitable. How can you expect this complex process to be understood by some common people if you do not make things easy?

Summing up, we hope this global D4D effort can help Ivory Coast in decision making for policy measures and ultimately will lead to perform some trustworthy plans to improve the quality of life of the local residents. Taking advantage of such tool will save them money and time.

Recommendations for further work

This project has reached interesting results, but it can still be developed in several ways. Due to not having enough enough, some ideas were only proposed and they could not be developed. They have been listed here below as an outline for those teams who agreed with us in seeing their potential related to the model.

a) **Particularization**: general findings are useful to know how to face a problem initially, but with concrete findings more adequate solutions will be reached. Specific conclusions and maps for different cities and regions could be calculated.

b) Clustering and/or filtering antennas: by traffic (using the 1st dataset), by location (city/field, latitude), etc.

c) Replicating the model and results with the 3rd dataset.

d) A 60-minute time span and a daily displaying approach have been used. However, both assumptions could be modified in order to explore data from another perspective: season patterns, overlapped or shorter time spans...

e) Looking for more understandable charts: normalizing time series by dividing values by the maximum volume of the day, and some other strategies to obtain relative magnitudes.

f) Trying to find correlation evidence between the amount of calls rate and prosperity indicators (business, grants, etc.).

g) Looking for additional socio-economic datasets to conduct new mixed analyses (e.g. weather).

h) Developing a new kind of maps based on tessellations (i.e., Voronoi diagrams), which have been proved to be really useful for this kind of studies.

i) Applying DTW[9] or LCS[10] algorithms to discover similarities among different traces series, in order to identify types of areas (residential, commercial, business) and to predict their evolution in mobility terms.

j) Trying to identify where people live and work.

k) **Outliers detection**: both in particular time zones and in specific regions, or also regarding patterns evolution and consolidation.

Moreover, some ideas and modules of this project are expected to be used into other completely different fields, so that original and novel conclusions can be reached with a high synergy value.

Eventually, we strongly believe that working with more detailed data (both Internet and Call/SMS communications) will allow researchers to find more accurate mathematical models and therefore, more precise and useful visualizations. Furthermore, having longer time span datasets available would assert the patterns and would allow to describe more reliable people dynamics trends.

[9] http://en.wikipedia.org/wiki/Dynamic_time_warping
[10] http://en.wikipedia.org/wiki/Longest_common_subsequence_problem

GIS tools are by themselves a fantastic way to research on geolocated data. Applying different GIS algorithms and methods could obtain new dynamics patterns, providing profitable new results to be taken into account when better development policies are decided in order to improve the people quality of life.

For instance, a traffic density estimation through user dynamics antennas comunications is proposed. The next picture shows an agregated value, however, a flow intensity related with the direction of each user trace can be applied too. A correlation of segment creation from each two antenna positions and the flow intensity between antennas taking into account the direcction, could be directly applied to a real time traffic manager.

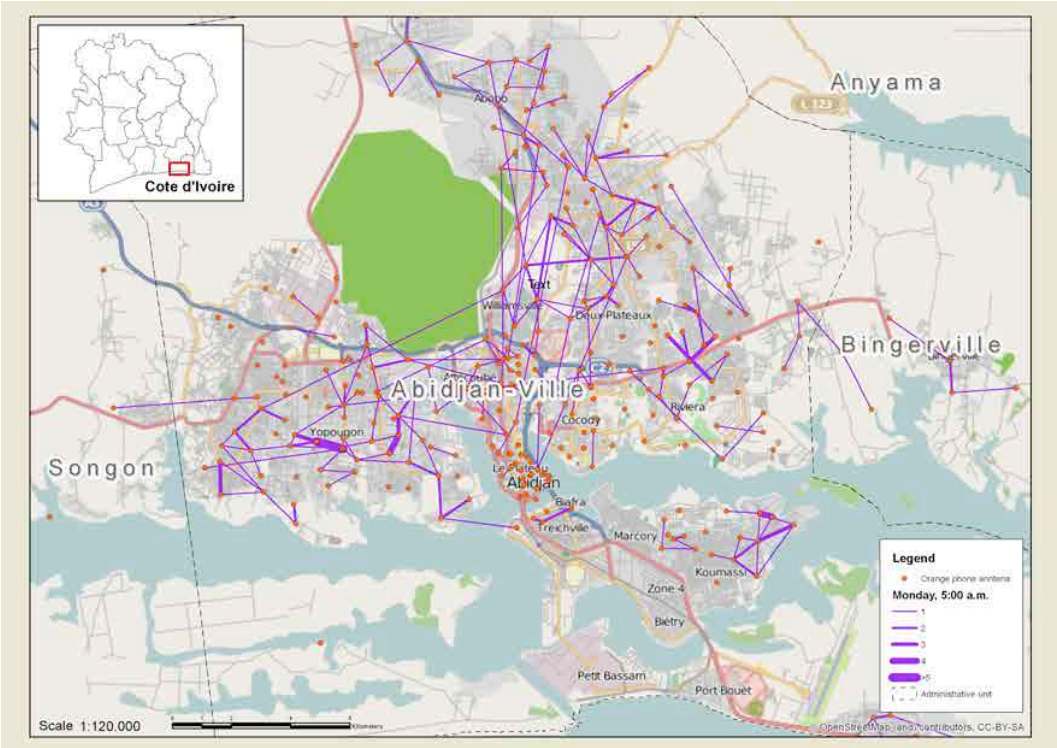


Figure 10: Traffic between antennas

Starting from a file generated from the dataset that contains the geographic coordinates (longitude, latitude) of an antenna origin and those of other target antennas, together with a field that measures the traffic of information between one and the other one ('weight' field), we have generated the corresponding linear entities that join both positions, and the value of the calculated weight has been assigned to them. The visualization is achieved classifying the numerical field (weight) for graduated symbology using an equal interval classification method that divides the range of attribute values into five equal-sized subranges.

Acknowledgement

To Professor Antonio Hernando Esteban[11] (Automatic & systems engineering depto at EUI-UPM) to review the mathematical formalism proposed in this paper.

Also Carmen Vidal, Operations Manager at Paradigma Tecnologico to support this project and the people at Paradigma Labs.

Special thanks to Chris Kostov for his valuable help reviewing this document, carefully spotting grammar, mechanics, and punctuation mistakes, and providing insights into writing in better English.

[11] http://www.sia.eui.upm.es/isa/doku.php?id=profesores:ahernando

References

A.C. Alegria, H. Sahli, and E. Zimanyi. Application of density analysis for landmine risk mapping. In *Spatial Data Mining and Geographical Knowledge Services (ICSDM), 2011 IEEE International Conference on*, pages 223 –228, 29 2011-july 1 2011. doi: 10.1109/ICSDM.2011.5969036.

J. F. Bithell. An application of density estimation to geographical epidemiology. *Statistics in Medicine*, 9(6):691–701, 1990. ISSN 1097-0258. doi: 10.1002/sim.4780090616. URL `http://dx.doi.org/10.1002/sim.4780090616`.

Vincent D. Blondel, Markus Esch, Connie Chan, Fabrice Clérot, Pierre Deville, Etienne Huens, Frédéric Morlot, Zbigniew Smoreda, and Cezary Ziemlicki. Data for development: the d4d challenge on mobile phone data. *CoRR*, abs/1210.0137, 2012.

David P Chrest and William D Wheaton. Using geographic information systems to define and map commuting patterns as inputs to agent-based models. *Methods report (RTI Press)*, 2009(12):906, 2009.

John D. Cook. Computing the distance between two locations on earth from coordinates, 2012. URL `http://www.johndcook.com/python_longitude_latitude.html`. [Online; accessed 4-January-2013].

W.J. Craig, T.M. Harris, and D. Weiner. *Community Participation and Geographical Information Systems*. Community Participation and Geographic Information Systems. Taylor & Francis, 2002. ISBN 9780415237529. URL `http://books.google.es/books?id=IePxXod45z8C`.

W. De Jong, M. Shaw, and N. Stammers. *Global Activism, Global Media*. Pluto Press, 2005. ISBN 9780745321967. URL `http://books.google.com.co/books?id=ny6DAAAAMAAJ`.

Nathan Eagle, Alex Pentland, and David Lazer. Inferring Friendship Network Structure by Using Mobile Phone Data. *PNAS*, 106(36), 2009.

N.G. Fielding, R.M. Lee, and G. Blank. *The SAGE Handbook of Online Research Methods*. SAGE Publications, 2008. ISBN 9781446206607. URL `http://books.google.es/books?id=EeMKURpicCgC`.

Carlos Llano Verduras. Localización residencial y movilidad laboral: un análisis del" commuting" de trabajadores nacionales e inmigrantes en la comunidad de madrid. *Cuadernos de economía: Spanish Journal of Economics and Finance*, 29(81):69–100, 2006.

Oscar Marin, Alejandro Gonzalez, Roberto Maestre, Julio Gonzalez, Marco Martinez, Ruben Abad, and Leonardo Menezes. 15th october on twitter global revolution mapped, 2012. URL `http://labs.paradigmatecnologico.com/2011/12/19/15th-october-on-twitter-global-revolution-mapped/`. [Online; accessed 23-January-2013].

A. Mitchell. *Zeroing In: Geographic Information Systems at Work in the Community*. Esri Press, 1997. ISBN 9781879102507. URL `http://books.google.com.co/books?id=Tj6AAAAAMAAJ`.

D. Moyo. *Dead Aid: Why Aid Is Not Working and How There Is a Better Way for Africa*. Farrar, Straus and Giroux, 2009. ISBN 9780374139568. URL `http://books.google.es/books?id=2T_RbtTslzEC`.

A.A. SEN, A.K. Sen, E. Rabasco, and L. Toharia. *Desarrollo y Libertad*. Colección Documento Series. GeoPlaneta, Editorial, S. A., 2000. ISBN 9788408035244. URL `http://books.google.com.co/books?id=Jk_bPQAACAAJ`.

Ana Solana and David Alonso. Self-organizing networks and gis tools cases of use for the study of trading cooperation (1400-1800). *Journal of Knowledge Management, Economics and Information Technology*, page pp. 402, 2012. ISSN 2069-5934. URL `http://www.scientificpapers.org/special-issue-june-2012/`.

Marius Thériault, Marie-Hélène Vandersmissen, Martin Lee-Gosselin, and Denis Leroux. Modelling commuter trip length and duration within gis: Application to an od survey. *Journal for Geographic Information and Decision Analysis*, 3(1):41–55, 1999.

Fahui Wang. Modeling commuting patterns in chicago in a gis environment: A job accessibility perspective. *The Professional Geographer*, 52(1):120–133, 2000.

Wikipedia. Commuting — wikipedia, the free encyclopedia, 2012. URL `http://en.wikipedia.org/w/index.php?title=Commuting&oldid=525136196`. [Online; accessed 4-January-2013].

F Benjamin Zhan, Xuwei Chen, Carol Lewis, Laura L Higgins, and Judy Perkins. Gis models for analyzing intercity commute patterns: A case study of the austin-san antonio corridor in texas. Technical report, Center for Transportation Research, University of Texas at Austin; Springfield, Va., 2008.

The geography and carbon footprint of mobile phone use in Cote d'Ivoire

Vsevolod Salnikov,[1] Daniel Schien,[2] Hyejin Youn,[3] Renaud Lambiotte,[1] and Michael T. Gastner[4]

[1] *naXys, University of Namur, Rempart de la Vierge 8, 5000 Namur, Belgium*
[2] *Department of Computer Science, University of Bristol, Merchant Venturers Building, Woodland Road, Bristol BS8 1UB, UK*
[3] *Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501, USA*
[4] *Department of Engineering Mathematics, University of Bristol, Merchant Venturers Building, Woodland Road, Bristol BS8 1UB, UK*

The newly released Orange D4D mobile phone data base provides new insights into the use of mobile technology in a developing country. Here we perform a series of spatial data analyses that reveal important geographic aspects of mobile phone use in Cote d'Ivoire. We first map the locations of base stations with respect to the population distribution and the number and duration of calls at each base station. On this basis, we estimate the energy consumed by the mobile phone network. Finally, we perform an analysis of inter-city mobility, and identify high-traffic roads in the country.

I. INTRODUCTION

The availability of mobile phone records has revolutionised our ability to perform large-scale studies of social networks and human mobility. Traditionally, researchers had to rely on a combination of surveys, census data and vehicle counting. These methods are costly and time consuming so that data were collected either infrequently or for small population samples only. In the last few years, while searching for innovative methods to circumvent these limitations, researchers have turned their attention to mobile phones as sensors to collect communication and mobility data [1]. The vast majority of studies were carried out in developed countries where mobile communication competes with established landline technologies. However, mobile phones are nowadays commonplace in developing countries too. Especially in Africa, mobile phones now provide affordable telecommunication where no alternative had previously existed [2, 3].

The data bases for Cote d'Ivoire, made accessible during the Orange D4D challenge [4], present the first opportunity to analyse a nationwide mobile phone network in Africa. The data are obtained from the so-called Call Detail Records (CDRs) which contain an approximate location of mobile phones every time they connect to a cell tower (e.g. due to a phone call). A growing body of research has shown that CDRs can accurately characterise many aspects of human mobility. Practical examples include the tracking of population displacements after disasters [5, 6], the estimation of traffic volumes in cities [7], the calculation of carbon emissions due to commuting [8] and transport mode inference [9]. Here we apply geospatial techniques to address several questions related to social and economic development. How is mobile phone infrastructure related to its use (Sec. II and III)? How much energy is needed to operate the network (Sec. IV)? Is the road infrastructure adapted to the population mobility patterns (Sec. V)?

II. MAPPING BASE STATION LOCATIONS WITH RESPECT TO POPULATION DENSITY

Where to place the base stations that house the antennas is a central decision for any mobile communication provider. It determines how many people can access the network, the quality of calls and the ease with which the provider can operate the facilities. Optimising the base station locations is a difficult task, complicated by spatially heterogeneous demand and topological obstacles such as tall buildings or mountains. As a rule of thumb, however, population density is a crucial factor: where there are more people, we expect a higher density of base stations. Conversely, if rural areas with lower population are served by a disproportionately low number of base stations, these communities would be left with little or no access to the network. As mobile communication has enormous potential to improve the lives of the rural population (e.g. by access to banking and real-time information about agricultural commodity prices), one development objective must be to provide a roughly equal per-capita number of base stations for the entire population of Cote d'Ivoire.

We map the 1238 base station coordinates given in the D4D file `ANT_POS.TSV` on a standard latitude-longitude projection (left map in Fig. 1). Since Cote d'Ivoire is close to the equator, such a projection is nearly distance-preserving. The base stations (coloured dots on the map) are spatially very unevenly distributed: in some parts of Abidjan there are more than ten base stations per square kilometre, whereas some subprefectures in the north of the country have no base station at all. That there should be many base stations in Abidjan is quite obvious because $\approx 20\%$ of all citizens live in the country's most populous city. However, whether the number of base stations is proportional to its population is not immediately apparent from the latitude-longitude projection.

We will thus have to combine the base station coordinates with information about the population distribution. Here we use census estimates from the AfriPop project (http://www.afripop.org) [10]. Based on these

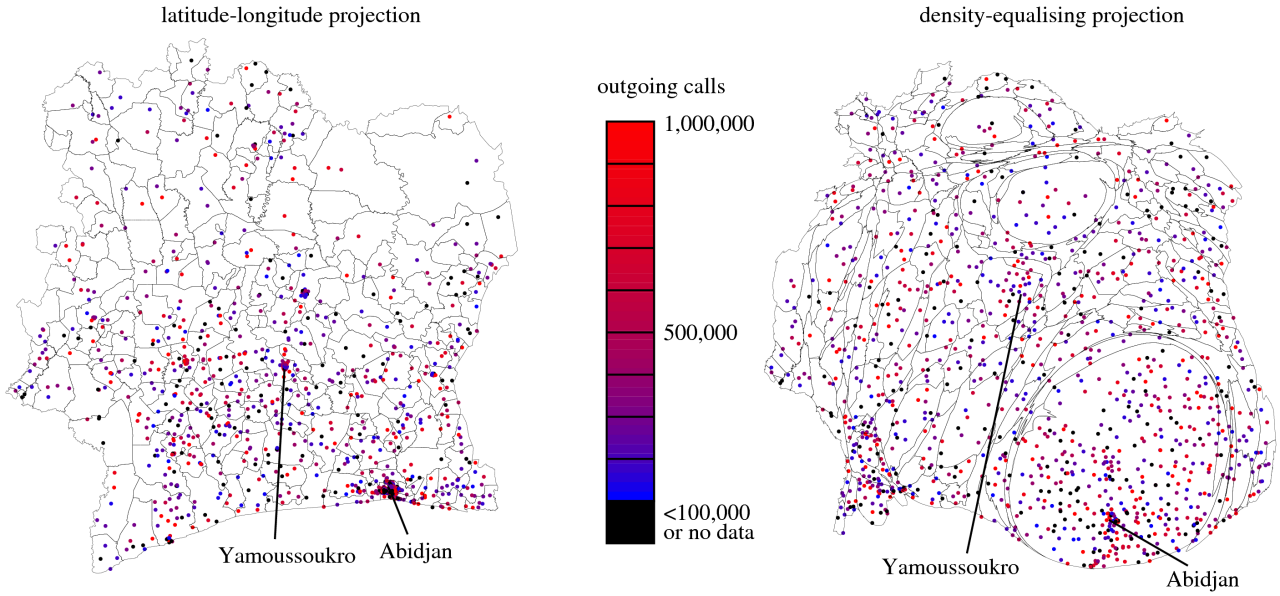


FIG. 1: Base station locations on a conventional longitude-latitude projection (left) and a cartogram where areas are rescaled to be proportional to the number of inhabitants (right). The colours of the dots indicate the number of outgoing calls at each base station. The boundaries of subprefectures are shown for ease of orientation. The geographic distribution of the base stations are largely explained by the heterogeneous population distribution so that the point pattern appears less clustered on the right than on the left. Still, regions of significantly higher per-capita base station density remain (see Fig. 2), especially in Abidjan, where even on the cartogram the dots are noticeably aggregated. The colours of the dots do not exhibit any clearly visible large-scale trends. However, a more careful statistical analysis shows that a significant correlation between traffic at nearby base stations exists (see Fig. 3b).

numbers, we project the map of Ivory coast so that all regions of the country are represented by an area proportional to its population [11]. Such a density-equalising map – also known as a cartogram – has become a popular tool to visualise inequality and development challenges [12]. Plotting the base station locations on the cartogram (right map in Fig. 1) reveals a nuanced picture. On one hand, the point distribution is much less aggregated on the cartogram and thus is indeed largely proportional to population. On the other hand, the points are far from a homogeneous pattern. In Abidjan, in particular, a dense cluster of base stations remains clearly visible, indicating a disproportionately high per-capita connectivity there.

We confirm this observation by calculating the population in the base stations' Voronoi cells. (The Voronoi cell of a given base station is the polygon that contains the area closer to this base station than to any other.) A population-proportional base station distribution would result in an equal population inside each Voronoi cell. A rank plot of population numbers (Fig. 2), however, has a clear S-shape: although most cells have a population around $10,000$, there are outliers in both directions. Interestingly, the 16 lowest ranked cells are all in Abidjan, making it by far the region with the highest per-capita base station density. By contrast, the Voronoi cells with the largest populations are in rural areas near inland borders (e.g. the second ranked base station at 7.267° N, 8.160° W is 20 km east of the Liberian border and the fifth ranked at 9.803° N, 3.303° W is 6 km south of the border with Burkina Faso) or near smaller cities (the top and third ranked base station are only a few kilometres outside Bouaké and the fourth and sixth ranked near Korhogo, the country's third and seventh largest cities respectively). Because many facility location models suggest that a fair distribution of resources should intentionally be skewed in favor of less populated areas [13, 14], our finding suggests these regions as targets for a future expansion of the network.

III. SPATIAL CORRELATION BETWEEN THE POPULATION DENSITY AND THE NUMBER OF CALLS

Recent studies of mobile phone records in developed countries [15] have argued that the number of human interactions in cities increases faster than linearly with the city population. This poses the question: does the number of calls in Cote d'Ivoire depend similarly on population density? Here we use the population within a 5 km radius around a base station as a proxy for the local population density. As an indicator of the local phone activity, we count all phone calls that were made between

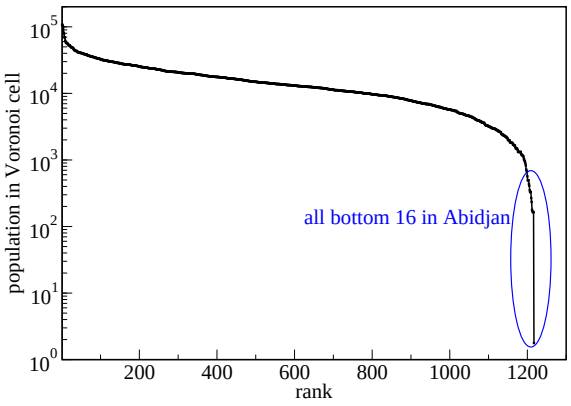


FIG. 2: Rank plot of the population inside the Voronoi cells of the 1217 distinct base station locations in the Orange D4D challenge data base. Although a majority of 616 cells are within 50% of the median population (12 897 inhabitants), there are significant outliers at both top and bottom ranks. While the bottom ranked cells are predominantly in Abidjan, the top ranks are in rural areas as well as smaller cities.

5 December 2011 and 9 April 2012 from base stations within the same 5 km radius. This number always includes calls at the focal base station itself, but may also include other nearby base stations. An ordinary least-squares fit of the form $\log(\text{traffic}) = a\log(\text{population}) + b$ yields a slope $a = 0.943$ with a 95% confidence interval $[0.919, 0.967]$. Consequently, there is no significant indication that mobile phone traffic scales superlinearly with population density. The same conclusion remains true if, instead of the number of outgoing calls, we consider the number of ingoing calls (95% confidence interval $[0.923, 0.970]$) or the total duration of out- ($[0.920, 0.967]$) or ingoing calls ($[0.908, 0.954]$).

The slightly sublinear relationship indicates that the spatial distribution of calls is not entirely random. We confirm this finding in Fig. 3(b) where we plot the number of outgoing calls at each base station versus the number of calls at its nearest neighbour. These two variables have a moderately high Pearson correlation coefficient $\rho = 0.329$. In order to establish the significance of the correlation, we performed a Monte Carlo simulation where we shuffled the number of calls and assigned each number randomly to one of the existing base station locations. In $240,000$ runs, we never found a correlation coefficient as large as the observed value, so that the correlation differs indeed significantly from zero. Similar results are true for the number of ingoing calls ($\rho = 0.325$) and the durations of out- or ingoing calls ($\rho = 0.3560$ and $\rho = 0.344$ respectively).

Thus, the phone call intensity is spatially autocorrelated and we tentatively conclude that there is no indication of superlinear scaling with population density. However, the interpretation of Fig. 3(a) is not straightforward because the data points are not independent: the population and calls within a 5 km radius overlap for nearby

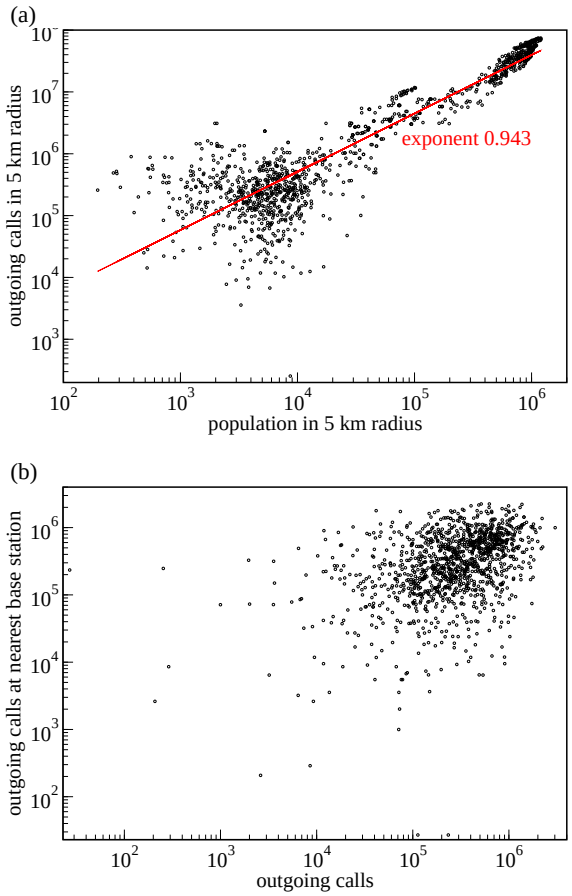


FIG. 3: (a) Scatter plot of the total number of outgoing calls versus the population. Both variables are measured inside a 5 km radius around each base station. An ordinary least-squares fit yields a slightly sublinear dependence of the calls on population. (b) Scatter plot where the horizontal axis is the number of outgoing calls at a base station and the vertical axis the number of outgoing calls at the nearest base station. These two variables are significantly correlated ($\rho = 0.329$).

base stations. In particular, an analysis based on a more careful socio-economic definition of "city size" [16] may still unearth more details.

IV. ENERGY AND CARBON FOOTPRINT OF WIRELESS CELLULAR NETWORKS IN COTE D'IVOIRE

In this section we estimate the energy and greenhouse gas (GHG) emissions, contributing to climate change, of the wireless cellular network in Cote d'Ivoire and compare its share of the national GHG emissions with that of wireless networks in other countries.

While mobile network operators are increasingly transparent about their environmental impact and GHG emissions, few data are available about the energy consump-

tion and resulting GHG emissions of mobile networks in developing countries. Moreover, the Orange Cote d'Ivoire (OCI) data permits discussing energy consumption of parts of the entire network in relationship to population density. Thus, this work contributes to ongoing research that investigates the direct environmental impact of ICT (Information and Communication Technology) of systems in general, independent of development contexts, by estimating an entire network's footprint from the number of its base stations.

Much research has shown that mobile technologies are an important instrument of current information and communication technologies for development (ICT4D) strategies, for example [17]. On the other hand, the increasing deployment of these technologies can result in increasing GHG emissions, sometimes labelled "footprint", which recently has also received increasing interest by the community of ICT4D researchers [18, 19]. It is our aim to contribute to a more informed discussion through provision of quantitative estimates of energy consumption and GHG emissions. We want to precede this analysis with a qualification: in or outside of a development context the analysis of environmental impact of a technical system and its results can stand separately from the interpretation of these results towards decision making for policy formation. In this text we estimate the annual GHG emissions of the mobile network in Cote d'Ivoire and suggest directions for existing or future development of these networks from the perspective of their technical operation. However, this analysis would only provide an incomplete basis for policy making towards a development strategy as it does not include an analysis of the social or economic impacts and benefits of the wireless network.

The goal of our assessment is to estimate the national energy consumption and GHG emissions using the number of base stations as an input parameter. This requires an estimate of the power consumption per base station and the overhead from the remaining parts of the network. Depending on its type, the power consumption of a base station can vary between 800 and 2800 W (estimations presented in [20]). Without additional information about the specific types of base stations, the OCI data can only be parameterised with average data. Additionally, an assessment of the energy consumption of a mobile network should include all relevant system parts in order to enable greater transferability of results. We assume the following composition of the wireless network: the base stations, which house the antennas and amplifiers, and auxiliary equipment for cooling and power transformation provide the radio signal to subscribers. They are controlled by several base station controllers and a few mobile service centres to which they are connected via a radio or fixed network. This network also provides connectivity with the Internet or networks of other operators. In our estimate of the GHG emissions we had to make some simplifying assumptions about the network infrastructure. We estimated the energy consumption for a single base station (including overhead for other system parts such as base station controllers) of around 2100 W based on similar assumptions made in [20] and [21] that are based on publicly available data by Vodafone. This value is a top-down estimate based on the total energy consumption of the network and the total number of base stations. The corporate responsibility report of the Vodafone Group states that in 2011 the company globally operates 224 000 base stations and that the energy consumption was 4117 GWh [22]. This value does not account for energy consumption in offices. Given that the average power consumption per base station is around 1.5 kW, the resulting value of 2100 W per base station is plausible and further corroborated by other studies such as [23] who state that the energy consumption of the base stations constitutes 60-80% of the total energy consumption of the network.

An estimate of the contribution of the remaining parts of a mobile operator's organisation to energy and carbon footprint can, for example, be based on corporate social responsibility reports by Vodafone and O2, which state that the network accounts for around 80% to 90% of an operator's energy consumption [24, 25] and constitutes a similar portion of its GHG emissions [26]. The GSMA Mobile Green Manifesto report [27] makes similar assumptions. We assume that these ratios also apply to the OCI network and networks of other operators in Cote d'Ivoire.

Based on the data inventory we have a precise count of mobile base stations (1238). In order to estimate the total annual national energy consumption by mobile networks we had to also estimate the number of base stations by competitors of OCI in addition to the power consumption by the other system parts. We assumed that all mobile operators deploy their network on average with similar density. Based on the market share of subscriptions (between 33% OCI [28] and 35% [29]) we estimate that the total number of base stations in Cote d'Ivoire is around 3700.

Given these assumptions, we estimate that the GHG emissions of the wireless mobile networks by all operators in Cote d'Ivoire amount to about 29.1 kilo tonnes carbon dioxide equivalent per year ($ktCO2e$). This is about 0.4% of the total annual carbon emissions of 6596.933 ktCO2e [30]. We further estimate the energy consumption to be 68 GWh which is about 1.9% of the total annual energy production of Cote d'Ivoire [31]. Compared to the pro-rata energy consumption and GHG emissions by mobile networks in Germany, this value is relatively high. Based on publicly available data by Vodafone Germany energy consumption by the network accumulated circa 600 GWh [25] in 2011. Assuming other wireless networks are equally efficient and Vodafone's market share of 32.97% [32], the energy consumption by mobile networks in Germany would be about 0.3% of the absolute energy consumption in the country in 2011 [33]. In [27] it is found that on global average, mobile networks result in 0.2% of all GHG emissions. Based on the Vodafone

data, however, the portion of German mobile networks of the national GHG emissions is only around 0.1%.

Given the lack of data on the power consumption by each base station, there is a relatively high uncertainty to the estimate of the total annual energy consumption by all networks. The estimate of the carbon emissions is further affected by uncertainty in the parameter for the carbon intensity of electricity. In OECD countries, base stations are typically operated with energy from the electrical grid. In developing countries, however, electrical energy is possibly supplied by diesel generators to a significant degree. Diesel generators result in a greater carbon intensity per generated kWh of electricity (0.788 kgCO2-eq/kWh [34], as compared to 0.426 kgCO2-eq/kWh of the average intensity of grid electricity).

In table I we evaluate the influence of these parameters to the estimates of GHG emissions and energy consumption. In scenario I, we consider how the energy consumption and GHG emissions would change if the average power consumption per base station was reduced by 25% relative to our base line. The resulting average power consumption per base station, including a portion for remaining network parts, is 1.58 kW. In this case the carbon footprint of the network is 0.33% and slightly closer to the global average value estimated by GSMA in [27] of 0.2%.

Secondly, we evaluate the scenario that half of the electricity consumed by the base station was provided by diesel generators and the other half by the electrical grid which would increase the carbon intensity from 0.426 kgCO2e/kWh to 0.602 kgCO2e/kWh. We include the assumption that this would free capacity in the electrical grid. In our scenario, the mobile network would consume 0.95% of CI's electrical energy and constitute 0.62% of the national GHG emissions. We also considered evaluating more complex assumptions about the types of base stations. Such a scenario would assume that network planners generate relatively precise predictions of demand in a cell. However, the number of outgoing calls that we plot in Fig. 1 together with incoming calls are only a possible proxy to overall demand of voice traffic. Data services and number of calls at peak time must both be considered to estimate the minimum capacity of a base station. We believe that the results of such a scenario would have too much uncertainty to bring significant value for our discussion.

Given this sensitivity analysis it remains clear, that mobile networks in Cote d'Ivoire contribute to a greater degree to the total GHG emissions of the country than those in Germany. One of the main reasons for this difference is likely to be the contrasting structure of the German and Ivorian economy to which the energy intensive manufacturing industry in Germany is likely to contribute. This assumption is also supported by a comparison of street lighting as another energy consuming infrastructure. A report by the World Bank mentions in passing that 400 000 public street lights are operated in Cote d'Ivoire [35]. Assuming that street lights have a power consumption between 35 and 400 W [36] each, they constitute a share of the total energy consumption in Cote d'Ivoire between 1.4 and 16 percent. In contrast, the street lighting in Germany constitutes only 0.56% of the total energy consumption [37].

	BASE	I	II
Carbon intensity of electricity (kgCO2e/kWh)	0.43	0.43	0.60
Average aggregate power consumption per BS (kW)	2.10	1.58	2.10
Total national energy consumption by mobile networks(GWh)	68.32	51.24	68.32
National energy consumption by mobile networks (percent of total)	1.90	1.43	0.95
Total national GHG emissions by mobile networks (ktCO2e)	29.11	21.83	41.13
National GHG emissions by mobile networks (percent of total)	0.44	0.33	0.62
Annual energy consumption per subscriber (kWh/sub)	3.83	2.88	3.83

TABLE I: Scenarios of alternative input parameters. Scenario 'base' assumes average values for all parameters, scenario 'I' assumes a value for the power consumption per base station compared to 25% from base. Scenario 'II' assumes that 50% of the electricity consumed by base stations is supplied by diesel generators.

Interestingly, if apportioned to each subscriber, the annual energy consumption of the OCI mobile network is 3.83 kWh/sub which is much lower than the same metric for customers of Vodafone Germany (16.5 kWh/sub). The value is also a lot lower than the values reported in [38] (values between 7 and 34 kWh/sub with an average of 16.7 kWh/sub). In the case of Cote d'Ivoire, this is likely to be partly the result of a sparser deployment of base stations, in particular outside of Abidjan as we illustrate in Sec. II. One contributing factor to this sparser deployment is likely to be the lower degree of urbanisation (52% compared to 74% Germany [30]). Another factor is the delayed introduction of data services to Cote d'Ivoire. Third generation services are only just being introduced to this market.

These figures have relevance to the ICT4D community because development in Cote d'Ivoire can be seen as indicative for many other African countries. Cote d'Ivoire currently is among the countries with the highest mobile phone penetration [39]. As our estimates illustrate, the uptake of mobile phone technologies is accompanied by an increase in energy consumption. For Cote d'Ivoire specifically, it is likely that the energy consumption by the network will increase in the near future with an increasing adoption of data services.

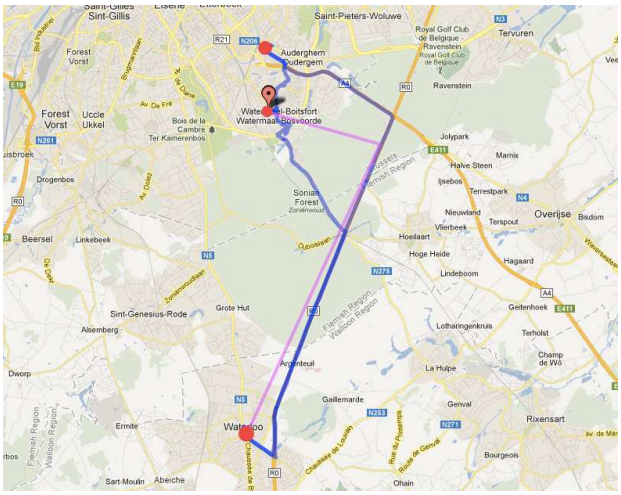


FIG. 4: To illustrate the different ways to uncover mobility patterns from CDRs, let us focus on the motion of an individual in Brussels, as measured by his GPS. The user took his car in Watermael and went to two shops, one in Auderghem and one in Waterloo. The three locations are plotted in red. Three phone calls were made. One at home, one on the highway, and one in Waterloo. An approach where a path is composed of successive position measurements is shown in pink. In contrast, an approach where paths are based on important locations would detect the stop in Waterloo, rightly discard the one on the highway, but would still be blind to the location in Auderghem.

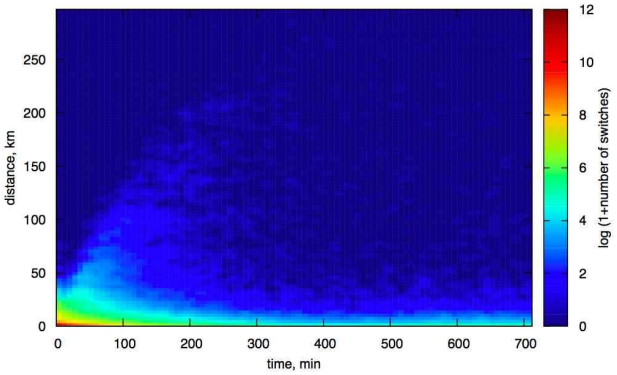


FIG. 5: Heat map of the distances and time intervals between consecutive CDRs. The burstiness of phone activity leads to a broad distribution over time. Keeping transitions from different regions in the two-dimensional space allows for the identification of different aspects of human mobility.

V. DETECTING IMPORTANT ROUTES FOR INTER-CITY MOBILITY

CDRs provide a cheap and efficient source of data to study human mobility patterns at a large scale [1]. Yet they suffer from limitations that need to be carefully considered, and in some case dealt with, to ensure the validity of the observations. A key limitation is the sparse and heterogeneous sampling of the trajectories, as the location is not continuously provided but only when the phone engages in a phone call or a text message exchange. Moreover, the spatial accuracy of the data is determined by the local density of base stations. When estimating mobility from CDRs, different approaches have been developed in the literature (see Fig. 4 for illustration).

First, researchers interested in statistical models of human mobility have adopted a Brownian motion approach [40], where each individual is considered as a particle randomly moving in its environment. Mobility is considered as a path between positions at successive position measurements. Authors have observed statistical properties reminiscent of Levy flights, together with a high degree of regularity. Yet, the usefulness of these observations is limited by the bursty nature of phone activity, as burstiness is expected to alter basic statistical properties of the jumps, such as their distance distribution (see Fig. 5). Even in studies where the positions are evaluated at regular intervals, the nature of the jumps remains unclear, as the method tends to detect short trips due to localisation errors, and is blind to the type of the places sampled from the real trajectory. As a side note, let us mention recent work using geo-localised web services, such as Foursquare, where users voluntarily check-in at places [41, 42]. Foursquare check-ins are also characterised by a bursty behaviour, but they provide a GPS accuracy, and semantic information (at the office, travelling, etc.) that might solve the aforementioned problems.

The second approach relies on the idea that mobility consists of moving from one place to another. The observation of mobility patterns thus requires one to define and identify important locations. A trajectory is seen as a set of consecutive locations visited by the user. Important locations can either be defined as a place where a user spends a significant amount of time, which he visits frequently, or where he has stopped for a sufficiently long time [1, 43–45]. This approach provides a more intuitive picture of mobility, where the sampling is determined by the periods of rest of the user. However, it is blind to the multi-scale nature of human mobility, as it requires the parametrisation of thresholds in time and in space to identify important locations. The value of the threshold and the corresponding granularity of the places depends on the system under scrutiny, say cities for international mobility or rooms for human mobility inside hospitals [46].

When measuring human mobility from CDRs, it is important to remember that mobility is about space and time. Both aspects must be carefully considered to provide a faithful description of human trajectories, especially in situations where the sampling of the data is heterogeneous. For this reason, each transition should be remembered as a jump in space over an interval in time and, if possible, be put in relation to the previous and following transitions. Contrary to the universality

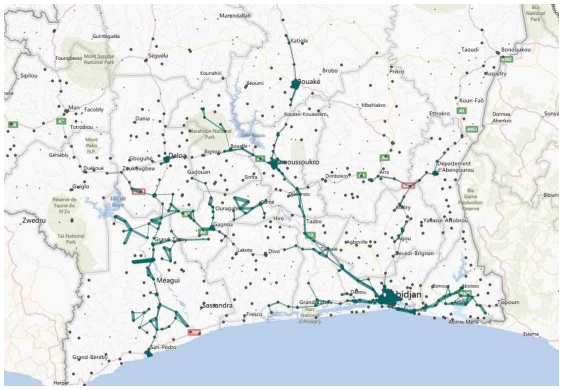


FIG. 6: High-traffic road detection, as obtained from CDR data.

FIG. 7: Roads that are identified from CDR data and are absent on a Microsoft map (upper panel) can be identified on an OpenStreetMap (lower panel).

viewpoint of [40], not all transitions are alike. On the contrary, it is possible to extract different information and different types of mobility patterns by focusing on different regions in space-time. This filtering has been adopted in various studies, but usually either in space or in time. Let us mention [45], where transitions between identified places are considered only if they are registered within two hours of each other; in [44] the daily range of mobility is calculated, and in [7] a trip is defined as a displacement between two distinct base stations occurring within one hour in each time period. More complex filters can be defined on so-called *handoff patterns*, that is a sequences of cell towers that a moving phone uses while engaged in one voice call, e.g. in [43] where only sequences of more than 5 cell towers are included. Let us note that a filtering in space and in time allows for the selection of a characteristic velocity and, if needed, of the removal of noisy transitions occurring at a small spatial scale, e.g. transitions between neighbouring cells of a static user, or long temporal scale, e.g. transitions over several days where several intermediate steps are expected to be missing.

This overview of recent research suggests direct applications that would be of particular interest in a developing country, where empirical data on human mobility tend to be lacking. Using the aforementioned methodologies, it would be possible, for instance, to identify and map nationwide commuting patterns. Traffic tracking and route classification would also be possible after additional data is collected from test drives or signal strength data collected by high-resolution scanners [43]. In this work, we illustrate the potential benefits of a CDR analysis by focusing on the detection of high-traffic roads between cities. Such a detection might help deploy new infrastructure where the population actually needs it, e.g. in regions where mobility is high but the infrastructure is poor. Finding high-traffic roads requires one to filter transitions in the two-dimensional space of Fig. 5. To do so, we apply the following procedure. We consider only transitions in a certain velocity range and occurring in less than one hour. Moreover, to remove noisy connections and to identify persistent motion, we only consider largest connected components in the corresponding graph. The velocity range is chosen to be $[15, 150]$km/h, as we are interested in car mobility. When applied to the `POS_SAMPLE_X.TSV` file provided by D4D, this simple procedure directly identifies the main road structure of the country (Fig. 6). More interestingly, it also allowed us to identify unknown roads, which we could validate a posteriori. Examples are shown in Fig. 7 and Fig. 8 where roads that were absent in Microsoft maps are identified in maps provided by OpenStreetMap and Yahoo respectively.

VI. CONCLUSION

In this article we have presented how an analysis of the Orange D4D mobile phone data base reveals important patterns of communication infrastructure and mobile phone use in Cote d'Ivoire. The placement of base stations is biased towards Abidjan so that one development goal is an enhancement of the network in smaller cities and rural regions. We estimate that the network currently consumes between 2.88 and 3.83 kWh of energy annually per subscriber. Although this figure is less than in an industrial country such as Germany, the fraction of the national energy consumption spent on mobile telephony (estimated between 0.95% and 1.90%) is actually higher. Finally, we argued that mobility data from CDRs need further filtering to extract truly meaningful commuting patterns. We used the mobility traces that were part of the Orange D4D database to demonstrate how the main roads in Cote d'Ivoire can be identified.

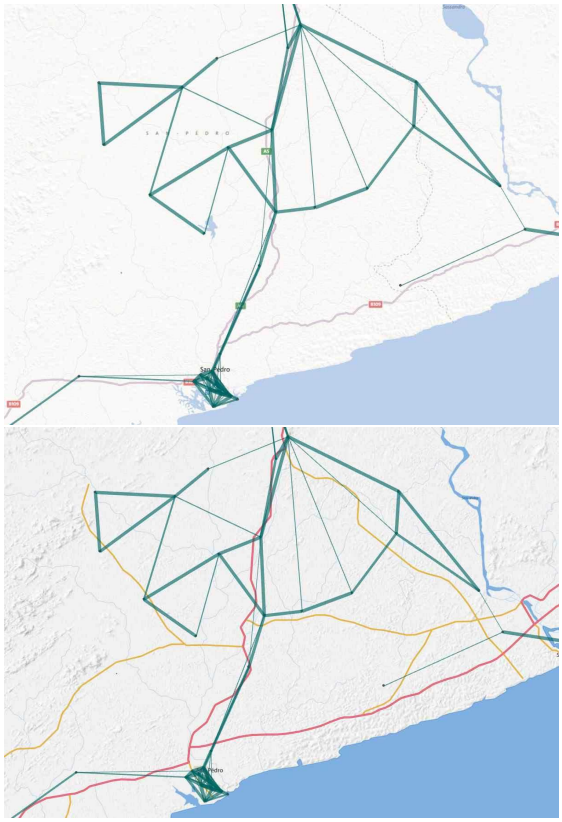


FIG. 8: Roads that are identified from CDR data and are absent on a Microsoft map (upper panel) can be identified on a Yahoo map (lower panel).

Acknowledgment

VS and RL acknowledge financial support from FNRS. MTG is grateful for financial support from the University of Bristol and the EPSRC Building Global Engagements in Research (BGER) grant. This paper presents research results of the Belgian Network DYSCO (Dynamical Systems, Control, and Optimization), funded by the Interuniversity Attraction Poles Programme, initiated by the Belgian State, Science Policy Office. HY acknowledges the support by grants from the Rockefeller Foundation and the James McDonnell Foundation (no. 220020195).

[1] R. Becker, R. Cáceres, K. Hanson, S. Isaacman, J. M. Loh, M. Martonosi, J. Rowland, S. Urbanek, A. Varshavsky and C. Volinsky, "Human Mobility Characterization from Cellular Network Data", Communications of the ACM **56**, 74–82 (2013)

[2] R. Heeks and A. Jagun, "Mobile phones and development", id21 insights **69**, 1–2 (2007), available at http://www.dfid.gov.uk/r4d/PDF/Articles/insights69.pdf

[3] R. Singh, "Mobile phones for development and profit: a win-win scenario", Overseas Development Institute Opinion 128 (2009), available at http://www.odi.org.uk/sites/odi.org.uk/files/odi-assets/publications-opinion-files/3739.pdf

[4] V. D. Blondel, M. Esch, C. Chan, F. Clerot, P. Deville, E. Huens, F. Morlot, Z. Smoreda and C. Ziemlicki, "Data for development: the D4D challenge on mobile phone data", arXiv:1210.0137 (2012)

[5] L. Bengtsson, X. Lu, A. Thorson, R. Garfield and J. von Schreeb, "Improved Response to Disasters and Outbreaks by Tracking Population Movements with Mobile Phone Network Data: A Post-Earthquake Geospatial Study in Haiti", PLoS Medicine **8**, e1001083 (2011)

[6] X. Lu, L. Bengtsson and P. Holme, "Predictability of population displacement after the 2010 Haiti earthquake", Proc. Nat. Acad. Sci. **109**, 11576–11581 (2012)

[7] P. Wang, T. Hunter, A. M. Bayen, K. Schechtner and M. C. González, "Understanding Road Usage Patterns in Urban Areas", Scientific Reports **2**, 1001 (2012)

[8] S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, M. Martonosi, J. Rowland and A. Varshavsky, "Identifying important places in people's lives from cellular network data", Proceedings of the 9th international conference on Pervasive computing, 133–151 (2011)

[9] H. Wang, F. Calabrese, G. Di Lorenzo and C. Ratti, "Transportation mode inference from anonymized and aggregated mobile phone call detail records", Proceedings of the 13th international IEEE conference on intelligent transport systems, 318–323 (2010)

[10] A. J. Tatem, A. M. Noor, C. von Hagen, A. Di Gregorio and S. I. Haym, "High resolution population maps for low income nations: combining land cover and census in east Africa", PLoS One **2**, e1298 (2007)

[11] M. T. Gastner and M. E. J. Newman, "Diffusion-based method for producing density-equalizing maps", Proc. Nat. Acad. Sci. **101**, 7499-7504 (2004)

[12] D. Dorling, M. E. J. Newman and A. Barford, "The atlas of the real world: mapping the way we live" (2nd ed.), Thames & Hudson (London, 2010)

[13] M. T. Gastner and M. E. J. Newman, "Optimal design of spatial distribution networks", Physical Review E **74**, 016117 (2006)
[14] M. T. Gastner, "Scaling and entropy in p-median facility location along a line", Physical Review E **84**, 036112 (2011)
[15] M. Schläpfer, L. M. A. Bettencourt, M. Raschke, R. Claxton, Z. Smoreda, G. B. West and C. Ratti, "The Scaling of Human Interactions with City Size", arXiv:1210.5215 (2012)
[16] E. Arcaute, E. Hatna, P. Ferguson, H. Youn, A. Johansson and M. Batty, "City boundaries and the universality of scaling laws", arXiv:1301.1674 (2013)
[17] J. C. Aker and I. M. Mbiti, "Mobile Phones and Economic Development in Africa", Journal of Economic Perspectives **24**, 207–232 (2010)
[18] H. Roeth and L. Wokeck, "ICTs and Climate Change Mitigation in Emerging Economies", Technical report (2011). available at http://www.niccd.org/RoethWokeckClimateChangeMitigationICTs.pdf
[19] D. I. Paul and J. Uhomoibhi, "Solar power generation for ICT and sustainable development in emerging economies", Campus-wide Information Systems **29**, 213–225 (2012)
[20] D. Schien, C. Preist, P. Shabajee and M. Yearworth, "Modeling and assessing variability in use phase energy of online multimedia services", Journal of Industrial Ecology, to appear (2013)
[21] D. Schien, P. Shabajee, S. G. Wood, and C. Preist, "A Model for Green Design of Online News Media Services", to appear in the Proceedings of the 22nd International World Wide Web Conference (2013)
[22] Vodafone Group, "Sustainability Report" (2011)
[23] E. Oh, B. Krishnamachari, X. Liu, and Z. Niu, "Toward Dynamic Energy-Efficient Operation of Cellular Network Infrastructure", IEEE Communications Magazine, 56–61 (June 2011)
[24] O2, "O2 Sustainability Report" (2011)
[25] Vodafone Deutschland, "Corporate Responsibility Report 2010/2011" (2011)
[26] Vodafone Group, "Sustainability Report" (2012), available at http://www.vodafone.com/content/dam/vodafone/about/sustainability/reports/2011_12/vodafone_sustainability_report.pdf
[27] GSMA, "Mobile's Green Manifesto", Technical report (2012)
[28] Abidjan.net, "Téléphonie mobile : Les parts de marché de chaque entreprise" (2012), available at http://news.abidjan.net/h/435550.html
[29] Liberation, "La Côte-d'Ivoire, un terrain trop mobiles" (2012), available at http://www.liberation.fr/economie/2012/05/28/la-cote-d-ivoire-un-terrain-trop-mobiles_822054
[30] World Bank, "World Development Indicators" (2009), available at http://data.worldbank.org/data-catalog/world-development-indicators
[31] data base "Electric power consumption (kWh) in Cote d'Ivoire" available at http://www.tradingeconomics.com/cote-d-ivoire/electric-power-consumption-kwh-wb-data.html
[32] Wikipedia, "Vodafone" (2011)
[33] Statistisches Bundesamt, "Erzeugung" (2011)
[34] B. K. Sovacool, "Valuing the greenhouse gas emissions from nuclear power: A critical survey", Energy Policy **36**, 2950–2963 (2008)
[35] World Bank, "Cameroun : Plan d'Action National Energie pour la Réduction de la Pauvreté", Technical report (2007)
[36] BBC News, "How much does it cost to operate a street light?" (2008), available at http://news.bbc.co.uk/1/hi/magazine/7764911.stm
[37] Bundesministerium für Wirtschaft und Technologie, "Energieverbrauch des Sektors Gewerbe, Handel, Dienstleistungen (GHD) in Deutschland für die Jahre 2007 bis 2010" (2011)
[38] J. Malmodin, Å. Moberg, D. Lundén, G. Finnveden and N. Lövehagen, "Greenhouse Gas Emissions and Operational Electricity Use in the ICT and Entertainment & Media Sectors", Journal of Industrial Ecology **14**, 770–790 (2010)
[39] GSMA, "Sub-Saharan Africa Mobile Observatory 2012", Technical report, 2012
[40] M. C. González, C. A. Hidalgo, A.-L. Barabási "Understanding individual human mobility patterns", Nature **453**, 779-782 (2008)
[41] A. Noulas, S. Scellato, C. Mascolo and M. Pontil, "Exploiting semantic annotations for clustering geographic areas and users in location-based social networks", AAAI Workshop - Technical Report, 32–35 (2011)
[42] A. Noulas, S. Scellato, R. Lambiotte, M. Pontil and C. Mascolom "A Tale of Many Cities: Universal Patterns in Human Urban Mobility", PLoS ONE **7**, e37027 (2012)
[43] R. A. Becker, R. Cáceres, K. Hanson, J. Meng Loh, S. Urbanek, A. Varshavsky, C. Volinsky "Route classification using cellular handoff patterns", Proceedings of Ubiquitous Computing, 13th International Conference (UbiComp 2011)
[44] S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, J. Rowland and A. Varshavsky, "A tale of two cities", Proceedings of the Eleventh Workshop on Mobile Computing Systems (HotMobile '10), 19–24 (2010)
[45] D. Quercia, G. Di Lorenzo, F. Calabrese and C. Ratti, "Mobile Phones and Outdoor Advertising: Measurable Advertising", IEEE Pervasive Computing **10**, 28–36 (2011)
[46] J.-C. Lucet, C. Laouenan, G. Chelius, N. Veziris, D. Lepelletier, A. Friggeri, D. Abiteboul, E. Bouvet, F. Mentré and E. Fleury, "Electronic Sensors for Assessing Interactions between Healthcare Workers and Patients under Airborne Precautions", PLoS ONE **7** (2012)

Analysis of New Strategies for Resources Allocation and Infrastructure Development in Côte d'Ivoire by Mapping Telecommunication Densities

Yuk Hui[1], Menjian Liu[2], Pan Hui[3]

Motivation of Research: Our research starts with a motivation to investigate the situation of poverties in Ivory Coast, and tries to find correspondent solutions by focusing on circulation infrastructures (including telecommunication and transport), and inter regional interactions and exchanges. We identify that, one of the problems in under-developed and developing countries is the misallocation of resources and the lack of circulation channels which can effective distribute resources. This two questions are significant in particular for Ivory Coast, since Ivory Coast has a quite high rate of poverty, according to an IMF report[1], the poverty rate has risen from 40% in 2002 to 43% in 2008. In particular in northern regions, 4 out of 5 people are living under poverties[2]The main reason according to IMF is because of agricultural policies, the taxation of prices of principle agricultural products has lowed the income of the farmers, while at the same time the price of primary materials have soared up. We believe that one way other than a large reform in agricultural policies to overcome this question is, to improve circulations and create a the market based on efficient exchange of information and goods. We understand here market as the mechanism to redistribute resources and products in a more flexible manner, and it demands circulation channels such as transport, telecommunication, power, to facilitate its functioning. The other reason dear to Ivory Coast is the war situation, for reasons of rescuing , fleeing, food and emergency aid supply, it is necessary to reconfigure the current infrastructures, interregional communications, inter-city communications in order to enhance mobility and exchange between local cities and regions. IDA (International Association of Development) and IMF all hold the view that the long-lasting situation of poverties is due to the fact that the current administrative decentralisation is lack of coherent strategy [1]. Therefore, we take this as the entry point for our research, and by the end we hope to show how these communication data from Orange can be used to help figuring out new strategies.

Difficulties and Solutions: The difficulties lie in two parts. Firstly missing data in the data sets from Orange (for example, in Data Set 1, 7 antennas don't not have exact GPS data, 15 antennas without records; in Data Set 3, 15 out of 250 sub-prefectures without data, especially a lot of data

1 Centre for Digital Cultures, Leuphana University Lüneburg, Germany

2 University of Electronic Science and Technology of China

3 Department of Computer Science and Engineering, The Hong Kong University of Science and Technology

marked with -1, etc) make it difficult for us to undergoes very accurate calculations and hence produce precise trends and predictions with the given data. Then the question turns to be: how could we find co-relations/comparable parameters between these sets of data and other data set which are produced rather on macro-levels, e.g. maps? The second difficulties which is also relevant to the former is that, due to the unstable political situation in Ivory Coast in the past 10 years, the statistics on various aspects of the regions has varied from year to year, and this makes comparison difficult due to the lack of consistencies, for example the amount of production of agricultural products. These two problems will last in the future concerning the use of communication datasets. Our intuitive solution is to visualize the data by mapping them with the exact cartography of Ivory Coast according to the GPS metadata provided. We find that through mapping, we can avoid the problem caused due to inexactitude of data on one hand (meaning that even with missing data, we can still have an overview of some interesting tendency and phenomenon); on the other hand, we can also compare them with some of the current mapping projects that allow us to arrive at some fruitful observations.

Methodology: In this research we mainly focus on two major datasets 1 and 3. We choose these two particular ones, since we are more interested in telecommunication density and how telecommunication density reveals certain *needs* and *saturations* of infrastructure in difference regions and cities. We are also interested in the connection between cities, and the amount of communication between them, we want to use them as indicators that allow us to re-imagine the flow of resources. We map the communication data according to three levels by using the GPS metadata : regions, sub-prefectures, antennas. Dealing with these two sets of data, we have already made two assumptions, firstly there is close relation between telecommunication density and poverty (we will go into more details in the following sections). Secondly we assume that the locations indicated in set 3 concerning the calls of an individual reflect the mobility of this individual. This second assumption is not really accurate, since the switch in locations doesn't necessarily means physical movement, but we tend to think that it indicates certain connection or affinity between these two locations/sub-prefecture.

For the perspective of development studies, it has been debated since wrong whether natural resources are good or bad for development partly due to civil war, regions with diverse natural resources become the targets of military siege[3]. We take a different perspective, we tend to believe that once if there is better circulation channels, alternative economies can be developed. This is rather an old Saint Simonian idea (that imagines the society as a totality of physiological circulations) rather than the current neo-liberal concept. Hence we believe that circulation channels

are crucial to the formation of a more transparent and effective market, by which resources, properties can be better distributed. We want to look closely at the current circulation routes especially road facilities within the country. New circulation routes will be able to vitalize some of the under-developed areas. We also believe that inter-cities/sub-prefectures interactions and exchanges must be deliberately encouraged and facilitated, this is like "urban acupunctures" after the main circulation routes are fixed, smaller veins between cities and areas will be "opened" to facilitate information flow, products/resources exchanges.

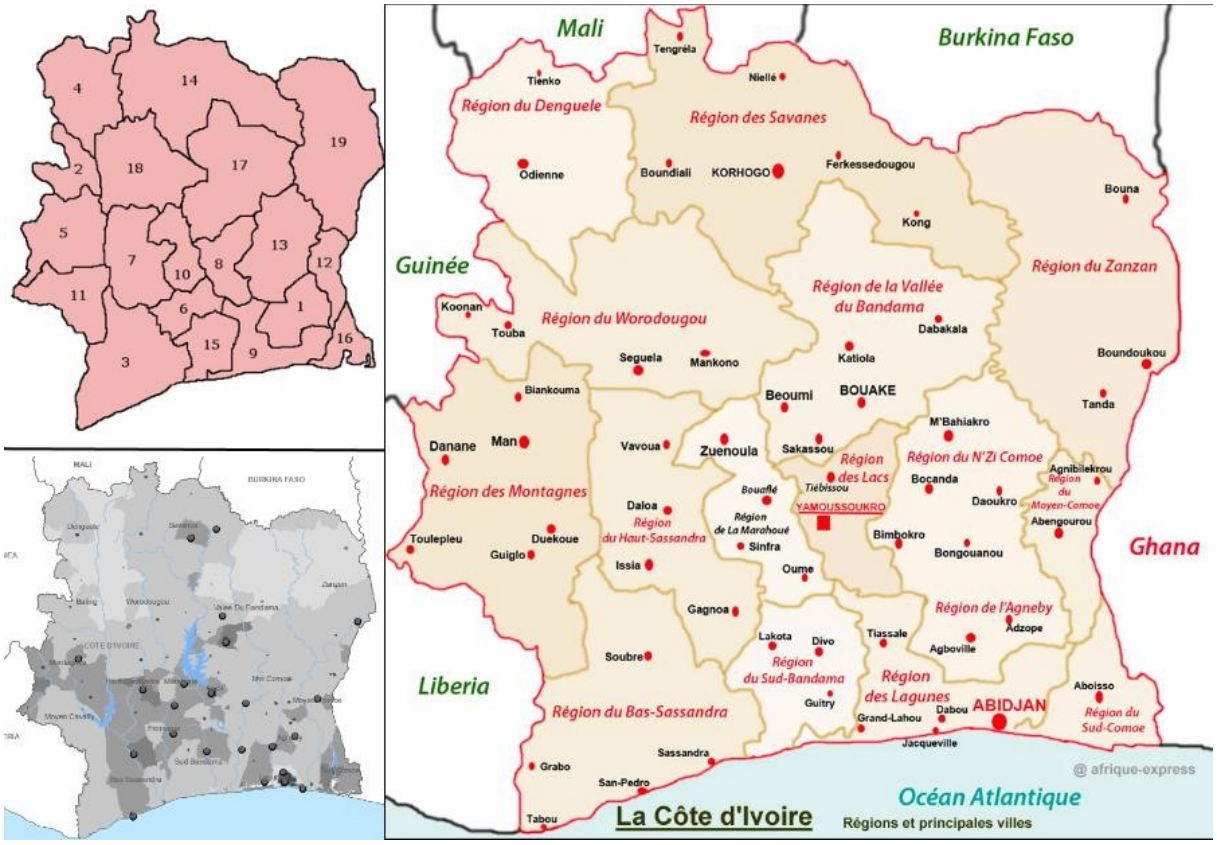


Figure 1: Maps showing (top left) number of regions(source: wikipedia), (lower left) population distribution(source:AICD), (right) name of regions and major cities (source:afrique-express.com) You will have to refer to these maps often.

Result and Discussions

1. Telecommunication Density and Distribution of Poverties

Based on Google map, we plot the geographical locations of 1216 antennas (originally 1231 but 15 of them don't have records). We divide the amount of communication into 20 levels, 20 the highest and 1 the lowest [Figure 2]. Our findings maps the telecommunication density in terms of the distribution of the antennas, and the amount of telecommunication processed by these antennas,

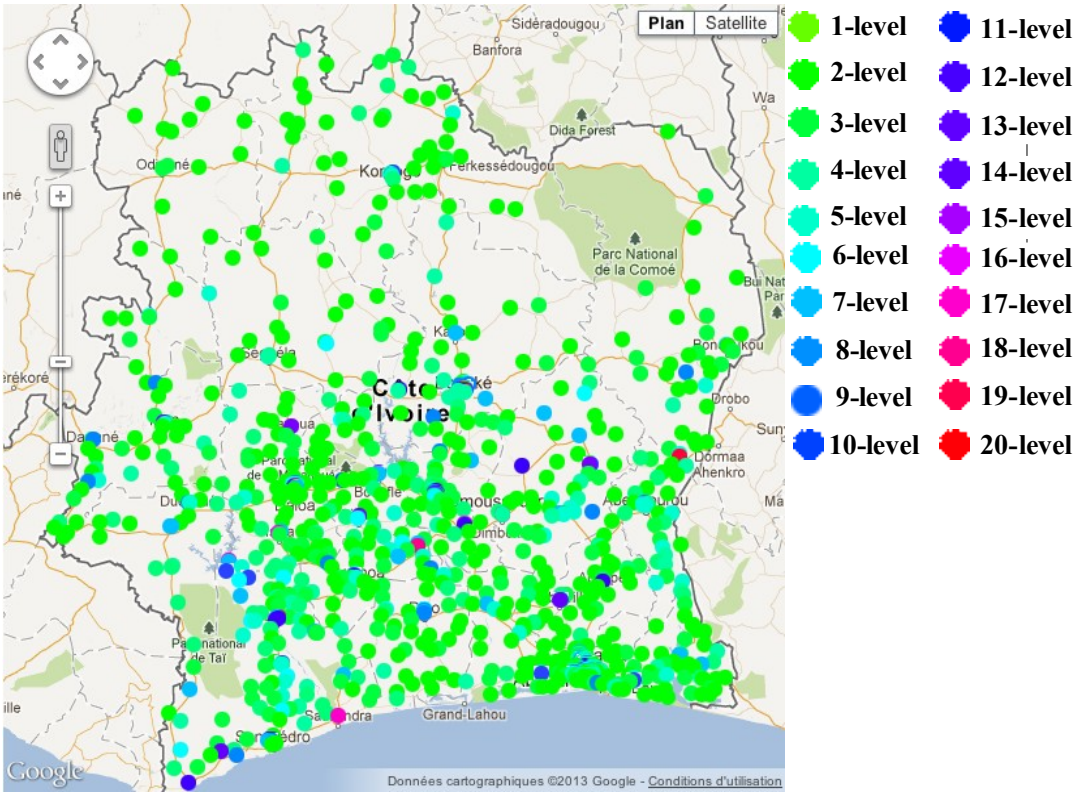


Figure 2: Distribution of antennas throughout the country

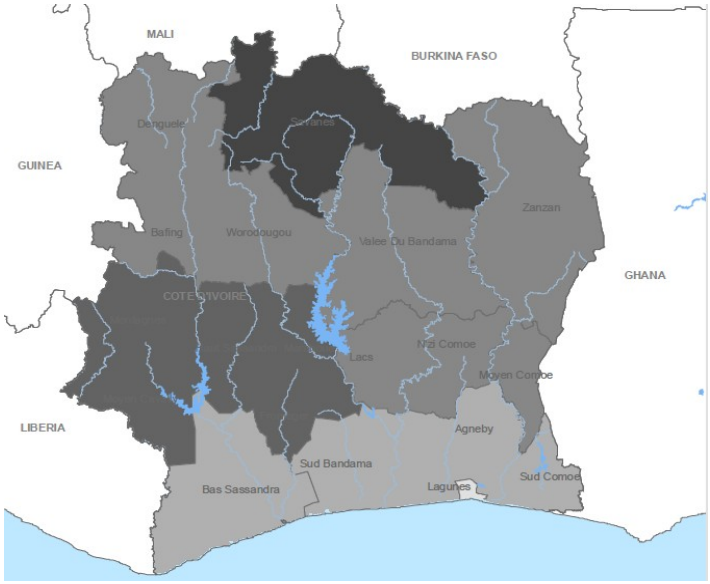


Figure 3: Distribution of poverties throughout the country

shows a similar pattern with the distribution of poverty in the country[Figure 3]. It comes naturally that the wealth centres in the south part of the country, especially around the major port city Abidjan, then diffuses slowly to the centre of the country (Region 7, 10, 8, 13), and drop sharply in the northern and eastern parts. This also shows that resources in this area is properly too condensed, since resources don't not easily flow to the northern part (Region 4, 14, 19).This can be understood that these northern regions has much lower population density compared with region 9 where Abidjan locates. For example region 14 has the lowest population density in the country (10.49 per km), while region 9 has population density up to 285.4 per km. Nearly 4 out 5 persons living in the northern area were below the poverty line in 2008[2]. This uneven distribution of population densities and resources though justifies each other, it also shows the general problem in the development of Ivory Coast. We listed 10 antennas with lowest usages[Table 1], and with surprise, we found out that *8 out of 10* are located in Abidjan, this could be only the problem of the previous investment of Orange, but it also seems to us that other development project may have the same tendency to excessively invest in Abidjan. This over-saturation of investment cannot be simply justified. We propose rather to look at other new routes and routing mechanisms, which can help to distribute resources in the country in a more meaningful sense.

ATN ID	Total No. of Calls	Total amount of Communication Duration	Geographical location
251	5675866	557613721	Bouaké, Region 17
481	3632735	539905914	Soubré, Region 3
44	4390603	513476137	Agnibilékrou, Region 12
331	4153570	499839206	Daloa,7 Region
860	3662156	496706698	Oumé,6 Region
268	4242147	469512528	Soubré, Region 3
230	3539302	453499074	Bondoukou, Region 19
956	3514524	452142814	Sassandra, Region 3
367	2689297	449102016	Dimbokro, Region 13
206	2826984	427626045	Toumodi, Region 8

Tableau 1: List of 10 least used antennas, 8 of them are in Abidjan

2. Development of New Circulation Channels and Roads

We also listed 10 antennas with highest usages, and we identifies that except ID 956 is located in city along the coast, the rest of them locate in land[Figure 4, Table 2]. We understand that there are high demands from the central part of the country, this to us characterizes some movements in terms resources and service demands from the south toward north. In particular we notice that there are 3 top antennas in region 3, this shows that it is an area of potential acting as hub for neighbouring regions.

ATN ID	Total No. of Calls	Total amount of Communication Duration	Geographical location
998	5626	351693	Abidjan
732	3996	332806	Abidjan
742	3796	314303	Abidjan
740	2058	204819	B 102 Road
1233	523	79365	A7 Road
737	555	55445	Abidjan
733	425	36762	Abidjan
846	73	8938	Abidjan
811	11	734	Abidjan
1234	1	94	Abidjan

Tableau 2: List of 10 antennas of highest usage

Figure 4: Distribution of the top 10 antennas on the maps

We further use the geographical location data provided by the antennas to look at the distribution in the sub-prefecture level. We try to calculate the amount of conversations happened in 255 sub-prefectures[Figure 5]. We listed 10 sub-prefectures with least communication usage, we found that *5 out of 10* locates in region 4, one of the poorest region we have identified above[Table 3]. Region 18 and 11 also sub-prefectures that have relatively low usage of telecommunication. We can observe from the map that further than region 4, 14, 19 at the north of the country, region 18 and 11 are relatively under-developed. This form an upside down inverted "L" shape on the map, and it indicates to us that we should pay special attention to these regions.

Prefecture ID	No. of Calls	Geographical Location
127	51206	Samatiglia - Kouto Rd, Region 4
157	48902	A701, Toulépleu, Region 11
45	46257	B412, Katiola, Region 17
206	43392	Séguéla, Region 18
132	41203	Samatiglia - Kouto Rd, Odienné, Region 4
241	37772	A7, Region 2
128	35881	A7, Region 4
161	23579	A701, Toulépleu , Region 11
130	20413	A7, Bako, Region 4
129	13772	Denguele, Region 4

Tableau 3: List of 10 sub-prefectures with lowest telecommunication density

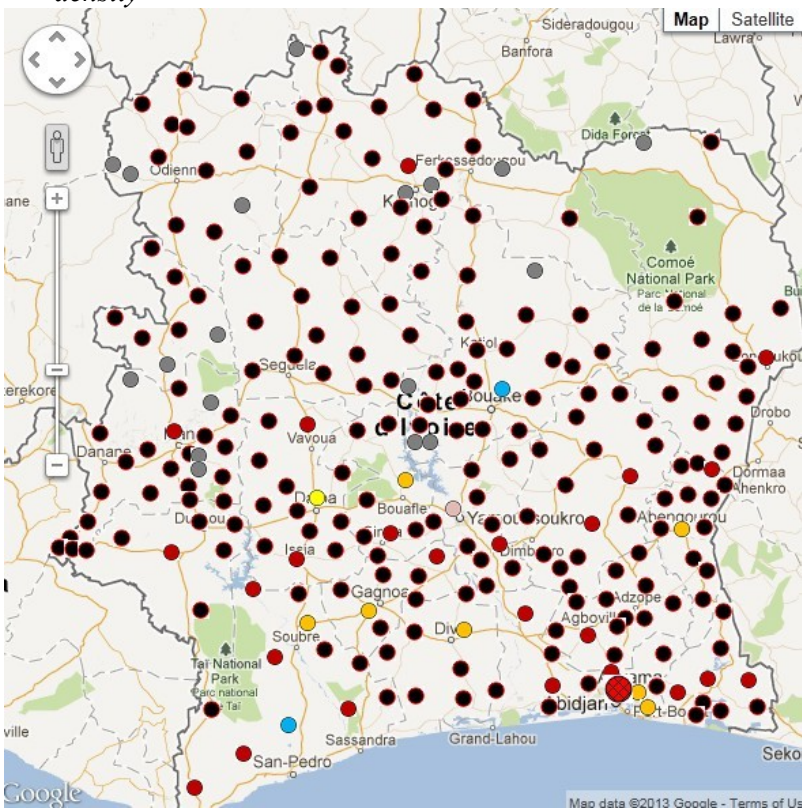


Figure 5: Distribution of telecommunication density in 255 sub-prefectures, color scheme in decending order , means 0 records, sub-prefecture 60 occupies 44% of the total call time.

In order to make use of these discoveries we compare our map with that of distribution and condition of roads and poverty provided by Africa Infrastructure Country Diagnostic (AICD), It seems that the connection between across the country is quite poor, while from central west (region 11) to region 4, there is very good connection, while the connection between region 4 and region 14

doesn't seem quite stable. We think that probably investment in the road infrastructure between region 4, 14, 19, can help to redistribute resources in a more productive way. Our speculation doesn't come from no where. If we look at some of the agricultural products, for example tomato[Figure 6], they mainly come from the region 14, 19, and most of these agricultural products go directly to Abidjan, we understand that is partly due to the fact that Abidjan has the largest port in the country for exportation. According to the statistics of usage, as we expect, the road condition from the north to Abidjan is rather congested even though the volume is considerably largest[Figure 8]. We try to identify new routes that can facilitate circulations throughout the country, that is from 14 and 19 to region 4 and then to region 3 via region 18 and 11. Since we have spotted that regions 3 has rather high demand of telecommunication, while region 11 next to it is rather poor, and between region 4 and 11 there is very good road infrastructure and condition[Figure 7]. In particular in the northern area, there are many natural resources such as mines. This seems to us, could be a new way to connect different regions without further excessive investment in road infrastructure from Abidjan to the south. Instead we suggest to think of the following cities as future hubs: Bouaké、Soubré、Agnibilékrou、Daloa、Oumé、Bondoukou、Dimbokro、Toumodi.

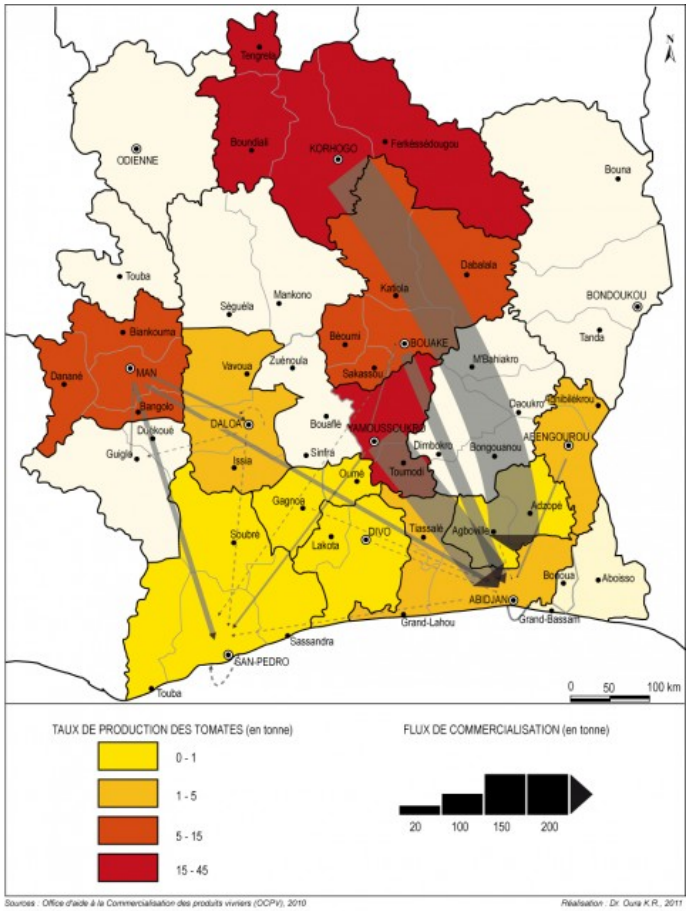


Figure 6: Map showing the distribution of tomato products and logistics(source:[3])

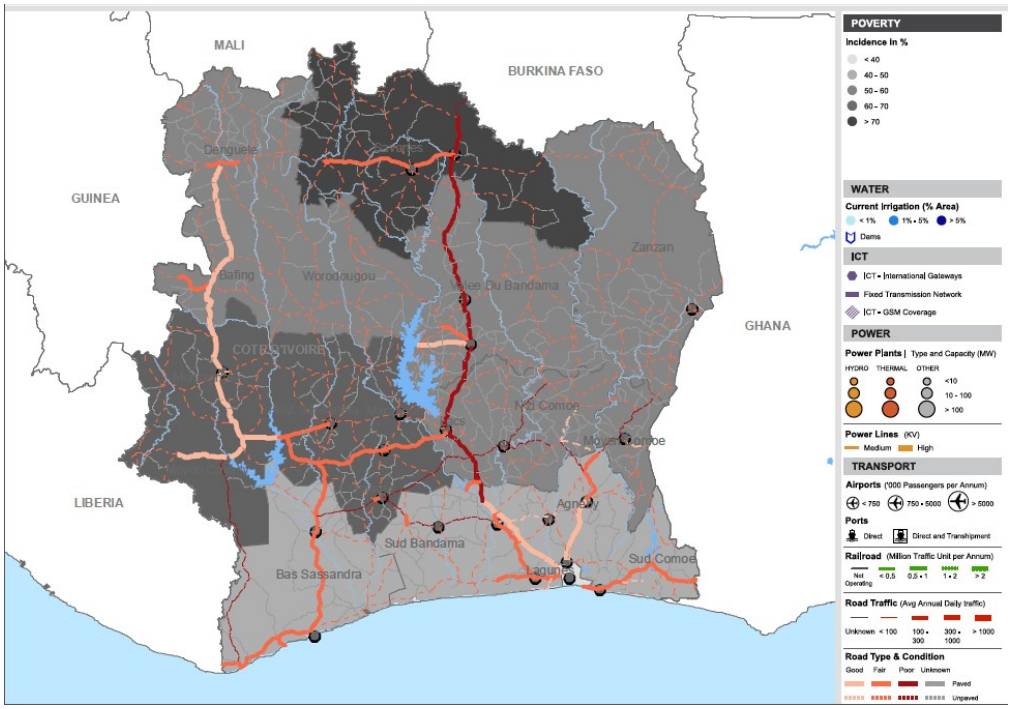


Figure 7: Map shows the conditions of the main roads and the location of airports (Source : AICD)

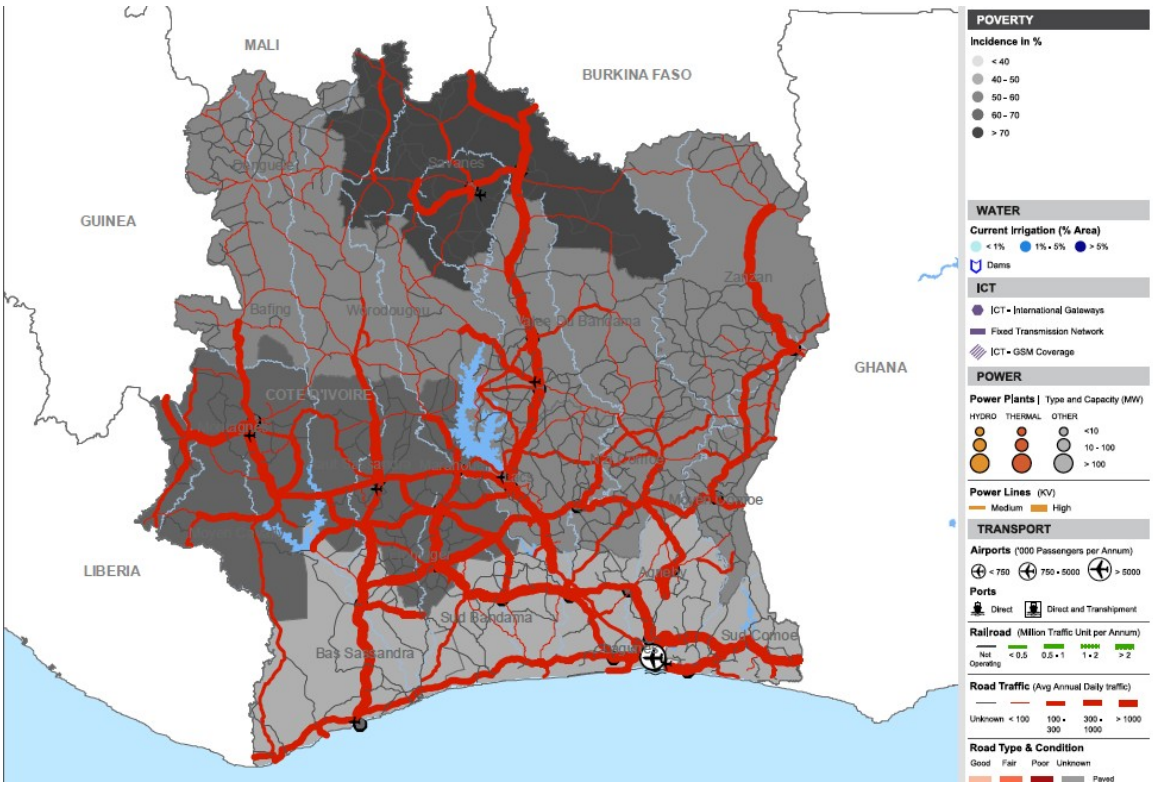


Figure 8: Map showing the usage of roads in the country

3. Inter-City/sub-prefecture Interactions and Mobilities

Besides of inter-regional infrastructures, we further think of interactions and communications between cities or sub-prefectures. We think that by analysing the conversational data (set 3) we will be able to identify some potential connections between nearby cities. For if there is already frequent telecommunication activities between cities, that means certain networks either social or communicational are already established, hence these cities will be able to exchange different materials including resources and information effectively. These are potential cities of collaboration and interchanges. Here is the top 10 inter-sub-prefecture communication list we have computed.

Prefecture	Geo-location	Prefecture	Geo-location	No. of moves
60	Cité Toucan	61	G 50	1211807
60	Cité Toucan	64	R8	625781
117	A4	123	Soubré	603364
60	Cité Toucan	63	A3	442928
60	Cité Toucan	198	A 100	417652
159	A7	237	A7	329841
138	A5	140	A5	187170
116	A5	122	San-Pédro	167933
22	A6	58	A6	155161
190	A 100	192	A100	148808

Tableau 4: Table shows top 10 sub-prefecture with highest frequency of communication

In order to further analyse the mobilities between sub-prefectures, we take each sub-prefecture as a vertices, and the telecommunication connection between them as edge, the value of the edge is indicated by the frequencies of moves between these two cities. The chart we obtained contains : 255 vertices, 18 isolated vertices, and 12438 edges. After removing 18 isolated nodes, we were able to produce a graph between consisting of the remaining vertices. We then are hence able to compute the degree of connections of each sub-prefecture, table 9 shows the top 10 corresponding sub-prefectures. It indicates that these 10 sub-prefectures are hugs of the current network of the country. For example sub-prefecture 60 is in Abidjan, and its degree value is 236, meaning this area is able to reach 236 other sub-prefectures. But besides of Abidjan, by using this graph we are able to find out more hugs that can be further developed in the future to facilitate circulations of resources, information, human mobilities.

Figure 9: Top 10 sub-prefecture with highest degree of connection

Prefecture ID	Degree
60	236
58	229
39	223
138	219
69	217
64	213
144	212
250	206
122	205
63	202

Tableau 5: Table lists 10 nodes with higest degree of connection

Conclusion and Future Researches:

By using the telecommunication data and existing map services, we are able to map the current situations and weakness in Ivory Coast. Dealing with these data, it seems to us that development in Ivory Coast is highly uneven, especially the distribution of wealth and resources, also telecommunication facilities. We proved our early hypothesis that telecommunication density reflects directly distribution of poverties and resources allocations. It seems to us that in order to find a better plan for infrastructure improvement, it is necessary to develop new circulation channels apart from the current situation that Abidjan as the centre of the country. As discussed in this article, 8 of

the 10 least used antennas locate in Abidjan, meaning that resources are mismatched and apparently it has been the case in Abidjan for long. We suggest hence to take a new route that runs through regions 19, 14, 4, 18, 11, 3, since the northern regions suffer a lot from poverties, while southern regions like 3 are experiencing more and more demand of communication, and provided that the road condition is excellent now, this new route will be able to lessen the burden of the existing infrastructure from north to Abidjan. We also identify some potential cities/hubs by using the antenna to antenna communication data to create a graph of the degree of connections of 255 sub-prefectures. With this we are able to suggest some cities that should be further invested and developed, inter-city/sub-prefecture activities has to be encouraged. We propose to specifically consider more intense investment in the following cities: Bouaké, Soubré, Agnibilékrou, Daloa, Oumé, Bondoukou, Dimbokro, Toumodi that shows great potentials from our maps. The limitation of our research is that we only focus on Data Set 1 and 3 due to the time frame of this research, we believe that there must be much more that one can learn with these data sets.

References:

[1] Côte d'Ivoire : Note consultative conjointe sur le rapport d'avancement du document de stratégie de réduction de la prauvreté

[2]Vivien Foster and Nataliya Pushak, Côte d'Ivoire's Infrastructure: A Continental Perspective, AICD Country Report, 2011

[3]Andrew Rosser, The Political Economy of the Resource Curse: A Literature Survey, The Institute of Development Studies, 2006

[4]Raphaël Kouadiov, Extension urbaine et protection naturelle : La difficile expérience d'Abidjan, in Vertig 0, La Revue électronique en sciences de l'environment, Volume 12 Numéro 2, Septembre 2012

Social, Disconnected or In between: Mobile Data reveals urban mood

Eiman Kanjo[1]Nour El Mawass[1] João Pedro Craveiro[2] Fernando M. V. Ramos[2]

College of Computer and Information Sciences, King Saud University (Riyadh, Saudi Arabia)
ekanjo.c@ksu.edu.sa

Universidade de Lisboa, Faculdade de Ciências, LaSIGE (Lisbon, Portugal)
ffvramos@di.fc.ul.pt

Abstract. Underdeveloped countries are facing serious urban issues that negatively affect urban life. The poor management of urban creation and low resources can cause many challenges for urban planners and construction actors. Life standards differ from country to country and even when using objective parameters like salary level, we are faced with the variant importance of these parameters depending on the country and the people in question; the design of a better city does not have a single global formula: planners have to understand what are the people needs in order to reach a suitable planning. In this sense, mood measuring can be both a parameter and an output of urban planning; it forms a continuous evaluation process, and serves as a feedback for future development. In this paper, we argue that the mood of the city, inferred from mobile datasets – the up surging replacement of traditional census methods, can be a strong reflection of the happiness of its population and their satisfaction toward their urban environment. We specifically explore the possibilities of modeling the public city mood based on the mobile datasets of Abidjan, economic capital city of Cote d'Ivoire. Preliminary results show collective human communications in selected areas of Abidjan reflecting social mood which could provide additional insights into how collective social activities shape urban systems.

Keywords *- Mobile phone data; Urban development; Mood, Affective Computing, data analysis.*

Introduction

Environmental psychology [1] is an interdisciplinary field focused on the interplay between humans and their surroundings. The field defines the term environment broadly, encompassing natural environments, social settings and built environments in a city. It's priority to solve complex environmental problems in the pursuit of individual well-being within a larger society. In a recent project [2] we studied the link between people emotion and outdoor places, using subjective as well as objective data – in particular, measurable, physiological response data. The ever-wider diffusion of mobile devices makes the idea of a 'smart city' become more and more realistic and concrete. Linked to this is the hope for a more livable, more efficient and greener city, which is increasingly oriented towards people's needs. However, our experiments are limited to the number of users we can recruit and it will be difficult to scale the experiment to cover a whole city or region.

Compared to our limited datasets, mobile datasets offer a wider coverage, which is easier to track in both time and space. Mobile phone data is ubiquitous, abundant and continuously generated; therefore, the analysis can reach a population that is difficult to reach in traditional data gathering methods, and can be easily expanded to future observations. Mobile operators understand the power of mobile phone data and they are willing to share their mobile data with researchers. Using mobile data as the substance of statistical studies is a growing trend; recent results are promising and cover a wide variety of fields traditionally reserved for census drawn data [3][4][5][6].

Sharing mobile data, however, poses security and confidentiality issues and although some operators are satisfied with anonymizing callers' Ids; [7][4] others believe this precaution is not enough; consequently, many of them prefer to provide broader Erlang data [3][8].

With the help of a network mobile dataset, it might be possible to look closely at patterns of happiness and mood swinging through a city. We will be interested in looking at call durations and times, and also to look for specific pattern of interaction related to emotions or particular places. The idea is to analyze activities within each network cell and see how they differ in terms of mood. These "mood signatures" might suggest a direct link between a certain mood and a place or time

of an event, which will lead to the detection and classification of Abidjan regions based on their mood signatures.

For example, the New Year's festivities in a stadium at *Le Plateau* of Abidjan ended up with a stampede. The tragic event could have been analyzed in real-time using mobile dataset. The unexpected number and duration of calls could have been automatically visualized and classed the area as "depressing".

Related Work to Urban Mood Monitoring

When solving problems involving human-environment interactions, whether global or local, one must have a model of human behavior or mood that predict the environmental conditions under which humans will behave in a decent and creative manner. With such a model, one can design, manage, protect and restore environments that enhance reasonable behavior, predict what the outcome will be like when these conditions are not met, and diagnose problem situations.

In this paper, we made simple premise based on Network theory when applied to social interaction:

> • *People are connected, therefore their health is connected.*
>
> • *People are connected, therefore city economic wellbeing is connected and thus, they are in good mood.*

In other words, the more people talk and interact over mobile phones the better their well-being.

In a recent paper [9], researchers performed a sentiment analysis of tweets publicly broadcasted on Twitter between August 1 and December 20, 2008. They speculate that twitter characteristics make it an interesting source to investigate public mood and emotive trends. They compare mood trends to fluctuations recorded by stock market and crude oil price indices and to major events in media and popular culture and found that social, political, cultural and economic events are correlated with significant fluctuations of public mood levels.

The current scope of the work is to propose a mood model and briefly explore how this model is evolving over Abidjan spatiotemporal map. Our aim is to measure level of social interaction - more broadly good mood - using a non-traditional objective measure easily inferred from mobile datasets. We therefore use the non-individual aspect of a mobile dataset as a tool to understand the punctual and broader mood status of the population it represents. Mood analysis over a long period of time is able to reveal patterns and routines of social activity, and later explain the link between special aspects of urban life and their mood signatures.

We argue that reports of good mood are the best clue of well-being. Normal work areas, normal education life, happy residential neighborhoods, crowded shopping areas, all have their unique signature. Comparing signatures between different areas that share the same main activity such as Business, Entertainment, Shopping, Relaxation, Social interaction, Residential...etc., can help urban planners and economists understand the impact of design on people's mood.

City of Abidjan

Usually seen as a haven of peace and prosperity in West Africa, Cote d'Ivoire - known for its cocoa industry and its high living standards, is lately facing major political and economic problems[10][11]. Abidjan, the former capital city of Cote d'Ivoire and its current economic capital, is the most modernized city in the country with an area of 2119 km² and 13 communes[12]. This western style city has benefited from good planning, housing policy and infrastructures achieved by French colonialists and later on by Ivorian governments.

With more than 160 different registered nationalities [13], Abidjan population estimated at 6,783,906 (2011) is rapidly growing and represents ¼ of the whole population of the country.

The main industrial and economical city (70% of the industries and some 60% of the country's workforce), is considered as an important place in West Africa based on economic, diplomatic and social roles. Many international bodies such as, UNICEF, WHO, WFP …etc. and worldwide companies such as, Microsoft, Alcatel …etc., established their regional headquarters there, in particular in *Le Plateau*.

Nevertheless, the 'Pearl of Lagoons', as it's referred to, was home for several civil wars in the last decade; the city is facing serious urban development issues that are urban environment degradation due to the overpopulation, social disorders, mismanagement and unplanned areas development. Overall, safety and security are what define wellbeing and the city will be much happier when less violence is triggered in its neighborhood. Clean water, streets, and houses, organized traffic, and safe school trips are probably what make people feel better about their city.

Abidjan offers an exceptional diversity of neighborhoods. Examining the mood differences between its different communes may help to understand how environment, urban design, and standard of living are related to social activity and

happiness. Linked to this is the hope for a more liveable, more efficient and greener city, which is increasingly oriented towards people's needs. The planning and design of urban spaces and cities can greatly enhance people's lives. They can also make people miserable. Researchers in urban economics, as elsewhere, need to get out into the field and directly assess the well-being and behavioural consequences of different places and spaces.

In this paper we examine people social interaction as indication of their mood the more they communicate the better is they wellbeing and hence their mood.

As we will see from the data, increment in phone calls and duration is directed to where people live.

Dataset

This project is a part of the D4D challenge [15]. In the context of the challenge, we received various datasets from Orange Cote d'Ivoire, and we selected the antenna-to-antenna traffic dataset that's relevant to our work. This dataset is based on anonymized Call Detail Records (CDR) of phone calls between five million of Orange's customers in Cote d'Ivoire in the 5 months period between December 1, 2011 and April 28, 2012 [8]. The call activity is provided on an hourly basis, and contains information on the total number and duration of calls exchanged between two antennas within an hour. Communication between Orange customers and customers of other providers have been managed by summing these calls in a single record for each hour and assigning -1 to the originating/receiving out-of-network antenna.

Another dataset provides Ids and geographic location (latitude and longitude) of 1231 antennas. The data related to some of these antennas has been wiped from the former dataset. Interpolating values in these regions is weakened by the fact that most of the wiped data is associated with regions of significance. This has led us to drop these regions from our analysis in order to avoid inaccurate data Interpolation.

Data Analysis

In this project, we carried out some analysis on the variability of the number of calls and call duration in 6 Abidjan communes, these are:

1. *Residential area in Marcoy*
2. *Residential area in Adjame*
3. *Marche de Marcoy (Market)*
4. *Super Hayat Shopping Centre, Cocody*
5. *Industrial Zone in Koumassi*
6. *Stade Félix Houphouët-Boigny, Le Plateau*

We tried to determine whether and how these data might provide new and useful understanding of urban social mood at different temporal scales.

Using the first CDR dataset on the 7th of December 2011, we have created a sequence of heat maps each illustrates the number of phone calls over one hour starting from 1am till 12 pm respectively. Selected hours from 1 am to 9 am are shown in Figure 1. Red color indicates more phone calls have been made in this hour. The heat maps yield to believe that phone calls are on increase from 5 am and at its peak is around 10am.

Figure 2 and 3. show an interesting observable fact about Abidjan which is the distribution of weekend traffic is less than weekday traffic by a small margin. It's usually expected to observe a noticeable drop in traffic drops during weekends.

As expected, graph 6 (*Le Plateau*) has shown less traffic out of working hours since this commune is considered as the business centre of Abidjan with tall buildings and it's more modern than other communes. The same pattern can be observed in graph 5 which illustrates traffic over a week in the industrial zone of *Koumassi*. In graph 1 and 2, it is visible that phone calls are on the increase from 5 am till about 11pm when it starts to drop.

Figure1. ***Heat maps each representing number of phone calls over one hour starting from 1am till 9am respectively. Red color indicates more phone calls have been made in this hour.***

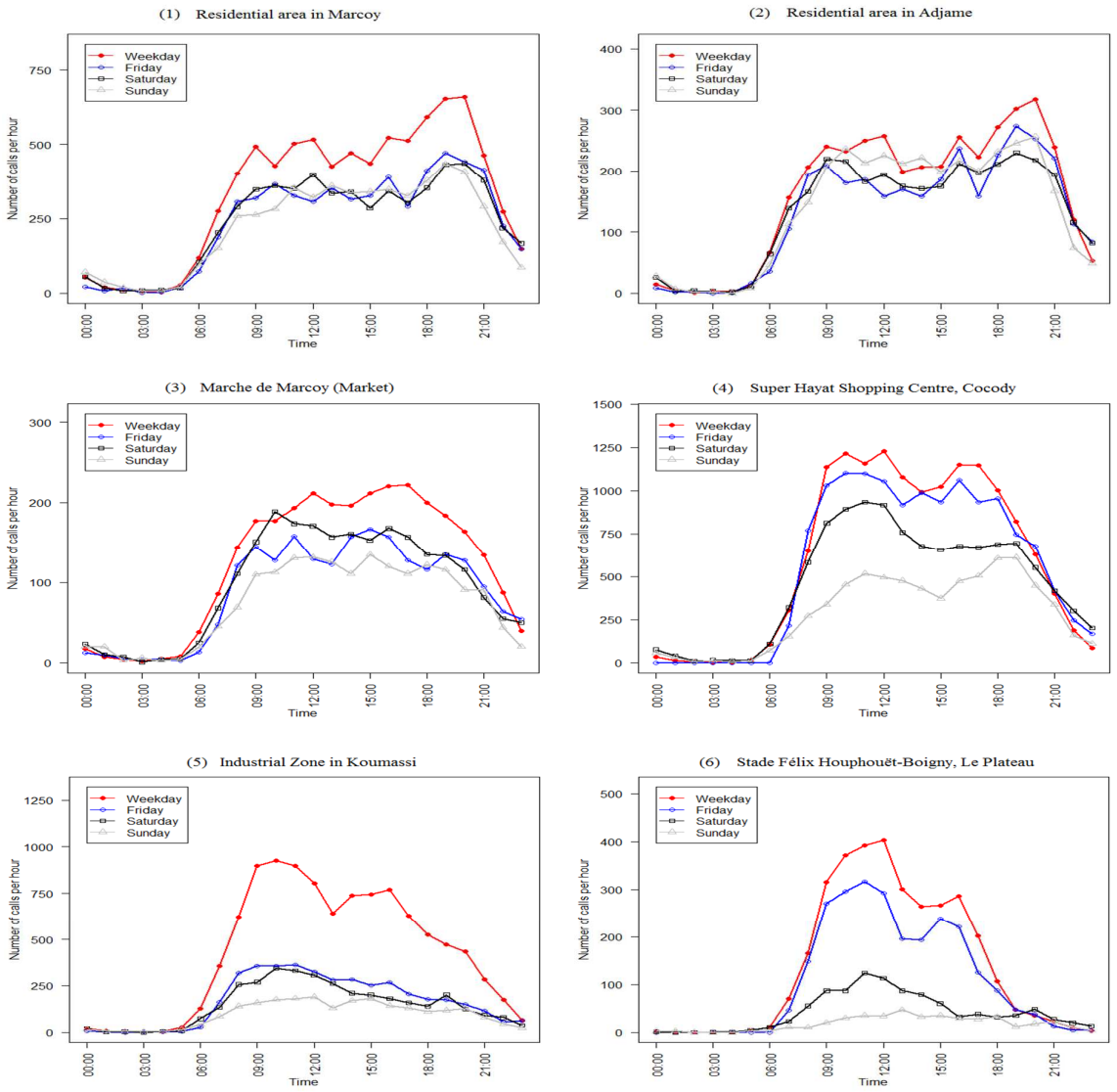


Figure2. ***Traffic plot of the 6 regions respectively. Each graph shows 4 plots representing weekday in red, Friday in blue Saturday in black and Sunday in grey in the week between 12/12/2011 and 18/12/2011***

To compare the number of calls between different regions on the same day, we use Abidjan map with overlay of circles and ovals, see figure 4. The radios of each circle and oval scales with the number of calls between *19/12/11* and *25/12/11*on a particular day. Each color corresponds to a day of the week. These circles are pin on the corresponding region of Abidjan. We find *Cocody* by far more active and vibrant despite the fact that less people live there due to the wealthy nature of this region.

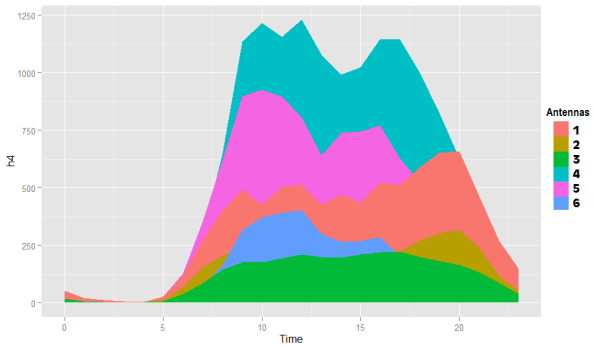


Figure 3. ***Distribution of traffic from the same 6 regions on Wednesday 14/12/11.***

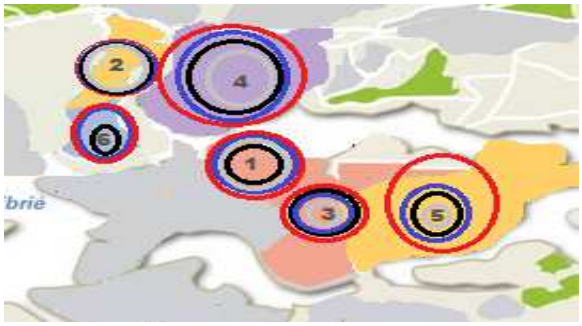


Figure 4. ***Map giving visual representation of traffic in the same 6 regions. Wednesday in red, Friday in blue, Saturday in black and Sunday in grey.***

Next, we look at call duration in relation to number of calls. We define the "ratio"r_i^t as the average duration of one call at antenna *i* during one hour. r_i^t is given by the ratio of the duration over the number of calls at hour t.

$$r_i^t = d_i^t / ncalls_i^t.$$

As shown in figure 5, the ratio between 22:00 and 6:00 is obviously higher than the rest of the day but it seems sporadic and does not have any clear pattern. This observation can be expanded to the majority of antennas. However, the number of calls during this period is very low compared to the day's activity, and it isn't possible to draw any pertinent conclusion based on the data collected at night. This data appear to solely reflect a fragment of the population with no lasting or repetitive effect. The shrinking of traffic window has been supported by the low signal over noise ratio during night's hours in other studies.

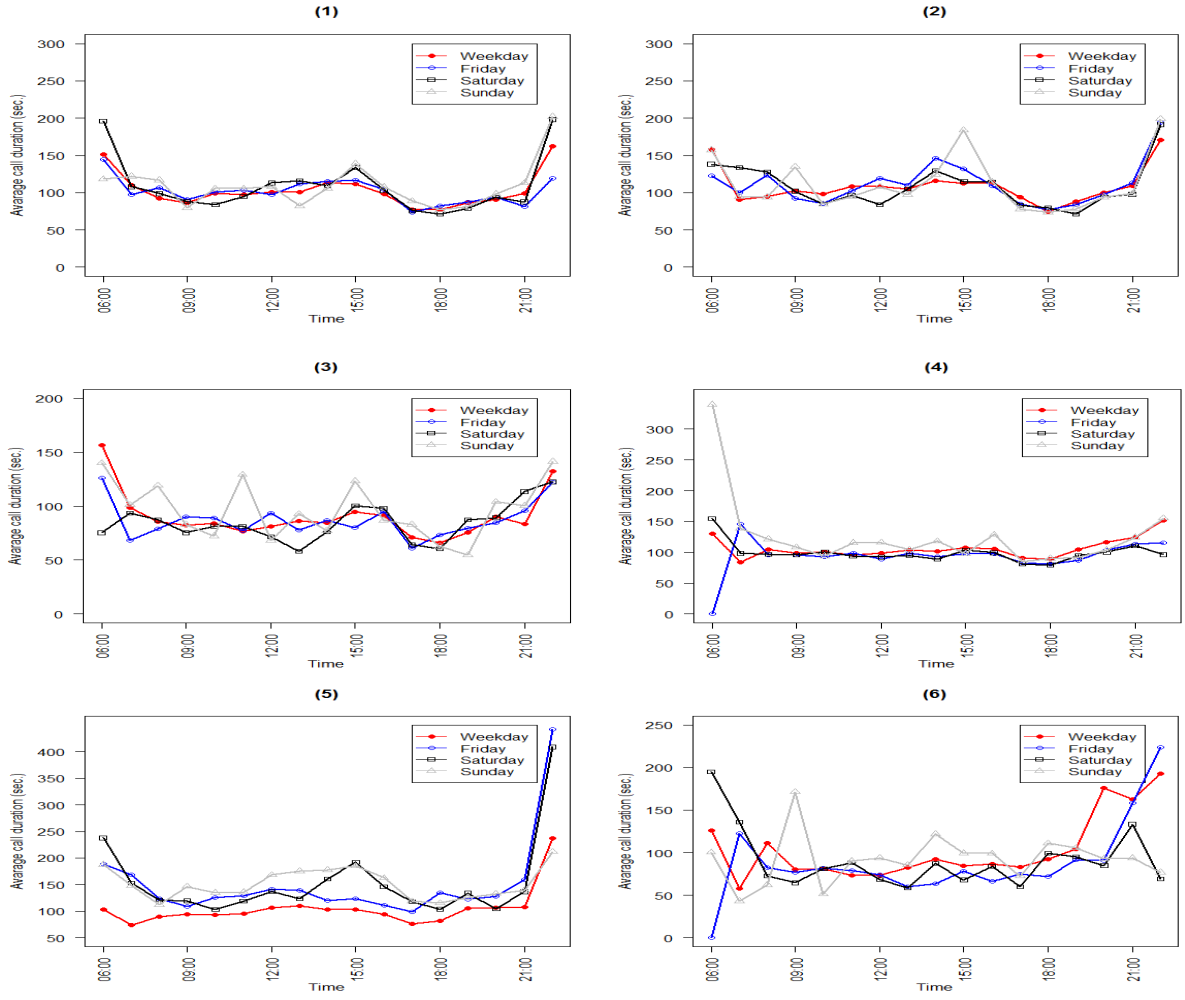


Figure5. Plots showing ratio of the duration over the number of calls in each region between *12/12/11 and 18/12/11.*

Mood Index and map

We define the mood index of the city as follows:

$$I_{\text{mood}}^{t} = \sum_{i=0}^{N} b_i c_i^t r_i^t$$

Where:

- I_{mood}^{t} is the function that gives the social mood in Abidjan for a given time t.
- t is the hour of day at which the function is calculated, $t \in \{0,..,23\}$.
- N is the total number of antennas in Abidjan. The sum is done over all the antennas a_i in Abidjan area where $\in \{1,..,N\}$.
- b_i is a discrete value indicating whether the antenna is located in a residential, a business or an entertainment area, $0<b_i<1$. We used Google maps and other sources from the web to come up with this value.
- c_i^t is the normalized value over the space of the number of calls of antenna a_i at time t;

$$c_i^t = \frac{ncalls_i^t}{mean_{j\in\{1,..,N\}}(ncalls_j^t)}$$

where $ncalls_i^t$ is the number of calls of $a_i$antenna during hour *t*.

- $r_i^t$is the ratio of the duration *d* over the number of calls at the hour *t*,

$$r_i^t = d_i^t / ncalls_i^t$$

We are still in the process of refining the model and illustrating the results in temporal and spatial visual representations. Figure 6 shows an example graph with 4 plots representing 4 days mood data across Abidjan as direct results of the mood function.

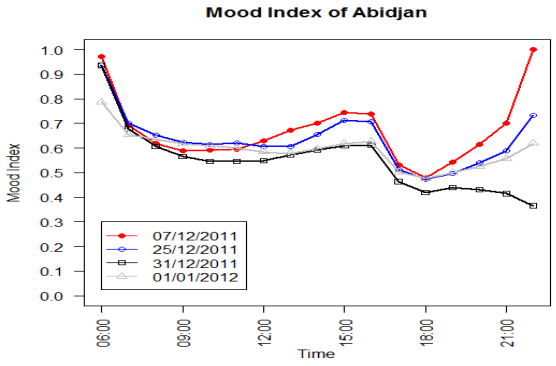


Figure 6. ***This graph illustrate the product of the I_{mood} function across a random day (07/12/2011), Christmas day, the New Year eve, and the New Year day.***

In order to show the a mood map representing our mood model we have used an advanced geostatistical procedure called Kriging [14] that generates an estimated surface from a scattered set of points with mood at mast locations z.

Kriging is similar to IDW in that it weights the surrounding measured values to derive a prediction for an unmeasured location. The general formula for both interpolators is formed as a weighted sum of the data:

$$\hat{Z}(s_0) = \sum_{i=1}^{N} \lambda_i Z(s_i)$$

where:

Z(s$_i$) = the measured mood value at the *i*th mast location.
λ_i = an unknown weight for the measured value at the *i*th location
s_0 = the prediction location
N = the number of measured values

However, we are still working on the spatial representations of the mood data, and we expect the outcome to be similar to figure 7. In addition, we are working on other map representation of mood data including Voronoi tessellation.

In this work, population is not equally represented in the data. Power law seems to fit in here. Wealthier districts are expected to have a larger penetration rate, hence be more represented in the data. The cost of communication is less an issue for wealthier people, thus their average communication time and numbers of calls are expected to be higher than the rest of the population, and they are expected to score higher on the mood index. This comportment is matching the index

concept, where a larger social activity is a reflection of well-being - for instance, better financial level.

The mood index that we calculated is a reflection of the social and communication behavior of the population. Communication is a healthy indicator of a vital population but it doesn't directly reveal if the population is happy, excited, outraged, working or enjoying the leisure time. In order to correctly label the output of the mood function, and to be able to unlock pattern of activity in the city, the index must be combined with other factors like direct expert information on the city routine. News, political situation, holidays, social networks.

Figure7 Example mood map that is still under development.

Conclusion and Future work

In this work, we have demonstrated the importance of happiness index and our plan is to extend our data analysis techniques and happiness hypothesis around Abidjan to come up with Happiness Management Model. We are trying to detect urban routines and mood cycles with the goal of a deeper understanding of the social activity rhythms and its causes in Abidjan. Our ultimate goal is to build a comprehensive model that combines several sources of information and successfully predicts next days' moods and detects constant routines. We hope that using these results will help policy makers and urban planners see the big view of the city and better understand how it dynamically evolves.

Finally, we speculate that collective affective trends can be modelled and predicted combining mobile data with user-generated content from our emotion sensing system. Results can then be discussed in terms of the social, economic, and cultural spheres in which the users are embedded. By looking into the sensor-based data of large populations that are increasingly easily and readily captured, we could get broad indications of developing relationships and overall trends of particular social phenomenon. In addition, by analyzing various combinations of such large datasets, we might also identify novel or interesting social phenomena that are worth pursuing in greater depth.

Acknowledgment

To be added later.

References

[1] H.M. Proshansky, "The field of environmental psychology: securing its future," in *Handbook of environmental psychology*, D. Stokols and I. Altman, Eds. New York: John Wiley & Sons, 1987, pp. 1467–1488.

[2] Eiman Kanjo, "mFeel: An affective mobile system," *IEEE Pervasive Computing*, vol. 11 , no. 3, pp. 43-45, 2012.

[3] Jonathan Reades, Francesco Calabrese, Andres Sevstuk, and Carlo Ratti, "Cellular Census: Explorations in Urban Data Collection," *Prevasive Computing*, vol. 6, no. 3, pp. 30–38, 2007.

[4] Edwin De Jonge, Merijn van Pelt, and Marko Roos. , "Time patterns, geospatial clustering and mobility statistics based on mobile phone network data," in *Paper for the Federal Committee on Statistical Methodology research conference*, Washington, USA, 2012.

[5] Rein Ahas, Margus Tiru, Erki Saluveer, and Christophe Demunter, "Mobile telephones and mobile positioning data as source for statistics: Estonian experiences," in *presentation for NTTS*, 2011.

[6] Andres Sevtsuk and Carlo Ratti, "Does Urban Mobility have a Daily Routine? Explorations Using Aggregate Mobile Network Data," in *Proceedings of the 11th International Conference on Computers in Urban Planning and Urban Management*.

[7] Sahar Hoteit et al., "Content Consumption Cartography of the Paris Urban Region using Cellular Probe Data," in *Proceedings of the first workshop on Urban networking*, 2012, pp. 43-48.

[8] Vincent B. Blondel et al., "Data For Development: The D4D challenge on mobile phone data," Orange Cote d'Ivoire, 2012.

[9] Johan Bollen, Alberto Pepe, and Huina Mao, "Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena," in *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, 2011.

[10] (2011, October) ICC to investigate Ivory Coast post-election violence. [Online]. www.bbc.co.uk/news/world-africa-15148801.

[11] (2013, February) Ivory Coast, wikipedia. [Online]. en.wikipedia.org/wiki/Ivory_Coast

[12] (2012, January) CIA World Fact Book. [Online]. https://www.cia.gov/library/publications/the-world-factbook/geos/iv.html.

[13] (2009, November) Abidjan District web site. [Online]. www.districtabidjan.org/page.php?id=44.

[14] Kyriakidis, P.C.: A geostatistical framework for Area-to-Point spatial interpolation. Geographical Analysis 36(3), 259–289 (2004)

[15] (2013, February) Orange D4D challenge official web site. [Online]. www.d4d.orange.com.

Studying Intercity Travels and Traffic Using Cellular Network Data

Wei Wu, Eng Yeow Cheu, Yuzhang Feng, Duy Ngan Le, Ghim Eng Yap, Xiaoli Li
Institute for Infocomm Research, A*STAR, Singapore

***Abstract*—We use the anonymized mobile phone CDR (call detail record) dataset provided by Orange D4D and crowd sourced OpenStreetMap data to study people's intercity travels and traffic speed on intercity highways in Ivory Coast. Knowledge about people's intercity travel behaviour and speed on highways would help government officers such as transportation strategic planners make informed development decisions. It is also very valuable to end users who would like to be informed about the traffic situation so as to avoid traffic jam. We design algorithms and implement a system that is able to discover people's intercity trips and extract intercity travel trajectories suitable for estimating traffic speed on intercity highways. The data and results are visualized using a visualization engine we develop for this project. Besides describing our system, we also present our analysis of intercity trips and highway speed. Video demo of our system is located at http://www1.i2r.a-star.edu.sg/~wwu/d4d/intercity.html**

I. Introduction

Mobile phones are increasingly a part of our life as we move around and stay in contact with one another. Data of mobile phone calls, messages, and Internet access can directly convey the approximate location of each phone user at a specific time, and allow for effective and automated collation of movement data for an entire population over large area [11], [14], [10]. Such data can be used to study social or economic behaviours and make high-impact contributions towards predicting epidemics [13], [15], detecting crisis [2], and urban planning [3], among others.

Taking on the opportunity provided by Orange "Data for Development" (D4D) challenge [1], we propose to use anonymized mobile phone (CDR) data collected in Ivory Coast and publicly-available geographical map data to study two important and interrelated problems. Firstly, the people's intercity travel pattern. For example, we would like to find out how often and which city people normally travel to. Secondly, the health of highways interconnecting the cities. It will be valuable to forecast the traffic condition of the supporting transport infrastructure at different hours of a day.

We believe CDR data can be used to study the above two problems. People normally use their mobile phones before their departure and after arrival. For example, they will contact their family to report safety. They also call or SMS during their intercity journey. These phone usages before, during, and after their intercity trips provide data for our study. Although the CDR data of a mobile user can be temporally sparse and spatially coarse, this limitation can be overcome by aggregating data from many users over a longer period of time.

The technical and data challenges that we address in our study include the following. How to find out whether a user is in a city? How to discover that a user has travelled from one city to another? How to integrate the intercity road network data of Ivory Coast with CDR data? What kind of intercity trip trajectories can be used to estimate highway traffic speed? How to estimate the highway traffic speed with such intercity travel trajectories? How to effectively visualize the data and results?

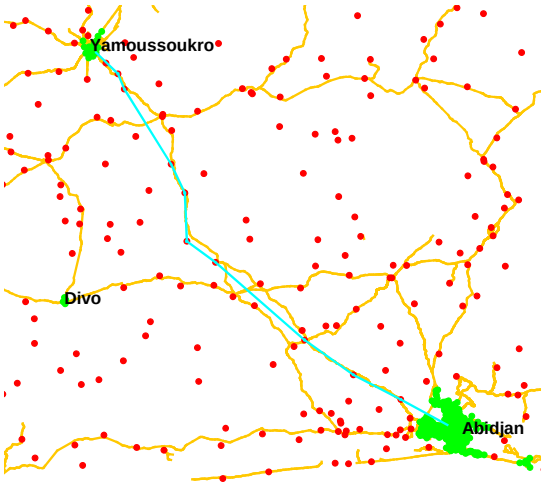
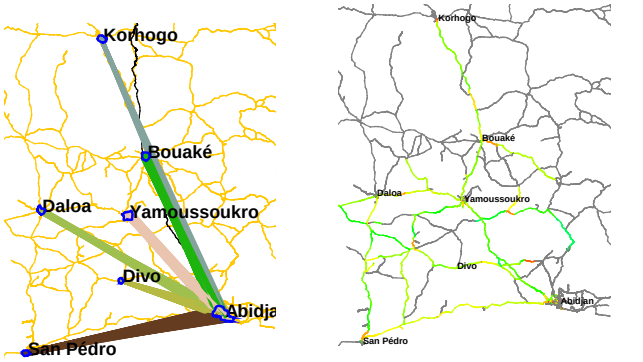


Fig. 1: An example of intercity trajectory (cyan) along highway (orange) between Abidjan and Yamoussoukro.

We design and implement a system that solves all the above challenges. Figure 1 illustrates our solution.



(a) Trips from Abidjan.

(b) Highway speed from 3 p.m. to 6 p.m..

Fig. 2: Examples of our study results.

We apply a clustering algorithm on the antenna position data (red and green points) to automatically discover antennas in the major cities (green points). From CDR data, we extract intercity trips of users (e.g., a user travels from Abidjan to Yamoussoukro). By looking at the detailed CDR records during the trip, we discover and select the intercity trajectories (cyan polyline) along the highways (orange lines). We estimate the traffic speed of highway segments by selecting suitable intercity trajectories and projecting them onto highways. City information and highway road networks are extracted from OpenStreetMap data [6].

The results are visually communicated to the decision makers using a visualization engine we develop for this project. Figure 2 shows two representative examples of our analysis and mining results. Figure 2a illustrates the volumes of intercity travels from Abidjan to other major cities. Lines in different colours are trips to different cities. Figure 2b is a visualization of the average highway speed during the 3 p.m. to 6 p.m. time interval. Red coloured segments are sections of the highways where traffic are very slow. Video showing the visualization capability of our system and the interaction of a user with our system is located at http://www1.i2r.a-star.edu.sg/~wwu/d4d/intercity.html.

Our system reveals some very interesting facts about people's intercity travel behavior and the state of highway travel in Ivory Coast. For example, a lot of people travel from Abidjan to Yamoussoukro in the middle of April 2012. As another example, our temporal analysis of the CDR data consistently reveals that the slowest commute speeds occur daily between 9 p.m. to 6 a.m. on most intercity highways over the five months of data. We are intrigued by this and conducted some information search online, which revealed the possible explanation: inadequate lighting make driving conditions hazardous [5], [8].

Such insights from data analytics demonstrate the immense potential of using the readily available mobile CDR dataset to reliably find out something that is typically much more costly to implement, such as transport network analysis using traditional measurement-intensive techniques like travel survey. Our solution is much faster and cheaper as it requires only anonymized mobile phone data and open source map data.

Our system can be used to make recommendation to the Ivory Coast government on the locations where the most benefits will be felt if they have to prioritize their infrastructure improvement works. Possible actions that can then be taken include short/medium term works like lane expansions and addition of slip roads, and longer-term investments such as commissioning the constructions of a new highway linking two major cities. The decisions shall be supported by scientific, data-driven observations from our fully-automated system and will be based on key considerations including intercity travel volume and highway speed limitation. The ability to recommend specific segments of each highway for immediate improvement possibilities is highly impactful for a country with limited resources and growing economy that depends a lot on the efficacy of the transportation network for the workforce and the businesses.

The rest of this report is organized as follows. Section II describes the details of our proposed solution. In Section III, we present the main results that demonstrate the usefulness of our proposed system. Finally, Section IV concludes this report.

II. Solution

The Orange "Data for Development" (D4D) [1] dataset (SET2) that we use consists of anonymized CDR records of phone calls and Short Message Service (SMS) exchanges among 500,000 of Orange's customers in Ivory Coast between December 5, 2011 and April 22, 2012. The dataset contains the coordinates of antennas and high resolution trajectories of 50,000 randomly sampled individuals in every two-week period.

Figure 3 depicts the steps involved in extracting intercity trips and estimating the speed profile of the highway segments using CDR data, antenna location data, and geographical map data.

Using the city locations of Ivory Coast in OpenStreetMap (OSM) (see section II-A) as the cluster seeds, we design a clustering algorithm that autonomously determines the subset of cellular antennas within each city

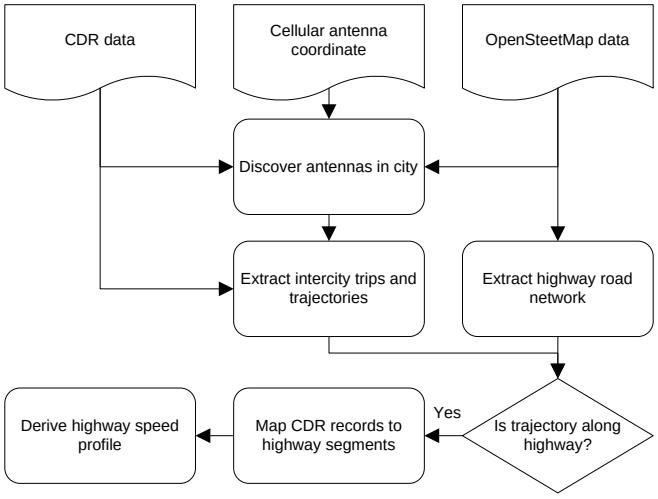


Fig. 3: Overview of our approach.

(see section II-B). After discovering the set of cellular antennas in every city, we extract all communication records which span over more than one city. These records are used to estimate the inter-city trips made by individuals (see section II-C). Using the cellular antenna positions associated with these inter-city communication records and highway geospatial information from OSM, we determine whether a trajectory is along highways and decide whether the trajectory can be used to estimate highway traffic speed. For those selected trajectories, we map their individual communication records to the nearest highway segment. We then estimate the speed profile of the interconnecting highway segments using the estimated distance which individuals travelled during the time between two communication records.

A. City and Road Network Extraction

OpenStreetMap (OSM) is a free worldwide map created and maintained by the public. Worldwide geographical data from OSM is available under the Open Data Commons Open Database License (ODbL) [6]. OSM map data is structured with basic data primitives: node, way, and relation. A node contains geospatial information of a single location. A road, stream or railway is defined by way data primitive by using an ordered interconnection of between 2 and 2000 nodes. A relation data primitive defines the relationships between other data primitives.

For this project, we study only the highways interconnecting the following major cities of Ivory Coast [7]. They are, Abidjan, Bouaké, Daloa, Yamoussoukro, Korhogo, San Pédro, and Divo. The city information, including name and a coordinate (probably city center), is described by node data primitives in Ivory Coast geographic data from OpenStreetMap. However, the OpenStreetMap data does not contain city boundary information. As a result, we need to devise a way to identify antennas in each city.

The geospatial information of Ivory Coast highways is described by way data primitives in OSM map data. Each highway is described by an interconnection of geospatial points. A highway segment is a section of the highway connecting two way points. A total of 356770 segments were extracted from the Ivory Coast geographic data from OpenSteetMap.

B. Antennas-in-City Discovery

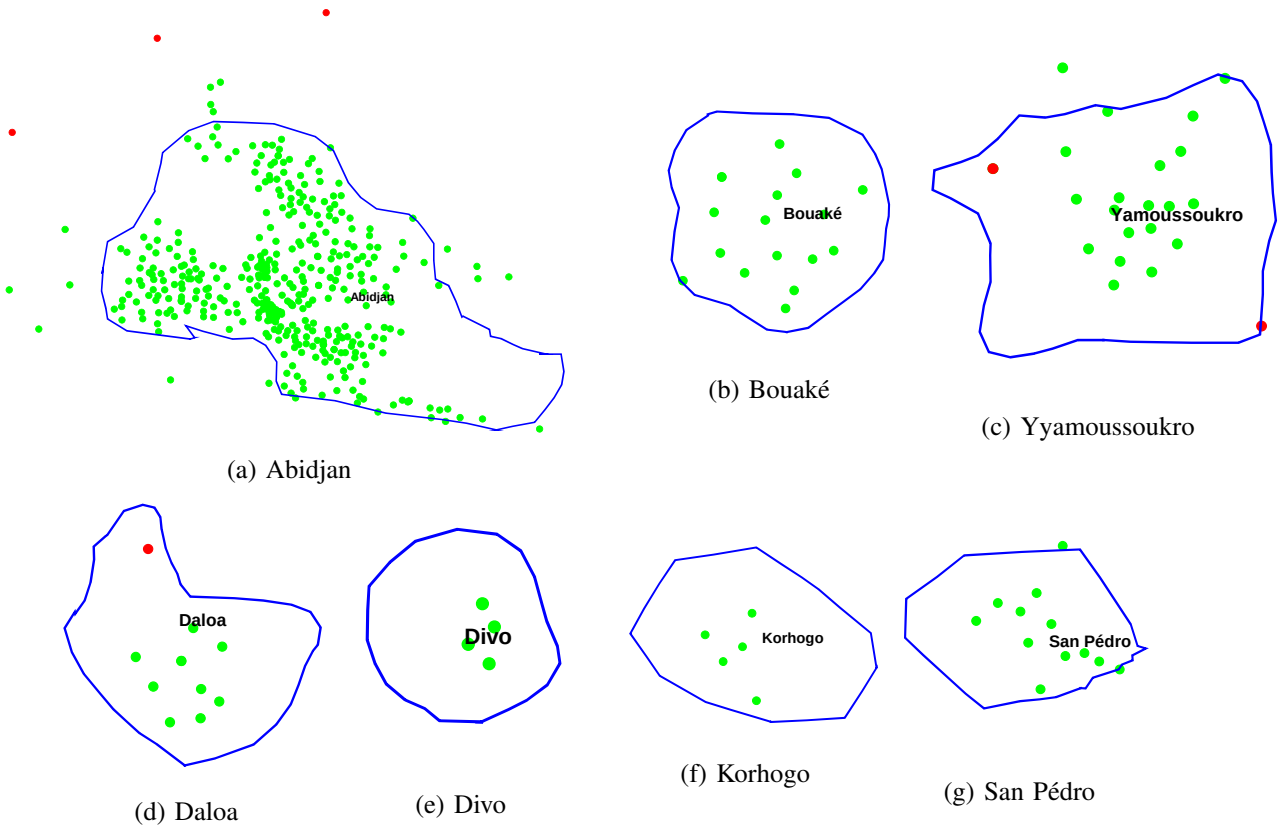
The objective of Antenna-in-City discovery is to *automatically* detect the set of antennas in each city using the cellular antenna data across the whole Ivory Coast and city information extracted from OpenStreetMap data.

We define the antenna-in-city discovery problem as follows. Given a set of antennas $A = \{a_1, ..., a_n\}$, and a set of points each representing a location in one of the major cities, the task is to identify the set of antenna in each city without knowing the city boundaries.

We propose a novel clustering algorithm for antenna-in-city discovery. To perform clustering, we first represent each antennas $a_i, i = 1, 2, ..., n$, into a feature vector $V(a_i) = (X(a_i), Y(a_i), Volume(a_i))$, where $(X(a_i), Y(a_i))$ represents a_i's geographical location coordinate and $Volume(a_i)$ denotes a_i's average call volume.

Intuitively, population density in city is higher than that in non-city areas. Therefore the antennas in city are geographically closer to each other and more densely distributed whereas the distances between antennas outside a city are relatively farther away from each other. At the same time, people in city normally will call more (due to business) than people in non-city areas, so antennas in city will have higher call volumes. As such, our feature representation takes both geographical location and average call volume information into consideration.

We modify a density based clustering algorithm [12] to group the antennas into a number of clusters/cities. To initialize clustering algorithm, we also select a seed data point for each city/cluster. In particular, we have searched OpenStreetMap using each city name as a query word to get a point which represents an internal position of the city. We then expand each individual seed nodes by including its nearby antennas to form city clusters. Particularly, for each seed, if its neighborhood contains



(a) Abidjan (b) Bouaké (c) Yyamoussoukro

(d) Daloa (e) Divo (f) Korhogo (g) San Pédro

Fig. 4: Antennas in cities discovered by clustering. Red and green points are antenna locations. Green points are the cluster of antennas discovered by our program for the city. Blue polygon is the city boundary drawn based on Google Maps.

more than a certain number of neighbors, then we expand all its direct neighbors into corresponding clusters. Similarly, for each of the newly added neighbors, we continue adding their neighbors if their neighborhoods are also located at the high density regions and exhibit high call volumes. This recursive clustering process stops until no more neighbors can be added into the clusters. Finally, we output all the clusters where each of them corresponds to a city.

Note our proposed density based algorithm can detect arbitrarily shaped and sized clusters as long as the antenna data points within each clusters are densely connected. This is especially important for the city boundary detection, as in our case cities can have totally different shapes and/or with different sizes.

We have compared our automatically discovered antenna clusters with the city boundaries approximated through Google Maps [4]. Figure 4 shows the antenna clusters and city boundaries. Our clustering results match very well with the sets of antenna in city boundaries, indicating that our clustering algorithm is very effective for antennas-in-city discovery.

There are some advantages to using an algorithm to automatically discover antenna-in-city over relying on manually collected city boundaries data. First, the process is automated. Second, the algorithmic solution can scale to many cities. Third, our solution adapts with data and therefore adapts with the development (e.g., expansion) of cities.

C. Intercity Trip Trajectory Extraction

After discovering antennas for each city, we associate antennas in city with the city they belong to. The CDR records are now also attached with city tags. A mobile user made an intercity trip if two of his/her CDR records are associated with different cities. The trajectory of the intercity trip is the sequence of CDR records from one city to another city.

For convenience, we define a city as the set of antennas in that city.

Definition 1 (city): Given a set of antennas $A = \{a_1, ..., a_n\}$ and a location in city i, a city c_i is defined as

the cluster of antennas our clustering algorithm outputs for city i.

In other words, the city c_i is defined as a set of m cellular antennas $\{a_{i_1}, a_{i_2}, ..., a_{i_m}\}$ which is a subset of the whole antenna set A.

Definition 2 (CDR trace): CDR trace ct_i of user i is the sequence of his/her CDR records $ct_i = \{\langle a_{i_1}, t_{i_1}\rangle \dots \langle a_{i_n}, t_{i_n}\rangle\}$ where $\langle a_{i_j}, t_{i_j}\rangle$ is a CDR record. a_{i_j} is an antenna and t_{i_j} is a timestamp.

Given a user's CDR trace, the user made a intercity trip if there exist two records in his/her trace whose antenna belong to different cities.

Definition 3 (Intercity Trip Trajectory): Given a CDR trace ct_i of user i, an inter-city trajectory tr of the user i is defined as the shortest subsequence of ct_i such that the first CDR record and the last CDR record are in different cities. $tr = \{\langle a_{i_p}, t_{i_p}\rangle \dots \langle a_{i_q}, t_{i_q}\rangle \mid a_{i_p} \in c_1, a_{i_q} \in c_2\}$.

Note that a user can have several intercity trip trajectories extracted from his/her CDR trace. Here "shortest" means that the number of records in that trajectory can not be smaller. More precisely, it means that the first CDR record of the trajectory is the last record of the user in the departure city and the last record of the trajectory is the first record of the user in the arrival city.

D. Trajectory Along Highway

Because one of our objectives is to study the highway transportation infrastructure amongst cities of Ivory Coast, we need to select intercity trajectories that use the highway. For this purpose we need a way of identifying whether an intercity trajectory is along highways.

Our basic idea is to find out the set of antennas that will serve the users on highways. If during a trip the user made all his/her calls through these antennas then with high probability the user was travelling along highways.

We apply the Voronoi Diagram technique [9] to find out the antenna along highways. Given a space and n points, called seeds, in the space, a Voronoi diagram is a way of dividing the space into n regions, called Voronoi cells, such that every cell covers exactly one seed and all points in the Voronoi cell are closer to the seed in this cell than to any other seeds in any other cell.

Definition 4 (Highway Antennas): Given a set of cellular antennas $A = \{a_1, \dots, a_n\}$, a function $V(a_i)$ which gives the Voronoi cell of a given antenna a_i, and a set of highway segments $S = \{s_1, \dots, s_m\}$, the set of highway antennas is a defined as $A' = \{a \in A \mid \exists s : S, V(a) \cap s \neq \emptyset\}$.

Basically a highway antenna is an antenna whose Voronoi cell spatially overlaps with at least one highway segment. Here we slightly misuse the symbol $\cap$ and $\emptyset$.

If all the CDR records in a trajectory are associated with highway antennas, we say this trajectory is a highway trajectory.

Definition 5 (Highway Trajectory): Given a set of highway antennas A' and an intercity trajectory $tr = \{\langle a_{i_p}, t_{i_p}\rangle \dots \langle a_{i_q}, t_{i_q}\rangle\}$, tr is a highway trajectory if and only if $\forall \langle a, t\rangle \in tr, a \in A'$.

Please refer to Figure 1 to see an example of highway trajectory.

E. Further Selection of Highway Trajectories

Not all highway trajectories are used to derive highway speed. We further select the trajectories by looking at trip's "observed" duration and the number of CDR records available during the trip.

The "observed" duration of a trip is the time from the traveller's last CDR record in departure city to the first CDR record in arrival city. If the "observed" duration is too long, for example more than 12 hours, the traveller might have breaks or even stopover during his/her trip, and therefore speed estimation based on such trips would be very erroneous.

CDR records during a trip are in fact the time and location samples of the traveller's true trajectory. If there is very few CDR records during a trip, we won't know have enough information to find out which highway the traveller had took and again the speed estimation can be very wrong.

For these two reasons, we further restrict the trajectories used in highway speed study to the highway trajectories with reasonable "observed" travel duration and at least a few CDR records.

F. Derive Highway Speed Profiles

From the previous steps, we obtain a set of intercity trajectories which are close to the highway and fulfill our other selection criteria. Now we use them to build speed profiles for different segments of the highway.

We project the trajectories onto the highway segments (using map-matching) and estimate the speed the user was travelling (assuming the user was travelling with constant speed between two consecutive CDR records). Then for each of the highway segments passed by the user, we use the estimated travel speed as observed traffic speed at the time the user passed that segment.

Note that a highway segment will have many speed observations because many users' highway trajectories

are mapped to that highway segment. These speed observations are timestamped. Although the speed estimation for each particular trajectory may not be very accurate (due to limited antenna spatial resolution), the statistical measures (e.g., the median) of these speed observations could be quite telling.

Definition 6 (Highway Segment Speed Profile): For each highway segment s_i, a time-dependent speed profile P_i is defined as $P_i = \{\langle v_{i,j}, t_{i,j}\rangle \mid j = 1 \dots n_i\}$ where n_i is the total number of trips recorded for the highway segment s_i, $v_{i,j}$ and $t_{i,j}$ are the speed and time stamp of the j^{th} trip over this highway segment respectively.

Recall that each highway trajectory is a set of tuples $tr = \{\langle a_1, t_1\rangle, \dots, \langle a_n, t_n\rangle\}$ where a_i is the antenna through which the call is made and t_i is the time of call. We perform the trajectory projection by pairs of consecutive calls. We define a function $N(a)$ which returns the nearest way point of a given antenna a. We also use a function $SP(p_1, p_2)$ which implements a standard Shortest Path algorithm and returns a set of highway segments for two given way points p_1 and p_2. Lastly we implement a function $D(p_1, p_2)$ which returns the road network distance of two given way points p_1 and p_2.

Now for each pairs of consecutive call records $\langle a_{i_1}, t_{i_1}\rangle$ and $\langle a_{i_2}, t_{i_2}\rangle$ in a highway trajectory, we compute the speed $v = \frac{D(N(a_{i_1}), N(a_{i_2}))}{t_2 - t_1}$. Then we update the speed profile of all highway segments in $SP(N(a_{i_1}), N(a_{i_2}))$ by appending a tuple $\langle v, t^*\rangle$ where t^* the estimated time by assuming that the user travels at constant speed from way point $N(a_{i_1})$ to way point $N(a_{i_2})$.

III. Results

A. City, Antennas-in-city, and Road Networks

We focus on the seven major cities in Ivory Coast. Figure 5 depicts our results of antennas-in-city discovery and road network extraction. Antennas are shown with red and green points. Antennas in cities are shown with green points. They are the dense clusters discovered by our program. Yellow lines are the highway networks extracted from the OpenStreetMap road data.

Note that if authentic data of city boundaries and highway road network are available (for example from Ivory Coast government), they can be easily incorporated into our system to improve data quality. For example, authentic city boundary data can be used to find out the accurate sets of antenna-in-city.

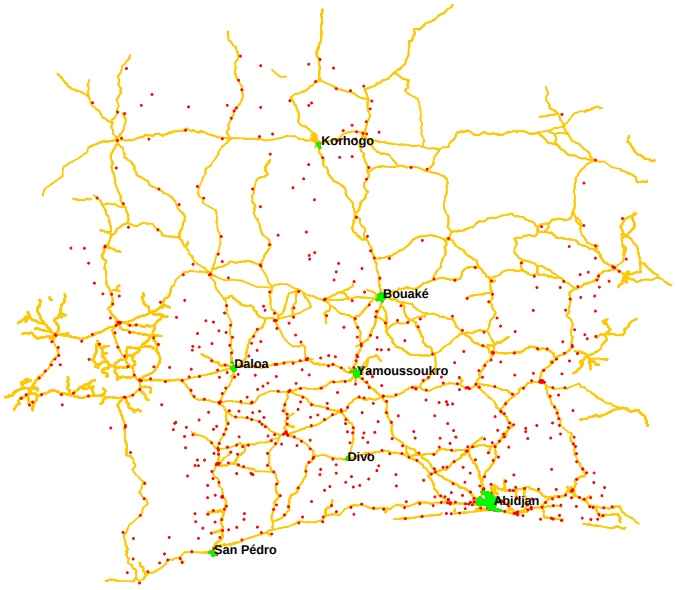


Fig. 5: City names (black), antennas (red and green), antennas in city (green), and highways (orange).

B. Intercity Trips

The CDR data set provided by Orange D4D contains 10 samples where each sample contains the records of randomly selected 50,000 users in 2 weeks. The time spans of the 10 samples do not overlap. The whole time span is 20 weeks: last four weeks of 2011, and the first sixteen weeks of 2012. We have CDR records of 500,000 users collected in 20 weeks (2011-12-05 to 2012-04-22) where data of each user are collected in only two weeks.

With this dataset, our system discovers that 15,800 users had intercity trips and they made 28,860 intercity trips. This means that about 3.2 percent (15800/5000000) of the sampled population made intercity trips during two weeks, and the average number of trips made by a intercity traveller is about 2. Note that travelling from city A to B and then travelling back from city B to A are counted as two trips.

Figure 6 shows the distribution of the number of intercity trips. Most of the travellers made one or two intercity trips within two weeks. From the chart, we also see that there are some users who travelled more frequently. For example, there are users who made more than six trips in two weeks.

Figure 7 and Figure 8 show the time series characteristics of the intercity trips by month and week respectively. The number of trips observed in different months are quite similar. Considering that the dataset includes 27 days of December and 22 days of April, then it seems that people travel a little more in December and April

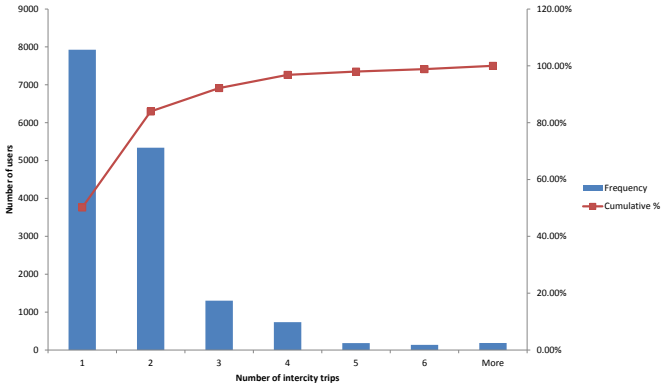


Fig. 6: Distribution of users by their number of intercity trips.

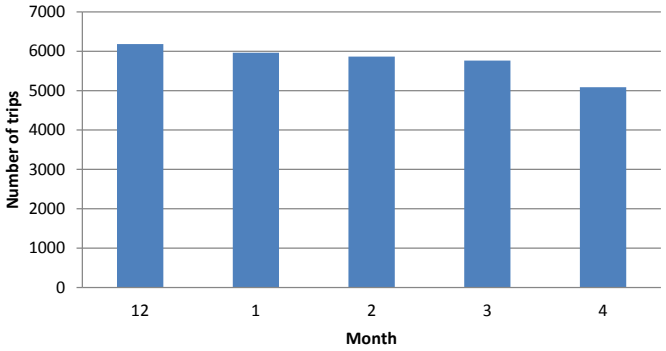


Fig. 7: The number of intercity trips in each month.

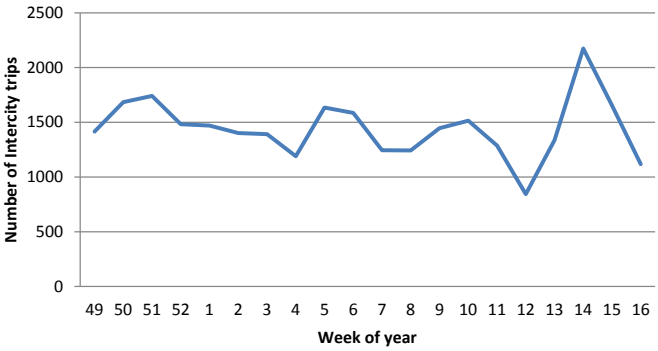


Fig. 8: The number of intercity trips in each week.

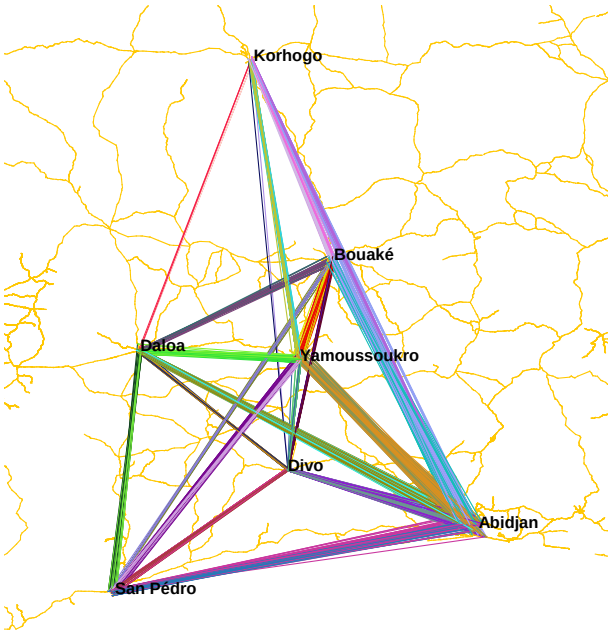


Fig. 9: Visualization of intercity trips.

than in the other months. The time series by week is more fluctuant, as shown in Figure 8. For example, we can observe a sudden drop in the 12th week of 2012 and a sudden increase in the 14 week of 2012.

1) Origins and Destinations: Figures 9 is a visualization of the intercity trips. Different colors are used to represent different trips between different cities. For example, there are more trips made between Abidjan and Yamoussoukro.

Table I lists the 10 pairs of cities with the most number of trips and their shares of intercity trips. We see that the shares of trips from city A to B and from B to A are quite similar.

Figure 10 illustrates the percentage of intercity trips from and to the major cities. For example, of all the intercity trips, 32% are from the city of Abidjan and 31% are to the city of Abidjan. Again, it can be observed that the percentage of trips from a city is approximately equal to that of trips to the city. It is also clear that Abidjan, Yamoussoukro, and Bouaké are the busiest cities in Ivory Coast.

Figure 11 depicts the number of intercity trips from Abidjan to other cities from the 49th week of 2011 to the 16th week of 2012. Although there are fluctuations in the number of trips to each city over the time period, there seems to be a consistent pattern amongst all the cities.

The observations above verify that our methodology of deriving the intercity trips from the user call data is sound. This allows us to use the intercity trips to study the highways among major cities in Ivory Coast.

Abidjan-Yamoussoukro	13.7%	Yamoussoukro-Abidjan	13.1%
Yamoussoukro-Bouaké	9.0%	Bouaké-Yamoussoukro	7.7%
Abidjan-Divo	6.4%	Divo-Abidjan	6.3%
Abidjan-Bouaké	5.2%	Bouaké-Abidjan	4.6%
Bouaké-Korhogo	3.9%	Abidjan-San Pédro	3.3%

TABLE I: Most travelled city pairs

(a) Starting from cities.

(b) Arriving at cities.

Fig. 10: Percentage of trips starting from and arriving at cities.

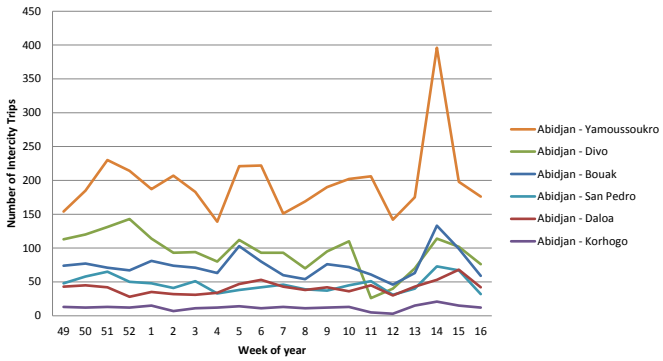


Fig. 11: Number of intercity trips from Abidjan to other cities by week.

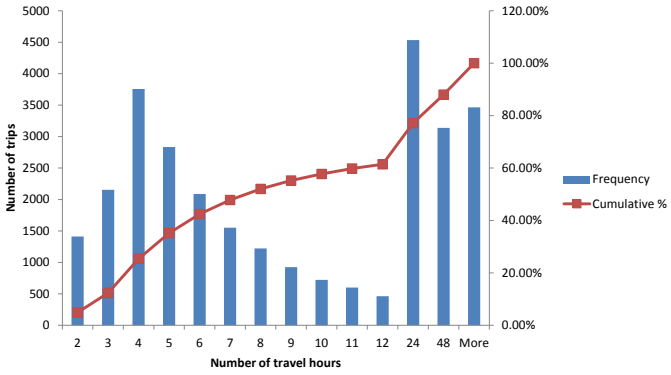


Fig. 12: The distribution of intercity trips based on duration of the trips.

2) Duration of Intercity Trips: Figure 12 shows the histogram of estimated durations of the intercity trips. The estimated duration of a trip is the time from the traveller's last CDR record in departure city to the first CDR record in arrival city. We have the following observations:

- Some intercity trips are quite short, e.g., 2 and 3 hours only. They probably are trips by flight.
- There are also many very long trips, e.g., above 12 hours, probably due to the following reasons. The user did not make call right before he/she leaves. The user did not make call right after he/she arrives. The user makes stopover during the trip.

When we use intercity trip trajectory data to study highway speed, we filter out those very short and very long trips.

3) Calls Made in Transit: Figure 13 shows the histogram of the numbers of CDR records of intercity trips. We observe that many people do not make many calls during their trips, probably because making calls during trips is expensive. We also see that about 40 percent of the trips have five or more CDR records. In our highway traffic study, we use only highway trajectories with at least five CDR records.

C. Highway Speed

As described in Section II, we use intercity trip trajectories along highways that fulfill our selection criteria to study the traffic speed on the highways. Our system selected about 4,000 intercity trip trajectories to derive highway segment speed profiles.

Figure 14 shows the average speed on the highways by hour of day and day of week. We see that it exhibits clear patterns. Speed varies in the range from 20 km/h to 50 km/h. Speed in night is much slower than speed in daytime. In particular, speed from 9 p.m. to 6 a.m. is very slow. A possible explanation (we found through Internet search) is that inadequate lighting make driving

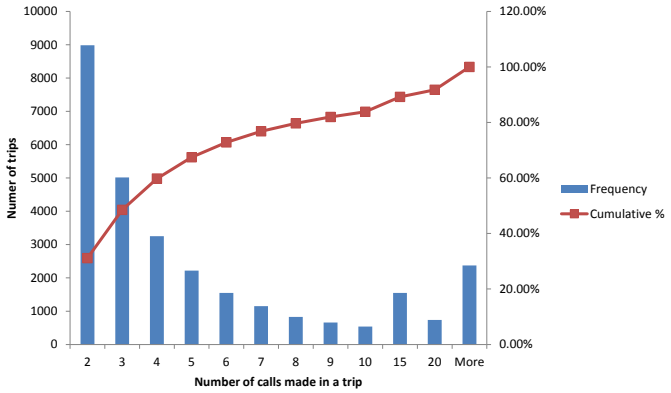


Fig. 13: The distribution of intercity trips based on the number of calls.

conditions hazardous [5], [8]. We also observe that speed in weekend nights is faster than that in weekday nights. It is probably because people travel less on weekends and therefore there is less traffic on weekends.

Figures 15a-15h show the estimated median travelling speed at different segments of Ivory Coast major highways at different time intervals on all Thursdays for the entire observation period. Segments in red indicate that commuters are travelling at slow speed. Segments in yellow indicate commuters are travelling at moderate speed. Segments in green indicate commuters are travelling at a faster speed. Segments in gray have no speed data available. Figure 15i shows the colour bar for the speed.

There are a number of key observations our system users can make on the intercity travel patterns:

- There are fewer coloured segments during night time, meaning less people travel at night.
- Driving at night is understandably slower.
- The highway segments nearer to the cities normally have slower traffic than the other segments.
- Majority of the major cities has travellers moving in and out throughout the day, but there are some particular cities such as Korhogo where there are not many travellers between 12 a.m. to 6 a.m..
- For pairs of cities, there are some highways that are utilized by travellers in the night time in a manner that is different from during the day.
- The network of highways right between the cities of Dalao, Yamoussoukro, Divo and San Pédro see substantial traffic movement only between 9 a.m. to 9 p.m., and few or no travellers actually use these highways from 9 p.m. to 9 a.m..

The above are examples of the most direct observations our automated system offers, and we see these as valuable insights to transport planners.

IV. Conclusion

In this project we study intercity travels and traffic speed on intercity highways using anonymized mobile phone data and publicly available map data. We tackle the technical and data challenges and implement a system that automatically discovers intercity trips and CDR trajectories that are suitable for highway speed estimation. We also analyze the trip and highway speed data to reveal interesting patterns and insights. The results are presented to the decision makers using a visualization engine we develop for this project.

Our solution has the following key advantages. Firstly, our solution is far more economical than travel survey and on-site speed sampling. This is because we use existing mobile phone data and no extra data collection is needed. Secondly, this approach is also faster for the same reason. Thirdly, our approach is generic and applicable to other countries, and even across countries. Lastly, although our method and system are designed for intercity travel and traffic study, it is extensible to study intra-city travel behaviour and traffic.

We understand that our proposed method of using mobile phone CDR data for intercity travel and traffic study could have the following limitations. Firstly, not all people have cell phones and people with cell phones may not make calls during travel. Therefore the sample could be biased. The number of trips could be underestimated. Secondly, due to overlapping coverage of adjacent cellular antennas, travel distance estimated with antenna locations can be inaccurate. However, this problem is serious only when the travellers have many CDR records from overlapping antennas. This problem can be mitigated by skipping some CDR records from overlapping antennas during the speed estimation process.

References

[1] D4D Challenge. www.d4d.orange.com/.
[2] EPIWORK Project. http://www.epiwork.eu/.
[3] Expanding the use of data analytics in city governments. http://www.mikebloomberg.com/index.cfm?objectid=8A9B9F00-C29C-7CA2-FA465E3F83FD7CD9.
[4] Google Maps. http://maps.google.com/.
[5] Ivory coast travel advice. http://www.smartraveller.gov.au/zw-cgi/view/Advice/Ivory_Coast.
[6] OpenStreetMap. http://www.openstreetmap.org/.
[7] Population of major cities in Ivory Coast. http://en.wikipedia.org/wiki/Ivory_Coast.
[8] *Benin Mineral, Mining Sector Investment and Business Guide*, volume 1. International Business Publications, USA, 2012.

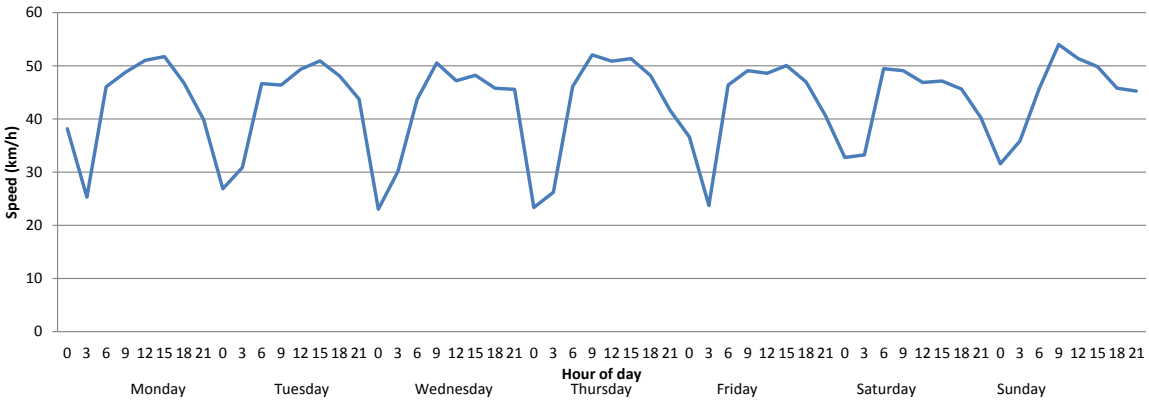


Fig. 14: Average traffic speed on highway by day of week and hour of day.

[9] Franz Aurenhammer. Voronoi diagramsła survey of a fundamental geometric data structure. *ACM Computing Surveys (CSUR)*, 23(3):345–405, 1991.

[10] Richard A. Becker, Ramón Cáceres, Karrie Hanson, Sibren Isaacman, Ji Meng Loh, Margaret Martonosi, James Rowland, Simon Urbanek, Alexander Varshavsky, and Chris Volinsky. Human mobility characterization from cellular network data. *Commun. ACM*, 56(1):74–82, 2013.

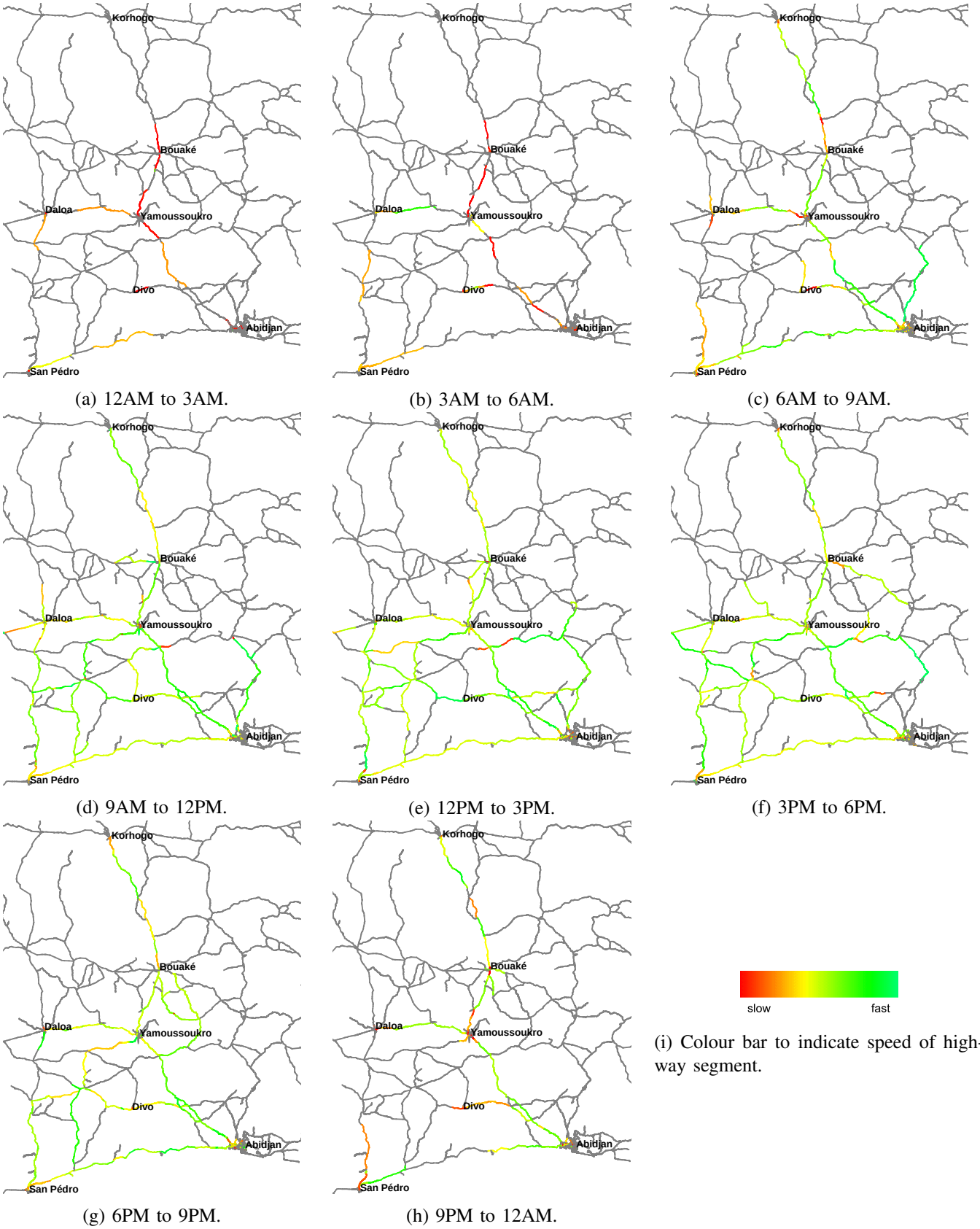
[11] Balázs Csanád Csáji, Arnaud Browet, Vincent A. Traag, Jean-Charles Delvenne, Etienne Huens, Paul Van Dooren, Zbigniew Smoreda, and Vincent D. Blondel. Exploring the mobility of mobile phone users. *CoRR*, abs/1211.6014, 2012.

[12] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data mining*, volume 1996, pages 226–231. AAAI Press, 1996.

[13] Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, February 2009.

[14] P. Glotz, S. Bertsch, and C. Locke. *The Thumb Culture: The Meaning of Mobile Phones for Society*. Kultur- und Medientheorie. Transcript Verlag, 2005.

[15] Wei Pan, Nadav Aharony, and Alex Pentland. Composite social network for predicting mobile apps installation. In Wolfram Burgard and Dan Roth, editors, *AAAI*. AAAI Press, 2011.



(a) 12AM to 3AM.

(b) 3AM to 6AM.

(c) 6AM to 9AM.

(d) 9AM to 12PM.

(e) 12PM to 3PM.

(f) 3PM to 6PM.

(g) 6PM to 9PM.

(h) 9PM to 12AM.

(i) Colour bar to indicate speed of highway segment.

Fig. 15: Highway median speed plotted with different colours over different time intervals of all Thursdays. Segments in red, yellow, and green respectively mean the corresponding speeds of vehicles at the segments are slow, moderate, and fast. Segments in gray colour indicate no reliable data is available.

Human Mobility Flows in the City of Abidjan

Diala Naboulsi, Marco Fiore, Razvan Stanica
INSA Lyon / Inria, Lyon, France
firstname.lastname@insa-lyon.fr

***Abstract*—The growing ubiquity of mobile communications has offered researchers new possibilities to understand human mobility over the last few years. In this work, we analyze Call Detail Records (CDR) made available within the context of the Orange D4D Challenge, focusing on calls of individuals in the city of Abidjan over a period of five months. Our results illustrate how aggregated CDR can be used to tell apart typical and special mobility behaviors, and demonstrate how macroscopic mobility flows extracted from these cellular network data reflect the daily dynamics of a highly populated city. We discuss how these macroscopic mobility flows can help solve problems in developing urban areas.**

I. Introduction

Understanding human movements is critical for different scientific domains. In order to deploy efficient networking solutions, a clear view of human mobility patterns is required. The same applies for urban planning, where the global mobility flows can determine the optimal deployment of infrastructure. Human mobility also plays a major role when analyzing the ways diseases can spread in a population.

Significant research efforts have been conducted in this direction, aiming at understanding how people move as a first step, and proposing models of such mobility as a second step. Recently, as people are more and more connected, network traces have received particular attention as a source of information about human mobility at large scales.

However, previous studies have focused on developed countries, and whether the observed patterns and models are applicable to developing countries remains an open question, due to differences in the lifestyle, country's infrastructure and modes of transportation. Indeed, a clear understanding of human movements would be crucial for the progress of such countries, especially in highly populated urban regions where new transportation infrastructures are being deployed.

In this paper, we explore Call Detail Records (CDR) of Orange customers in Abidjan, the economic capital of Ivory Coast. The dataset, made available within the context of the D4D Challenge, provides the position of each caller – approximated as the base station's location – at every time he/she initiates a call or sends an SMS. We start by analyzing the temporal, spatial and geographical characteristics of the calls, which allows us to capture differences between distinct times of the day and different days of the week over multiple geographical regions of the city. We propose a method to distinguish between typical and outlying behaviors in the CDR dataset, enabling the detection of special events such as the New Year's Eve and football games played during the Africa Cup of Nations. Our approach also allows us to infer which moments can be aggregated in order to characterize macroscopic mobility flows that provide a view of the global and local mobility flows in Abidjan, as well as of their daily evolution.

The rest of the paper is organised as follows. Sec. II discusses previous studies focused on human mobility traces obtained from wireless network data. In Sec. III, we analyze the CDR dataset. We introduce our methodology to detect typical and special behaviors in Sec. IV. We then aggregate typical behaviors and extract the global mobility flows in Sec. V. Finally, we draw conclusions in Sec. VI.

II. Related work

Human mobility has been drawing significant research efforts over the last few years. Previous work mostly explores real-world movement traces in a variety of contexts and for diverse goals. Kim et al. [1], use wireless network traces from WiFi access points at Dartmouth College to extract a human mobility model. Their study stays valid within the limited context of a university campus, thus cannot be generalized to the level of a city where more complex human movements emerge. More recent works consider social networking platforms. Data from Foursquare is analyzed in [2] to propose a mobility model reflecting movements within a city. Girardin et al. [3] analyze geographically tagged photos from Flickr to uncover the movements of tourists in Rome. Similarly, in [4] the authors separate the behavior of tourists and that of residents in the city of New York. These studies are capable of capturing users' most preferred locations [2] and the main paths followed by tourists [3] [4], however they do not account for the temporal properties of human movements, while in our work we generate O/D matrices capable of capturing both temporal and spatial characteristics of human movements.

Analyses based on CDR have offered studies on human mobility a much wider perspective, uncovering important characteristics at large scales. González et al. [5] analyze the CDR obtained from a European mobile operator providing data on 100,000 mobile phone users over a 6 month period. Their aim is to understand individual users' movements, proving that these mobility patterns present high temporal and spatial regularities. In [6], the authors check the level of movement predictability that can be achieved knowing the history of individual movements. Both of these papers focus on the individual human mobility, while in our study we are interested in working on a more macroscopic level.

Isaacman et al. [7] consider the analysis of CDR by considering a population as a whole. They use the CDR of more than 100,000 individuals randomly chosen in Los Angeles and New York and track their movements over a period of 4 months. The authors compare the moving patterns between population in both cities. The same dataset is also applied in [8] to build a mobility model. Although in these studies the authors consider the metrics at the level of populations, their

main goal in [7] is to check the difference between population behaviors in both cities, and generate synthetic CDR in [8], while we are interested in detecting mobility flows. Closer to our study, the work by Pulselli et al. [9] shows how general trends of movements through the city of Milan can be detected based on call volume variations. However, our methodology, leveraging individual call records, allows to obtain a more accurate description of the global mobility flows.

In conclusion, all previous works consider urban areas of developed countries: their results do not necessarily reflect the mobility observed in cities of developing countries, which is our goal. As a corollary contribution, we focus on separating typical CDR behaviors from outlying ones – a subject that has drawn minor attention to date.

III. Dataset characterization

In our analysis we study datasets provided by Orange within the context of the D4D Challenge, based on the Call Detail Records of 5 million anonymized Orange customers in Ivory Coast. The information obtained span over 5 months, from December 5th, 2011 until April 22nd, 2012. We focus on two of the four datasets provided, detailed below.

Dataset D1: Antenna-to-antenna traffic. This dataset includes the call traffic volume exchanged between any two base stations in Ivory Coast on an hourly basis. It provides both the number of calls and their total duration for the whole observation period.

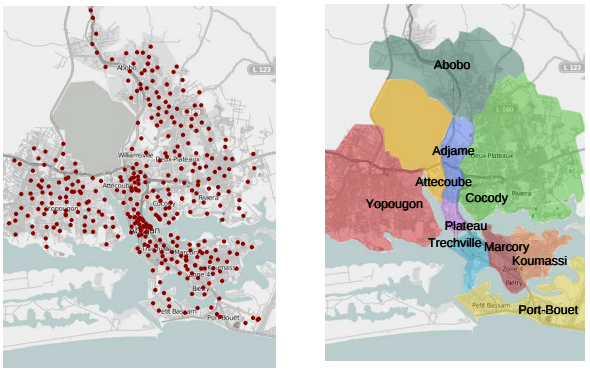
Dataset D2: Individual trajectories. For each two weeks of the observation period, this dataset provides the CDR of 50,000 individuals randomly chosen over the whole Ivory Coast Orange customer population.

Our study focuses on the city of Abidjan, the economic capital of Ivory Coast. Thus, we filter both D1 and D2 by keeping only the information involving the antennas in Abidjan. This leaves us with information about 364 antennas out of the 1231 antennas covering the whole country. In Fig. 1(a) we show the position of these base stations together with the city's street layout obtained from OpenStreetMap [10]. Fig. 1(b) presents the different *communes* of the city, providing a reference for discussions in the rest of the paper. We remark that, in the following, we will use the term *snapshot* to refer to data aggregated over each one-hour interval, and the term *call* to denote a call or an SMS indifferently.

A. Temporal and spatial inter-call properties

We start our analysis by considering the temporal and spatial properties of the second dataset. Specifically, our goal is to determine whether the D2 dataset can be reliably leveraged for our objective of characterizing human movements patterns. To that end, we generate the distributions of two relevant metrics: the inter-call duration that we define as the time elapsed between two consecutive calls made by the same person, and the inter-call distance defined as the distance separating the position of the person between two consecutive calls.

Fig. 2(a) displays the Probability Distribution Function (PDF) of the inter-call duration. We can clearly notice the concentrantion of high probability values for small inter-call durations, and the exponentially decreasing trend with peaks on a daily basis. The plot outlines the heterogeneity in the inter-call duration, as we can observe that, despite the low



(a) (b)

Fig. 1. (a): Distribution of the base stations in Abidjan over the street layout. (b): The 10 communes of Abidjan.

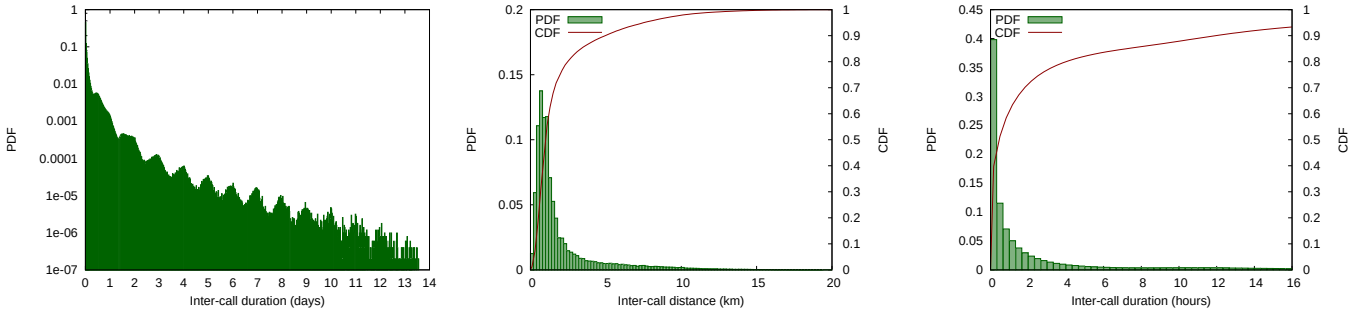
probability, very high values of inter-call duration can appear. Moreover, the probability peaks indicate the periodicity in calling patterns as successive calls are mainly separated by days. More importantly, 90% of the calls occur at one hour distance or less: the high time granularity makes the D2 dataset fit to our analysis from a temporal viewpoint.

After understanding the global temporal trend of calls, we check the distributions of the inter-call distance in Fig. 2(b). For a clearer representation, we remove from this figure the consecutive calls that occur from the same base station, i.e., data that cannot be exploited from a mobility point of view. We remark that this concerns around 70% of the overall records, and we thus focus on the remaining 30% of the data, still statistically sufficient for our study. We can observe in this figure that more than 70% of the inter-call distances are concentrated within a 2km range, which means that when the position of an individual changes between two succesive calls, he/she tends to move within a limited area. Again, short traveled distances between calls positively affect the reliability of human movements inferred from the dataset.

Finally, in Fig. 2(c) we focus on inter-call durations corresponding to successive calls occuring from different locations only. This figure indicates that movements can still be captured with a high temporal precision, as 60% of the calls yielding a physical user movement occur within intervals smaller than one hour. Overall, our analysis shows that the D2 dataset can effectively be leveraged towards the characterization of user mobility, since a significant portion of the records yield low inter-call distances and durations.

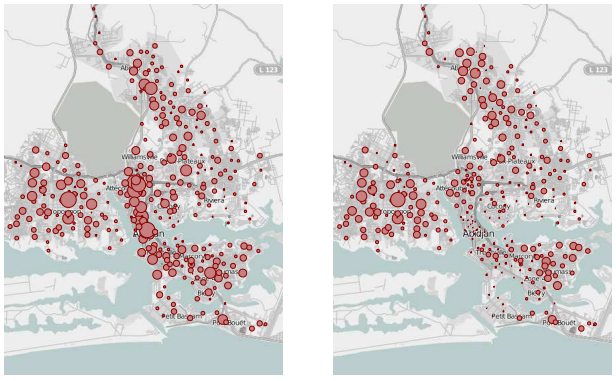
B. Geographical call volume diversity

Although the inter-call duration and the inter-call distance are capable of capturing the global temporal and spatial characteristics of the dataset, they cannot reveal the behavior in different regions of the city. Thus, we examine the distribution of calls over the city obtained from the first dataset D1, through a set of geographical plots. These geographical plots are capable of capturing what happens in different regions for different times of the day at an hourly basis. In Fig. 3, we present an extract of these plots, in which we represent the number of calls at each base station with a disk, whose radius



(a) (b) (c)

Fig. 2. Distributions of inter-call duration and inter-call distance. (a): PDF of inter-call duration for the whole observation period. (b): Distributions of the inter-call distance excluding consecutive calls from the same base station. (c): Distributions of the inter-call duration excluding consecutive calls from the same base station.



(a) Thursday, April 12th at 10:00 (b) Thursday, April 12th at 20:00

(c) Sunday, April 22nd at 10:00 (d) Sunday, January 1st at 0:00

Fig. 3. Geographical distributions of calls.

is proportional to the number of calls detected there. Fig. 3(a) and Fig. 3(b) present the geographical distribution of calls on Thursday, April 12th, 2012 respectively at times 10:00 and 20:00. We pick these times as typically time 10:00 reflects the concentration of individuals at the working and studying areas while at time 20:00 individuals are mostly present at their home locations.

At 10:00, we can clearly see the explosion of cellular traffic at the city center, the region in which most of the working acitivities take place. In some other places, such as the region of Cocody, we can still observe a certain number of base stations with a high density; these are mostly the regions where different universities of Abidjan are located. As for the commune of Yopougon, we can see it presents a heterogeneity at the level of base stations in terms of cellular traffic: base stations with high and low traffic coexist. That is due to the fact that this region is a mix of both residential and industrial areas.

On the other hand, if we consider the same day at time 20:00, we can observe that, in the city center, the traffic strongly decreases; the same applies for the university campuses. As for some residential areas in Abobo, they encounter a traffic increase, since people will be present at their homes, while Yopougon still presents a heterogeneity in the base stations in terms of traffic.

Fig. 3(c) illustrates the distribution of calls on Sunday, April 22nd, 2012 at time 10:00. Compared with the previous 2 figures, we can observe a behavior diverging from the typical behavior of a normal week day at the same time, and resembling more to what can be seen at time 20:00 of a normal weekday. This comes confirming the fact that on Sunday morning people mostly stay at their homes.

As for the New Year's Eve portrayed in Fig. 3(d), first we remark that some of the base stations do not appear due to technical problems that we will discuss in more detail in the following section. Second, for the existing antennas, we can see that the volume of all the base stations explodes except for the working areas.

Based on these observations, we conclude that different hours of the same day show different patterns for the traffic on the cellular network, and these patterns can be linked with the human geographic distribution and mobility. More interestingly, different days at the same time present different behaviors. Finally, some special days can present strongly diverging behaviors. In the next section, we shed some light on the detection of typical days as well as special days and anomalies that we call outliers.

IV. Typical behavior and outliers

In this section, we first discuss the reliability of individual snapshots, and then explain how we tell apart typical and special behaviors in the CDR.

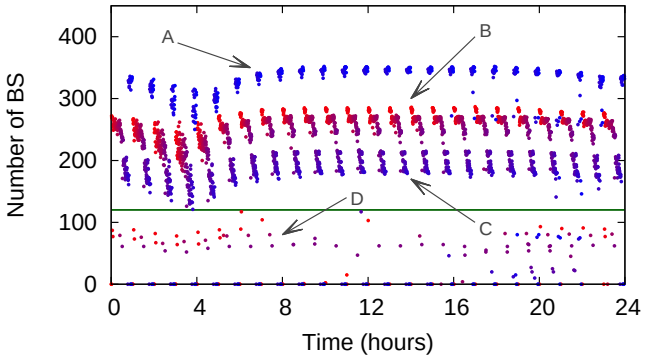


Fig. 4. Number of antennas over the whole observation period for the different hours of the day.

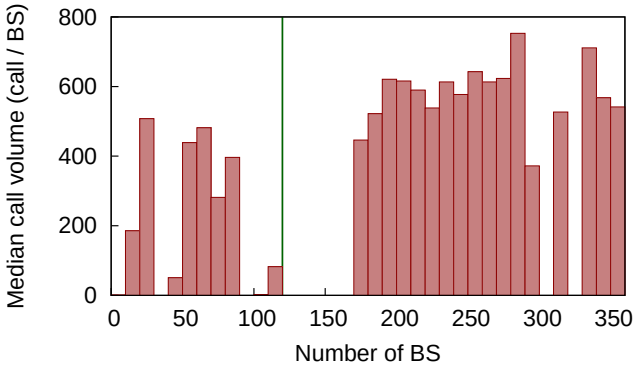


Fig. 5. Median volume per base station as a function of the number of existing base stations, aggregated over the entire observation period for high traffic hours: between 10:00 and 20:00.

A. Determining the snapshot reliability

As we mentioned in the previous section, we notice that some snapshots do not include all the antennas. That is due to technical problems encountered by Orange as well as electricity failures that can occasionnally occur in Ivory Coast. Thus, we now focus on the evolution of the number of antennas throughout the 5-month observation period.

In Fig. 4, we plot the number of antennas detected in D1 with respect to the time of the day. Each point on this figure represents the number of antennas included in the dataset at a specific hour of one day. We distinguish between the different days with a colour degradation from red to blue, such that the red colour maps to the first day in the dataset and the blue colour maps to the last one.

Three different behaviors are detected: the first one (labeled as A), includes the highest number of base stations – around 350 – and goes from March 28th, 2012 until April 22nd, 2012; the second one (B), with almost 250 antennas, goes from February 22nd, 2012 until March 27th, 2012; and the third behavior (C), featuring the smallest number of base stations – around 170 – and goes from December 7th, 2011 until February 21st, 2012.

We also point at the appearance of fluctuations in each of these behaviors, where local minima appear in the night hours between times 3:00 and 6:00, while the number of antennas

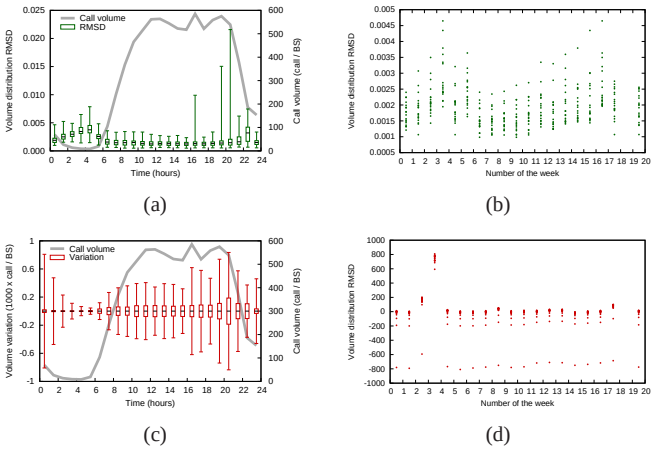


Fig. 6. (a): The volume distribution RMSD for all Sundays at different times of the day. (b): The volume distribution RMSD for all Sundays at time 0:00. (c): The volume variation for all Sundays at different times of the day. (d): The volume variation for all Sundays at time 0:00.

stays almost stable for the rest of the day. The presence of these minima can be explained by the fact that during night hours a very small number of individuals use the cellular network, which means that there is less chance to detect calls from base stations all over the city.

However, we observe that a limited number of points (tagged as D) with a very small number of antennas also arise. These are the result of missing information or malfunctioning of the majority of base stations. We verify that such a significantly reduced number of base stations has an impact on the reliability of the call traffic information in Fig. 5. There, we draw the evolution of the median per-base station call traffic volume with respect to the number of base stations, for all the snapshots with notable call traffic, i.e. between times 10:00 and 20:00, over all the 5-month observation period. The three behaviors A, B and C map to a consistently high median volume per base station. The two intervals centered around the values of 305 and 325 antennas present null median volume values as no snapshot exists with a number of base stations lying in one of these intervals. Conversely, for the snapshots with less than 120 antennas, falling in the D case above, the behavior differs: we observe highly variable median values with a lower average. This leads us to exclude from our analysis snapshots that fall in the D category, i.e., for which less than 120 antennas are recorded, since they yield irregular traffic volume information, due to technical issues in the network and that would risk to bias our analysis.

B. Identifying outlying behaviors

After having pruned the D1 and D2 datasets from unreliable snapshots, we focus on separating typical and special snapshots. To that end, we compare different snapshots through two metrics, defined next.

Let us use $C_i(t)$ to refer to a cell i appearing in the 1-hour interval from time t until time $t+1$. $\mathbb{C}(t) = \{C_i(t)\}$ represents the set of cells detected between time t and $t+1$. We denote the volume of each cell $C_i(t)$ as $v_i(t)$. We use $\mathcal{V}(t)$ to refer to the total volume of calls obtained between time t and $t+1$. We define the intersection of the appearing

antennas between the two 1-hour intervals at times t and $t+k$ as: $I(t,t+k) = \mathbb{C}(t) \cap \mathbb{C}(t+k)$ We define the root mean square deviation of the volume distribution distance, or *volume distribution RMSD*, between two different 1-hour intervals at times t and $t+k$ as:

$$\mathcal{D}(t,t+k) = \frac{1}{\sqrt{|I(t,t+k)|}} \sqrt{\sum_{C_i \in I(t,t+k)} (\frac{v_i(t)}{\mathcal{V}(t)} - \frac{v_i(t+k)}{\mathcal{V}(t+k)})^2}.$$

The $\mathcal{D}$ metric captures the distance between two snapshots, in terms of distribution of volumes among the base stations present in both snapshots.

We also account for the overall volume variations, by calculating the *volume variation* per base station as:

$$\mathcal{DV}(t,t+k) = \frac{\sum_{C_i \in I(t,t+k)} (v_i(t) - v_i(t+k))}{|I(t,t+k)|}$$

The $\mathcal{DV}$ metric captures large positive or negative variations in the calling volume.

We leverage the two metrics above to compare the same hour of the same days of the week at different dates: for example we compare every Saturday at 7:00 with all the other Saturdays at 7:00.

In Fig. 6(a) and Fig. 6(c), we show the evolution of $\mathcal{D}$ and, respectively, $\mathcal{DV}$ for all the Sundays for different times of the day. The candlesticks show the median, minimum and maximum values, together with the first and third quartile of each of the two metrics. We also show on each of these figures the median volume detected per base station. It is clear from Fig. 6(a) that $\mathcal{D}$ stays mostly small, but still presents some relatively high outlying values. We can also notice the candlesticks positioned slightly higher for the night hours. If we map that with the variation of the median volume per base station, we notice that this shift is caused by the variation of traffic: in the case of small traffic, it is more probable to detect variations in the volume distribution RMSD. It is noteworthy that these variations are not necessarily linked to the overall cellular traffic volume: different volumes distributed in a similar geographic manner result in small $\mathcal{D}$, while a similar total volume distributed differently can produce high $\mathcal{D}$ values.

As for the volume variation in Fig. 6(c), we can see the larger fluctuations are detected in the day hours, when high volumes are reached; this can be the result of large increase in the calling volumes on special events or large decrease in the calling volume due to the technical problems.

In the next step, we detect outliers based on the boxplot technique [11] such that: if Q1 and Q3 represent the first and third quartile, and IQR = Q3 - Q1 the interquartile range, then every value that is smaller than Q1-1.5*IQR or greater than Q3+1.5*IQR is detected as an outlier. We eliminate every snapshot causing outliers detected based on $\mathcal{D}$ since it reflects untypical cellular traffic distribution. As for the outliers detected based on $\mathcal{DV}$, we distinguish between negative and positive variations. The negative variations represent moments of relatively very low traffic, which can be due to technical problems in producing the datasets. As these problems are not uniformly distributed over all the base stations, such negative variations usually produce outliers on the $\mathcal{D}$ set; therefore these snapshots are removed from the aggregation process described in the next section. However, a different trend can be observed for positive variations of the total volume, usually produced by special events when people use their phones simultaneously. Our analysis shows that these snapshots do not necessarily result in outliers on the volume distribution RMSD, meaning that the use of the cellular network changes (as more people make a call), but the geographical distribution of the users does not modify.

For example, this technique allows us to detect special days such as the New Year's Eve, which happened to be on a Sunday and it is therefore shown in Fig. 6. This outlying value is detected by a high positive volume variation and a high value of $\mathcal{D}$, which means that it does not only imply a volume variation, but also a different volume distribution RMSD. This can be caused by the fact that possibly on such events, people gather in special places to celebrate. This outlying value can be better noticed in Fig. 6(b) and Fig. 6(d). In these figures, we show the obtained $\mathcal{D}$ and, respectively, $\mathcal{DV}$ values at time 0:00 for the different Sundays of the dataset. We refer to every Sunday based on the number of the week to which it belongs such that the first Sunday of the observation period is designated by 0. In these plots, every point represents the value of $\mathcal{D}$ and $\mathcal{DV}$ obtained by considering the snapshot at time 0:00 on the Sunday specified on the x axis, with another Sunday snapshot at the same time. We remark that The New Year's Eve is the fourth column on these figures, and the high relative traffic volume and RMSD distance can be distinguished.

However, the volume variation $\mathcal{DV}$ also allows us to detect 5 out of 6 football games of the Africa Cup of Nations 2012, hosted by Gabon and Equatorial Guinea, in which the Ivory Coast participated, with the last one left undetected because it presents less than 120 antennas and thus is excluded of this part. As an example, we detect the final game that took place on Sunday, February 12th, 2012, causing the positive outlying values at 23:00 in Fig. 6(c). Nevertheless, these events do not present an abnormal distribution of the cellular traffic, despite the increased overall volume, therefore they are not detected as outliers on the $\mathcal{D}$ set. This is an interesting result, as it shows that, despite the common practice in other studies which recommend excluding high traffic moments when studying human mobility based on CDRs, these snapshots can still be useful from a mobility point of view when they present classical geographical distribution patterns. Moreover, their use can even bring benefits, as the increased traffic on the cellular network allows us to detect more individual trajectories during these moments.

Finally, our results allow us to conclude that it is possible to aggregate the same times of the typical days presenting acceptable values of the volume distribution RMSD and the volume variation.

V. Mobility vectors

In this section, we construct a series of origin-destination (O/D) matrices that accurately represent the human mobility flows over the Abidjan urban area and discuss the major mobility flows detected using this procedure.

A. Building O/D matrices

While the analysis proposed in the previous section is based on the first dataset, D1 can not be used as the main source for an O/D matrix. The reason is that, although it provides us with a view of the entire traffic over the cellular network, the dataset D1 is focused on values aggregated per base station, without any information on the individual trajectories. Therefore, in order to construct the mobility flows in the Abidjan area, we make use of the dataset D2 which contains individual information for a subset of the users.

However, using only a subset of the individuals, sampled by the operator following an unknown distribution, can reduce the accuracy of the mobility matrices. To alleviate this problem, we decided to aggregate multiple days, which allows us to follow a larger set of individuals; for example we aggregate multiple Tuesdays at time 10:00, therefore creating a *classical 10am Tuesday*. In order to establish which days can be aggregated, we use input from the detailed analysis of the dataset D1, as follows:

i) Because in dataset D2 the position of every individual is associated to the geographical location of the serving base station, to accurately represent the mobility flows in the city of Abidjan it is preferable to use the snapshots containing the highest number of antennas. We therefore chose the consecutive days from April 2nd, 2012 until April 15th, 2012 as they include the complete set of antennas, capable of offering a clear picture of the movements all around the city.

ii) Even during this period, during some time intervals the cellular traffic might be affected by special events or situations. As our goal is to detect the mobility patterns in a *classical* day, we use the outliers detection technique described in the previous section to eliminate these special moments from the aggregation process. Overall, this leaves us with more than 32k individual trajectories, representing almost 1% of the entire Abidjan population.

In the following, we denote as $\mathcal{OD}(t)$ the O/D matrix representing the mobility flows between time t and $t+1$. $\mathcal{OD}(t)$ is a square matrix of order n_a, where $n_a = \max_t |\mathbb{C}(t)|$ represents the highest number of base stations simultaneously present in the dataset. Every line and column i of the matrix represents the outgoing, respectively the incoming, human mobility flow in the geographical area covered by base station C_i during the time interval between t and $t+1$. We denote as Δt the *time step* of the O/D matrix, practically the time duration between t and $t+1$. As D1 provides data on per-hour basis, we decided to keep this time granularity in our mobility analysis, therefore $\Delta t = 1$ hour. This means that every element o_{ij} from $\mathcal{OD}(t)$ represents the number of trips with a starting point at base station C_i and arriving in the area covered by C_j during a Δt time interval.

However, building the O/D matrix from dataset D2 is not a trivial task. Indeed, when using CDR to infer human trajectories, the accuracy of an individual movement is given by the frequency of the CDR data. As shown in Fig. 2(a), the inter-call duration presents a very heterogeneous distribution, which makes it difficult to map the detected displacements to a mobility matrix using a fixed time-step. For example, an important question is how to include in $\mathcal{OD}(t)$ a movement that spans over multiple hours. In this sense, we distinguish among three types of movements:

i) A movement with a duration smaller than Δt where both the starting and end time are situated between t and $t+1$. An example from this category is a person who calls two times between 9am and 10am, the first time using base station i and the second time while being associated to base station j. We consider such a movement adds a weight of 1 to the element o_{ij} from $\mathcal{OD}(9)$.

ii) A movement with a duration smaller than Δt spanning over two snapshots. This is the case of a person making a call from base station i at 9:45am and another one at 10:20am using base station j. The problem in this case is that such a displacement needs to be shared by the two matrices $\mathcal{OD}(9)$ and $\mathcal{OD}(10)$. While different strategies could be applied in such a scenario, in our study we decided to add a weight of 1/2 to the element o_{ij} in both matrices.

iii) A movement with a duration higher than Δt. Such an event clearly covers multiple snapshots, therefore we denote as n_s the number of O/D matrices that need to take into account the movement. For example, a person detected near base station i at 9:10 and who makes the next call at 15:30 using base station j has obviously moved between these two moments and needs to be considered in at least one of the $n_s = 7$ Δt intervals covering this time period. Although it is highly probable that the person in question did not continuously move for the entire interval, and the movement could actually be restricted to a smaller number of snapshots, the limited resolution of CDR data does not allow this operation. Therefore, we decided to add a weight of $1/n_s$ to o_{ij} in each of the n_s matrices covering the detected time interval. However, we need to point out that, according to Fig. 2(c), the probability of high inter-call durations is relatively low, meaning that the impact of these *long movements* on the mobility flow is not very important.

B. Identifying major mobility flows

The O/D matrices obtained using the process described above are an essential input for studies in different scientific areas, such as intelligent transportation systems, urban infrastructure planning, or cellular network deployment. However, presenting numerical values in a row/column format is not the best choice for visualization purposes. The solution we adopted to solve this issue consists in associating a vector $\vec{V}_{ij}(t)$ to every value o_{ij} in $\mathcal{OD}(t)$. The origin of $\vec{V}_{ij}(t)$ is the location of base station C_i, while its direction is given by the segment connecting base stations C_i and C_j. Finally, the length of the vector $|\vec{V}_{ij}(t)| = o_{ij}$.

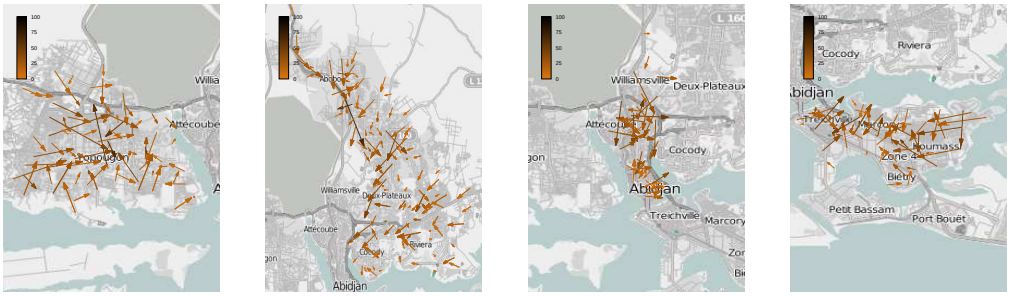
As each pair of antennas results in a vector, this implies that every base station C_i has a number of $n_a - 1$ associated vectors created this way. In the next step, we compute the resultant vector $\vec{V}_i(t)$ for every C_i, as follows:

$$\vec{V}_i(t) = \sum_{j \in \mathbb{C}(t), j \neq i} \vec{V}_{ij}(t)$$

While $\vec{V}_i(t)$ practically represents an aggregation of the human flows, we believe it can represent very well the general trends of the urban mobility. Moreover, as it can be noticed in Fig. 8, even after this aggregation the number of vectors on a figure

(a) 6:00 (b) 8:00 (c) 13:00 (d) 19:00 (e) 22:00

Fig. 7. Mobility vectors of each group of base stations for different times of the day for Thursdays.



(a) Yopougon (b) Cocody and Abobo (c) Plateau and Adjame (d) South of Abidjan

Fig. 8. Mobility vectors of each base station for different regions of Abidjan for Thursdays at 8:00.

is high, which makes the mobility flows difficult to follow. Therefore, we use these per base station vectors only when we want to zoom on a geographical area to uncover local flows. For city-scale mobility, we add another aggregation step, where we divide the entire map in squares with an area of 1.2 km^2, and we group together the base stations in every square, by summing all the associated vectors.

In Fig. 7 we represent the global mobility flows in Abidjan, for different daytimes of a classical Thursday, while in Fig. 8 we represent the mobility flows per base station at time 8:00 for separated regions of the city. These vectors are represented with arrows whose sizes map to their lengths, while the colors map to the total volume of outgoing movements. It is noteworthy that these two metrics are not necessarily correlated, as a geographical area with a high number of outgoing trips uniformly distributed in the space would result in a high total volume, but a small length of the resultant vector.

In Fig. 7(a), at time 6:00 we can observe a general trend of vectors oriented towards the city center, however their sizes and colors reflect the presence of only minor movements.

At time 8:00, represented in Fig. 7(b), we notice that the vectors present an increased length, meaning that a higher level of mobility is detected, mainly due to the movement of people between their homes and their working or studying areas. Most of the vectors are oriented towards the Plateau, as it constitutes a working and studying area itself, and a bridging region between the north and south of Abidjan.

At time 13:00, we observe a different behavior in the multiple regions of the city, shown in Fig. 7(c). Important flows emerge from the center; this pattern is probably related to the part time jobs that a part of the population has, which means that people in this category go back to residential regions at this time of the day, such as in Cocody. More interestingly, at time 19:00 presented in Fig. 7(d), the behavior detected is almost the opposite of what can be seen for the case of 8:00 as this is the time when most of the people go back to their homes using the same paths taken in the morning, but in the opposite direction. Finally, at time 22:00, Fig. 7(e) clearly shows that the city calms down: less movements can be detected, it is the time when most people are already back at their homes.

While these general trends are very clear, it is interesting to observe local mobility flows, using a per base station representation. If we focus on each commune of Abidjan separately at time 8:00, we can observe the following:

Yopougon: In this commune, represented in Fig. 8(a), we observe important flows directed towards the center of the commune. This can be explained by the fact that the bus station is located in that area, with public transportation offered to the residents to reach the other regions of Abidjan. In addition to that, we notice the presence of arrows directed towards the center of the city. This behavior is expected as people will mostly go to this region for their daily activities or transit to the communes of Cocody or Marcory for the same reason, as already observed on the general flows in Fig.7(b). We can also detect some flows oriented towards the north of the commune, going in the direction of the highway that links Yopougon to the other regions of the city. Vectors pointing at the south east of Yopougon are also detected, and they could be explained by the presence of the boat station, where another public transportation service is offered.

Abobo and Cocody: The arrows in these regions, drawn in Fig. 8(b), are mostly oriented towards the south following the direction of the main streets and the highway leading towards the Plateau.

Plateau and Adjame: In the Plateau we can observe from Fig. 7(b) a small dark arrow, which means that in this region an important volume of movement is detected but spread in different directions. This is clearly confirmed by the individual base station's vectors in Fig. 8(c), showing that in this area the destinations of the individual trips are much more heterogeneous. The same effect holds as well for the commune of Adjame represented on the same figures.

Marcory, Treichville, Koumassi: In these communes we can see from Fig. 7(b) that the largest arrows are mostly oriented towards the center of the southern part they form. In this case, it can be due to the working nature of Marcory or the fact that in this central area people can reach the highway to go to the north of Abidjan. We can also observe a large flow oriented towards Cocody, which is mostly representative of the students' flow. The vectors per base station in Fig. 8(d) come to confirm this behavior.

Our observations reflect the high commuting activity between Northern Abidjan and Southern Abidjan, notably at peak traffic hours such as 8:00 and 19:00. While the two parts of the city are joined by two bridges, they both seem to suffer from an insufficient capacity regarding the high traffic congestions during rush hour times. Thus, our results confirm the need for a third bridge linking both parts to reduce the level of traffic congestions.

VI. Conclusion and future works

In this paper, we analyze Call Detail Records of mobile phone users in the city of Abidjan. We introduce a method that allows to distinguish between typical and special calling behaviors of the population. We extract the global mobility flows across the whole city, which we prove to reflect the dynamics of the lifestyle in Abidjan. We believe that our results can help solve important problems in the city. They can be exploited to improve the public transportation services, adapting the paths taken by buses and taxis as well as their number to the mobility flows. They can also be used when considering traffic problems and the road infrastructure, showing where new roads are needed: our results clearly confirm the need for the construction of the third bridge linking the northern and southern parts of the city. Mobile phone services can be ameliorated as well based on our results, by adapting them to the macroscopic movements detected taking into account the geographical locations where people mostly cluster at different times of the day. As for the accuracy of the obtained flows, we remark that it can be improved by considering larger, more complete datasets. Finally, our results are limited to one city in Ivory Coast, a similar study can be achieved for the other areas of the country, also it can be interesting to check the mobility flows between different regions of Ivory Coast when considering the improvement of services in the country as a whole.

References

[1] M. Kim, D. Kotz, and S. Kim. Extracting a mobility model from real user traces. In *IEEE INFOCOM*, 2006.

[2] A. Noulas, S. Scellato, R. Lambiotte, M. Pontil, and C. Mascolo, A tale of many cities: universal patterns in human urban mobility, In *PLoS One 7*, 2012.

[3] F. Girardin, F. Calabrese, F. Dal Fiorre, A. Biderman, C. Ratti, and J. Blat. Uncovering the presence and movements of tourists from user-generated content. In *Intnl Forum on Tourism Statistics*, 2008.

[4] F. Girardin, A. Vaccari, A. Gerber, A. Biderman, and C. Ratti. Towards estimating the presence of visitors from the aggregate mobile phone network activity they generate. In *Intl. Conference on Computers in Urban Planning and Urban Management*, 2009.

[5] M. C. González, C. A. Hidalgo, and A.-L. Barabási. Understanding individual human mobility patterns. *Nature*, 453, June 2008.

[6] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási. Limits of predictability in human mobility. *Science*, 327, February 2010.

[7] S. Isaacman, R. Becker, R. Cceres, S. Kobourov, M. Martonosi, J. Rowland, and A. Varshavsky. Ranges of human mobility in los angeles and new york. In *Eighth IEEE Workshop on Managing Ubiquitous Communications and Services*, 2011.

[8] S. Isaacman, R. Becker, R. Cceres, M. Martonosi, J. Rowland, A. Varshavsky and W. Willinger, Human mobility modeling at metropolitan scales. In the *10th ACM Conference on Mobile Systems, Applications, and Services*, 2012.

[9] R. Pulselli, P. Ramono, C. Ratti, and E. Tiezzi. Computing urban mobile landscapes through monitoring population density based on cellphone chatting. *Int. J. of Design and Nature and Ecodynamics*, 3, 2008.

[10] OpenStreetMap, http://www.openstreetmap.org.

[11] J.W. Tukey, Exploratory Data Analysis (limited preliminary ed.), Reading, MA: Addison-Wesley.

Revealing the pulse of human dynamics in a country from mobile phone data

Sanja Scepanovic*[1], Pan Ben Hui†[2] and Antti Yla-Jaaski‡[1]

[1]Department of Computer Science and Engineering, Aalto University
[2]Department of Computer Science and Engineering (SyMLab), HKUST

Abstract

We use mobile phone dataset provided by the Orange operator, Cote d'Ivoire, for Data for Development Challenge (D4D). From the frequency of antenna communication we discover different regions that exhibit more internal communication. From frequency of user calls at certain subprefectures (departments), we infer home departments for the users. This division of users to their home departments then enables us to analyse the different patterns for the different regions of the country and compare how they relate to data from other sources. By analysing users locations at the times of calls, the dynamics of people in the country is revealed. We analyse the user movements, commuting patterns, and time and frequency of calls in different regions of the country.

1 Introduction

Human dynamics is a research branch of complex systems that was particularly influenced by Barabasi, from his 2005 seminal paper which gave an explanation for the bursty nature for many of human activities [1], to his later works, where limits to predictability of human behavior are analysed [13] and, in another, a universal model for mobility and migration patterns is proposed [12]. Availability of variety of large datasets nowadays and the novelty of the field inspired many researchers from different domains, particulalry computer science, to focus on analyzing human dynamics. Mobile phone records have shown to be particularly useful and popular as they provide temporal and spatial information on a scale that was not available to researchers before. In this work, we use a large mobile phone dataset provided by Orange Cote d'Ivoire, a 20 million people African country. The dataset is preprocessed from a 5 million users communication through 5 months period, thus covering significant portion of

*sanja.scepanovic@aalto.fi
†panhui@cse.ust.hk
‡antti.yla-jaaski@aalto.fi

the population and a long enough period to observe patterns in user dynamics. The processed data is provided to us in 4 different datasets. We focus on two of those datasets:

1. **Antenna communication dataset:** geolocation coordinates of antennas that belong to the Orange operator in the country are provided together with their hourly communication (comprising the number of calls and their total duration from one antenna to the other). We model this dataset as a network of communication with antennas as nodes. In this way, we can assign the weight to the edges to be the number of calls between two antennas, or, in another case, the total duration of the calls. The modularity based network community detection algorithm [3] is applied on such networks, and we report results depending on the time period for which we aggregate the statistics as well as depending on the different algorithm parameters we choose.

2. **Low resolution user trajectories dataset:** for half a million users we have the time when a call is made and the subprefecture (department) in which the the call is placed. This dataset enables us to define the home subprefecture for a user and thus divide the half a million users base into groups by subprefectures. Our approach for finding home is stohastic and simple, but validation based on commuting patterns indicates good accuracy of the approach. After such division, we calculate different types of statistics for users from different subprefectures and compare these statistics. We also tackle this dataset from another point: by extracting and aggregating commuting patterns between different subprefectures. In this way we can form a directed network with subprefectures as nodes having the links between them with a weight based on the number of found commuting patterns. We apply a community detection algorithm [3] on this network to discover the commuting regions in the country. Analysis of different call timings in the different regions is also presented.

2 Related Work

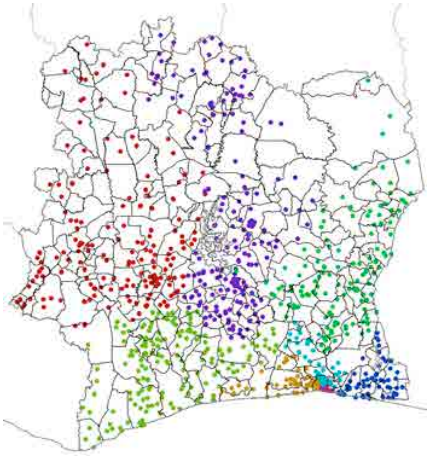
As mentioned previously, mobile phone datasets proved to be very fruitful in the human dynamics research. The authors in [11] give example on how using location-based services (LBS), in particular from mobile phone datasets obtained in the city of Milano, has potential for urban studies and planning. In [6] the authors show how emergency situations can be detected by only observing normal collective calling patterns and alerting those patterns that exceed threshold around mean activity.

On the individual level, they find that the average travel distance of users remains constant during the day as well as the fraction of traveling users. Radiation model, a new model for mobility and migration patterns presented in [12] is based on the observations from U.S. census data on commuting as well as on mobile communication data. The authors explain that the new model avoids many limitations and provides many improvements compared to the gravity model that was used in variety of fields that require predicting population movement earlier. An example where the modularity based community detection algorithm [3] is applied on a mobile network based on frequency and duration of communication is presented in [4]. The communities found by the algorithm show to follow the linguistic borders in Belgium, and perhaps more interestingly, the grouped municipalities are always geographically neighboring. This confirms previously assumed idea: that people communicate more to the people who are close than to those on larger distances.

3 Antenna communication dataset

The number of antennas owned by Orange in Cote d'Ivoire and provided in the dataset is 1231. The communication between those antennas is aggregated on hourly basis during the whole period in the dataset, from 5th December 2011 until 28th April 2012. We create a directed graph of communication with an edge from antenna A to antenna B if there exists a communication initiated from A to B even once during the period. Thus we work with a very dense network, having density 0.898. We assign edge weights based on the total count or total duration of communication between A and B. The modularity based community detection algorithm [9], particularly the fast heuristic approach described in [3], is implemented in Gephi [2], the software that we use. We tried running the algorithm on the network with the described initial edge weights as well as with weights scaled by the total number of calls that participating antennas exhibit during the whole time period. Precisely, for a link between antennas A and B of weight w, the scaled weight is: $w_s = w\ \sqrt{A_{totO} * B_{totI}}$, where A_{totO} is the total weight of outgoing edges from antenna A, and B_{totI} is total weight of incoming edges of antenna B. Running the community detection algorithm gives quite similar output when we use frequency of calls and when we use total duration of calls as link weights, so we present only the output based on frequency. We also test the algorithm in which the holiday period (19th December to 2nd January) is omitted because the data in this period differed from the rest. The communities detected by the algorithm are quite resilient to changes in the parameters. We report the result in Figure 1a.

Figure 1: Antenna communities and language groups



(a) Communities based on frequency of communication

(b) Language groups in Cote d'Ivoire [7]

Similarly to the results reported in [4], we find that the algorithm groups antennas that are geographically adjacent in the communities. However, unlike to the case of Belgium, we do not find that antenna regions follow a simple language border. There are 4 main language groups in Cote d'Ivoire as show in the map 1b. A couple of reasons explain why the communities detected are not strictly following presented langauge borders. First, it is estimated that over 70 different languages are spoken in the country. Linguists agree on the the historic language groups that we show here; however, the borders of the different groups are not always clear in the literature. Second, due to high immigration rate, as well as population migrations, the population is quite diverse, particularly in the cities. Finally, the official language spoken in the country is French and that is the only language that should be spoken by most of the adult (mostly male) population. It is particularly noticeable how the algorithm detects couple of communities in the relatively small area where Abidjan is situated. This might be explained by so called Lagoon complex of ethnicities [7]. It is worth noting that our results do not indicate any ethnic division along the north-south divide of the country during the recent armed conflicts and political elections.

Language Groups	Jaccard index
Mande	0,59
Kru	0,68
Senufo and Baoule	0,68
Kwa and Gur (Zanzan region)	0,93

In order to measure how similar the result of our algorithm is to the linguistic regions, we apply Jaccard similarity index [8]. It is defined as $\frac{|C_i \cap C_j|}{|C_i \cup C_j|}$, where C is a community and $|C|$ is the size of the community. In our case, we use the number of antennas in the given region as the size of the community.

From Figure 1 we can see which output communities in a best way correspond to which language group, and we calculate the similarity index on such pairs. For example, the community output in red color captures mostly Mande Language group, and the community in light green corresponds to Kru languages. We also group our results in Lagoon region, Akan complex and Zanzan and treat them as one big region. The table shows similarity found for all the 4 main groups. Overall, the results show some similarity with the language groups, but the present communities do not clearly match the historic areas.

4 Low resolution user trajectories dataset

4.1 Division of users to subprefectures

Based on obtained user division, we calculate some basic measures that define human dynamics:

1. frequency of calls
2. average trajectory length
3. radius of gyration
4. number of distinct visited places (subprefectures)

There are 255 subprefectures in Cote d'Ivoire. From user calling data, we find a home subprefecture for each user in the following way. We rank subprefectures for each user based on the number of calls user makes during the non-working hours on weekdays (19h to 5h) and during the whole days on the

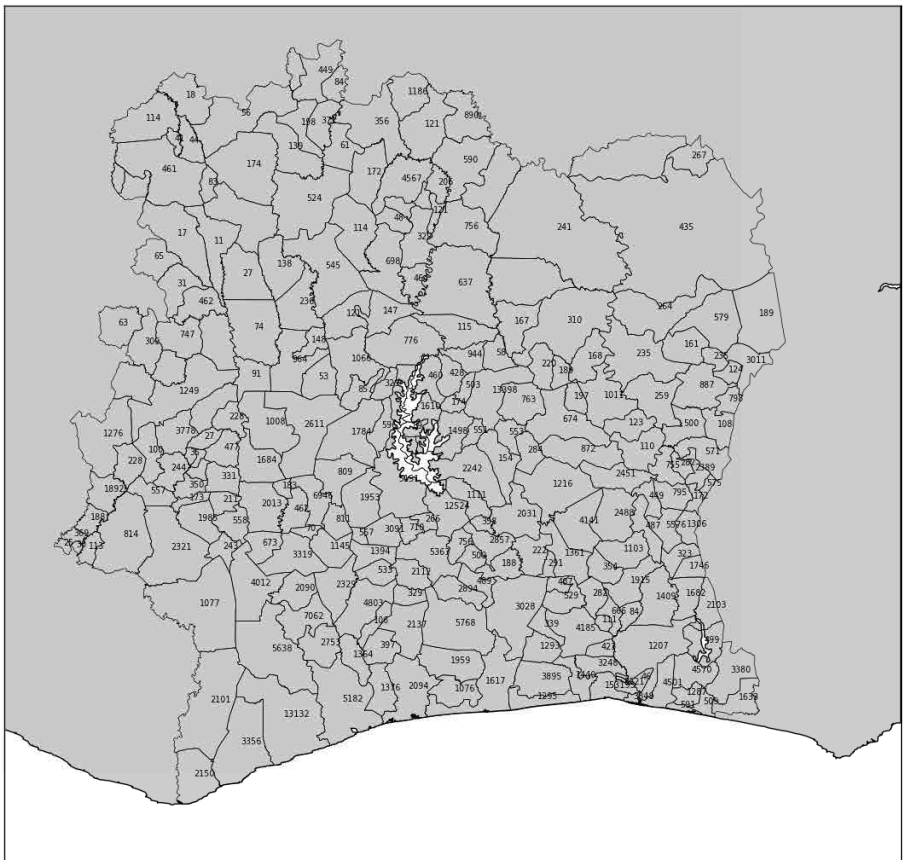


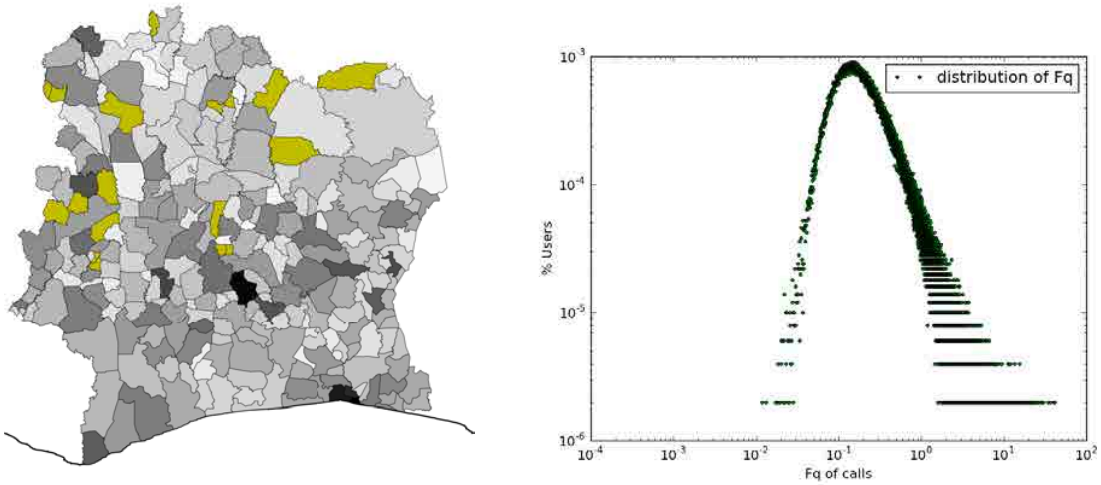
Figure 2: Distribution of number of users per found home subprefecture

weekends. The subprefecture with most recorded calls is defined to be the home. This approach is approximate but we validate it based on the commuting patterns: in at least 85% of commuting trips, the users start from the places we found to be their homes. The distribution of the number of users per subprefecture (Figure 2) is not even, a result we would expect, since the population density and mobile phone penetration are not homogenous in the country. The regions in the north are sparsely populated and in those subprefectures we find the smallest number of users. In some of them, Orange does not have any antennas, so there are no users found.

On the grayscale maps below, a subprefecture is shown in darker color if the value of the variable represented on the map is higher. The subprefectures colored yellow are those for which no users are found.

1. **Frequency of calls:** The grayscale map with average frequency of calls (3a) shows little variation. As we would expect, the frequency of calls is to some extent larger in the subprefectures

Figure 3: Frequency of calls statistics



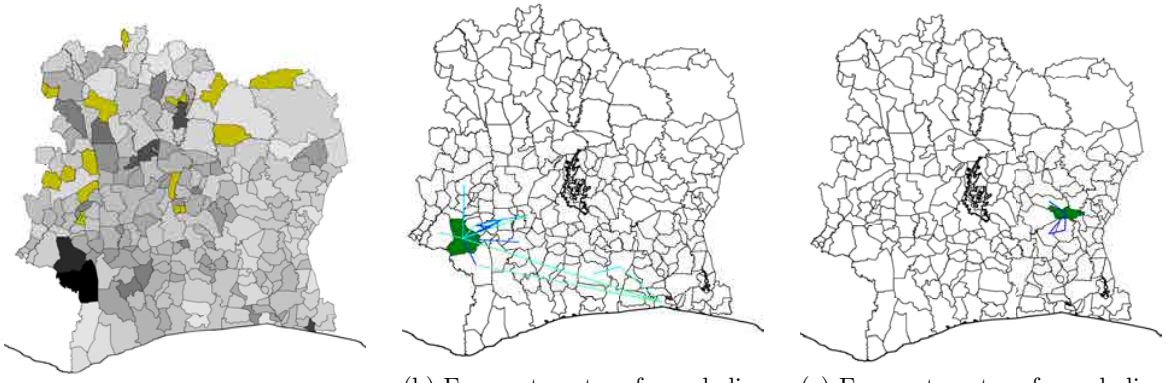
(a) Grayscale map (b) Distribution

hosting larger cities, such as Abijan, Yamoussoukro, Tabou, Odienne, Touba etc. If we, however, look at the distribution of different calling frequencies users exhibit, we can notice large variation. While most users place one call in two hours, we find users with a call per minute! At the same time, this distribution shows many infrequent mobile phone users.

2. **Average trajectory length:** On the grayscale map with average trajectory length, Figure 4a, the two subprefectures in Moyen-Cavally Region are interesting to note. As the map shows, the average trajectory length in those regions in more that 3 times higher than the average in the country. The northern of the two subrefectures hosts the city of Guiglo which is a market centre of the Guere, Yacouba and Mossi people. Guiglo also serves as a depot for timber and coffee that are taken to the ports. Thus the trajectory length of the people living in those subprefectures might be explained by them being traders or truckers. However, we cannot make any strong conclusion because significant amount of the traffic for those people seems to come from commuting between two neighboring regions. Also, in one of those regions, a national park is located.

   In order to gain additional insight into this phenomenon, we apply the following approach. For all the users living in a subprefecture of interest, we observe their *movement graph* obtained through the 5-months period. The movement graph for a user is created from the subprefectures in which the user makes consecutive calls. With this statistics, we cannot precisely follow how

Figure 4: Trajectory length statistics



(a) Grayscale map

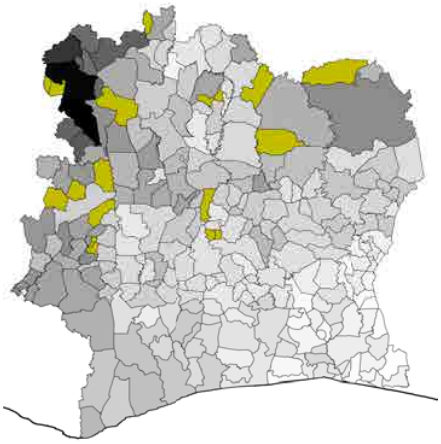
(b) Frequent routes of people living in Guiglo

(c) Frequent routes of people living in Daoukro

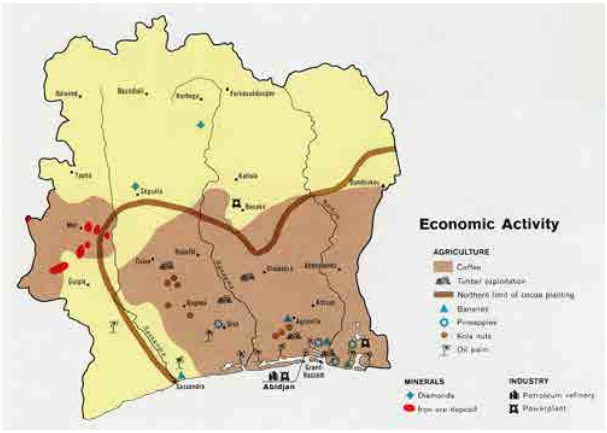
the user travels between two places, but we are sure that he has moved in some way between them. Frequent subgraph mining tool gSpan [14] is then applied on the set of user movement graphs. In figure 4b, we show frequent patterns that emerge with threshold 10% from the users who live in Guiglo. In figure 4c, we show a similar statistics for users in Daoukro, a subprefecture with similar number of users found to live there, but lower average trajectory length. As a conclusion, clear differences in people trajectories in different regions are present, but to understand the reasons behind them, additional information about the people living in these areas would be needed.

3. **Radius of gyration:** The grayscale map 5a presents another interesting pattern of human dynamics in Cote d'Ivoire. Namely, not only that the north-western subprefectures of Odienne and Minignan have average radius of user gyration more that 3 times higher than the average in the country, but also the people living in the south-eastern part of the country, in Lagunes and Sud-Comoe regions, have the same radius 2 times lower than the average in the country. The historical economic activity map of the country (5b), while not up to date, shows strong similarity. The regions shown in brown color are those where cocoa and coffee are grown and where most of the industry is concentrated. On the other hand, regions of Savanes, Denguele, Worodougou, Bafing, Moyen Cavaly, part of Zanzan and part of some other regions are less developed. In these regions, even though the economic situation has started improving lately, people sometimes still do not have the basic educational and health related services available. Thus, the people living in the less developed parts of the country need to travel on a wider

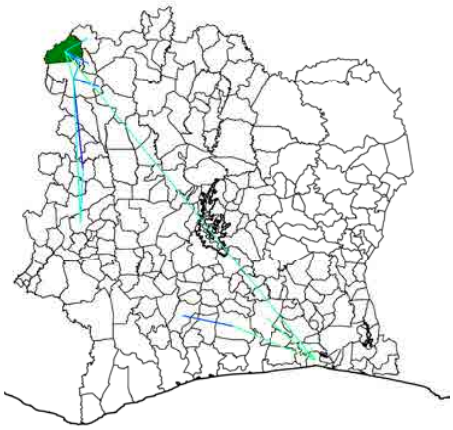
Figure 5: Radius of gyration statistics



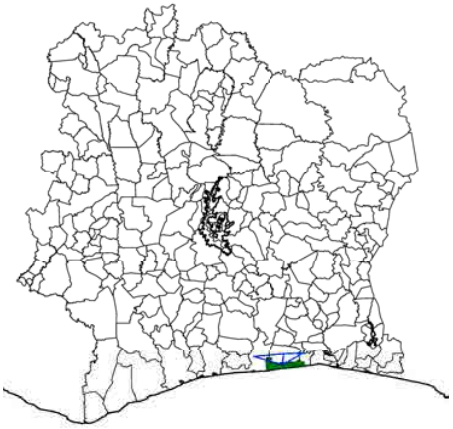
(a) Grayscale map



(b) Historical economic activity map of Cote d'Ivoire [10]

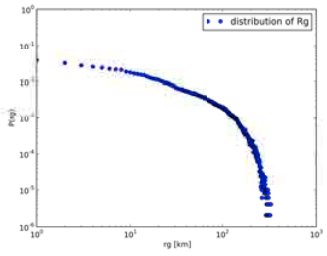


(c) Frequent routes from Odienne

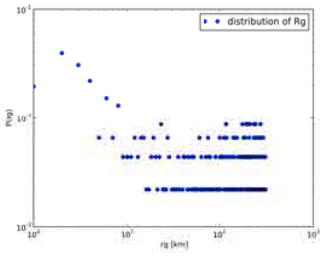


(d) Frequent routes from Jacqueville

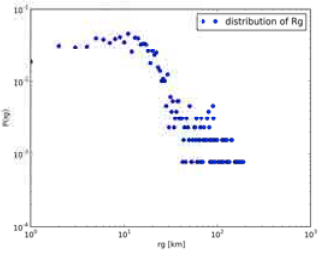
Figure 6: Distribution of length of radius of gyration



(a) Total user base

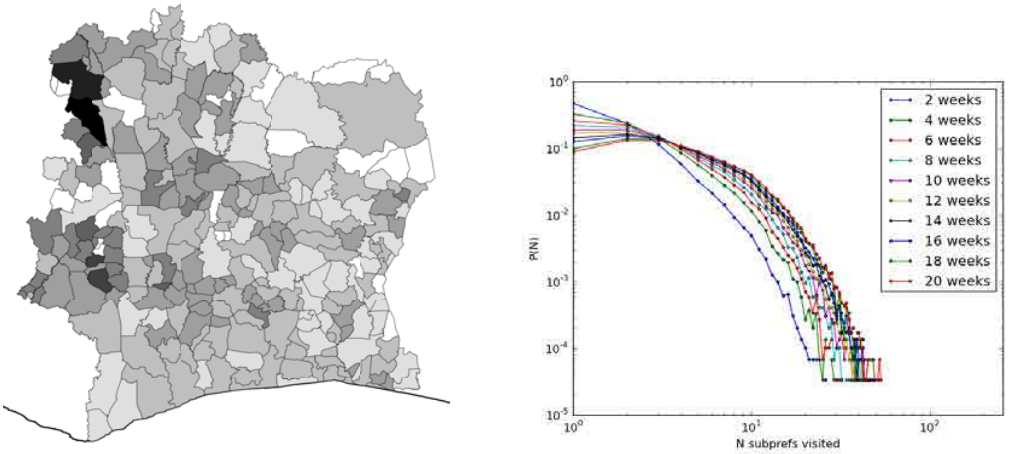


(b) Users in Odienne



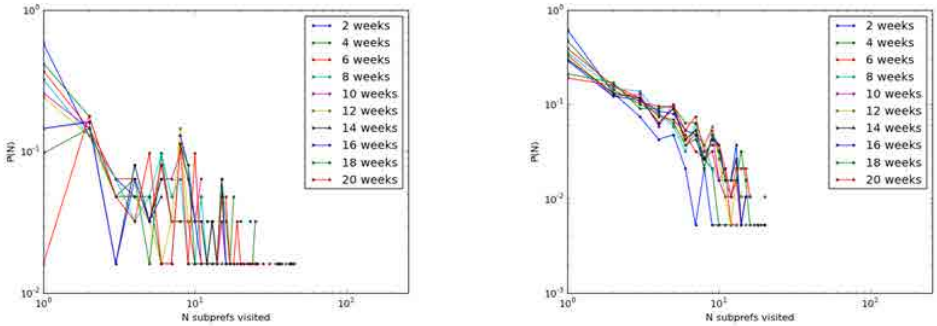
(c) Users in Jacqueville

Figure 7: Statistics on average number of visited places



(a) Grayscale map (b) Distribution per 2-week periods

Figure 8: Distribution of average number of visited places in

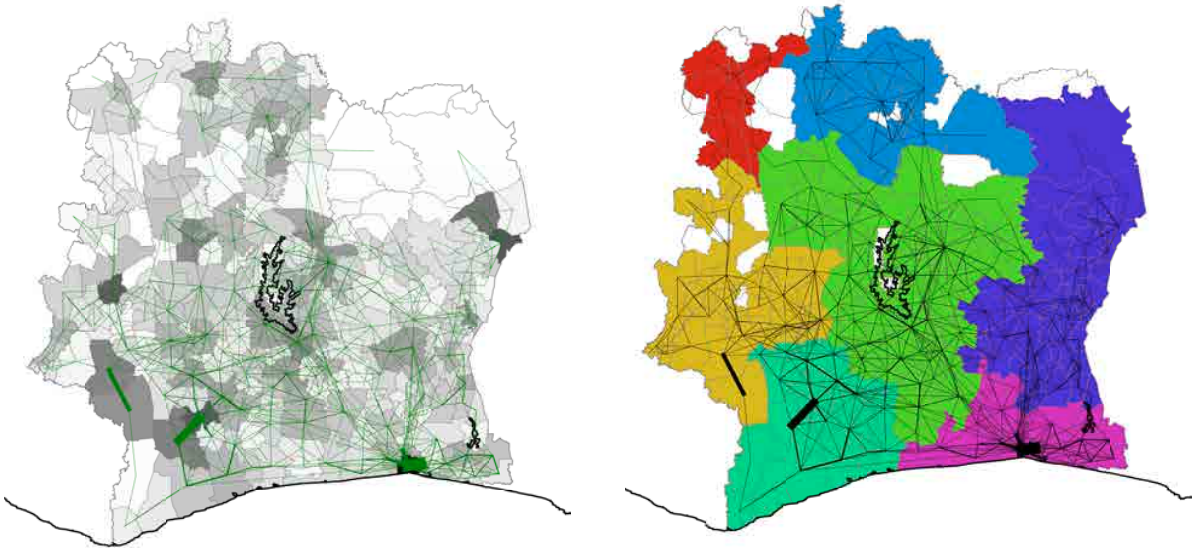


(a) Odienne (b) Bocanda

radius to the industrial and developed regions in order to fulfill their basic needs. At the same time, it is rather interesting how the inhabitants of wealthy southern regions, where also ports and larges cities are located, exhibit a smaller average radius of travel, showing lack of need to travel to the northern parts of the country in general.

We apply frequent graph mining to the user movements graphs in this case as well. At least 10% of people living in Odienne have frequent travels to the south of the country and Abijan, resulting in their large average radius of gyration (Figure 5c). If we make a comparison to a subprefecture in south, Jacqueville, we see from their frequent routes that the people who live there have considerably smaller radius of gyration (5d). The same differences are captured by the distribution of length of radius of gyration in Figure 6. These distributions also show that

Figure 9: Regions of commuting



(a) Grayscale with importance of the destination

(b) Regions of commuting patterns

the people who live in Odienne are specific in their dynamics, having different radius of gyration distribution from the from the total population.

4. **Number of distinct visited places (subprefectures):** When it comes to the average number of visited places by the people living in different subprefectures, we observe a bit more diverse picture. However, the people from the north-east, who travel larger radius on average, naturally do also visit more places (Figure 7a). The distribution for the average number of visited places taken over the various regions is shown on Figure 7b.

   However, as shown in Figure 8a, this distribution is very different when it comes to the people living in Odienne. Here we find a peak at around 9 visited places. For comparison, we also show the same type of distribution for a subprefecture with the number of visited places close to the population average, Bocanda (8b).

4.2 Extracting commuting patterns

In this part of analysis, we focus on the whole set of users and the whole time period. We define *commuting trip* to be a trip taken during one day starting from a certain subprefecture, going through any set of subprefectures and coming back to the starting subprefecture by the end of the day.

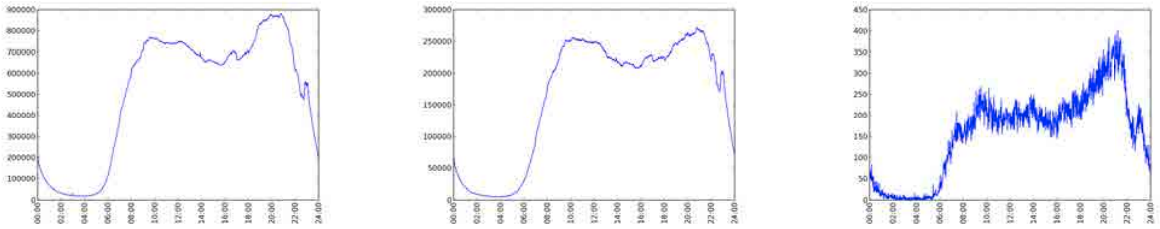
Additionally, such trip must be taken at least 3 times during a week by the user, in order to be counted. With this definition, we obtain a graph with 1379 commuting links between 234 subprefectures. Running the PageRank algorithm [5] on this network ranks the subprefectures by the importance in the commuting pattern of the country, show in Figure 9a. We can see how the cities are ranked higher, as well as couple of subprefectures where intensive commuting happens between (examples are two subprefectures in Moyen-Cavally Region and Soubre and Meaguy in the north of Bas Sassandra). Abijan is obviously the most important commuting center in the country. The thickness of lines between different subprefectures scales with the number of observed trips. Precisely, if a trip is taken more times (no matter whether by the same user or different users) this link is shown thicker. Thus from Figure 8, the frequencies of the trips between subprefectures are observed, which can provide important initial insight for the traffic planners in the country.

Applying the community detection algorithm [3] on this type of network results in a set of regions of commuting shown in 9b. The obtained regions follow nicely the borders of administrative regions (or groups of those). For example, the commuting region show in red corresponds with Denguele region (we point out out that the subprefectures left in white are those for which we did not find users to have calls in them, so they are left out from the network). The cyan color region corresponds entirely to Bas Sassandra and the blue region in the north to Savanes. However, in the region of Abijan, the pink commuting region, the regional borders are blurred and we have parts of Sud-Comoe, Lagunes and Agneby captured. This can be explained by the importance of Abijan as commuting center for the whole country and particularly for the close areas. Similar happens with Yamasuokro, another important commuting center, capturing parts of Lagunes and N'Zi Comore in the large green region in the middle. However, the overall picture shows that the administrative regions in Cot d'Ivoire are as well also important factors when it comes to human commuting.

4.3 Analysing call timings

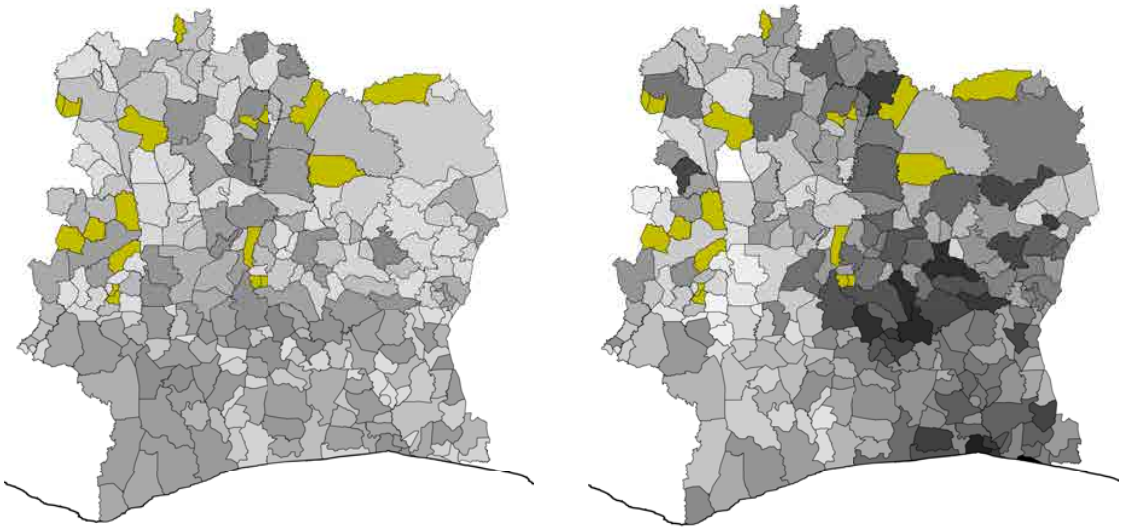
The number of calls made at different times of the day shows to follow certain patterns, for the whole population of the country (Figure 10a) and for different subprefectures (as examples, we provide Figures 10c and 10b). Analysing these calling patterns as signature shows that they are quite resilient over weekdays and weekends, both in different regions and for the whole population. While most of the subprefectures exhibit similar pattern with sharp increase in calling activity after 6h in the

Figure 10: Call timings patterns for



(a) Whole population (b) Region of Abijan (c) Region of Nassian (ZanZan)

Figure 11: Grayscale maps with call timings



(a) Times when 10% of subprefecture average calling frequency is reached

(b) Percent from average calling frequency that happens at midnight

morning and with two peaks, in the morning and evening, we also observe some specific patterns. Some regions exhibit sharper wake-up rise in calling activity, and for some we see prolonged activity during night hours. For the best representation of the diversity of the patterns, we again turn to a grayscale map of the country.

In Figure 11 we can see these patterns. Figure 11a shows how the rural and northern areas reach morning frequency increase to 10% of the average calling frequency earlier compared to the cities and the south. A possible explanation is that the people in those areas live more according to the sunrise, perhaps also as they lack electrical energy. Figure 11b captures night-life pattern, representing percent of its average calling frequency that the subprefecture has at midnight hour. As expected, the cities

have the larger percent. Perhaps less expected, our analysis also indicates that the eastern part of the country tends to have more activity at night hours compared to the western.

5 Conclusion

We show how using a 5-month period mobile communication dataset from a country can reveal many interesting insights about the country, with particular focus to the human dynamics. It is possible to find regions of the country where people tend to communicate more internally, as well as to discover commuting regions (which show high match with administrative regions in this case). From the users mobility, we measure the scale of commuting between departments. Comparing aspects of human dynamics in different subprefectures reveals interesting regional differences and some subprefectures that are outliers.

References

[1] A. Barabasi. The origin of bursts and heavy tails in human dynamics. *Nature*, 435(7039):207–211, 2005.

[2] M. Bastian, S. Heymann, and M. Jacomy. Gephi: An open source software for exploring and manipulating networks, 2009.

[3] V. Blondel, J. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[4] V. Blondel, G. Krings, and I. Thomas. Regions and borders of mobile telephony in belgium and in the brussels metropolitan zone. *Brussels Studies*, 42(4), 2010.

[5] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1):107–117, 1998.

[6] J. Candia, M. González, P. Wang, T. Schoenharl, G. Madey, and A. Barabási. Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical*, 41(22):224015, 2008.

[7] GeoCurrents. The peoples, places and languages shaping current events, 2013.

[8] P. Jaccard. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin del la Société Vaudoise des Sciences Naturelles*, 37:547–579, 1901.

[9] M. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

[10] U. of Texas. Cote d'ivoire (ivory coast) maps, 2013.

[11] C. Ratti, S. Williams, D. Frenchman, and R. Pulselli. Mobile landscapes: using location data from cell phones for urban analysis. *ENVIRONMENT AND PLANNING B PLANNING AND DESIGN*, 33(5):727, 2006.

[12] F. Simini, M. González, A. Maritan, and A. Barabási. A universal model for mobility and migration patterns. *Nature*, 484(7392):96–100, 2012.

[13] C. Song, Z. Qu, N. Blumm, and A. Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

[14] X. Yan and J. Han. gspan: Graph-based substructure pattern mining, 2002.

Profiling workers' activity-travel behavior based on mobile phone data

Feng Liu[a], Davy Janssens[b], Geert Wets[b], Mario Cools[c]

[a,b] Transportation Research Institute (IMOB), Hasselt University, Wetenschapspark 5, bus 6, B-3590, Diepenbeek, Belgium

[c] TLU+C (Transport, Logistique, Urbanisme, Conception) 1, Chemin des Chevreuils Bât B.52/3, 4000 Liège, Belgium

E-mail address: feng.liu@uhasselt.be (F. Liu), davy.janssens@uhasselt.be (D. Janssens), geert.wets@uhasselt.be (G. Wets), mario.cools@ulg.ac.be (M. Cools)

[a] Corresponding author: Tel: +32 0 11269125 fax: +32 0 11269199

Abstract

Activity-based micro-simulation models typically predict 24-hour activity-travel patterns for each individual in a study area. These patterns reflect the characteristics of the available transportation infrastructure and land-use system as well as individuals' lifestyles and needs. However, the lack of a reliable benchmark to evaluate the generated patterns has been a major concern. To address this issue, we explore the possibility of using mobile phone data to build such a validation measure.

Our investigation consists of three steps. First, the daily trajectory of locations, where a user performed activities, is constructed from the mobile phone records. To account for the discrepancy between the movements revealed by the call data and the real traces that the user has made, the daily trajectories are then transformed into travel sequences. Finally, all the inferred travel sequences are classified into typical activity-travel patterns which, in combination with their relative frequencies, define a profile. The established profile represents the activity-travel behavior in the study area, and thus can be used as a benchmark for the validation of the activity-based models.

By comparing the benchmark profiles derived from the call data with statistics that stem from activity-travel surveys, the validation potential is demonstrated. In addition, a sensitivity analysis is carried out to assess how the results are affected by the different parameter settings defined in the profiling process.

1. Introduction

1.1 Micro-simulation model of travel behavior

The main premise of *activity-based micro-simulation models* is the treatment of travel behavior as a derived demand of activity participation. In this modeling paradigm, travel is generally analyzed through daily patterns of behavior related to and derived from the context of the land-use and transportation network in a study region and of the personal characteristics such as social-economic background, lifestyles, and needs of the individuals in the area (e.g. Axhausen & Gärling, 1992; Bhat & Koppelman, 1999; Davidson et al., 2007; Lemp et al., 2007). As such, the modeling system is calibrated using land-use and transportation network information as well as a dataset stemming from household travel surveys which document the full daily activity-travel sequences of individuals during one or a few days. All the input data are analyzed and translated into heuristic decision making strategies which represent the scheduling of activities and travel by the individuals (e.g. Arentze & Timmermans, 2004). Once established, these strategies are used as the probabilistic basis for a micro-simulation process, in which complete daily activity-travel sequences for each individual in the whole population of the region are synthesized, using Monte Carlo simulation.

The synthesized individual activity-travel sequences are afterwards aggregated into origin-destination (OD) matrix, i.e. *a matrix that represents the number of trips between all the different locations of the region*. This matrix, after being assigned to road network, can subsequently serve as input for travel analysis in the region, such as travel demand forecasting, emission estimates and evaluation of emerging effects caused by different transportation policy scenarios. Figure 1 illustrates the entire process of a micro-simulation model.

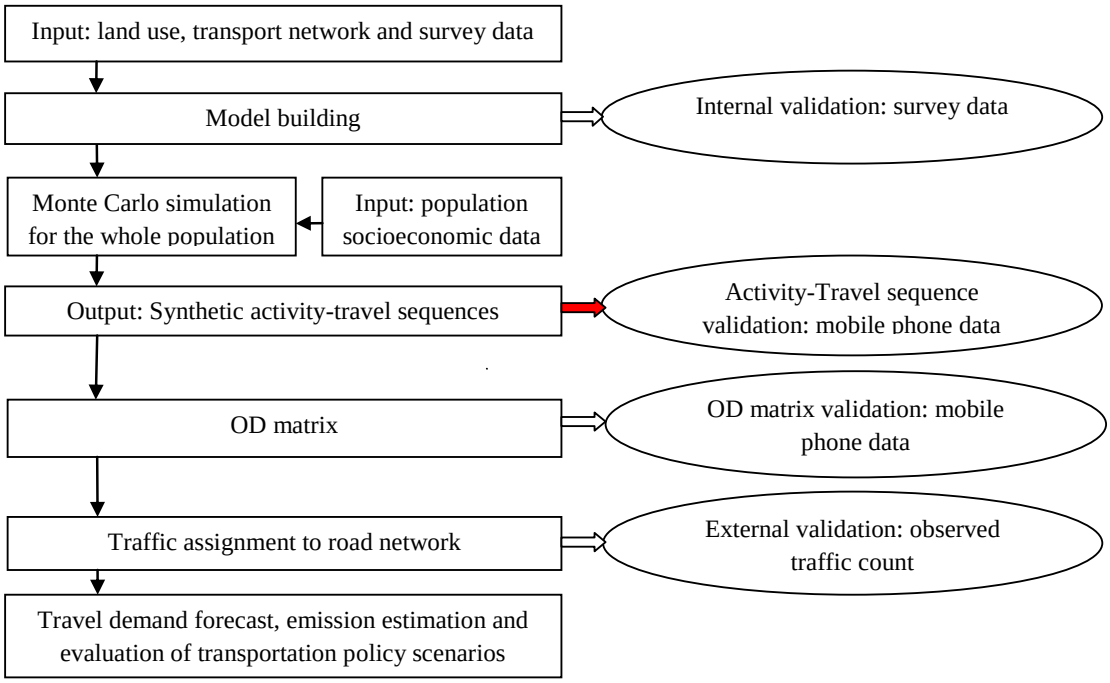


Figure 1. The entire process of a micro-simulation model

1.2. Problem statement

Despite comprehension and advancement of the activity-based modeling system, the lack of reliable data in sufficient size does not enable one to have a decent benchmark and evaluation criterion of the model output (e.g. Cools et al., 2010a; Cools et al., 2010b). Typically, for this purpose, one examines the results of the model both internally and externally at different stages of the simulation process, as indicated in Figure 1 (e.g. Bellemans et al., 2010; Yagi and Mohammadian 2007; Yagi & Mohammadian 2010). The internal validation involves the comparison of the estimation results with expanded survey data which are not used in the training phase of the model but usually collected in the same survey period. However, the process involved in the development of the model, from initial data gathering to exploitation and validation of the first results, is lengthy and may take years, imposing a time lag between the data initially obtained and the data that are required for an objective and up-to-date validation measure. In addition to this time limitation, the issue of budgetary constraints related to the financial cost associated with travel surveys, make it a challenge to collect samples that are sufficiently large to provide a good representation of the activity-travel behavior of a population. Moreover, travel surveys usually query information of only one or two days, to limit the negative effects associated to the respondent burden that is imposed by this type of surveys. This tends to obfuscate the less common activities which occur with a lower frequency (e.g. once a week or once a month), such as sports or telecommuting activities. These shortcomings have been well reported in the literature (e.g. Asakura & Hato, 2006; Cools et al., 2009; Wolf et al., 2001).

In contrast to the internal validation, the external validation consists of indirect evaluations of the model output at a later phase, i.e. traffic assignment stage (see Figure 1). The traffic volumes estimated by the model and assigned to transport network are compared against

commonly available external information sources, such as traffic counts collected by inductive loops detectors deployed on the road network.

However, this external validation process encompasses an aggregation step to compose the OD matrix which is assigned to the road network. Valuable information may be lost in this process. A major limitation that results from this loss of information is that positive outcomes of the comparison might be artifacts of the validation process itself, and thus provide no real guarantee of the accuracy of the model. Moreover, when mismatches are found, there exists no clear procedure to identify the causes, thus limiting remedies to improve model construction. Despite such limitations, at the present, indirect evaluation is essentially the only option for model quality assessment, as no well-established methods are known for operating closer to the model itself. This is a problem that seriously hampers further model development and model application. Having useful and reliable benchmark and evaluation criteria for activity-based micro-simulation models thus is a major concern. Nonetheless, to a large extent, this aspect is neglected in currently available benchmarking standards.

The wide deployment of mobile phone devices provides a very promising source of information on measuring people's transfer phenomenon. Mobile phone data reflect up-to-date travel patterns on significantly large samples of population – in terms of both spatial coverage and temporal extension, making them a natural candidate for the analysis of activity-travel behavior. The importance of mobile phone data in traffic related researches has been manifested by extensive studies on the development and application of the data (e.g. Hansapalangkul et al., 2007; Liu et al., 2013; Ratti et al., 2006; Rose 2006; Steenbruggen et al., 2011). Especially, OD matrices have been constructed based on mobile phone data in a number of regions and countries (e.g. Calabrese et al., 2011; Sohn & Kim, 2008; White & Wells, 2002), and they can be used for travel demand analysis after being allocated to a specific road network. Besides, these matrices can also be utilized for the examination of ODs generated by the simulated travel sequences, as indicated in Figure 1. The feasibility of the benchmarking approach at the ODs level based on mobile phone data has been explored in a recent European project 'DATA science for SIMulating the era of electric vehicles', namely DataSim (http://www.datasim-fp7.eu/). However, while moving the validation process one step closer to the simulated travel sequences, the comparison with ODs still involves an extra procedure of the calculation of the OD matrices. Consequently, the validation is unable to provide a direct assessment on the simulated sequences themselves, and therefore still does not fully address the problems which are related to the external validation measures.

1.3. Research contributions

Extending the current research on the application of the massive mobile phone data in traffic demand analysis, and particularly addressing the above mentioned limitations in having reliable evaluation measures for travel behaviour simulation models, our study proposes a new approach which is to build a profile of workers' activity-travel behavior, i.e. the relative frequency of each typical pattern which represents a certain class of activity-travel sequences, based on the mobile phone data. This profile can be used to directly evaluate the sequences yielded from the simulation models by comparing it against the frequencies of the corresponding pattern classes obtained from the simulated sequences (see Figure 1). This comparison is done at the level of the generated activity-travel sequences, thus capable of

detecting problems that are directly caused by the model itself and providing immediate feedback for the enhancement of the model.

Compared with existing validation measures, this approach offers the following advantages. (i) It monitors current activity-travel behavior in a large proportion of population and provides a more representative and up-to-date validation measure. (ii) Through a long recording period of the mobile phone data, inter- and intra- personal variations of travel behavior as well as weekday/weekend and seasonal deviations can be more efficiently captured. (iii) It can offer immediate response to problems directly linked to the model system, allowing problems to be addressed at an earlier stage of the modeling process before they are propagated into further analyses. (iv) It aims at generating a novel measure for evaluating and benchmarking activity-based micro-simulation models, filling in the gap between the development of the comprehensive model system and the lack of a good and widely accepted evaluation procedure. (V) Apart from the above described technical aspects, the mobile phone data is a by-product of phone companies, requiring no extra cost for data collection, thus providing another appeal in terms of financial consideration.

The remainder of this paper is organized as follows. Section 2 describes the typical patterns which characterize workers' activity-travel sequences. Section 3 introduces the mobile phone data and Section 4 details the construction process of location trajectories based on the data. The call location trajectories are then transformed into complete travel sequences by a method proposed in Section 5. Section 6 classifies both the call location trajectories and the travel sequences into the typical patterns which have been established in Section 2, and the profiles which describe the relative frequency of each pattern class are drawn. A case study is subsequently conducted in Section 7, and a comparison of the results against the outcome of real travel surveys is carried out in Section 8. An in-depth analysis on the sensitivity of this approach is further performed in Section 9. Finally, Section 10 ends this paper with major conclusions and discussions for future research.

2. Activity-travel sequence classification

Individuals make choices about the different activities being pursued, and travel may be required to participate in these activities. Traditionally, all activities performed at home are considered as *home* activities; while the remaining ones conducted outside home are categorized into *mandatory* activities e.g. working or studying, and *non-mandatory* activities that include maintenance activities e.g. shopping, banking or visiting doctors as well as discretionary activities e.g. social visit, sports or going to restaurant (e.g. Arentze & Timmermans, 2004; Bradley and Vovsha, 2005). The home, mandatory and non-mandatory activities are represented as 'H', 'W' and 'O', respectively.

The sequence of activities and travel that a person undertakes during a day is referred as the individual's *activity-travel sequence* for that day. A critical difference is imbedded in activity-travel sequences between workers and non-workers: the sequences of workers mostly rely on the regularity and the fixity of the work activity. In contrast, no such obvious periodicity is present in the case of non-workers (Spissu et al., 2009). This motivates the development of separate representations for these two types of individuals' behavior. In this study, only the activity-travel behavior of workers is analyzed. The representation of their daily sequences is described in Figure 2. In this representation, an activity-travel sequence is divided into four

different parts, including: (i) before-work sub-sequences which represent the activities and travel undertaken before leaving home to work as indicated in arrows 'a', e.g. HOH; (ii) commute sub-sequences which account for the activities and travel pursued during the home-to-work and work-to-home commutes (in arrows 'b' and 'd'), e.g. HOW or WOH; (iii) work-based sub-sequences which accommodate all activities and travel undertaken from work (in arrows 'c'), e.g. WOW; (iv) after-work sub-sequence which comprises the activities and travel engaged after arriving home from work (in arrows 'e'), e.g. HOH.

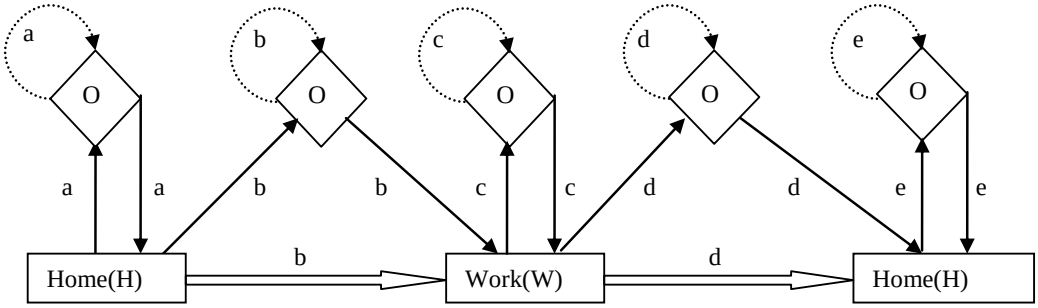


Figure 2. The representation of workers' activity-travel sequences
Note: Each 'rectangular' indicates the home or work location, while the 'diamond' represents a non-mandatory activity location. Each 'arrow' from a home, work or non-mandatory activity location to the other represents the related travel, and the 'arrow' from a non-mandatory activity location to itself indicates the chain of consecutive visits to different non-mandatory activity locations.

According to the above characterization, a home-based tour, comprised of a chain of trips (locations) that start and end at home and accommodates at most two work location visits, can be classified into the following patterns: HWH, HOWH, HWOH, HWOWH, HOWOH, HOWOWH, HWOWOH, HOWOWOH, where each H or W stands for a home or work location while each O represents one or a chain of visits to several non-mandatory activity locations. The days when an individual does not go to work, can be characterized with 2 additional patterns, namely H and HOH. In total, 10 classes are formed to identify each home-based tour in a worker's daily activity-travel sequence, and they are defined as *home-based-tour-classification.*

All the above pattern classes (excluding H) are then merged in pair, leading to 81 combinations which represent daily sequences accommodating maximum 2 home-based tours. For instance, the combination of HWH and HOWH results in the sequence HWHOWH. In addition these pairwise combinations, sequences that contain more than 2 home-based tours, e.g. HWHWHWH, or those that have more than 2 work activity locations in a home-based tour, e.g. HWOWOWH, are each assigned into one additional category. By contrast, a daily sequence can also accommodate only a single home-based tour, e.g. HWH. All these scenarios lead to a total of 93 patterns which underlie workers' activity-travel behavior, and which are denoted as the workers' *daily-sequence-classification.* Given a group of individuals, their activity-travel sequences can be attributed to the corresponding pattern classes. The relative frequency of each of the pattern classes over the total number of activity-travel sequences forms the *profile* of activity-travel behavior among these people.

3. Mobile phone data description

The mobile phone data was collected by a mobile phone company for billing and operational purposes. The dataset consists of full mobile communication patterns of around 5 million users in Ivory Coast over a period of 5 months between December 1, 2011 and April 28, 2012 (Vincent et al., 2012). The data contain the location and time when each user conducts a call activity, including initiating or receiving a voice call or message, enabling us to reconstruct the user's time-resolved call location trajectories. The locations are represented with the identifications of base stations (cells) in a GSM network; the radius of each of the stations ranges from a few hundred meters in metropolitan to a few thousand in rural areas, controlling our uncertainty about the user's precise location. Despite the low accuracy of users' exact locations, the massive mobile phone data represents a significant percentage (i.e. 25%) of this country's population, providing a valuable source and opportunity for the analysis of human travel behavior and for the drawing of relevant inferences that can be statistically sound and representative.

In order to address privacy concerns, the original dataset has been split into consecutive two-week periods. In each period, 50,000 of all the users are randomly selected and assigned to anonymized identifiers. New random identifiers are chosen for re-sampled users in different time periods. The data process results in totally 10 randomly sampled datasets, each of which contains communication records of 50,000 users over two weeks. One of the datasets is selected for this study. Table 1 illustrates typical call records of an individual identified as User2 on Monday, December 12th, 2011.

Table 1. The typical call data of an individual[a]

Time	11:57:00	13:40:00	16:59:00	17:43:00	21:28:00
Antenna_id	898	1020	972	926	926

[a] The 'time' represents the moment when this individual was connecting to the GSM network and the 'Antenna_id' as the cell area where he/she is located.

4. Construction of call location trajectories

A *raw_call_location_trajectory* from a mobile phone user during a day is defined as a series of locations where the user makes calls when traveling or doing activities, as the day unfolds. It can be formulated as a sequence of $l_1 -> l_2 -> ... -> l_n$, where n is the *length* of the sequence, i.e. the total number of locations that the user has reached when using his/her phone on that day, and $l_i (1 \le i \le n)$ is the identification of the locations, e.g. cell IDs in this study. At each l_i, there could be multiple calls, referred as *call_frequency,* denoted as $k_i (k_i \ge 1)$; the time for each of the calls is as $T(l_i,1), T(l_i,2), ..T(l_i,k_i)$, respectively. The time interval between the first and the last call time in the set of consecutive calls, i.e. $T(l_i,k_i) - T(l_i,1)$, is defined as *call_location_duration*. Accommodating the time signatures of the multiple calls, a *raw_call_location_trajectory* can be represented as $l_1(T(l_1,1),T(l_1,2),...,T(l_1,k_1)) -> ... -> l_n(T(l_n,1),T(l_n,2),...,T(l_n,k_n))$, simplified as $l_1(T(1),T(2),...,T(k_1)) -> ... -> l_n(T(1),T(2),...,T(k_n))$.

Given the *raw_call_location_trajectories* constructed from the mobile phone data, the home and work locations are first predicted. This is followed by the identification of stop locations where activities are being carried out.

4.1 Prediction of home and work locations

Various methods have been proposed to derive home and work locations from mobile phone data (e.g. Becker et al., 2011; Calabrese et al., 2011), mainly based on the visited frequency of a location during a particular time period. However, different time windows have been specified in these studies, depending on the context of the study area. In this paper, a similar approach is adopted, but the time windows are empirically estimated from the mobile phone data as follows. The time period when call activities start to increase considerably in the morning during weekdays is chosen as the work start time, denoted as w*ork_start_time.* Secondly, the moment when the second peak of call activities start to appear in late afternoon is considered as the work end time, referred as w*ork_end_time*. Around this time, it is assumed that people start to communicate for off-work activity engagement.

Based on these two temporal points, a location is defined as the home location if it is the most frequent stop throughout the weekend period as well as during the night-time interval on weekdays between w*ork_end_time* and w*ork_start_time*. On the contrary, a location is considered as a work place if it satisfies the following criteria. (i) It is the most common place for call activities in the perceived work period between w*ork_start_time* and w*ork_end_time* on weekdays. (ii) It is not identical to the previously identified home location for the user. (iii) The calls at the location are not limited in only one day, they should occur at least 2 days a week.

With the identification criteria, we assume that people have only one home location and at most one work location. The additional occasionally accessed places for home or work activities are regarded as a stop for non-mandatory activities. In addition, only individuals who work at different locations than their home location areas and who work at least two days per week are included for the analysis of workers' travel behavior.

4.2 Identification of stop locations

After the identification of the distinct home and work locations for each worker, the remaining locations in the *raw_call_location_trajectories* are either *stop locations* where people pursue activities, i.e. non-mandatory activities, or non-stop ones. Each of these non-stop locations could be either a *trip location* where the user is traveling, or a location that is wrongly documented due to location update errors. The location update errors normally occur when call traffic is busy in the user's real location area, and consequently this location is shifted to less crowded cells for short time periods, causing location area updates, without the users' actual moving (e.g. Calabrese et al., 2011; Schlaich et al., 2010).

In addition to the locations which are neither home nor work locations in the *raw_call_location_trajectories* and which need to be differentiated between stop and non-stop visits, the identified home or work locations are also not constantly reached for activity purposes, some occurrences of these locations could be caused by the non-stop reasons. The necessity to identify stop location from non-stop ones can be illustrated with the call records of two typical users.

The trajectory from the first user identified as User265 on a Friday is $l_1(17:06pm,17:43pm)->l_2(17:51pm)->l_3(17:56pm,19:41pm)->l_4(21:55pm)$, where 4 locations are observed, with each lasting 37, 0, 105 and 0 minutes respectively. From this trajectory, a distinction needs to be made to identify stop visits from possible trip visits.

The location update errors can be demonstrated using the call location trajectory of a second user of User72, which is $l_1(13:21pm,20:11pm)->l_2(22:00pm)->l_3(22:02pm)->l_4(22:06pm)->l_2(22:21pm,23:12pm)$. This user has 5 location updates, with the *call_location_duration* as 410, 0, 0, 0 and 51 minutes respectively. However, the time interval between the first visit and the second one to location l_2 is only 21 minutes. The temporary interruption of l_2 by the extra locations l_3 and l_4 in such a short interval most likely resulted from the location update errors. Consequently, locations l_3 and l_4 are falsely connected to the user's mobile phone at 22:02pm and 22:06pm although he/she had been actually remaining at location l_2 during this period.

4.2.1 Identification process

Schlaich et al. (2010) have proposed a method to distinguish a stop visit from a momentary access due to traveling or due to location update errors. In their approach, the interval between the first login of the location l_i under investigation and that of the next one l_{i+1}, i.e. $T(l_{i+1},1)-T(l_i,1)$, is examined. If this interval is longer than a time limit, e.g. 60 minutes in their experiment, l_i is considered as a stop location. However, this method is likely to overlook stop locations where calls are made just before the departure of the locations. In this situation, the time interval can be very short, despite the possibility that users may spend a considerable time period at the locations. This can be further illustrated with the case of User265. The interval between the two first time signatures of locations l_1 and l_2 is 45 minutes, shorter than this 60-minute limit, suggesting that the location l_1 would be for trip purposes. This may be true if this individual has made a long trip of at least 37 minutes within l_1 and made calls at the start and end of this travel. However, if this individual has stayed there doing activities for a long time, e.g. a few hours, and he/she made calls later in this sojourn period, the location l_1 is misclassified by the existing method.

In order to accommodate all the possible stop locations, we propose a new approach consisting of the following steps. (1) For each location visit l_i, the *call_location_duration* is first examined. If it is longer than a certain time limit, denoted as $T_{call\_location\_duration}$, this location is considered as a stop location. (ii) Otherwise, if the condition does not hold e.g. when only a single call being made at the location, and if the location occurs in the middle of a daily sequence of n, i.e. $1<i<n$, a second parameter, namely *maximum_time_boundary*, defined as the time interval between the last call time at the previous location and the first call time of its next location, i.e. $T(l_{i+1},1)-T(l_{i-1},k_{i-1})$, is computed. If this time period is longer than a threshold value, defined as $T_{\max imum\_time\_boundary}$, the location l_i is perceived as a stop visit. (iii) When the location is in the first or last position of a trajectory and the *call_location_duration* is shorter than $T_{call\_location\_duration}$, there is no sufficient information to estimate the maximum possible time for this visit. Thus, all the distinct locations where the user has stayed at least once for carrying an activity, are collected. These locations are considered as potential stop locations that are on the individual's daily activity agenda and that are visited routinely or once in a while. If the first or last visit of a day is to

these locations, it is assumed to be a stop for activity purposes. On the contrary, if this visit is to the place where the individual has not been observed doing activities, it is considered as a passing-by place or being recorded as a localization error and therefore removed.

To exemplify the procedure, we return to the examples of User265 and User72. For User265, based on the parameters of $T_{call\_location\_duration}$ and $T_{\max imum\_time\_boundary}$ which are set up as 30 and 60 minutes respectively in our experiment, l_1 and l_3 are predicted as stop locations, while l_2 is as a trip location due to the short *call_location_duration* (0 min) and *maximum_time_boundary* (13 min). Although only a single call is made at the last location l_4, knowledge gathered from other days has shown that this location has been a regular activity place for this individual. Consequently, this location is labeled as a stop visit. The finally obtained trajectory of stop locations for this user is $l_1 -> l_3 -> l_4$. For User72 this would imply that the locations l_3 and l_4 are deleted as a result of the identification process, and that the divided parts of location l_2 are merged together into a stop location. In comparison, using the existing method which only considers the first temporal logins of two consecutive locations (Schlaich et al., 2010), only one single location would be derived for each of these users, which is l_3 for User265 and l_1 for User72.

After the removal of locations that are either trips or stemming from localization errors, all the remaining locations reached by an individual on a day are formed into a *call_stop_location_trajectory*. Each location l_i in these trajectories is complemented with its function, categorized into home, work and non-mandatory activities, denoted as $activity(l_i)$. Travel is implicit in between each two consecutive locations of these sequences.

5. Transformation of call location trajectories

The considered mobile phone dataset is event driven, in which location measurements are only available when the devices make GSM network connections. Consequently, users' call behavior can affect the possibility of capturing a larger or smaller number of trips and/or activity locations. In general, the more active a user is in communicating electronically with others, the better his/her activity-travel behavior is revealed by his/her call records. The call locations can be seen as the observed behavior at certain temporal sampling moments during a day, and the characteristics of the real travel behavior must be deduced. A transformation therefore should be made from the previously derived *call_stop_location_trajectories* into the sequences that mirror the real picture of people's activity-travel behavior.

During this transformation, we first derive for all the users the actual activity duration as well as the call rate at each minute. These two variables are then translated into the call probability at each location, which describes how likely the individuals make at least one call when they visit the location and which thus indicates to what extent their call records reveal their actual movement. Given a real daily activity-travel sequence, various *call_stop_location_trajectories* could be possibly observed from call data. Next, the probability under which a certain *call_stop_location_trajectory* is generated from the original travel sequence is calculated based on the call probabilities at these locations in the travel

sequence. Finally, given the observed frequencies of the *call_stop_location_trajectories*, a linear equation is built and the frequencies of the original travel sequences are inferred.

5.1 Call rate and actual location duration

Call_intervene for an individual measures the time interval between each two calls, and it is calculated as the ratio between the total number of calls each day, denoted as *total_number_calls(individual, day),* and the time span of the day (measured in minutes), denoted as *time_span(day),* as follows.

$$call\_\text{int}ervene(individual) = \frac{\sum_{day} time\_span(day)}{\sum_{day} total\_number\_calls(individual, day)}$$

The average call intervene across all the users is obtained as

$$average\_call\_\text{int}ervene = \frac{\sum_{individual} call\_\text{int}ervene(individual)}{total\_number\_of\_individuals}.$$

Based on the *average_call_intervene,* the variable of *call_rate* which describes the probability that the individuals makes calls each minute, can be calculated as

$$average\_call\_rate = \frac{1}{average\_call\_\text{int}ervene}$$

The other important variable, defined as $actual\_location\_duration(individual, l_i)$, specifies the actual activity duration *(in minutes)* at a location l_i for an individual. This variable is simplified by the average duration over all individuals across all locations with the same activity purposes as follows.

$$average\_actual\_location\_duration(activity) = \frac{\sum_{individual} \sum_{l_i=activity} actual\_locationy\_duration(individual, l_i)}{\sum_{individual} total\_number\_visit(individual, activity)}$$

Where the *total_number_visit(individual,activity*) represents the total number of actual visits by the individual to the locations with the particular *activity* purposes, such as home, work or non-mandatory activities.

5.2 Call probability at a location

Given a user's call rate and the duration of a location l_i where the individual has actual spent, the probability of making at least a call during the entire period of the visit to the location, defined as $CallP(l_i)$, can be estimated in the following manner. The location duration is first divided into episodes with an equal interval referred as *episode_length*, e.g. 5 min, each of which can be regarded as an experiment. Under the assumption that the user makes calls (including both initiating and receiving voice calls and messages) independently in each

episode, and that the probability of making calls across different episodes at the location is identical, $CallP(l_i)$ can then be modeled as Binomial distribution. The actual location duration delimits the total number of episodes, i.e. the number of independent experiments. While the call rate provides the probability of success for each experiment result, that is the probability of making a call in each episode. This leads to the final estimation of the probability $CallP(l_i)$ as the probability of having at least one success (making calls) over the total number of experiments, in this case, over the total location duration.

In this study, the previously derived two variables including the *average_call_rate* and the *average_actual_location_duration(activity)* are used as the approximation of the call rate for each individual and the duration for a location with a particular activity purpose, respectively. The probability $CallP(l_i)$ is then obtained as follows.

$$CallP(l_i) = CallP(activity) =$$

$$1-\{1-episode\_length \times average\_call\_rate\}^{average\_actual\_location\_duration(activity)/episode\_length}$$

5.3 Sequence conversion probability

After the probability of making calls at a location of home, work or non-mandatory activities is known, the likelihood that a call location trajectory is generated from an actual activity-travel sequence can be derived. In addition to the assumption that users make calls independently in each episode during a location visit, we also hypothesize that they make calls independently across each location visit. The sequence $l_1 -> l_2 -> \ldots -> l_n$ is defined as the *actual_travel_sequence*, and the call probability at each location l_i as $CallP(l_i)$. In constrast, $\overline{CallP(l_i)}$ is used to denote the probability that no calls are made at location l_i, $\overline{CallP(l_i)} = 1 - CallP(l_i)$. Based on these probabilities, the likelihood of various *call_stop_location_trajectories,* that could be observed from the *actual_travel_sequence*, defined as *ConversionP*, can be calculated as follows. The probability that the original full travel sequence can be revealed by the call records is

$$\text{ConversionP}(actual\_travel\_sequence, call\_stop\_location\_trajectory)$$

$$= \text{ConversionP}(l_1 -> l_2 -> \ldots -> l_n, l -> l_2 -> \ldots -> l_n) = \prod_{i=1}^{n} callP(l_i). \quad (1)$$

While the probability that only a part of the travel sequence is observed, is

$$\text{ConversionP}(l_1 -> l_2 -> \ldots -> l_n, l -> \ldots l_{i-1} -> l_{j+1} \ldots -> l_n)$$

$$= \prod_{m=1}^{m-1} callP(l_m) \times \prod_{m=i}^{j} \overline{callP(l_m)} \times \prod_{m=j+1}^{n} callP(l_m). \quad (1)$$

Where we assume that locations from x_i to x_j ($i \le j$) are missing since no phone communications have been made during the visits to these locations.

Suppose that the probabilities to make at least one call at the locations of home, work and non-mandatory activities are 0.805, 0.903 and 0.424, respectively. For the sequence of HWOH which represents the actual activity-travel behavior of a user identified as User121, there could be various location traces generated by this original travel sequence under certain

probabilities. For instance, the possibilities to emanate trajectories HWOH, HWH and H are as follows:

$$ConversionP(HWOH,HWOH) = CallP(H) \times CallP(W) \times CallP(O) \times CallP(H) = 0.248$$

$$ConversionP(HWOH,WH) = \overline{CallP(H)} \times CallP(W) \times \overline{CallP(O)} \times CallP(H) = 0.082$$

$$ConversionP(HWOH,H) = CallP(H) \times \overline{CallP(W)} \times \overline{CallP(O)} \times CallP(H) +$$
$$2 \times CallP(H) \times \overline{CallP(W)} \times \overline{CallP(O)} \times \overline{CallP(H)} = 0.054$$

5.4 Derivation of activity-travel sequences.

Based on the previously obtained conversion probabilities and the frequencies of the observed call location trajectories, the occurrences of original activity-travel sequences can be ultimately derived. Suppose that *m* different *call_stop_location_trajectories* $s_1, s_2, \cdots s_m$ are constructed from a user's call records, sorted by the length of these sequences, i.e. $length(s_1) \geq length(s_2) \geq \ldots \geq length(s_m)$. Let the frequencies of these observed trajectories as $y_1, y_2, \cdots y_k$ respectively; the original occurrences of the corresponding travel sequences, denoted as $x_1, x_2, \cdots x_k$, can be estimated by the following linear equation.

$$x_1 \times \mathrm{ConversionP}(s_1, s_1) = y_1$$

$$x_1 \times \mathrm{ConversionP}(s_1, s_2) + x_2 \times \mathrm{ConversionP}(s_2, s_2) = y_2$$

...

$$x_1 \times \mathrm{ConversionP}(s_1, s_k) + x_2 \times \mathrm{ConversionP}(s_2, s_k) \ldots + x_k \times \mathrm{ConversionP}(s_k, s_k) = y_k$$

From the above equation, the variables $x_1, x_2, \cdots x_k$ can be solved as follows.

$$x_1 = \frac{y_1}{ConversionP(s_1, s_1)}$$

$$x_2 = \frac{y_2 - x_1 \times ConversionP(s_1, s_2)}{ConversionP(s_2, s_2)}$$

...

$$x_k = \frac{y_k - x_1 \times ConversionP(s_1, s_k) - \ldots - x_{k-1} \times ConversionP(s_{k-1}, s_k)}{ConversionP(s_k, s_k)}$$

In the case of the User121, apart from a daily sequence of HWOH, four other *call_stop_location_trajectories* are revealed by this user's call records, including WH, OH, W and H, with the occurrences as 3, 2, 1 and 3 respectively. The original frequencies of these sequences, i.e. $x_1 - x_5$, can be solved in the following equation:

$$x_1 \times ConversionP(HWOH, HWOH) = 1$$

$$x_1 \times ConversionP(HWOH, WH) + x_2 \times ConversionP(WH, WH) = 3$$

$$x_1 \times ConversionP(HWOH, OH) + x_3 \times ConversionP(OH, OH) = 2$$

$$x_1 \times ConversionP(HWOH, W) + x_2 \times ConversionP(WH, W) + x_4 \times ConversionP(W, W) = 1$$

$$x_1 \times ConversionP(HWOH, H) + x_2 \times ConversionP(WH, H) +$$
$$x_3 \times ConversionP(OH, H) + x_5 \times ConversionP(H, H) = 3$$

From this equation, we obtain $x_1 = 4.03, x_2 = 3.67, x_3 = 5.78, x_4 = 0.30, x_5 = -0.23 \cdot$

The obtained results then undergo two further processes. First, a zero is assigned to the variables which have negative values, e.g. x_5 for the sequence H in the above case. The

negative frequency for a travel sequence suggests that the actual occurrence probability of the potential travel sequence could be very low, and that the corresponding observed identical call location trajectory, e.g. H in this example, is likely to be generated by other longer travel sequences, such as HWOH, WH and OH. These negative frequencies are thus dismissed by setting the corresponding variables to zero.

The second process is to normalize the obtained results for each individual, such that the total frequency of the derived travel sequences amounts to the observed sum of the call location trajectories. For the User121, the sum of the observed trajectories is 10, but that of the derived ones reaches 13.78, a ratio of these two numbers is used as the scaling factor, leading to the final solution as $x_1 = 2.92, x_2 = 2.67, x_3 = 4.19, x_4 = 0.22, x_5 = 0$.

From the call location trajectories for this user, a total of 10, 5 and 3 location visits for home, work and non-mandatory activity purposes respectively, have been observed; while for the derived travel sequences, the corresponding number changes to 12.7, 5.8 and 7.1, respectively. The ratio of the total locations between these two types of sequences is 0.79, 0.86 and 0.42 for these three activity classes respectively, close to the call location probabilities which are initially used for this derivation process. This further demonstrates that the derived travel sequences not only maintain the sequential order of the activity locations which are imbedded in the call location trajectories, but that they also preserve the call probabilities at individual locations as a whole.

It can be noted that during the entire procedure of seeking the solutions, we assume that the original travel sequences could only occur within the space of observed call location trajectories $S = \{s_1, s_2, \ldots s_m\}$. In theory, however, there could be a chance that an observed call location trajectory is produced by many other potential travel sequences, rendering the solution space to become infinite. However, for a possible travel sequence s_p which is not in the observed sequence space S, i.e. the frequency of the corresponding call location trajectory y_p being zero, a value less than or equal to zero would be obtained as the actual frequency x_p of this travel sequence. This implies that the positive frequencies of a travel sequence can only be found if this sequence is within the limited space S. For instance, for the User121, if the potential travel sequence is longer than any trajectory in S, i.e. $length(s_p) \geq length(s_1)$, assume $s_p = \text{HWOWH}$, we obtain the following equation: $x_p \times \text{ConversionP}(\text{HWOWH}, \text{HWOWH}) = 0$. From this equation, we have $x_p = 0$. Otherwise, if the length of this travel sequence is shorter than certain observed trajectories in S, e.g. $s_p = \text{HWO}$, we have $x_1 \times \text{ConversionP}(\text{HWOH}, \text{HWO}) + x_p \times \text{ConversionP}(\text{HWO}, \text{HWO}) = 0$, from which a value of $x_p < 0$ would be derived.

6 Classification

All the obtained *call_stop_location_trajectories* and *actual_travel_sequences* are subsequently classified according to the *home-based-tour-classification* and *daily-sequence-classification*, which have been previously established for workers' activity-travel behavior. During this classification, a home location H is added at the end of a sequence if it is absent

from this sequence, based on the assumption that each individual starts and ends a day at home. For each of these two types of sequences, two corresponding profiles are obtained and they are stored into matrices, namely *home-based-tour-profile* and *daily-sequence-profile*.

The Pearson correlation coefficient is used to measure the relation of the corresponding profiles between these two types of sequences. The correlation coefficient, denoted by *r*, is a measure of the strength of linear relationship between two variables. It takes on values ranging between 1 and -1, with 1 indicating a perfect positive linear relationship: as one variable increases in its values, the other variable increases as well. The closer the value is to 1, the stronger the relationship is.

For two matrices, denoted as *A* and *B*, and let *d* as the total number of the matrix elements, the *r* is computed as follows:

$$\overline{A}=\frac{\sum_{i=1}^{d}A_i}{d},\ \overline{B}=\frac{\sum_{i=1}^{d}B_i}{d},\ S_A=\sqrt{\frac{\sum_{i=1}^{d}(A_i-\overline{A})^2}{d}},\ S_B=\sqrt{\frac{\sum_{i=1}^{d}(B_i-\overline{B})^2}{d}},$$

$$r=\frac{\sum_{i=1}^{d}\left(\frac{A_i-\overline{A}}{S_A}\right)\left(\frac{B_i-\overline{B}}{S_B}\right)}{d-1}.$$

7. Case study

In this section, adopting the proposed profiling approach and using the mobile phone data described in Section 3, we carry out an experiment. In this process, a set of *call_stop_location_trajectories* are first constructed, followed by the translation of the trajectories into *actual_travel_sequences*. Each step of this process is highlighted with the examination of some particular parameters.

7.1 Construction of *call_stop_location_trajectories*

7.1.1 *Work_start_time* and *work_end_time*

Figure 3 describes the distribution of the frequency of calls made in each hour of the day during weekdays, showing that from 9am in the morning, calls reach to their peak level; while from 18pm in the late afternoon, a second climax of call activities start to occur. These two temporal points are chosen as the w*ork_start_time* and w*ork_end_time*, respectively.

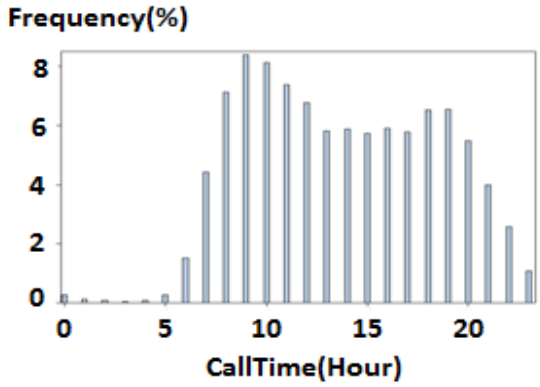


Figure 3. The distribution of the time of calls

Based on the pre-defined criteria for home and work location identification, 49436 (98.9% of the total) users have their home locations discovered. The remaining 1.1% are those who made no calls at weekend or in the night period from 18pm to 9am across the two surveyed weeks, and as a result their homes cannot be spotted by these rules. Meanwhile, 9,458 users (18.9% of the total) are screened out as employed people, if they work between 9am and 17:59pm at least two weekdays per week. By contrast, those who work at night shifts or at weekends, or who work less than two days a week, or who make few calls at work, are not identified as workers.

For those who have both predicted home and work locations, we further remove nearly 15% of the individuals who have unknown cell IDs for the identified home or work locations due to technical reasons that occur in the mobile phone data collection process. This results in a final dataset of 8,027 workers who represent 16% of the total users in the selected dataset. All the call records of these individuals during weekdays are extracted, and the consecutive calls made at a same location are aggregated. This leads to 69,578 *raw_call_location_trajectories* constructed for further analysis.

7.1.2 *Call_location_duration* and *maximum_time_boundary*

Two parameters characterize the stop location identification process. The first one, *call_location_duration,* determines the time limit above which the location is defined as a stop. This parameter depends on the minimum time required to possibly pursue an activity as well as the time period needed for traversing across the area. The other parameter, *maximum_time_boundary,* measures the time interval between the last call time at the previous location and the first call time at the next location, relative to the current place under investigation. Similar to *call_location_duration,* this parameter must be longer than a combination of the possible activity duration and the travel time needed going from the previous cell, passing the current one, and to the next area. In addition, it should also be able to detect location update errors which usually occur in a short time interval.

In this experiment, $T_{call\_location\_duration}$ and $T_{\max imum\_time\_boundary}$ are set as 30 minutes and 60 minutes respectively. Under these thresholds, 40.3% of all locations from the *raw_call_location_trajectories* are removed; the remaining locations in these sequences form the set of *call_stop_location_trajectories*. The average length of these trajectories is 3.3. In comparison, using the existing method which defines as a stop location if the interval between the first login of the location and that of its next location is longer than 60 minutes, 67.6% of all the raw call locations are dismissed, with the average length of the retained sequences as 2.33.

7.2 Conversion from *call_stop_location_trajectories* into *actual_travel_sequences*

7.2.1 *Average_call_intervene* and a*verage_call_rate*

When estimating these two variable values, all the calls made by the identified workers, including the ones that may be made on a road or have false location IDs due to localization errors, are all considered. This results in the *average_call_intervene* as 192 min over a full day of 24 h. However, as demonstrated in Figure 3, the occurrence of calls is not equally distributed, more calls are observed during the day than at night, the inclusion of the night

period would bias the real call intervene time during the daytime period. In this study, only the period of 6am-12pm is thus taken into account. This reduces the call intervene to 137 min; accordingly, the *average_call_rate* is 0.0073.
In an existing study (Calabrese et al., 2011), however, a 260 min of call intervene is derived; this difference could be caused by the following factors. (i) Only workers are considered in our study. (ii) The mobile phone data in this experiment is more recent than the data used in the existing study. (iii) People could make more calls in Ivory Coast than in Massachusetts in the United states where the existing study is performed.

7.2.2 *Actual_location_duration*
This variable value is approximated by a real activity-travel behavior survey that was conducted in Belgium which will be described later. From this survey, the average location duration $average\_actual\_location\_duration(activity)$ are 222, 317 and 75 min for home, work and non-mandatory activity locations, respectively.

7.2.3 E*pisode_length*
This variable specifies the time window by which the location duration is split into a number of episodes, i.e. experiments. The length of this window is decided such that the call behavior of users in an episode should be independent of that in its next episode. To obtain such an episode length, the average voice call duration of users is considered, which is derived from an additional dataset that records the duration for all voice calls between each two cells in Ivory Coast. The resultant average call duration is 1.92 min, a 2-min interval is thus taken as the estimation of this episode length.
Based on all the above parameter settings, the call probability at a location is derived, and it is 0.805, 0.903 and 0.424 for home, work and non-mandatory activity locations, respectively. These obtained probabilities, combined with the observed frequencies of the *call_stop_location_trajectories* for each user, lead to the prediction of the number of the *actual_travel_sequences* for the individual, using the method described in Section 5.3 and 5.4.

8 Comparison of results from mobile phone data with real activity-travel diary data
To illustrate the practical ability of our approach to really serve as a benchmark method, we compare the results derived from the mobile phone data with the statistics drawn from real activity-travel surveys. Unfortunately, no official activity-travel surveys have been documented in Ivory Coast. Therefore, data stemming from other countries, including South Africa and Belgium, have been adopted for this purpose. The authors acknowledge that the real travel behavior in Ivory Coast most likely is considerably different to the one reported in South Africa and Belgium. Consequently, the illustration serves to underline the applicability of the approach, not to infer the travel behavioral relationships in this particular case. The comparison is carried out in two aspects, including the aspect of individual locations, e.g. the average number of locations visited each day, and the sequential aspect of the activity locations, e.g. the *home-based-tour-profile* and the *daily-sequence-profile*.

8.1 Travel survey in South African

The South Africa National Household Travel Survey (NHTS) was the first national survey of travel habits of individual and households, aimed at making significant improvements in public transport services. The survey was based on a representative sample of 50,000 households throughout South Africa and undertaken between May and June in 2003 (http://www.arrivealive.co.za/pages.aspx?nc=household).

The information recorded by the survey includes the travel time to various public transport services, e.g. trains and buses, as well as to activity services, e.g. shops and post offices. The number of trips and the purposes for these trips are also documented for each individual on a typical weekday. The survey results reveal that the majority of the respondents can access to most of the activity services within half an hour (i.e. the travel time), and the average activity location visited by a worker on a weekday is estimated at between 3.46 and 4.06 (http://www.arrivealive.co.za/document/household.pdf).

8.2 Travel survey in Belgium

Despite the relative geographic proximity between South Africa and Ivory Coast, the information on the NHTS is nevertheless limited. Moreover, the detailed travel patterns for each individual are not accessible for us. This necessitates the use of a second survey that provides activity-travel sequences on entire days and will be used as a reference for the illustration of the derived profiles.

The survey, namely SBO, stems from a large scale **S**trategic **B**asic **R**esearch project on transportation modeling and simulation, and it was conducted on 2500 households between 2006 and 2007 in Belgium. In the survey, the respondents recorded trip information during the course of one week, such as trip start time and end time, purpose of the trip (e.g. activity type), and trip origin and destination (e.g. activity location). The average travel time is 24 min, comparable to the 30 min for a typical travel in South Africa.

In the SBO survey, activity locations are represented with statistical sectors, each of which ranges from a few hundred meters to a few thousands in radius, similar to the spatial granularity level of cell locations in GSM network. Table 2 illustrates a typical diary of respondent identified as 'HH4123GL10089' on May, 9th, 2006. Only the variables that are relevant for the current study are presented in this table; a more detailed variable list and elaboration on this survey can be found in (Cools et al., 2009).

TABLE 2. Travel Diary Data

Respondent ID	Date	Trip Start Time	Trip End Time	Trip Origin	Trip Destination	Trip Purpose
HH4123GL10089	09/05/2006	07:45:00	08:00:00	34337	34345	Work
HH4123GL10089	09/05/2006	17:00:00	17:15:00	34345	34349	Shopping(non-mandatory)
HH4123GL10089	09/05/2006	17:40:00	17:30:00	34349	34337	Home

From the dataset, the diaries on weekdays from 372 individuals who work at least two days a week are extracted. Activity duration at the destination of a trip is estimated as the time interval between the end time of the trip and the start time of its next trip, if the travel is not the last movement of a day. Otherwise, for the last trip, the activity end time at the travel destination is unknown. Another unknown factor is the activity start time at the origin of the first travel of a day. These two times are thus approximated by the typical time for getting up in the morning and going to sleep in the evening in Belgium, which are estimated as 6am and

12pm, respectively (Hannes et al., 2012). The average of all the obtained duration at locations with an identical activity motivation over all the individuals is stored in the variable *average_actual_location_duration* which has been previously used in the experiment to derive the *actual_travel_sequences*.

8.3 Statistics on the average length of sequences

Table 3 summarizes the statistics on the average number of locations visited each day, i.e. the average length of sequences, derived from the sequences of *raw_call_location_trajectories*, *call_stop_location_trajectories* and *actual_travel_sequences* which have been previously built based on the mobile phone data. The results drawn from both the NHTS and SBO surveys are also presented alongside as a comparison.

Table 3. Statistics on the average length of sequences[a]

Sequences	RCLT	CSLT	ATS	NHTS	SBO
Average length of sequences	5.69	3.30	4.02	3.46-4.06	3.96

[a]The columns from left to right represent the raw_call_location_trajectories (RCLT), call_stop_location_trajectories (CSLT), actual_travel_sequences (ATS), NHTS and SBO surveys, respectively.

It was noted from Table 3 that the average length of sequences first drops from initial 5.69 for the *raw_call_location_trajectories* to 3.3 for the *call_stop_location_trajectories*, and then rises again to 4.02 for the estimated travel sequences which is the closest to the number observed in both NHTS and SBO surveys. In addition, the differences in this variable value imply the importance of the process from the identification of stop locations to the inference of complete travel sequences proposed by our approach, when analyzing activity-travel behavior based on the mobile phone data.

8.4 Home_based_tour_profile

Table 4 shows the relative frequency of each pattern class in the home_based_tour_classification, obtained from the *call_stop_location_trajectories*, the *actual_travel_sequences* and the SBO diaries, respectively. The differences in the percentages of corresponding pattern classes between these each two types of sequences are also listed.

Table 4. Home_based_tour_profile (%)[a]

Pattern	CSLT	ATS	ATS - CSLT	SBO	ATS - SBO	CSLT - SBO
H	9.0	4.4	-4.6	6.4	-2.0	2.6
HWH	50.3	39.1	-11.2	42.9	-3.8	7.4
HOH	18.0	26.3	8.3	32.5	-6.2	-14.5
HOWH	5.1	6.7	1.6	3.1	3.6	2.0
HWOH	8.2	10.3	2.1	10.8	-0.5	-2.6
HWOWH	3.4	3.8	0.4	1.6	2.2	1.8
HOWOH	2.5	4.1	1.6	1.9	2.2	0.6
HOWOWH	0.7	1.0	0.3	0.2	0.8	0.5
HWOWOH	1.4	2.1	0.7	0.5	1.6	0.9
HOWOWOH	0.5	0.8	0.3	0.1	0.7	0.4
More than 2 work activities	1.0	1.3	0.3	0.2	1.1	0.8

[a] The columns from left to right represent the typical patterns, the call_stop_location_trajectories (CSLT), the actual_travel_sequences (ATS), the differences between ATS and CSLT, the SBO diaries (SBO), the differences between ATS and SBO, and the differences between CSLT and SBO, respectively.

Table 4 indicates that, when the *call_stop_location_trajectories* are converted into the *actual_travel_sequences*, the percentage of shorter patterns, e.g. H and HWH, increases; while that of longer patterns, e.g. HWOWOH, decreases. During this sequence conversion process, an observed call location trajectory is expected to be generated not only from an travel sequence that is identical to this observed trajectory, but more likely from a sequence that is longer than this observed one due to the fact that not every visited location is exposed by the mobile phone data. For instance, although 9.0% of the total call locations trajectories belong to the pattern of H, only 4.4% is estimated to be the days when the individuals do not make any trips but staying at home. The remaining 4.6% is probably generated from other longer travel sequences where the missing locations are as a result of the nature of the mobile phone data.

Another feature in the conversion process is that, the lower the probability that people make calls at a location, the higher the frequency of the derived travel sequence that contains this location, tends to be, in order to give rise to the call location trajectories that amount to the observed frequency of the trajectories. This can be further illustrated by the pattern HOH. Although this pattern is as short as HWH, the probability at a non-mandatory activity location O, e.g. 0.424 in this experiment, is the lowest among all the three activity types. This leads to a prediction of high frequency of this pattern for the derived travel sequences.

When the patterns obtained from the derived travel sequences are compared with the ones drawn from the SBO diaries, it was observed that the major contrast resides in the difference between the group of short sequences and the other group accommodating long patterns. The SBO data has higher frequencies in short sequences while lower occurrences for long patterns, than the derived travel sequences. This tendency remains when the SBO data is compared with the *call_stop_location_trajectories*. While apart from the likelihood that people in Belgium may conduct less activities on average than in Ivory Coast, this also demonstrates the possibility that the diaries under-represent people's activity-travel behavior, especially for short period of activities. The shortcoming has been well documented in literatures (e.g. Cools et al., 2009).

8.5 Daily_sequence_profile

Figure 4(a) depicts the correlation between the relative frequency of each pattern class in the daily_sequence_profile obtained from the *call_stop_location_trajectories* and the *actual_travel_sequences*. It was noted that, the majority pattern classes follow a similar distribution in relative frequencies between these types of sequences. The few outliers can be divided into two groups: the group of HWH, H and HWHWH which are 14.1%, 5.9% and 1.4% higher for the *call_stop_location_trajectories*, and the other group consisting of HOH, the patterns with more than 2 home-based tours, and HOWOH, which show a 3.7%, 2.5% and 2.1% higher frequency for the *actual_travel_sequences*, respectively. This further demonstrates that, compared to the *call_stop_location_trajectories*, the derived travel sequences tend to have a high proportion for long patterns and for patterns which accommodate locations with low call probabilities, e.g. non-mandatory activity locations O. In contrast, a lower percentage is anticipated for short patterns and for patterns containing locations with high call probabilities, e.g. work places W, after the sequence conversion process.

In Figure 4(b) which describes the correlation between the daily_sequence_profiles obtained from the *actual_travel_sequences* and the SBO diaries, we found that most patterns have a moderately higher frequency for the *actual_travel_sequences* than the SBO data. However, a few outliers show remarkably higher occurrences for the SBO diaries, e.g. HWOH and HWHOH accounting for a 7.3% and 7.1% higher percentage, respectively. It suggests that compared to Ivory Coast, people in Belgium may carry out more non-mandatory activities on the way from work back to home as well as in the evening period after arriving at home. In addition, a further examination reveals that out of all 93 pattern classes in the *daily_sequence_profile*, 59 (63.4%) are zero frequencies for the SBO data; while for the *call_stop_location_trajectories* and *actual_travel_sequences*, only 16 patterns (17.2%) are not represented. It reflects that the sequences derived from the mobile phone data are more representative in travel behavior than the survey data, further underlying the significance of using mobile phone data to explore the characters of travel behavior.

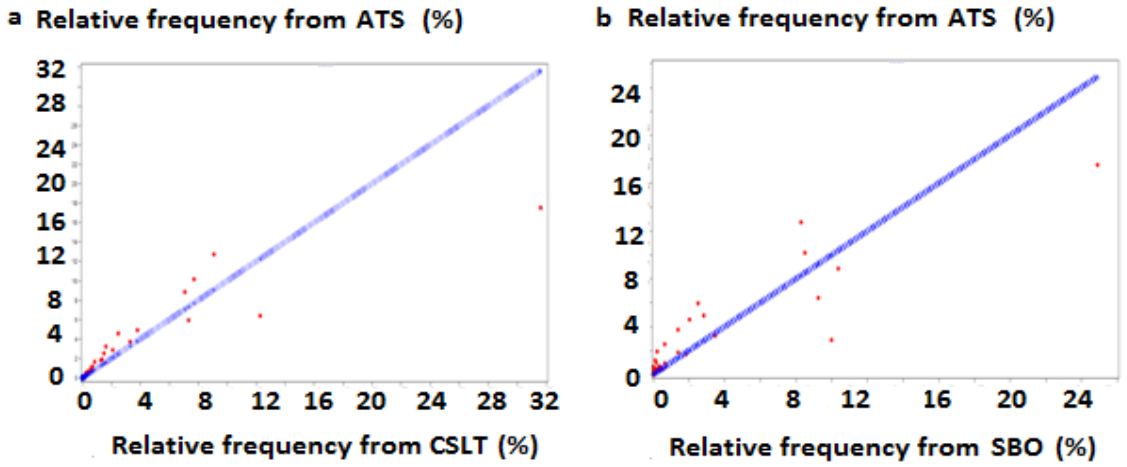


Figure 4. Correlation between the relative frequency of each corresponding pattern class
Note: x- and y-axis represent the relative frequency of each corresponding pattern class obtained from the call_stop_location_trajectories (CSLT) and the actual_travel_sequences (ATS) (a), and the SBO diaries and the actual_travel_sequences (b). The line of y=x is also presented as a reference line.

The correlation r between the *call_stop_location_trajectories* and the *actual_travel_sequences* as well as between the *actual_travel_sequences* and the SBO data is 0.91 and 0.89, respectively. The high correlation shows that the profile derived from the estimated travel sequences has an overall close relationship to that obtained from the call stop location trajectories, and in the meantime the profile of the travel sequences also accounts for the deviation in frequencies for each particular pattern which are caused by the discrepancy between the call behavior and the actual activity-travel behavior. In addition, the derived profile also resembles the frequency distribution of travel sequences from a real travel behavior survey. These results suggest the derived profile of travel sequences can properly represent workers' travel behavior in a studied area, and therefore capable of being used to validate the simulated sequences generated from travel behavior models.

Nevertheless, in this case study, we used the surveys conducted in South Africa and Belgium as an illustration for the results derived by our approach. However, variation exists across different regions and countries. As described in the introduction, travel behavior is shaped by

the conditions of land use and transportation network as well as the social-economic background of individuals. Besides, several years of time differences when these datasets were collected, as well as the fact that the surveys, especially the SBO survey, were based on a small set of samples, all contribute to the deviation exposed in this experiment results between the derived travel sequences and the survey data. With a real travel survey conducted in the same or similar context to where the mobile phone data is obtained, it is believed that the activity-travel behavior profiling approach based on the mobile phone data would bring even better results to current experimental outcome.

9. Sensitivity analysis

Throughout the profiling process, several parameters including *call_location_duration*, *maximum_time_boundary* and *actual_location_duration*, have been defined. This prompts to have a final investigation into how the thresholds of these parameters affect the predicted results, including the average length of *call_stop_location_trajectories* and the *actual_travel_sequences*, referred as CSLT_*length* and *ATS_length* respectively, as well as the coefficients between the *call_stop_location_trajectories* and the *actual_travel_sequences* as well as between the *actual_travel_sequences* and the SBO diaries, simplified as *r1* and *r2*, respectively.

9.1 Call_location_duration and maximum_time_boundary

In the process of stop location identification, when the threshold $T_{call\_location\_duration}$ for the parameter *call_location_duration* increases, the minimum time duration required to consider a location as a stop becomes longer, leading to a decrease in the number of daily location visits. This is well reflected in Figure 5(a). However, the rate of reduction is very slow; particularly, when this parameter reaches a certain threshold, e.g. 30 minutes set up in this experiment, the lengths of both types of sequences enter into a nearly constant level. A similar stabilization is observed in Figure 5(b) when $T_{call\_location\_duration}$ passes the 30 minutes threshold.

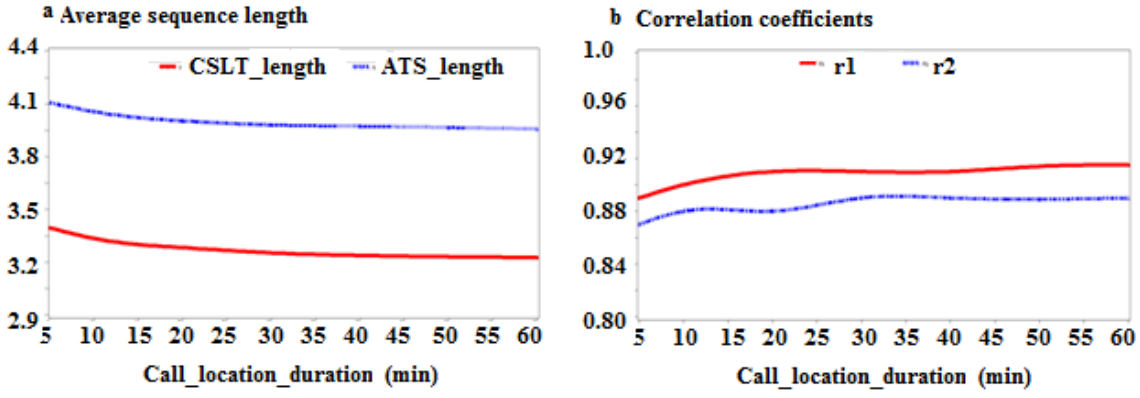


Figure 5. Correlation between the threshold of call_location_duration and the results
Note: x-axis stands for the threshold of call_location_duration, and y-axis for the sequence length of CSLT_*length* and *ATS_length* respectively (a) and the coefficients *r1* and *r2* respectively (b).

Figure 6(a) and 6(b) show how the results evolve with the threshold $T_{\max imum\_time\_boundary}$ for the parameter of *maximum_time_boundary*. As expected, when the maximum available time needed for a possible stop location sets longer, the number of identified stop locations drops, as shown in Figure 6(a). However, this does not bring about the same amount of change to the coefficients; especially when $T_{\max imum\_time\_boundary}$ increases to a certain value, e.g. 60 minutes adopted in our experiment, both *r1* and *r2* develop into a stable level. This suggests that, although the number of disclosed stop locations diminishes as this duration limit becomes stricter, the disregarded potential stop locations are likely distributed randomly across various types of pattern classes. As a result, the coefficients which reflect the relative frequency of these patterns remain almost the same, regardless of the minor changes that could arise from these parameter settings.

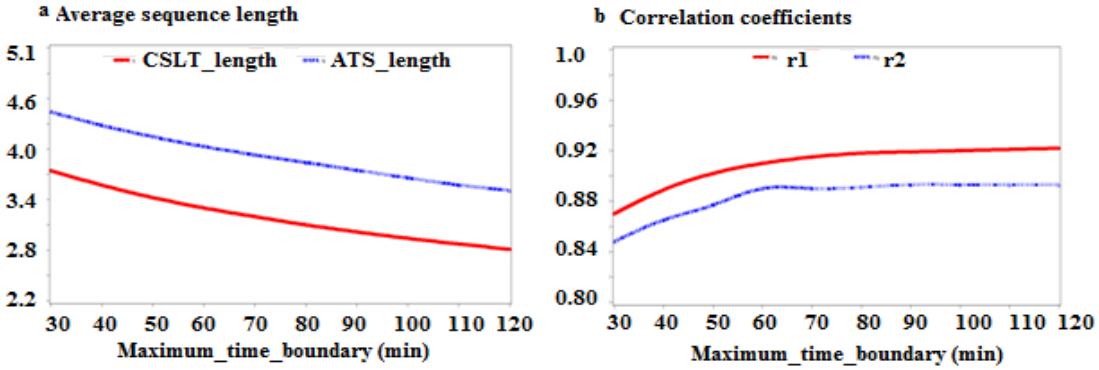


Figure 6. Correlation between the threshold of maximum_time_boundary and the results
Note: x-axis stands for the threshold of maximum_time_boundary, and y-axis for the sequence length of CSLT_length and ATS_length respectively (a) and the coefficients r1 and r2 respectively (b).

9.2 Actual_location_duration

Figures 7 describes the relation between the parameter *actual_location_duration* for work activities and the estimated results. It indicates that, as this duration becomes longer, the ATS_length2 for the derived travel sequences decreases while r1 and r2 increases, but these changes disappear when this duration pass a certain point, e.g. 240 minutes.

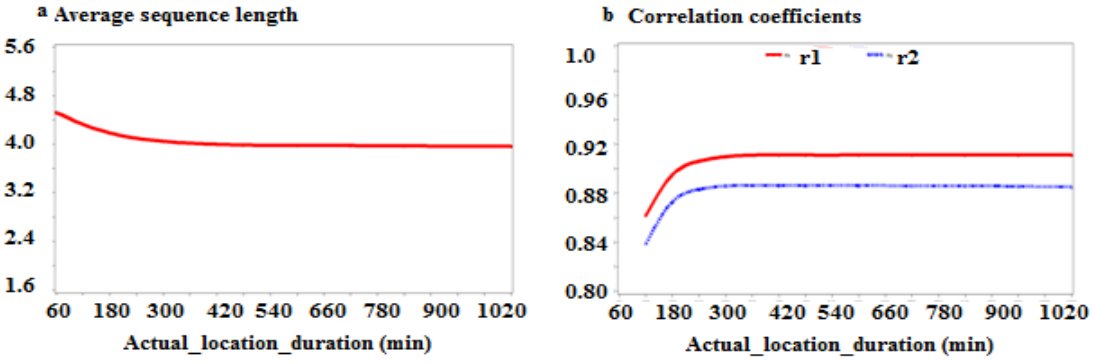


Figure 7. Correlation between the actual_location_duration for work activities and the derived results
Note: x-axis stands for the actual_location_duration for work activities, and y-axis for the sequence length of ATS_length (a) and the coefficients r1 and r2 (b).

This phenomenon can be explained by the binomial model employed to estimate the call probability at a location. According to this model, when the *actual_location_duration* is longer, the probability at a location $CallP(l_i)$ becomes higher, as demonstrated by Figure 8(a). However, the amount of increases in the probabilities as the activity duration extends e.g. one hour longer, is not evenly distributed, which can be manifested in Figure 8(b). It shows that: as the activity duration becomes longer, the amount of growth in the location probabilities diminishes until to a nearly zero level. This explains the occurrence of the flat curves observed in Figure 7.

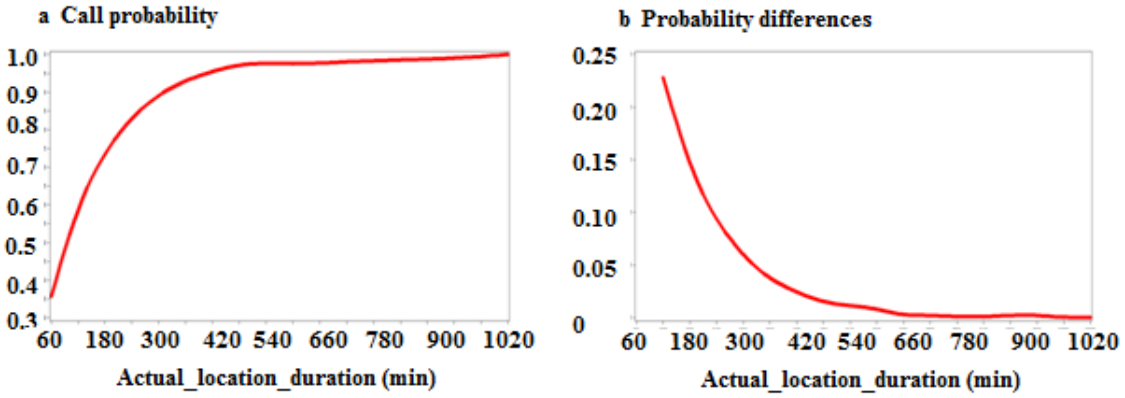


Figure 8. Correlation between the actual_location_duration and the call probability at a location
Note: x-axis stands for the actual_location_duration for work activities, and y-axis for the call probability at a location (a) and for the difference between the probability obtained from the corresponding actual_location_duration and the other probability derived from a duration which is 60 min longer than this current duration (b).

All these above analysis shows that, except that the increase in $T_{\max imum\_time\_boundary}$ reduces the number of identified stop locations, a certain amount of changes in these parameters do not incur a significant deviation in the results of both the average length of sequences as well as the profiles. This suggests that the profiles built upon the mobile phone data are stable and consistent in revealing people's activity-travel behavior; a minor change in these parameters that are required in the profiling process will not lead to a substantially different outcome.

10. Conclusions and discussion

The approach of profiling workers' travel behavior based on mobile phone data is both unique and important in that it builds a new measure which can be used to directly evaluate the simulated activity-travel sequences yielded from micro-simulation models of travel behaviour. The advantage of using this method is that it does not depend on conventional diaries, the data requirement is fairly simple and its collection cost is low. More importantly, the massive mobile phone data monitors current activity-travel behavior in a large proportion of population expanding over a long time period, the profile derived from the data is thus capable of providing a more representative and objective validation measure.

Experiments on this approach by using data collected from people's natural mobile phone usage have demonstrated an overall high correlation coefficient between the profiles derived from the observed call location trajectories as well as from the derived travel sequences. The

relative frequency of each corresponding pattern class between these two profiles, however, shows a certain level of differences, a reflection of the deviation between the movement revealed by the call behavior and the real path that the individuals have experienced. In addition, the derived travel sequences also show reasonable outcome when they are compared to the statistics drawn from real travel surveys conducted in South Africa and Belgium, respectively. Furthermore, the examination into this method's sensitivity demonstrates its consistence and stability in drawing the real picture of activity-travel behavior over various parameter settings.

Beyond the initial goal of building a new measure for travel behavior simulation models, the proposed method for stop location identification and subsequent actual travel sequences derivation provides a broad use for the application of the massive mobile phone data. For instance, in the process of building OD matrices (Calabrese et al., 2011), only the stop locations revealed by the phone data are used; places where no calls are made are thus ignored. The results in our experiments suggest of an average of 21.8% increase from the initially obtained call stop locations to the derived complete location visits. Consequently, the OD matrices based on the mobile phone data reflect only a part of the whole picture of people's transfer phenomena, as acknowledged by the authors of the study. Based on our method, the real travel sequences could be derived first and a more accurate OD matrix could be anticipated.

The proposed approach can also be adopted for the characterization of non-workers' travel behavior. No work activities dominate the individuals' activity-travel sequences, but more home-based tours for non-mandatory activity purposes could be considered.

Nevertheless, despite the promising experiment results, there are still certain areas which need to be improved in the future research. First, when calculating the call probability at a location, we simply use a universal call rate which is derived from the mobile phone data across all types of activity locations and all individuals. But people communicate with others not at a same pace, and they may also call at different frequencies depending on what they are doing. Like the call rate, the use of an average actual location duration for each activity purpose across all users leaves a second possibility for improvement, as the activity duration across different individuals is likely to differ. The proposed method will be undoubtedly strengthened if both the call rate and the activity duration is considered at individual level and across each category of activity locations. Third, the method to identify home and work places could also be enhanced through machine learning techniques, as explored by a recent study (Liu et al., 2013).

While being faced with the challenge of acquiring both the mobile phone data and the real travel survey from a same or similar study region, in this study we use the travel surveys which are conducted in different environments than the phone data, as the reference to compare and illustrate the results. Nevertheless, in the future research, the proposed method must be applied to a real travel survey which is sampled in a similar context to where the phone data is obtained. Such surveys thus provide another possibility of enhancement by bringing more relevance to this method in terms of tuning up the parameters as well as validating the results.

11. References

Arentze, T. A., & Timmermans, H. J. P. (2004). A learning-based transportation oriented simulation system. Transportation Research Part B: Methodological, 38(7), 613-633.

Asakura, Y., & Hato, E. (2006). Tracking individual travel behavior using mobile phones: recent technological development. Paper presented at 11th International Conference on Travel Behaviour Research, Kyoto.

Axhausen, K., & Gärling, T. (1992). Activity-based approaches to travel analysis: conceptual frameworks, models and research problems. Transport Reviews, 12, 324-341.

Becker, R., Cáceres, R., Hanson, K., Loh, J.M., Urbanek, S., Varshavsky, A., & Volinsky, C. (2011). A tale of one city: Using cellular network data for urban planning. IEEE Pervasive Computing, 10(4), 18–26.

Bellemans T, Janssens D, Wets G, Arentze T, & Timmermans H. J. P. (2010). Implementation Framework and Development Trajectory of Feathers Activity-Based Simulation Platform. Transportation Research Board, 2175, 111-119.

Bhat C. R. & Koppelman F. S. (1999). A Retrospective and Prospective Survey of Time-Use Research. Transportation, 26(2), 119-139.

Bradley M. & Vovsha P. (2005). A model for joint choice of daily activity pattern types of household members. Transportation, 32, 545-571.

Calabrese, F., Di Lorenzo, G., Liu, L., Ratti, C. (2011). Estimating Origin-Destination Flows Using Mobile Phone Location Data. IEEE Pervasive Computing, 10(4), 36-44.

Cools, M., Moons, E., Bellemans, T., Janssens, D. & Wets G. (2009). Surveying activity-travel behavior in Flanders: Assessing the impact of the survey design. In C. Macharis, and L. Turcksin (eds.), 369 Proceedings of the BIVEC-GIBET Transport Research Day, Part II, VUBPress, Brussels, 370, 727-741.

Cools, M., Moons, E., & Wets, G. (2010a). Calibrating Activity-Based Models with External Origin-Destination Information: Overview of Possibilities. Transportation Research Record: Journal of the Transportation Research Board, 2175, 98–110.

Cools, M., Moons, E., & Wets, G. (2010b). Assessing the Quality of Origin-Destination Matrices Derived from Activity Travel Surveys: Results from a Monte Carlo Experiment. Transportation Research Record: Journal of the Transportation Research Board, 2183, 49–59.

Davidson, W., Donnelly, R., Vovsha, P., Freedman, J., Ruegg, S., Hicks, J., Castiglione, J., & Picado, R. (2007). Synthesis of first practices and operational research approaches in activity-based travel demand modeling. Transportation Research Part A: Policy and Practice, 41(5), 464–488.

Hannes, E., Liu, F., Vanhulsel, M., Janssens, D., Bellemans, T., Vanhoof, K., & Wets, G. (2012). Tracking Household routines using scheduling hypothesis embedded in skeletons (THRUSHES). Transportmetrica, Special Issue "Universal Design", 8(3), 225-241.

Hansapalangkul, T., Keeratiwintakorn, P., & Pattara-Atikom, W. (2007). Detection and estimation of road congestion using cellular phones. In Proceedings from 7th International conference on intelligent transport systems telecommunications, 143–146.

Lemp, J., McWethy, L., & Kockelman, K. (2007). From Aggregate Methods to Microsimulation: Assessing Benefits of Microscopic Activity-Based Models of Travel Demand. Transportation Research Record: Journal of the Transportation Research Board, 1994, 80–88.

Liu, F., Janssens, D., Wets, G. & Cools, M. (2013). Annotating mobile phone location data with activity purposes using machine learning algorithms. Expert Systems with Applications. http://dx.doi.org/10.1016/j.eswa.2012.12.100.

Ratti, C., Pulselli, R. M., Williams, S., & Frenchman, D. (2006). Mobile landscapes: Using location data from cellphones for urban analysis. Environment and Planning B—Planning and Design, 33(5), 727–748.

Rose, G. (2006). Mobile phones as traffic probes: Practices, prospects and issues. Transport Reviews, 26(3), 275–291.

Schlaich, J., Otterstätter, T., & Friedrich, M. (2010). Generating Trajectories from Mobile Phone Data, TRB 89th Annual Meeting Compendium of Papers, Transportation Research Board of the National Academies, Washington, D.C., USA.

Sohn, K., & Kim, D. (2008). Dynamic origin-destination flow estimation using cellular communication system. IEEE Transactions on Vehicular Technology, 57(5), 2703–2713.

Spissu, E., Pinjari, A. R., Bhat, C. R., Pendyala, R. M., & Axhausen, K .W. (2009). An analysis of weekly out-of-home discretionary activity participation and time-use behavior. Journal Transportation, 36(5), 483-510.

Steenbruggen, J., Borzacchiello, M. T., Nijkamp P., & Scholten, H. (2011). Mobile phone data from GSM networks for traffic parameter and urban spatial pattern assessment: a review of applications and opportunities, GeoJournal.

Blondel, V. D., Esch, M., Chan, C., Clerot, F., Deville, P., Huens, E., Morlot, F., Smoreda, Z., & Ziemlicki, C. (2012). Data for Development: the D4D Challenge on Mobile Phone Data. Computer Science.

White, J., & Wells, I. (2002). Extracting origin destination information from mobile phone data. In Proceedings from 11th international conference on road transport information and control, 486, 30–34.

Wolf, J. L., Guensler, R., & Bachman, W. H. (2001). Elimination of the travel diary: an experiment to derive trip purpose from GPS travel data. In Journal of the Transportation Research Record, 1768, 125-134.

Yagi, S., & Mohammadian, A. (2010). An Activity-Based Microsimulation Model of Travel Demand in the Jakarta Metropolitan Area. Journal of Choice Modeling, 3(1).

Yagi, S. & Mohammadian, A. (2007). Validation of an Activity-Based Microsimulation Model of Travel Demand. Proc. of 11th World Conference on Transport Research Society, (DVD), Berkeley, CA.

Extracting Large Scale Social Relational Dynamics from Mobile Communications Data

Jonny Huck
School of Computing and Communications
InfoLab21, Lancaster University
Lancaster, LA1 4WA
j.huck2@lancaster.ac.uk

Paul Coulton
Lancaster Institute for the Contemporary Arts
LICA Building, Lancaster University
Lancaster, LA1 4YW
p.coulton@lancaster.ac.uk

Duncan Whyatt
Lancaster Environment Centre
Lancaster University
Lancaster, LA1 4YQ
d.whyatt@lancaster.ac.uk

ABSTRACT
Given the vast amounts of data generated daily by our burgeoning communication systems there are new opportunities emerging to examine one of the traditional concerns of geographers; that is the movement of people, goods and services through space and time. Such data are not restricted to simply revealing the movements of individuals but also the space-time trajectories of the information flows between individuals and places. Such flows have the potential to reveal new insights into the relational dynamics between people within cities, between cities, between regions and even between countries. To enable such insights we require new techniques to generalise and visualise these flows. In this paper we use of 3D visualisation techniques that originate from the field of terrain modelling, to summarise and communicate complex flow information to a wide audience. Furthermore, recognising the ability of physical models to enhance the communication and educational experience we propose to use laser cutters and 3D printers to turn these visualisations into such models.

Keywords
Mobile data, space-time, generalisation, visualisation

1. INTRODUCTION
The discipline of Geography has a long tradition of analysing movement of individuals through space and time, but this research has often been restricted to relatively small numbers of individuals. Despite the pioneering efforts of Torsten Hägerstrand, and in particular his work relating to space-time trajectories and prisms [1], little progress has been made towards exploring large volumes of spatio-temporal data using novel, alternate forms of visualisation, meaning that the information content of these data are seldom fully exploited. Given the considerable volumes of space-time data generated by mobile communication systems on a day to day basis it would seem appropriate to address this issue. Furthermore, these data do not simply provide the opportunity to examine individual space-time behaviour, but has the potential to provide significant insights into social network dynamics [2,3] and in particular the relational dynamics of large numbers of individuals at city, district and countrywide levels.

Whilst modern GIS tools provide a range of visualisation techniques in this research we also draw from Henrí Lefebvre's concept of Rhythmanalysis [4] in which he states that "Everywhere there is an interaction between a place, a time and an expenditure of energy there is a rhythm" to develop novel 3D visualisations techniques in order to represent and ultimately interpret flows of information (phone calls) between large numbers of individuals that relate to these social relational dynamics.

It is known that a 3D physical model communicates much more information than a flat screen image or paper printout as our everyday human experience means that we are innately able to estimate distances and evaluate terrain. Additionally the haptic experience (relating to an object through the sense of touch) of physical models is known to provide a powerful communication and education experience [5].

The aim of this research therefore is to produce digital and physical representations that will encourage other researchers to consider new techniques for the representation of large volumes of space-time data, and engage the wider public in the value that such data can provide.

2. DATA ANALYSIS
The data used within this research is supplied as part of the Orange Data For Development Challenge and relates The Orange phones calls in Ivory Cost (Republic of Côte d'Ivoire) during 5 month period from December 2011 to April 2012. The original dataset contains 2.5 billion records, calls and text messages exchanged between 5 million anonymous users.

Initial exploratory analysis was conducted upon the data in order to identify spatio-temporal patterns that could be represented within in a physical 3D model. In Figure 1 and 2 we show the total number of calls per hour and the average duration of the calls average for all days.

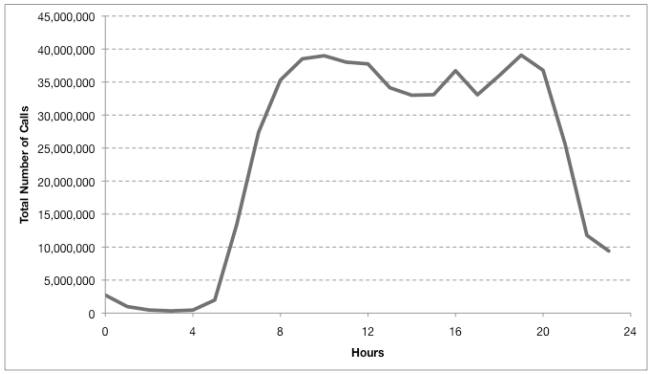


Figure 1. Total number of calls per hour during the average day

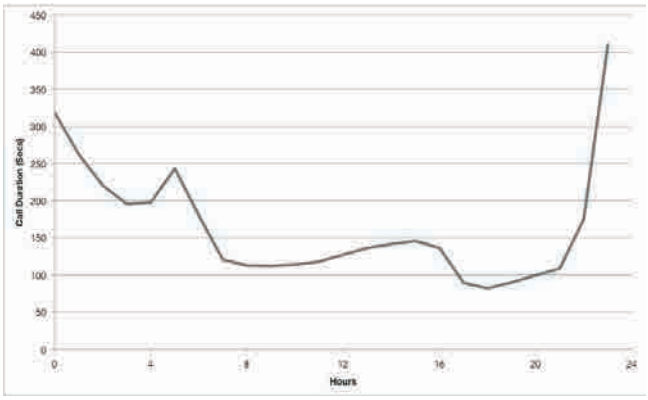


Figure 2. Average call duration during the average day

Using this information in the following figures, the data are aggregated by prefecture, and broken down by time in order to infer the number of business calls (Monday to Friday within the period 8AM-6PM), home calls (Monday to Friday outside the period 8AM-6PM), and weekend calls (Saturday and Sunday). We have used nested proportional polygons to illustrate the proportion of total incoming (Figure 3) and outgoing (Figure 4) call volumes associated with each prefecture for each of the call types previously defined. Note the outgoing is very similar to incoming so we have shown one of our early attempts towards 3D visualisation using Processing[1]. This technique represents each prefecture with 3 nested polygons, with the surface area of each polygon representing the proportion of calls within that prefecture falling into each call category.

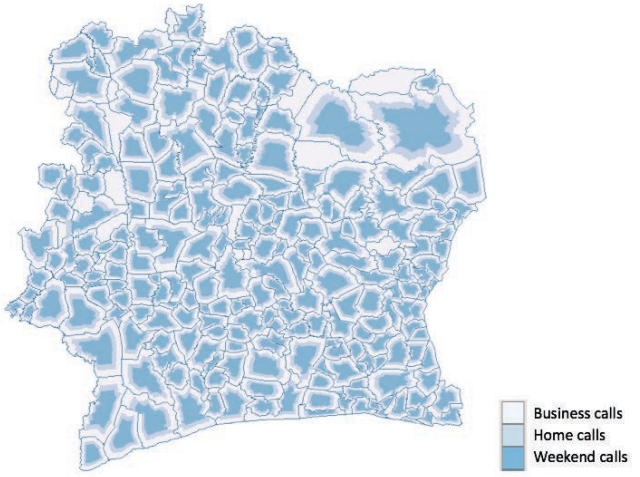


Figure 3. Classified incoming call volumes per prefecture

Whilst these plots are interesting in that they indicate the proportion of each defined call type, they appear to represent very similar social relational patterns across the prefectures and do not demonstrate differing levels of call volume between prefectures. In order to fully understand the flow of data around the country, it is therefore necessary to normalise call volumes by estimated population per prefecture in order to allow patterns of particularly high or low data flow to be identified. Therefore, in the following figures we illustrate the total volume of incoming and outgoing call duration per capita for each prefecture.

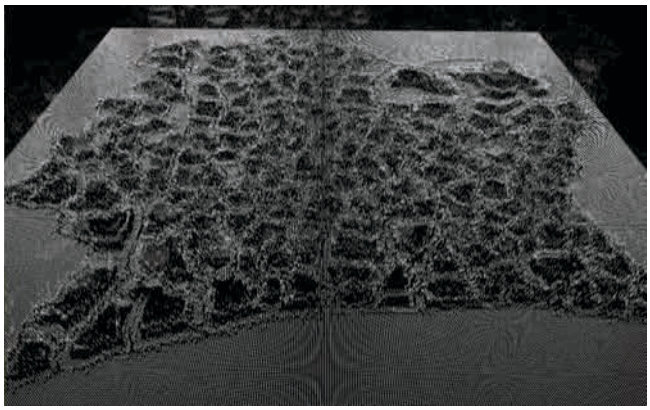


Figure 4. Classified outgoing call volumes per prefecture represented as 3D terrain plot

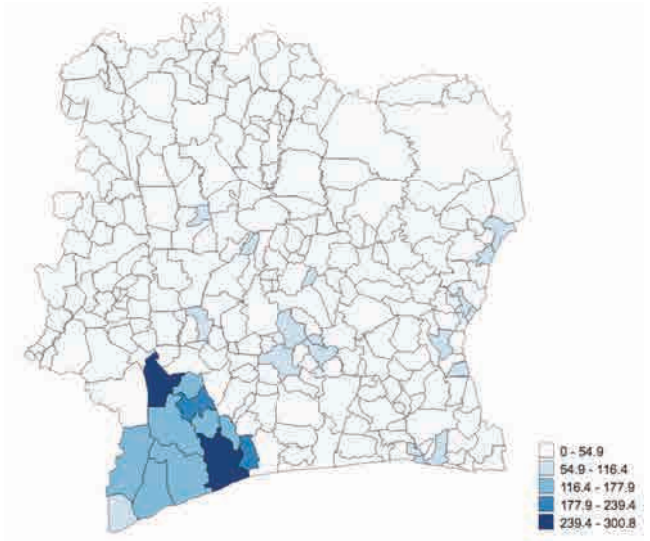


Figure 5. Incoming call volumes per capita by prefecture

The incoming and outgoing call volume per capita shows an unexpected level of activity in the south west of the country (a region known as Bas-Sassandra) which does not encompass the major population areas. This anomaly becomes further apparent if we add base station locations. These are shown on an individual basis and aggregate basis in Figures 7 and 8 respectively. It is interesting to note that the main cities in the south east of the country (Abidjan and Abobo) have the largest number of base stations, followed by Bouaké in the central region.

As a mobile phone infrastructure is typically put in place to meet the demands of potential subscribers, we can see from this figure that provision in the Bas-Sassandra region is relatively low, and would therefore suggest that something else is the cause for the increased activity.

[1] http://processing.org

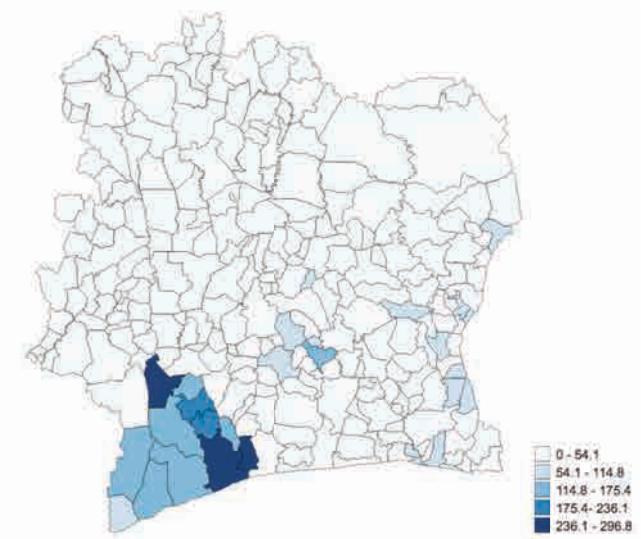


Figure 6. Outgoing call volumes per capita by prefecture

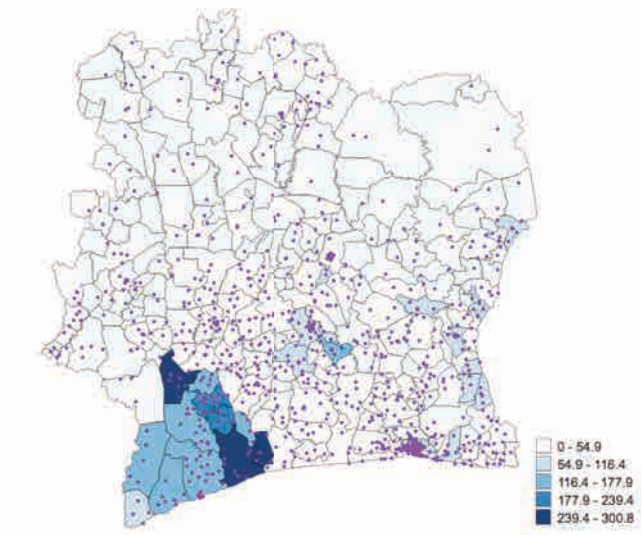


Figure 7. Outgoing calls volumes per capita by prefecture including base station locations.

To explore this apparent anomaly further, Figure 9 was constructed to reveal the top 5% connections between antennae (based on call duration). The figure suggests that the calls are being relayed through certain base stations. Examining general maps of the Bas-Sassandra region reveals that it contains the Taï National Park which covers an area of 4,540 km² of tropical evergreen forest. Our visualisations suggest the networks skirts around this region. There are similar patterns around Mont Péko National Park and the Marahoué National Park.

These anomalies highlight that any inferences developed through an analysis of large scale social relational data are likely to be affected by the physical topography of the mobile communications network infrastructure and this must be somehow accounted for within the data analysis. One possible solution is to only consider the point of origin and destination of the call and remove all the connections in between.

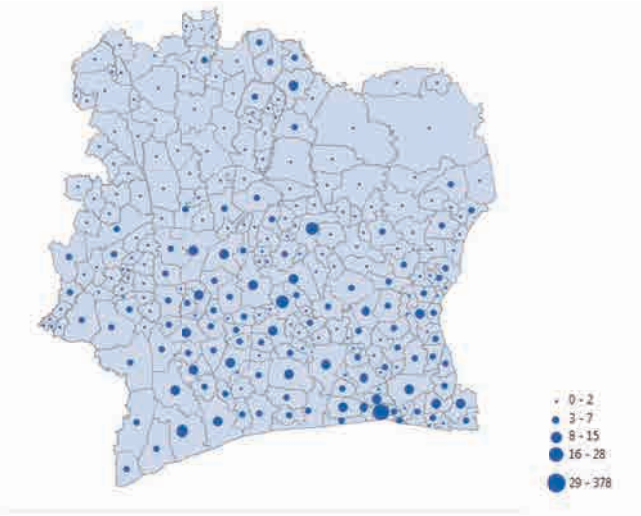


Figure 8. Number of base stations per prefecture

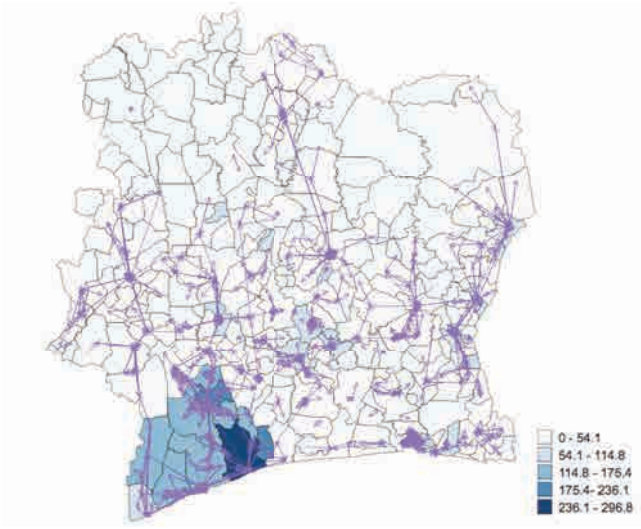


Figure 9. Top 5% High traffic connections between base stations.

3. CONCLUSIONS

The research conducted to date has been exploratory in nature, nevertheless, through generalisation (aggregation) and alternate forms of visualisation we have been able to identify some interesting results and highlight some of the many challenges to be overcome if complex spatio-temporal data are to be exploited to reveal their true information content.

In the development of a 3D visualisation of these data, we aim to combine the information contained within these exploratory analyses in order to produce a single visualisation that provides the user with an accessible, but information rich view of information flows throughout the Republic of Côte d'Ivoire.

4. ACKNOWLEDGMENTS

The authors would like to express their gratitude for the provision of the data presented in this paper through the Orange “Data for Development” initiative (www.d4d.orange.com/home).

5. REFERENCES

[1] Hägerstrand, T., 1970. What About People in Regional Science? Papers in Regional Science, Volume 24, Issue 1, 1970. pp 7-24.

[2] Eagle,N. Pentland, A. and Lazer, D. (2009), "Inferring Social Network Structure using Mobile Phone Data", Proceedings of the National Academy of Sciences (PNAS) 106(36), pp. 15274-15278.

[3] Blumenstock,J. Gillick, D. and Eagle, N. (2010), "Who's Calling? Demographics of Mobile Phone Use in Rwanda", AAAI Spring Symposium 2010 on Artificial Intelligence for Development

[4] Lefebvre, Henrí. (2004). Rhythmanalysis: Space, Time and Everyday Life, London: Continuum.

[5] Surdu, John R., and Udo W. Pooch. "Simulations technologies in the mission operational environment." Simulation 74.3 (2000): 138-160.

Mobility and communication patterns in Ivory Coast

Marija Mitrović,[1,*] Vasyl Palchykov,[1,2,*] Hang-Hyun Jo,[1,*] and Jari Saramäki[1,†]

[1]*Department of Biomedical Engineering and Computational Science (BECS), Aalto University School of Science, P.O. Box 12200, FI-00076, Finland*

[2]*Institute for Condensed Matter Physics, National Academy of Sciences of Ukraine, UA79011 Lviv, Ukraine*

(Dated: February 14, 2013)

We have studied patterns of human mobility and communication in Ivory Coast. We show that the characteristic mobility patterns, in terms of distributions of radius of gyration and distance travelled, are invariant in time and do not depend on the different samples in the source data. Thus there is a baseline mobility pattern, against which future deviations may be assessed; major sudden differences with respect to this baseline could be interpreted as warning signs. Mobility traces of individuals were also used to construct and visualize a weighted mobility network, depicting the flows of individuals between areas covered by different antennas. For studying communication patterns, we constructed a weighted network where the nodes represent antennas, and link weights denote call frequencies between antennas. Similarly to the mobility patterns, we confirmed that the overall characteristics of the call network are stable in time. We then addressed the question of the similarity of the mobility and call networks, with the result that these two weighted networks are strikingly similar. This points towards the possibility of estimating the overall flows of individuals between antenna positions from the antenna-level call network, without having to track the positions of individuals. We also constructed and visualized flow networks depicting the average directions of the flow of individuals and communication to and from antennas; it was seen that these networks display more diversity in densely populated areas.

Keywords: human mobility, communication networks

I. INTRODUCTION

A. Motivation

Mobile phone call records, especially when augmented with information on antenna positions, can be used to study societal-level patterns of human behaviour. Examples include structure of large-scale social networks [1, 2], the geography of calls and communication [3–5], characteristic patterns of mobility [6–11], population displacement in response to disaster [12], and the impact of human mobility on epidemics of infectious disease such as malaria [13]. The goal of the Data For Development (D4D) challenge was to contribute to the socio-economic development and well-being of populations. To this end, data sets on communication and mobility of mobile phone users in Ivory Coast were released, with the aim that researchers would extract information from the data that is relevant to development and improved life quality.

Our focus was on the mobility and communication patterns in Ivory Coast. Regarding mobility patterns, we wanted to establish a baseline picture of human mobility flows under normal circumstances. For this, we investigated the stability of statistical characteristics of mobility (distributions of the radii of gyration and travelled distances) and their invariance on the sample used. The aim was to provide an overall view that might be of use for *e.g.* infrastructure planning or geographic targeting of health interventions, and to visualize the flows of movement on the map of Ivory Coast. In addition to this, our motivation for establishing such a baseline was that a clear picture of country-wide mobility patterns under normal circumstances may be of crucial importance if some disaster strikes in the future that results in population displacement. First, any possible future deviations from this usual pattern may help to identify early signs of disasters. Second, this baseline mobility pattern can also be predictive of the targets of migration flows if disaster strikes: in a recent cell-phone based study of the mobility patterns of the population before and after the Haiti earthquake of 2010 [12], it was found that the destinations of people who fled from the heavily-affected capital were highly correlated with their mobility patterns during normal times.

Similarly to the mobility patterns, we investigated the characteristic communication patterns, aggregated at the antenna level, and their time invariance. We focused on the geospatial aspects of communication patterns, and investigated the radii of gyration of communication and average communication distances; the distributions of both measures were found to be stable in time. Hence, similarly to the mobility patterns, there is a baseline communication pattern of call frequencies between towers.

Having established the baseline mobility and communication patterns, we focused on investigating their relationship and similarity. It is known that the mobility patterns of individuals and their social network structure are intertwined to such an extent that human mobility patterns can be used to infer social ties or predict tie formation in the future [14, 15]. Here, instead of the level of individuals, we wanted to look at the big picture and use

*These authors have contributed equally to this work.
†Electronic address: jari.saramaki@aalto.fi

the available data on calls between antennas, comparing this geospatial network of information flows to the network of flows of individuals at the same level of coarse graining. We observed that these two networks are strikingly similar.

Finally, we also wanted to provide a visualized overview on the average directions of the flows of information and individuals in Ivory Coast; this was achieved with the help of flow networks where for each node, flow vectors representing the average communication/mobility directions are calculated. The direction of these flow vectors was seen to be related to geography: flow vectors from peripheral areas tend to point towards areas of high population density (as estimated by call volumes). Closer to the population center of mass, there is more diversity in the flow vector directions.

B. Data description

The data set from Ivory Coast was collected and provided by Orange Group [16]. It contains Call Detail Records (CDR) of phone calls between Orange's customers in Ivory Coast between December 1, 2011 and April, 2012. In this paper we consider two data sets: the *antenna-to-antenna* data set and the *high spatial resolution data set of individual trajectories*. The *antenna-to-antenna* data is used here to study aggregated call networks between antennas. It contains the number of calls as well as the their total duration between all pairs of 1231 antennas, *i.e.* mobile base stations. Antennas are uniquely identified by their id and geographical location. The temporal resolution of the data set is one hour, and calls spanning multiple time slots are associated with the time slot of their beginning.
The *high spatial resolution data of individual trajectories* is divided to 10 data sets that contain individual trajectories of 50,000 randomly chosen Orange customers (antennas where calls were made and the times of calls); these can then be linked to information on the location of each antenna. Each data set corresponds to one of the consecutive two week periods starting from December 5, 2011. This allows us to examine the time-invariance of mobility patterns and their independence on the particular sample.

For a significant number of calls in both data sets, the antenna identifiers were missing due to technical reasons. Such data were omitted from the analysis. Due to technical reasons, some of the data are missing in the data sets [16]. In our analysis we noticed certain periods of *low activity*, where the overall recorded daily activity is around an order of magnitude lower than during *normal* days. These low activity days were 15th of February, 24th of March, and the 10th, 15th and 19th of April.

II. ESTABLISHING THE BASELINE FOR MOBILITY AND COMMUNICATION PATTERNS

A. Trajectories of individuals: baseline statistics and time-invariance

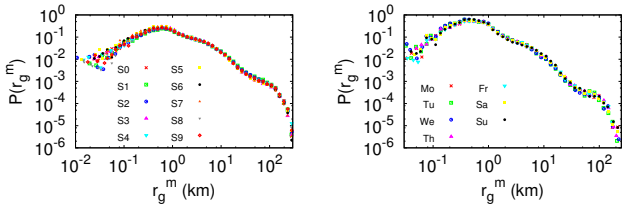


FIG. 1: Distribution of the radius of gyration calculated for different data sets for a period of two weeks (left) and for different days of the week (right).

Earlier studies of mobility patterns from cell-phone data sets have followed the same group of individuals for long time periods [6, 12, 13, 17]. Often, such detailed information on individuals' trajectories for long time periods can not be provided due to either technical or privacy reasons. As the source data here consists of several data sets where the trajectories of different individuals are followed, we first wanted to establish that the fundamental measures of human mobility remain consistent and sample-independent. At the same time, by verifying that the observed mobility patterns are time-invariant, we establish a baseline of typical patterns of human movement in Ivory Coast under normal circumstances.

We begin with the *radius of gyration* r_g^m of the mobility patterns that captures the average spatial extent of trajectories and can be interpreted as reflecting the characteristic distance traveled by an individual in the observed time period. It is defined as

$$r_g^m = \sqrt{\frac{1}{N_t}\sum_k |\vec{r}_k - \vec{r}_{CM}|^2}\ , \qquad (1)$$

where N_t is the number of recorded positions of an individual, $\vec{r}_k$ is the kth position of the mobility trace, and $\vec{r}_{CM}$ is the *center of mass* of the trace. The radius of gyration is different from the average distance travelled, since it accounts for immobility and weighs the contribution of each covered distance with a number of times an individual has traveled it. A person who spends most of the time at one location and travels only infrequently to distant parts of the country has a smaller radius of gyration than one who regularly covers these distances. Earlier studies [6] have indicated that the distributions of radii of gyration are typically broad; e.g. for customers of a European mobile phone company this distribution can be approximated with truncated power-law for $r_g \geq 1$ km.

For the Ivory Coast data, the distributions of radii of gyration for each two-week data set follow a fat-tailed

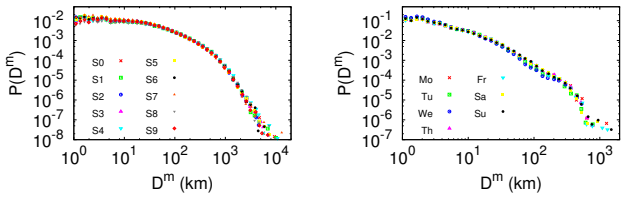


FIG. 2: Distribution of distances that humans cover in a two week period (left) and day (right).

distribution for values larger 1 km with three prominent peaks around 800 m, 5 km and 100 km (Fig. 1, left); the peaks reflect the spatial layout of the antenna network. These distributions are stable in time and appear similar for each data set, even though the samples track different individuals.

Earlier studies have indicated that human mobility patterns exhibit temporal periodicity [18, 19] on daily and weekly scales. Here, however, we find that for the Ivory Coast, in terms of the distributions of radii of gyration, no such differences can be detected (Fig. 1, right). Instead, $P(r_g^m)$ has a similar shape when calculated for a day or for a period of two weeks. As mentioned earlier in Sect. I B, there are in the data some days with very low activity levels. The average daily radius of gyration has a smaller value for these days; the functional form of $P(r_g^m)$ is qualitatively similar as for normal days except for a shift towards lower values.

In addition to the radii of gyration that characterize the overall spatial extent of the trajectories, we calculated aggregated daily and biweekly distances travelled by individuals, D^m, for each of the 10 data sets (Fig. 2). Although the shape of $P(D^m)$ differs for daily and biweekly aggregation windows, it is stable over time. The distribution is broad for both cases; however, the movement patterns are more heterogeneous for distances that individuals travel per day, Fig. 2 (right). We also calculated $P(D^m)$ for the low-activity days. These distributions are less broad and distances are on average shorter.

Although each data set follows the trajectories of a different sample of individuals, the distributions of both the radii of gyration and travelled distances, $P(r_g^m)$ and $P(D^m)$, appear very similar for each data set. This indicates that under normal circumstances the distributions for Ivory Coast are stable, forming a baseline against which possible deviations can be compared. Sudden future deviations from this baseline distribution may be indicative of *e.g.* natural disasters resulting in population movements; this was observed for the 2010 Haiti earthquake [12].

B. Mobility network

In order to provide a bird's-eye view to the mobility patterns of Ivory Coast residents, we aggregated the individual trajectories into a weighted mobility network. In this network, the antennas are represented as nodes, such that the weight of a link between two antennas is equal to the number of individuals' trajectories that join the two antennas, *i.e.* the number of times any individuals have moved between the antennas. Specifically, for each individual in the data set, one can create a sequence of consecutive positions, antennas, from which that individual made a call. If an individual at time t_i made a call that was originated at antenna i and, during the same day, that individual's next call at time t_j was from antenna j, we add one unit of weight to the directed link from i to j in the daily network. This way, we construct a directed network of human mobility between the antennas. The network does not contain self-loops, meaning that consecutive calls made in geographical proximity of the same antenna are not taken into account.

As the baseline distributions characterizing mobility were seen to be stable in time, we wanted to inspect whether there are differences in detailed mobility pat-

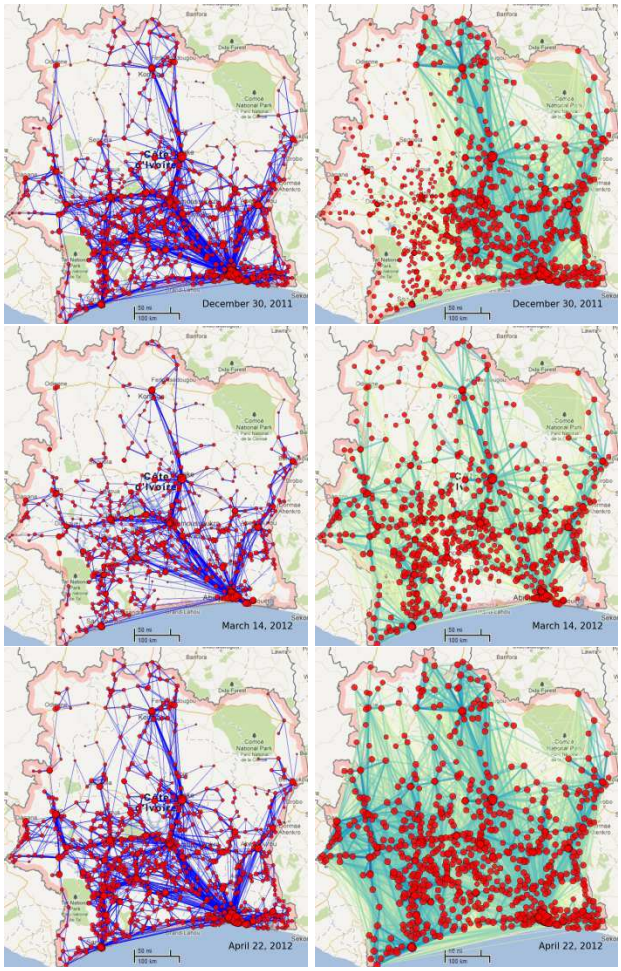


FIG. 3: Visualization of daily mobility (left) and communication (right) networks. For the communication networks, only links with >10 calls are shown. The map in the background is a screen capture from Google Maps.

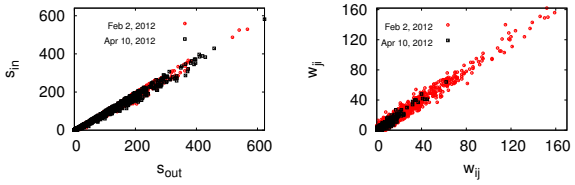


FIG. 4: (left) Scatter plot of number of incoming and outgoing individual trajectories for antennas during a period of one day (left). Comparison of links in terms of weights associated with different directions (right).

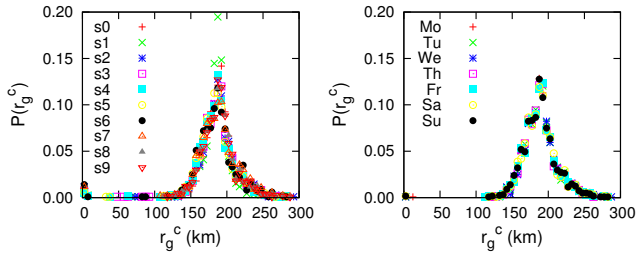


FIG. 5: Distributions of the radius of gyration r_g^c for communication networks for different two-week period datasets (left) and aggregated over different weekdays (right). Spatial resolution has been set to 5 km.

terns. To this end, we constructed daily mobility networks for each day in the period from December 4, 2011 till April 22, 2012. Fig. 3 shows three such networks obtained for different samples of individuals and different time periods. Although there are some differences in the daily networks, their overall structure appears rather similar, showing frequent movement of individuals between the largest cities. For the spatial communication networks (right panels), see Sect. II C.

Figure 4 (left) shows a comparison between the incoming and outgoing strength (sum of link weights) of each antenna. The number of individuals that arrive and make calls around an antenna is approximately the same as the number of individuals leaving it during the day. It follows from this that the daily mobility network does not have typical sources or sinks of trajectories, which is very feasible. A comparison of links in terms of their weights associated with different directions, Fig. 4 (right) shows that there is generally no preferential direction when it comes to trajectories between two nodes. Similar results were obtained for both normal and low-activity days.

C. Spatial communication patterns: baseline distributions and time-invariance

Similarly to the mobility network, we construct from the antenna-to-antenna data a geospatial communication network, where antennas are represented as nodes and the weight w_{ij} of the link $i-j$ equals the number of calls between the two antennas i and j. Our focus is on the geospatial aspects of this network, as well as its stability in time.

To characterize the geospatial diversity of calls — the area covered by communication froml antennas — we analogously to the mobility trajectories define the radius of gyration r_g^c for the communication patterns of the antennas. First, we calculate the position of the center of mass of communication for node i as

$$\vec{r}_{CM,i} = \frac{1}{s_i}\sum_j w_{ij}\vec{r}_j, \qquad (2)$$

where j runs over all nodes $j = 1 \dots N$, s_i is the strength of node i (the sum of the weights of its links), w_{ij} is the number of calls from node i to node j and $\vec{r}_j$ defines the geographical position of node j. Then we define the radius of gyration $r_{g,i}^c$ for the communication pattern of node i as

$$r_{g,i}^c = \sqrt{\frac{1}{s_i}\sum_j w_{ij}|\vec{r}_j - \vec{r}_{CM,i}|^2}. \qquad (3)$$

The difference to the mobility radius of gyration, r_g^m, is that this quantity characterizes antennas instead of individual subscribers. Thus, $r_{g,i}^c$ reflects the characteristic communication distance of the subscribers calling from the area covered by antenna i. The distributions $P(r_g^c)$ are shown in Fig. 5 aggregated for each two-week period (left) and for each week day (right). They are characterized by a well-defined maximum at ~ 180 km, independently of time window. The only notable difference between distributions is observed only for dataset between December 19, 2011 and January 01, 2012, where the width of the distribution is smaller as compared to other two-weeks periods and the maximum is more pronounced. This period covers public holidays. However, the overall shapes of the distributions overlap and this indicates that the overall pattern of communication is rather time-invariant similarly to the mobility network. As with the mobility characteristics, no significant differences between different days of the week are detected.

Similar conclusions may be drawn from the communication distance distributions $P(d)$ shown in Fig. 6, aggregated for two-week periods (left) and days (right). The distribution $P(d)$ gives the probability that a randomly selected call occurs between the antennas located at a distance of d km. This distribution demonstrates a high frequency of short-distance communication. Its steep decrease up to ~ 20 km can be attributed to the spatial extent of the biggest city, Abidjan. For the distances between 20 and ~ 500 km, the communication probability is slightly distance-dependent and demonstrates some local maxima around 290 km and 215 km. 290 km corresponds to the distance between Abidjan and Buoake, the second largest city; 215 km reflects the distance between Abidjan and Yamoussoukro, the capital of Ivory Coast. For distances larger than ~ 500 km the communication probability decreases steeply. The distributions

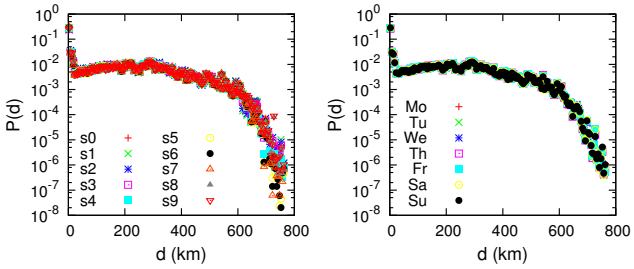


FIG. 6: Distributions $P(d)$ of communication distances d for different two-week periods (left) and aggregated over different weekdays (right). The spatial resolution has been set to 5 km.

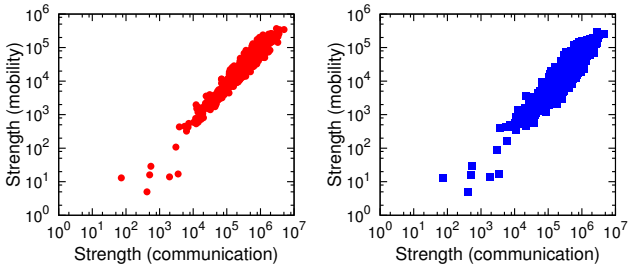


FIG. 7: Scatter plot for strength-strength correlations in the communication network (x-axis) and mobility network (y-axis) including and excluding self-loops (left and right panels, respectively). Node strengths in both networks are characterized by a high degree of correlations with the Pearson correlation coefficient $r = 0.934$ (with self-loops) and $r = 0.915$ (without self-loops).

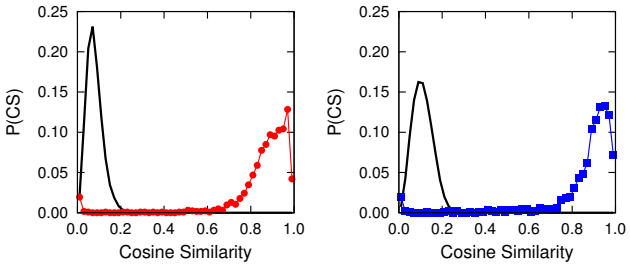


FIG. 8: Distribution of cosine similarities (4) between the link patterns of individual antennas in communication and mobility networks, with self-loops (left) and without self-loops (right). For most nodes, the similarities are very high, indicating a close match between the mobility and call networks. The black line denotes cosine similarities for a random reference, where the weights of the mobility network have been randomly reshuffled.

again show a high level of overlap, confirming the overall temporal stability of communication patterns.

III. SIMILARITY OF COMMUNICATION AND MOBILITY NETWORKS

Next, we turn to a comparison of the antenna-level mobility and call networks, and investigate whether their properties are similar, such that knowledge of one might be used to estimate the properties of the other. More specifically, the aim is to see whether the antenna-aggregated call network reflects the properties of the mobility network, whose construction requires tracking the trajectories of individual users. Similarity of the two networks would indicate that one can estimate mobility patterns from the call network, aggregated at the antenna level, without having to track the locations of subscribers. As the statistical characteristics of both the mobility and communication patterns appear stable during the investigated period, we can perform a direct comparison of the corresponding networks at the aggregated level, where link weights denote either the number of individuals that have moved between antennas, or the number of calls between antennas. For the following, we use undirected versions of both networks.

A. Correlations between mobility and call frequencies of antennas

First, we test whether high-strength antennas in mobility network, associated with high level of flow of individuals into and out of their areas, coincide with the highly-active antennas in the communication network. Fig. 7 displays the relationship between node strengths in communication and mobility networks, such that the strength of a node in the communication network represents the total number of calls made from and to the area of the respective antenna, whereas in the mobility network the strength represents the number of individuals moved from and to the area of that antenna. This comparison shows strong correlations between the two networks, with the Pearson correlation coefficient $r = 0.934$ if the network includes self-loops and $r = 0.915$ if self-loops are excluded. Thus, as expected, antennas associated with high call volumes are also associated with a high level of mobility. Here, a self-loop in the communication network indicates calls that originate and are received at the same antenna; for the mobility network a self-loop indicates immobility, *i.e.* two consecutive calls by the same individual at the same antenna.

B. Similarity of detailed antenna-level call and mobility patterns

In order to compare in detail the mobility and call patterns of antennas, we focus on their connectivity patterns in terms of link weights that characterize flows of mobility and calls. For this, we first assign to each node $i = 1 \ldots n$ two vectors $\vec{w}_i = (w_{i1}, w_{i2}, \ldots, w_{in})$ and $\vec{m}_i = (m_{i1}, m_{i2}, \ldots, m_{in})$. Here w_{ij} and m_{ij} are the link

weights between nodes i and j in the communication and mobility networks, respectively. Then, structural similarity of the links of node i and their weights for the two networks can be measured with the cosine similarity between vectors $\vec{w}_i$ and $\vec{m}_i$, defined as

$$CS_i = \frac{\vec{w}_i \cdot \vec{m}_i}{|\vec{w}_i||\vec{m}_i|}. \quad (4)$$

Here $\vec{w}_i \cdot \vec{m}_i$ is the scalar product $\vec{w}_i \cdot \vec{m}_i = \sum_j w_{ij} m_{ij}$ and $|\vec{w}_i|$ and $|\vec{m}_i|$ are the norms of the vectors $\vec{w}_i$ and $\vec{m}_i$, correspondingly. The value of cosine similarity can vary in the range $CS_i \in [0, 1]$ since link weights are always positive. If $CS_i = 0$, the two vectors $\vec{w}_i$ and $\vec{m}_i$ are orthogonal and the links of node i in the two networks do not overlap. In contrast, if $CS_i = 1$ the vectors $\vec{w}_i$ and $\vec{m}_i$ are parallel: not only are i's links the same in both networks, but their weights are proportional too, and thus node i's connectivity patterns in both networks are the same up to a scaling factor. The closer the value of CS_i to 0, the more different are the links, and the closer the value of CS_i to 1, the more similar they are.

Fig. 8 shows the distributions of cosine similarities between mobility and communication networks if self-loops are allowed (left) and if self-loops are excluded ($w_{ii} = 0$, $m_{ii} = 0$ for $i = 1 \dots n$) (right). The black line denotes cosine similarities for a random reference case, where the weights of the mobility network have been randomly shuffled (averaged over 100 runs). In both cases, it is clear that for most nodes, the values of cosine similarity are close to the maximum value of unity and very far from the random reference. Thus, even at the level of the links of individual nodes and their weights, the mobility and call networks are very similar. This points towards the possibility of estimating the flows of individuals between antennas simply by monitoring the numbers of calls between the antennas, without necessarily having to track the positions of individuals: the antenna-level call network can be used to estimate mobility patterns.

IV. FLOW NETWORKS

A. Flow network visualization

For an overview on the main directions of the flow of individuals as well as the flow of information, we have constructed flow networks on the basis of the mobility and call networks discussed above. These networks show the average direction of flow from and to antennas. For this, we denote the location of antenna i with the two-dimensional vector $\vec{r}_i$, obtained from the latitude and longitude in the dataset by means of the normal Mercator on the ellipsoid (NME) method [20]. Then, for each antenna i, we calculate the outgoing and incoming mobility and communication flow vectors as the weighted sum of unit vectors among antennas:

$$\vec{c}_{i,\text{out}} = \sum_{j \neq i} w_{ij} \hat{e}_{ij}, \quad (5)$$

$$\vec{c}_{i,\text{in}} = \sum_{j \neq i} w_{ji} \hat{e}_{ji}, \quad (6)$$

where w_{ij} is the number of calls from the antenna i to j, and $\hat{e}_{ij} \equiv (\vec{r}_j - \vec{r}_i)/|\vec{r}_j - \vec{r}_i|$ denotes the unit vector pointing to $\vec{r}_j$ from $\vec{r}_i$. For the mobility pattern, $\vec{m}_{i,\text{out}}$ and $\vec{m}_{i,\text{in}}$ are similarly defined with w_{ij} as the number of cases where one individual user moves from the area of antenna i to the area of antenna j, indicated by successive calls made from these antennas.

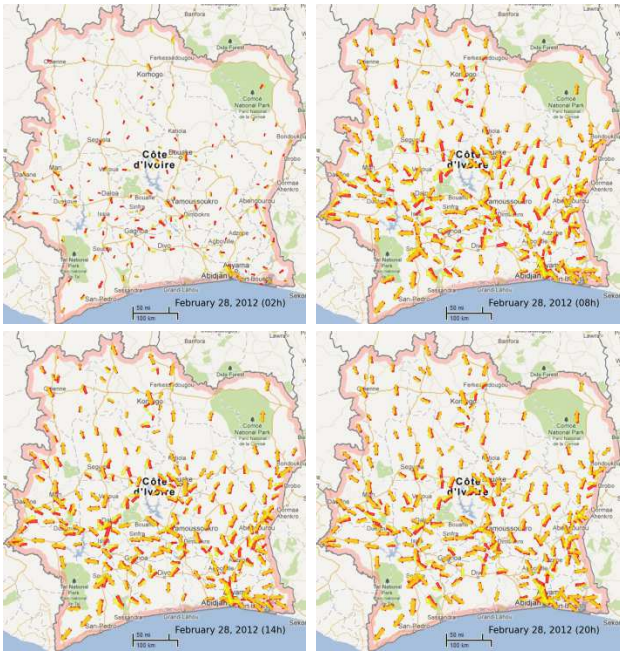


FIG. 9: Visualization of outgoing (yellow) and incoming (red) communication flow vectors. The size and length of arrow are logarithmic functions of the length of the vectors with positive lower bounds.

For visualizing the flow networks, we cluster the antennas at the level of sub-prefectures of Ivory Coast with the help of information on the administrative boundaries of sub-prefectures. We then calculate the outgoing and incoming communication flow vectors for each sub-prefecture and visualize them as yellow (outgoing) and red (incoming) arrows in Fig. 9. This visualization shows that there is a general tendency for the flow vectors of peripheral areas to point towards more central areas and areas of high population density, whereas in those areas, flow vectors display more diversity.

B. Similarity of communication and mobility flows

Next, we compare call and mobility flow vectors, as well as in- and out-flow vectors in both of these networks.

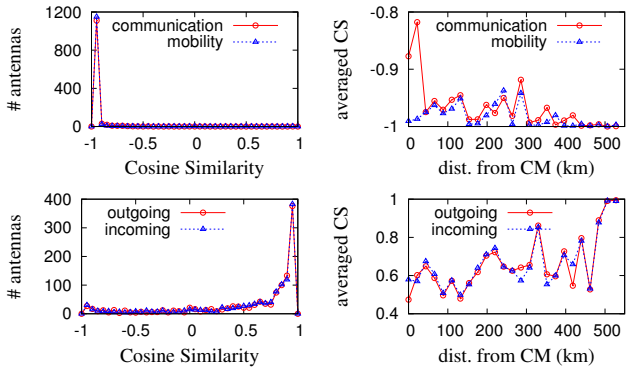


FIG. 10: (Left) Distributions of cosine similarities between outgoing and incoming flow vectors (top) and between communication and mobility flow vectors (bottom). (Right) The dependence of cosine similarity on the distance of the antenna from the country-level center of mass.

For this, we use the cosine similarity, defined similarly as in Eq. (4). First, we focus on possible differences between in- and out-flows, and see whether these vectors point in opposite directions for individual nodes or whether there are resultant net flows. This is done separately for the communication and mobility networks. Our results show (Fig. 10 (top left)) that in- and out-flows are typically antiparallel; this is in line with the observation that there are no sinks or sources in the networks. When cross-comparing the two different flow networks (Fig. 10 (bottom left)), we see that the flow vectors of both networks are very similar, which is again in line with observations on the detailed networks.

In order to investigate whether the structure of mobility and communication flows depends on geography, *i.e.* whether there are local differences, we compare the results against a geographic point of reference. For this, we calculate the country-level communication center of mass (CM) as

$$\vec{r}_{\rm CM} = \frac{\sum_i s_i \vec{r}_i}{\sum_i s_i}, \qquad (7)$$

where s_i is the total number of calls at antenna i and r_i its position. If we assume a correlation between the number of calls and population density, this center of mass reflects the position of the population center of mass. First, we investigate the cosine similarities between in- and out-flows in both call and mobility networks. Fig. 10 (top right) shows that the farther the antenna is from the CM, the more anti-parallel the outgoing and incoming flow vectors are on average to each other; close to the center of mass, there is a resultant communication flow as the vectors do not cancel out. When comparing the communication and mobility networks (Fig. 10 (bottom right)), it is seen that the farther the antenna is from the CM, the more parallel the mobility and communication flows are. Thus, proximity to the CM implies larger diversity in the flows.

V. SUMMARY

To summarize, we have analyzed human mobility and communication patterns in Ivory Coast in order to establish a baseline picture of human behavior under normal circumstances. Our findings show that the statistical characteristics of both mobility and geospatial communication patterns exhibit temporal stability and are sample-independent. Thus they can be considered as a baseline representing mobility and communication activity under normal circumstances. Surprisingly, the characteristic distributions look similar even at the daily level, and no significant differences between different weekdays are detected.

The distribution of cosine similarities between weighted connectivity patterns of antennas indicates a striking similarity between mobility and communication networks at the level of individual antennas. This result points towards the possibility of estimating large-scale mobility patterns — flows of individuals between antenna positions – from the antenna-level call network, without having to track the positions of individuals. The similarity of call and mobility patterns is also evident in flow networks depicting average directions of the flow of individuals and communication; these networks also reveal differences between central and peripheral areas.

Acknowledgments

VP acknowledges support by the Finland Distinguished Professor (FiDiPro) program of TEKES. JS acknowledges support by the Academy of Finland, project n:o 260427. HJ acknowledges financial support by the Aalto University postdoctoral program. We thank Arnab Chatterjee, Lauri Kovanen, and Gerardo Iñiguez for comments.

[1] W. Aiello, F. Chung, and L. Lu, "A random graph model for massive graphs," in *Proceedings of the 32nd annual ACM symposium on Theory of Computing*, (New York, NY, USA), pp. 171–180, ACM, 2000.

[2] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A.-L. Barabási, "Structure and tie strengths in mobile communication networks," *Proceedings of the National Academy of Sci-*

ences, vol. 104, no. 18, p. 7332, 2007.

[3] R. Lambiotte, V. D. Blondel, C. de Kerchove, E. Huens, C. Prieur, Z. Smoreda, and P. Van Dooren, "Geographical dispersal of mobile communication networks," *Physica A: Statistical Mechanics and its Applications*, vol. 387, no. 21, pp. 5317–5325, 2008.

[4] G. Krings, F. Calabrese, C. Ratti, and V. D. Blondel, "Urban gravity: a model for inter-city telecommunication flows," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2009, p. L07003, 2009.

[5] V. D. Blondel, G. Krings, and I. Thomas, "Regions and borders of mobile telephony in Belgium and in the Brussels metropolitan zone," *Brussels Studies*, vol. 42, no. 4, 2010.

[6] M. C. González, C. A. Hidalgo, and A.-L. Barabási, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, 2008.

[7] C. M. Song, T. Koren, P. Wang, and A.-L. Barabási, "Modelling the scaling properties of human mobility," *Nature Physics*, vol. 6, p. 818, 2010.

[8] C. M. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of predictability in human mobility," *Science*, vol. 327, p. 1018, 2010.

[9] F. Simini, M. C. González, A. Maritan, and A.-L. Barabási, "A universal model for mobility and migration patterns," *Nature*, vol. 484, p. 96, 2012.

[10] J. P. Bagrow and Y.-R. Lin, "Mesoscopic structure and social aspects of human mobility," *PLoS One*, vol. 7, p. e37676, 2012.

[11] B. C. Csáji, A. Browet, V. Traag, J.-C. Delvenne, E. Huens, P. Van Dooren, Z. Smoreda, and V. D. Blondel, "Exploring the mobility of mobile phone users," *Physica A: Statistical Mechanics and its Applications*, vol. 392, pp. 1459 – 1473, 2013.

[12] X. Lu, L. Bengtsson, and P. Holme, "Predictability of population displacement after the 2010 Haiti earthquake," *Proceedings of the National Academy of Sciences (USA)*, vol. 109, pp. 11576–11581, 2012.

[13] A. Wesolowski, N. Eagle, A. J. Tatem, D. L. Smith, A. M. Noor, R. W. Snow, and C. O. Buckee, "Quantifying the Impact of Human Mobility on Malaria," *Science*, vol. 338, p. 267, 2012.

[14] D. Crandall, L. Backstrom, D. Cosley, S. Suri, D. Huttenlocher, and J. Kleinberg, "Inferring social ties from geographic coincidences," *Proceedings of the National Academy of Sciences (USA)*, vol. 107, p. 22436, 2010.

[15] D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A.-L. Barabási, "Human mobility, social ties, and link prediction," in *KDD '11 Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, (New York, NY, USA), p. 1100, ACM, 2011.

[16] V. D. Blondel, M. Esch, C. Chan, F. Clérot, P. Deville, E. Huens, F. Morlot, Z. Smoreda, and C. Ziemlicki, "Data for Development: the D4D Challenge on Mobile Phone Data," *arXiv:1210.0137*, 2012.

[17] P. Wang, M. C. González, C. A. Hidalgo, and A.-L. Barabási, "Understanding the Spreading Patterns of Mobile Phone Viruses," *Science*, vol. 324, no. 5930, pp. 1071–1076, 2009.

[18] R. Schlich and K. Axhausen, "Habitual travel behaviour: Evidence from a six-week travel diary," *Transportation*, vol. 30, pp. 13–36, 2003.

[19] N. N. Eagle and A. S. Pentland, "Eigenbehaviors: Identifying Structure in Routine," *Behavioral Ecology and Sociobiology*, vol. 63, pp. 1057–1966, 2009.

[20] http://mercator.myzen.co.uk/mercator.pdf

Detecting Mobility Patterns in Mobile Phone Data from the Ivory Coast

Matthew F. Dixon, Spencer P. Aiello, Funmi Fapohunda, William Goldstein

Graduate Program in Analytics
University of San Francisco
2130, Fulton Street, San Francisco, CA 94117

This paper investigates the Data for Development (D4D) challenge [3], an open challenge set by the French mobile phone company, Orange, who have provided anonymized records of their customers in the Ivory Coast. This data spans a 5 month (150 day) horizon spread across 4 different sets containing antenna-to-antenna traffic, trace data for 50,000 customers at varying spatial resolution, and social graphs for 5,000 customers. By leveraging cloud-based and open-source analytics infrastructure to (1) merge the D4D datasets with Geographic Information System (GIS) data and (2) apply data mining algorithms, this paper presents a number of techniques for detecting mobility patterns of Orange customers in the Ivory Coast.

By applying a k-medoid clustering algorithm to the antenna locations and their average distance to nearby attennae, we show how the high spatial resolution mobile phone dataset reveals a number of daily mobility patterns and properties, including trends in week-day versus weekend, public holiday mobility behavior, and distributional properties of daily trip distances across each cluster. With a view towards providing tools to assist with transport infrastructure planning, we combine the high spatial resolution D4D dataset with GIS data for transport infrastructure and demonstrate an approach for detecting whether a mobile phone user is traveling on a segment of transport infrastructure. This work culminates in a preliminary cloud-based GIS tool for visualizing mobility traces.

1 Introduction

The Ivory Coast is a developing West African country renound for their cocoa and coffee. It faces disease outbreaks at its West, North, and coastal borders, suffers drought caused by the seasonal Harmattan winds, and is divided north-south by ethnic and religious tensions, occasionally encountering out-bursts of civil war [5]. The Ivory Coast is an important subject of study for economists attempting to quantify reactions to Harmattan winds and civil war outburst, for epidemiologists' models of human migration to guide the dispursing of resources for disease eradication, and for sociologists considering all aspects of social policy.

The emergence of vast quantities of communication and movement data generated by mobile phone customers has the potential to positively impact urban planning, enconomic development, and public health decisions [4]. While the validity of such data to answer broad questions of human behavior remains dubious [2], it is possible to harvest meaningful insights from mobile data, even if it only spans a 5-month horizon [1].

One promising area of application in particular is the use of high spatial resolution mobile phone data to study mobility patterns for the ultimate purpose of mitigating congestion in urban roads, urban planning, traffic prediction and the study of complex networks. Such studies are conventionally based on primitive travel surveys and typically fail to support transport planners with information needed to design future road networks able to withstand modern mobility demand [9]. In fact, across the world, our understanding of the origins and destinations of commuters through a particular segment of road or railline remains for the most part poorly understood and unquantified.

Recently Wang, Hunter, Bayen, Schechtner Gonzlez [9] presented an approach for understanding road usage patterns in urban areas through the integration of mobile phone data and GIS data. This approach revealed hidden patterns in road usage in the San Francisco Bay and Boston areas and demonstrated a basis for more informed and cost effective transport planning and congestion mitigation. This study used three-week-long mobile phone billing records generated by 360,000 San Francisco Bay Area users (6.56% of the population, from one carrier) and 680,000 Boston Area users (19.35% of the population, from several carriers) respectively [9]. Given the high density of service towers (there are 892 antenna service areas in the San Francisco Bay Area), the authors are able to perform a detailed study of road usage patterns.

Mobility patterns Motivated by the study of Wang et al. [9], this paper sets out to characterize and quantify the daily

distances travelled by mobile phone users. In order to do this, we begin in Sections 2 and 3 with a description of the data from which it becomes apparent that the spatial resolution of the antennae is not only much lower than in the study by Wang et al. [9], but is considerably more varied depending on proximity to large cities and its region. Variable resolution challenges the interpretation of mobility distance studies and in Section 4 we turn to k-medoid clustering to partition the antennae into artificial regions in which antennae density are more uniformly distributed. By partitioning the antennae into a small number of clusters, we proceed to characterize the distribution and time history of daily trip distances and identify trends in mobility behavior and anomalies which are unencountered for by religious events or national holidays.

Route detection An additional challenge to interpretating the mobility patterns is that commuters clearly do not follow the shortest path through a sequence of antennae, but travel on road and railroads whose routes may be significantly different to the mobility trace. Section 5 demonstrates a simple methodology for detecting whether a user has travelled on a segment of transport infrastructure. This methodology is based on integrating the D4D datasets with Geographic Information System (GIS) data and open-source software infrastructure. Using this methodology, we show the full mobility patterns of users who use the railroad on a particular day and reveal their source and end destinations. Such insight provides a basis for better informed transportation planning, including targeted strategies to mitigate congestion.

Visualization tool An effective and low-cost mechanism for transferring analytics research to application domains such as transport planning, is to provide an open source web-based prototype visualization tool which enables the user to study individual user trajectories and observe how they interact with the transport infrastructure. Section 6 describes the infrastructure used to create this tool and provides a URL which the reader can use to access this prototype visualization tool. The site is password protected and reader should contact the corresponding author for login credentials.

2 Data Description

The data, spanning December 2011 to April 2012, provided by Orange has been released to research teams in order to examine developmental questions in new ways [3].

There are four sets of data provided for the D4D project:
SET1: Antenna-to-antenna, number of calls as well as the duration of calls between any pair of antennas aggregated hour by hour. The antenna positions are given by longitude/latitude pairs.
SET2: Individual Trajectories, High Spatial Resolution Data movement trajectories for 50,000 users. The sub-prefectures are given by longitude/latitude pairs.
SET3: Individual Trajectories, Low Spatial Resolution Data movement trajectories for 50,000 users over the entire observation period at lower spatial resolution (phone calls aggregated by prefecture, rather than antenna position). The sub-prefectures are given by longitude/latitude pairs.
SET4: Communication Subgraphs, communication subgraphs for 50,000 randomly selected individuals

In this paper we use the first two datasets for detecting daily mobility patterns. We additionally use country and administrative subdivision outlines provided by GDAM [7] (an open geospatial datasource), and road and railroad vectors from the Digital Chart of the World, both of which can be accessed through the open GIS software and data content provider DIVA-GIS [6].

3 Route Visualization

In order to gain some initial insight into the mobility patterns of users, it is useful to visualize their mobility traces over the course of a day. Appendix A provides the details of how individual mobility traces are aggregated from the call data. Figure 1 shows the antennae (black dots), the major cities (black circles) with a population of at least 1M and the mobility trace of all users (red translucent lines) travelling on the 6th of December 2011, the second day of the D4D dataset. The department boundaries are also shown on the map of the Ivory Coast. Figure 2 shows the Ivory Coast road network and by comparing this figure with Figure 1, we are able to explain many of the features in the mobility traces. Hubs of mobility activity are observed to coincide with major city locations and the most intense mobility activity is in the Abidjan area in the south of the Ivory Coast. The rest of the southern half of the country has a fairly even distribution, while the western and northern areas of the country lag behind in terms of mobile capacity. This distribution of antenna intensity is further observed to be commensurate with the population density map shown in Figure 3.

4 Daily Trip Distances

There are many challenges in using call data in SET2 to study daily mobility patterns. Aside from the obvious fact that the 50k mobile phone users included in the dataset represent a small fraction of the estimated 20M people (source: World Bank, 2011) who reside in the Ivory Coast, there are more technical challenges. There are only 1231 antennae listed in the SET2 dataset to cover a country with an area of 124,500 sq miles. Furthermore, these antennae are non-uniformly located, thus introducing variation in the minimum threshold distance that a user needs to travel in order for that trip to be detected. In Abidjan, for example, the antennae density is higher and thus there is higher fidelity in trip detection. In contrast, antennae are sparsely located over the north of the Ivory Coast and shorter trips may not be detected.

In order to study the distribution and time series of daily mobility distances using all available call data recorded in SET2 , we use a clustering algorithm to separate the antennae into a small number of distinct clusters. The purpose of creating clusters is to partially address the above concern regarding the variation in minimum threshold distance for trip

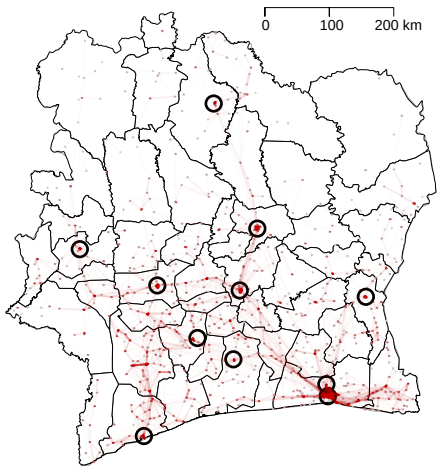


Fig. 1. A map of the Ivory coast showing the department borders, major cities with a population in excess of 1M (black circles) , antenna locations (black dots) and the mobility traces of all users on the 6th of December, 2011 (translucent red lines).

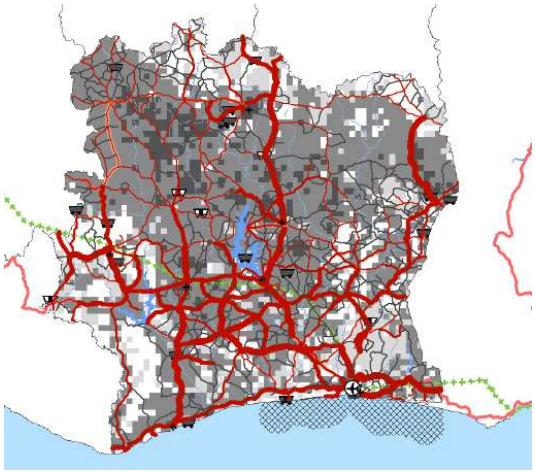


Fig. 2. The Ivory Coast Road Network. Source: AICD Interactive Infrastructure Atlas for The Ivory Coast downloadable from http://www.infrastructureafrica.org/aicd/system/files/civ_new_ALL.pdf

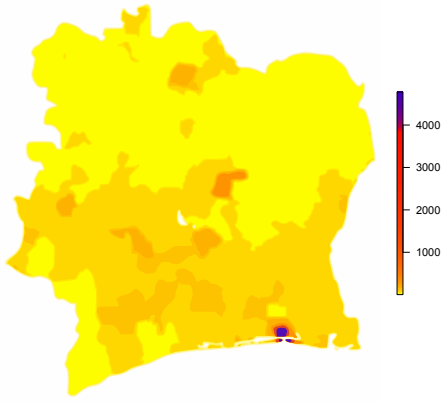


Fig. 3. A map of the population density (per sq. km) over the Ivory Coast.

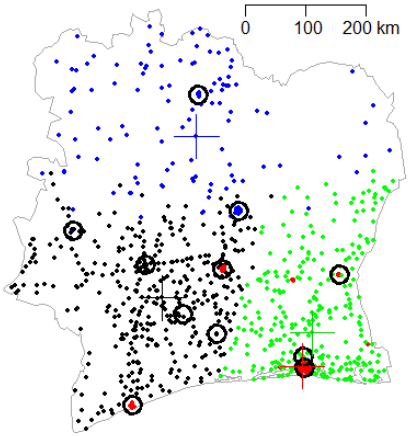


Fig. 4. This map show the clusters of antenna locations. The centroids of each cluster are shown by large colored crosses. Cities are shown with black circles. Key: Cluster 1 (black), Cluster 2 (red), Cluster 3 (green) and Cluster 4 (blue).

detection which is apparent by viewing the antenna locations in Figure 1. This direction is predicated on the notion that antenna service areas are approximately equal in radius and that the impact of terrain topology on coverage can be considered secondary. We introduce an antenna distance dispersion measure over an area A on the map. For each antenna location $p_i \in A$, we define the set of other antenna locations P_i which are in the neighborhood of p_i as

$$P_i := \{p_j \ : \ \text{dist}(p_i, p_j) \leq C_0, p_i \neq p_j\}.$$

Defining the average distance between p_i and all other points $p_j \in P_i$

$$d_i = \frac{1}{|P_i|} \sum_{p_j \in P_i} dist(p_i, p_j),$$

and our antenna distance dispersion measure as the standard

deviation of the average antenna distances over the area A

$$\sigma^2(D) = \frac{1}{|D|-1} \sum_{d_i \in D} (d_i - \mu(D))^2,$$

where D denotes the set of d_i corresponding to all points p_i and $\mu(D)$ is the mean of D.

Clusters are formed using the Partitioning Around Medoids (PAM) algorithm [8] applied to the standarized antenna longitude and latitude co-ordinates, **and** the standardized average distances between antennae D. C_0 is chosen to be 25km which is found by trial and error to yield a sufficiently high number of antennae with at least one other nearby antenna, while still resolving urban areas of dense antenna locations. A is chosen to cover the entire region of the Ivory Coast. This choice of features for clustering is based on a competing desire to preseve the contiguous geographic regions on the map but also partition the dataset by the average distance between antennae.

PAM is a type of k-medoids algorithm which attempts to minimize the distance between points labeled to be in one of k clusters and a point designated as the center of each cluster. k was set arbitrarily to 4 so that the granularity of the mobility study is coarser than a regional study but sufficiently granular to reduce dispersion in the average distance between antennae. Table 1 shows the details of the cluster properties - the number of points in the cluster (size) and the first two moments of the average distance between antennae over the cluster. The color codes listed in the table correspond to the colors of labelled antennae shown in Figure 4. The bottom row shows the moments of the average distance between antennae over the entire dataset, without clustering.

We note in particular that Cluster 2 (shown in red) represents some of the most significant dense urban areas, including Abidjan, and is characterized by more dense and uniformly distributed antenna locations. Clusters 1 and 3 have similar density and uniformity characteristics, with the average of antenna distances being higher than the national average and the level of dispersion being lower. Hence, Clusters 1-3 are observed to exhibit more uniform antenna locations. Cluster 4, representing the north region, on the other hand exhibits a smaller set of antennae and a high level of dispersion. For this reason, we approach the measurement of daily commute distances of users associated with Cluster 4 with more caution.

Data preparation The daily distances travelled by users can be associated with each cluster by the location of the antenna which they are most frequently closest to over the two-week period. This of course is not a reliable indicator of where a user resides, but in most cases it serves as a starting pointing for approximating a user's central place of calling activity and hence from in which most trip distance estimates are made. In the proceeding analysis below, it is important to note that we first applied two filters: (1) exclusion of users who do not travel on any day over two week period and (2) exclusion of daily distances for a user if at any point during the day the nearest antenna is listed as '-1'. The motivation for the first filter is to focus on the mobile cohort of customers in each cluster and the effect is to reduce the number of users included in the study by approximately 30%. The second filter removes daily distances which may be flawed and the effect is to remove a further 10% of all daily distances per user logged.

cluster	color	size	μ	σ
1	black	380	16.15	2.85
2	red	384	9.06	1.59
3	green	320	16.35	2.85
4	blue	147	15.22	5.54
Orig.		1231	13.84	4.41

Table 1. The characteristics of each cluster are shown for comparison with the original dataset (bottom row).

Times series of user mobility The top left graph in Figures 5 and 6 shows the historical time series of the average of the daily commute distances over each cluster between December the 5th, 2011 and February the 15th, 2012, and between February the 16th and April the 22nd, 2012 respectively. The splitting of the time series into two components is purely to improve the readibility of the graphs. The start and end of weekends are shown in the time series plot as two vertical gray lines. The bottom axis shows the monthly, weekly and daily markers, and the top axis shows the periods representing each two-week period in the dataset (every other marker indicates the start of each two-week period).

Weekends On first glance, it would seem apparent that the average daily distance travelled by users is generally lower at the weekend. However, this is a misleading interpretation of the mobility traces. We see in the top right graph in Figures 5 and 6 the corresponding ratio of 'undetected commuters' on any given day to the cluster cohort size. Undetected commuters are any combination of the following: users who either do not travel, are confined to the service area of one antenna, or make less than two calls on that day. We observe that the least sparse cluster (Cluster 2), representing dense urban areas, generally exhibits the lowest ratio of undetected commuters on any given day. In any two week period, we further observe that the Cluster 2 ratio of undetected commuters generates fluctuates the least between the period up until the 3rd weekend in March, after which point, ratios fluctuate considerably.

Although the user cohort is constant over any two-week period, the average number of calls that a user makes in a day (shown in the bottom left graph of Figures 5 and 6) varies over time. We observe a general trend of call volumes per user being lower at weekends, although this is less pronounced than in the ratio of undetected commuters. In each cluster, average call volumes are found to be correlated with average daily distances, $\rho = (0.42, 0.76, 0.67, 0.58)$.

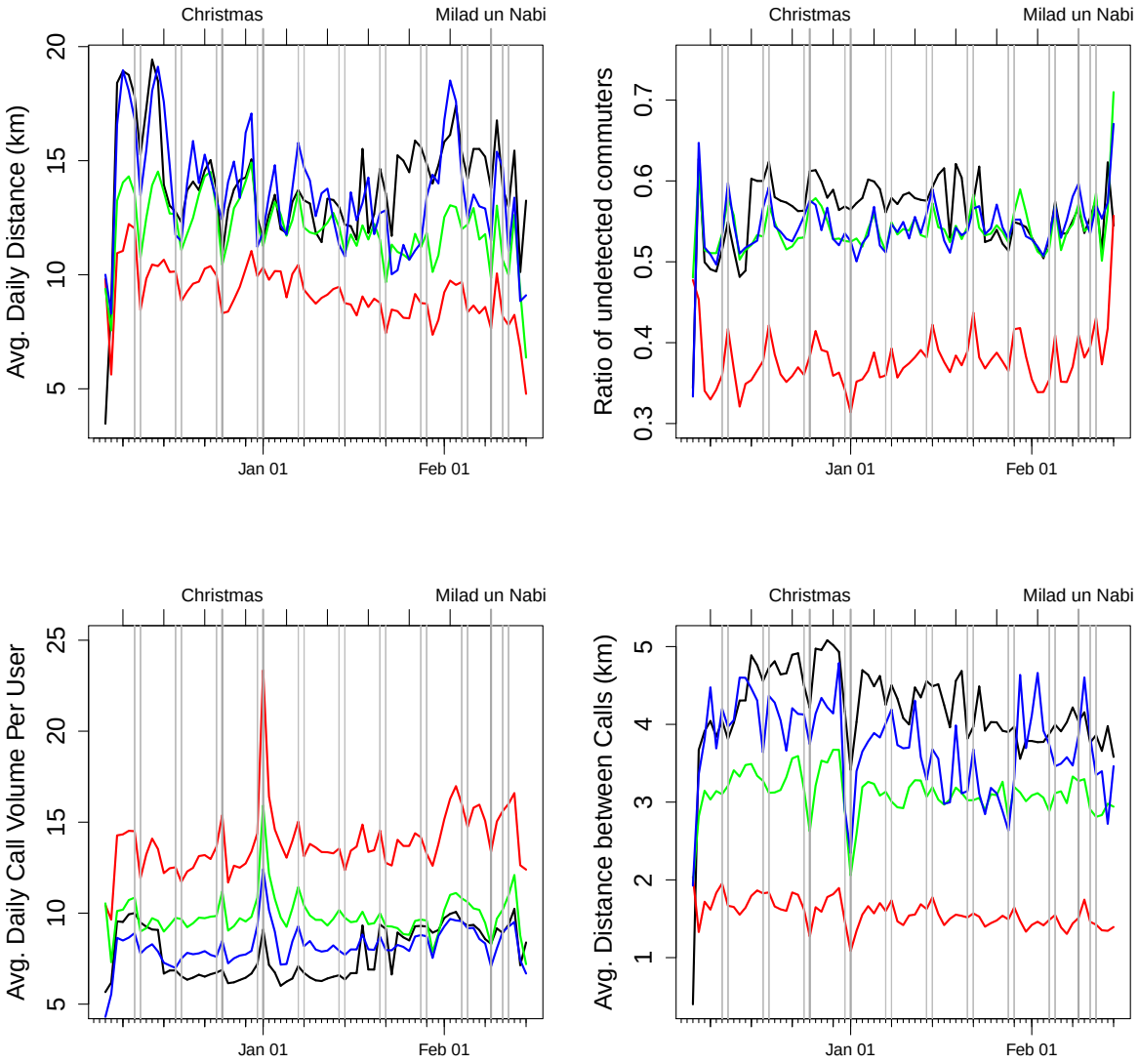


Fig. 5. This figure shows the historical time series of the daily average distance (top left), the ratio of undetected commuters (top right), the average daily call volume per user (bottom left) and the average distance between calls (bottom right) over the period of December the 5th, 2011 and February the 15th, 2012. The start and end of weekends are shown in the time series plot as two vertical gray lines. The bottom axis shows the monthly, weekly and daily markers, and the top axis shows the elapsed weeks since the date of the first call in the entire dataset (every other marker indicates the start of each two-week period). Key: Cluster 1 (black), Cluster 2 (red), Cluster 3 (green) and Cluster 4 (blue).

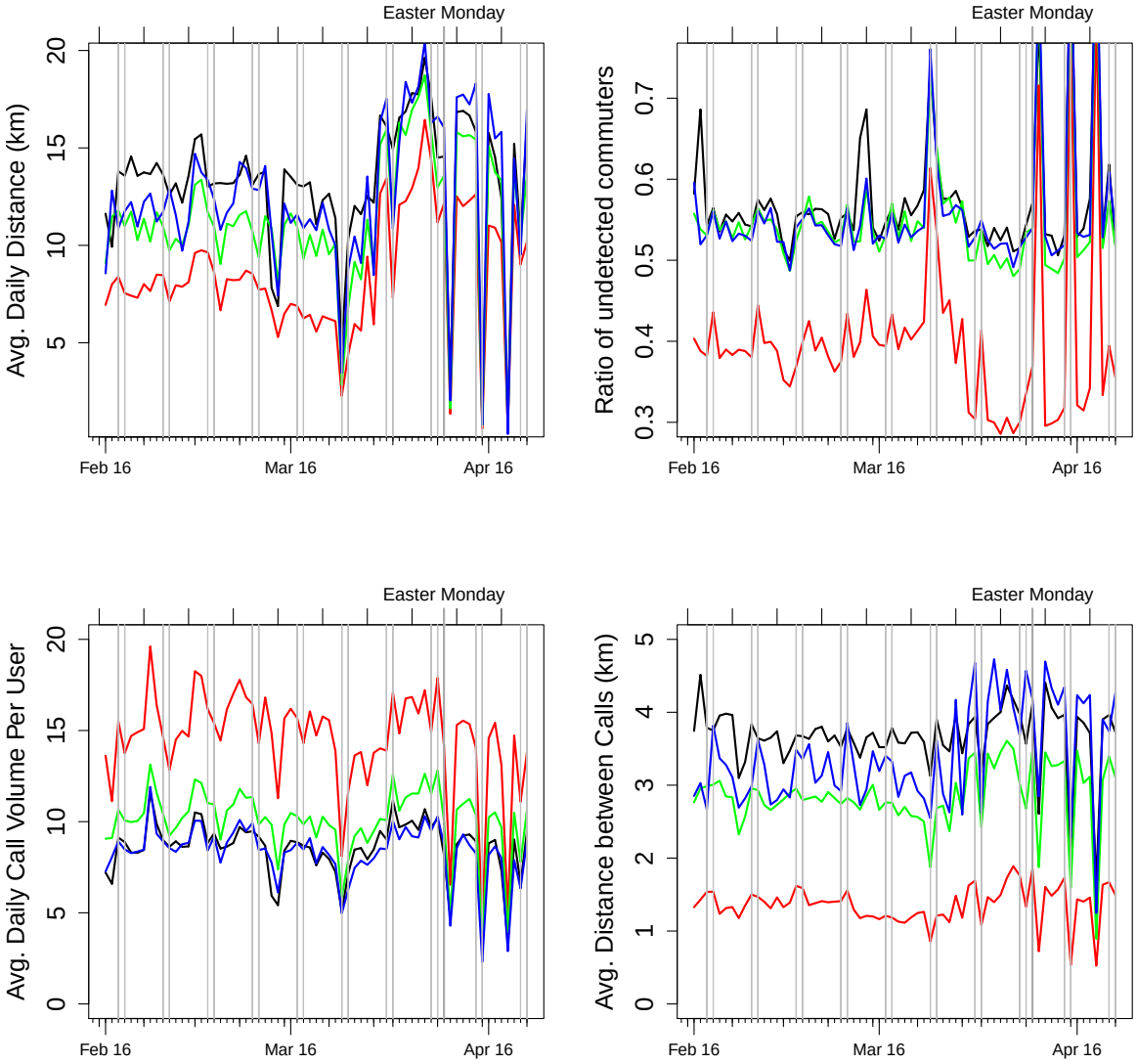


Fig. 6. This figure shows the historical time series of the daily average distance (top left), the ratio of undetected commuters (top right), the average daily call volume per user (bottom left) and the average distance between calls (bottom right) over the period of February the 16th, 2012 through to April the 22nd, 2012. The start and end of weekends are shown in the time series plot as two vertical gray lines. The bottom axis shows the monthly, weekly and daily markers, and the top axis shows the elapsed weeks since the date of the first call in the entire dataset (every other marker indicates the start of each two-week period). Key: Cluster 1 (black), Cluster 2 (red), Cluster 3 (green) and Cluster 4 (blue).

In contrast, average call volumes are found to be negatively correlated with the ratio of undetected commuters with $\rho = (-0.71, -0.69, -0.77, -0.65)$ for each cluster.

We further note in passing that users in Cluster 2 (dense urban areas) make more calls per day, and one might speculate that this is a function of better mobile phone network coverage perhaps. The challenge that we thus encounter in interpreting whether users, in general, actually travel less at weekends is obscured by the drop in call volumes at weekends and a corresponding increase in the ratio of undetected commuters. In other words, if users actually travel less at weekends, as the average distance time series suggests, then we would need to look to another measure in order to confirm this.

In the bottom right graph of Figures 5 and 6), we normalize the daily distance travelled with the number of calls that the user makes to yield a measure of the average distance travelled between calls. We posit that this normalized distance is a preferable indicator of daily trip distances since it attempts to counteract the effect of call volume variation on the distance estimates. This normalized distance is much lower in Cluster 2 as we would expect for a dense urban environment. There is some indication of increased average distance travelled between calls at the weekends in Cluster 2 but the effect is less pronounced compared to the average daily distances. To reiterate, the suspected causation is that the drop in average distances travelled on weekend days is due to the higher number of undetected commuters because call volumes are lower.

Holidays We do observe a notable drop in the distance per call on Christmas and New Year's day in each cluster. We note a surge in call volume on New Year's day, likely causing a drop in the ratio of undetected commuters and a misleading increase in the detected average distance travelled. We further observe a mid-week drop in average call volume across all clusters on February the 9th with coincides with 'Milad un Nabi' - the Suni celebration of the birth of Prophet Mohammed.

In the months of December 2011 and January 2012, there is evidence of regular peaked ratios at weekends (especially in Cluster 2), with the exception of the weekend before Christmas, where the peak is observed on the day after Christmas. There is also a marked drop in the Cluster 2 ratio on January 1st which is a national holiday. Other noteworthy holidays include Easter Monday when the average call volume is observed to drop in the call volume but with no significant change in the average distance between calls. We speculate that the remaining very large drops in the average volume, such as on February the 15th and April the 10th are caused by nationwide power outages since the call volume drops significantly over all clusters.

Distribution of daily distances Figure 7 shows the histogram of daily distances travelled by users over the entire twenty week period, partitioned by cluster. The histograms exclude users who are not detected as travelling on a given day, so that the distribution represents only non-zero daily commute distances. The variance of each distribution is significantly larger than the mean, an effect referred to as 'overdispersion'. Because of this property, the distribution can not be accurately described by a Poisson distribution and we fit instead a negative binomial distribution to the daily distances travelled by each user. The distribution can be expressed as a mixture of Poisson distributions with the mean distributed as a gamma distribution with scale parameter $(1 - prob)/prob$ and shape parameter *size*. The fitted negative binomial distribution together with the parameters *size* and *prob* are shown in each plot and observed to vary between cluster. Consistent with the increased antenna density, we observe that Cluster 2 exhibits a higher proportion of daily distances in the semi-open interval $(0, 10]km$ and thus the fitted density function shows a sharper decay rate with increasing distance than for the other clusters.

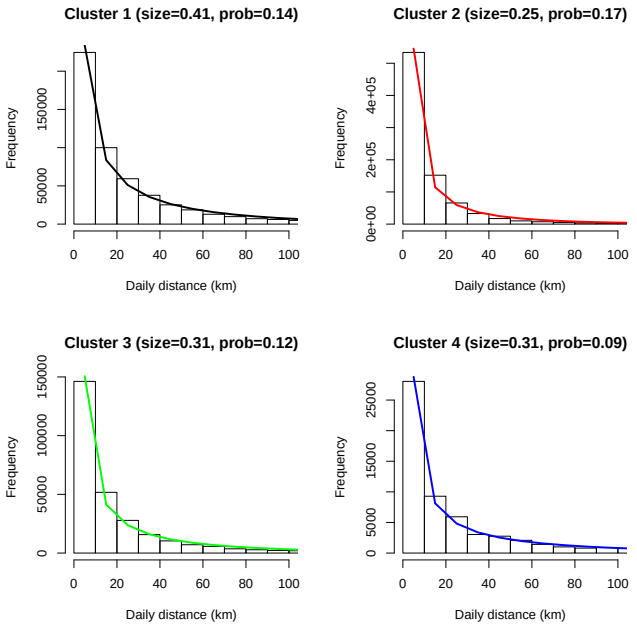


Fig. 7. Histograms of daily distances together with the fitted negative binomial distributions of daily commute distances over the entire twenty week period, partitioned by cluster.

5 Route Segment Detection

This section partially addresses the problem of how to detect that a mobile phone user is travelling on a particular segment of transport infrastructure such as a railroad or road based on the call data provided in SET2. By detecting that a mobile phone user is travelling on a particular segment of transport infrastructure, it is then possible to more accurately view their mobility traces, estimate their daily commute distances and determine the origins and destinations of their trips. Such insight may be useful in infrastructure extension projects such as new sections of railroads and new roads.

The proposed algorithm for detecting whether a user's trajectory interacts with a route segment in a particular interval consists of the following two steps:

Step 1: Associate antennae with route segments Identify the set of antennae whose minimum Euclidean distance is within a threshold of the target route segment.

Assume that a route segment S_j is a curve in a two dimensional plane. Let P_j denote the set of antenna locations $p_i \in \mathbb{R}^2$ whose minimum Euclidean distance with the j^{th} route segment S_j is within C_0 units of distance so that

$$P_j := \{p_i \ : \ \text{dist}(p_i, S_j) \leq C_0\}.$$

Denote the corresponding set of antenna ids as $A_j := (a_1, a_2, \ldots, a_m)$, where m denotes the cardinality of the set P_j. Define the point of intersection with the shortest path from p_i and the curve S_j as x_{ij}, which for ease of exposition, we shall assume is unique. Adopting a convention for determining which end of the curve is the start and end point based on its latitude and longitude, we may denote the normalized distance along the curve x_{ij} lies from the starting end point as $d_{ij} \in [0,1]$. Further denote the corresponding set of m normalized distances as D_j.

Step 2: Detect trip interaction with route segment Denote the chronologically ordered sequence of antennae id which a user calls on a given day as $A' := (a'_0, a'_1, \ldots, a'_N)$ with corresponding times $T := (t_0, t_1, \ldots, t_N)$

A user is identified as travelling on a route segment S_k at any period within the time interval $\tau := [t_1, t_2]$ if there exists at least one pair of antennae (a'_i, a'_j) where $a'_i \neq a'_j$, $t_i, t_j \in \tau$, $t_i < t_j$, $a'_i, a'_j \in A_k$, $|d_{i'k} - d_{j'k}| > \varepsilon$ and i', j' denote the local indices of a'_i, a'_j in A_k.

This algorithm is parameterized by two constants C_0 and $\varepsilon << C_0$. We demonstrate this algorithm by detecting which users travel on a segment of the main railroad which runs north-south from Adbijan. Figure 8 shows the detection of antennae within $C_0 = 10$km of the rail line. The `ST_Distance` function provided in postGIS is used to determine the set of normalized distances D_j.

We then apply Step 2 to detect which users' trajectories interact with the railroad at any point over the day using the minimum normalized threshold distance $\varepsilon = 0.1$. The sign of $d_{i'k} - d_{j'k}$, where $|d_{i'k} - d_{j'k}| > \varepsilon$, determines which direction the user is travelling. Figures 9 and 10 show the entire trip of all users who are travelling northbound or southbound on the railroad respectively.

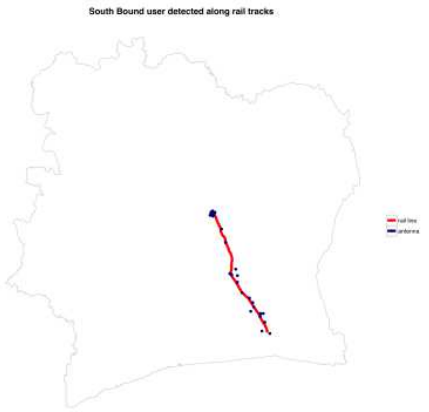


Fig. 8. This figure shows the antennae which have been detected as being located less than 10km from the illustrated railroad segment.

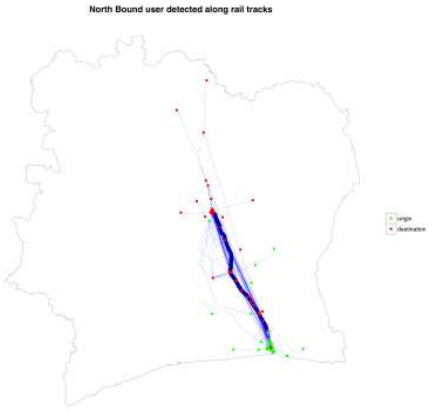


Fig. 9. This figure shows the mobility traces of all users who travel north on the segment of the railroad on a particular day.

6 Web-based Individual Trajectory Visualization Tool

In this section, we outline a visualization tool which has been developed for the purpose of enabling the community to interact with SET2 and visualize the mobility traces in combination with GIS data. The user is able specify date ranges, user IDs, and toggle topographical and infrastructural elements. Our tool dynamically loads the D4D data in real time, and allows granular playback of call events, exposing commute and travel patterns.

There are three ways in which we've made the data more accessible for user interaction. First, we allow for several static toggles that display topographic and infrastructural information for roads, railways, water areas, rivers and antenna locations. Second, we enable the tool user to choose between a more narrow study of particular individual customer trajectories or a more exploratory study of customer samples. By inputting a user ID and Week ID, a user's mobility trace for the 2 week period is plotted on the map. For more open-ended data exploration, we've included the ability to selectively plot all data for a given day or randomly plot customer data for a particular or random date. Lastly, we allow for trajectories to be explored by cluster using the k-medoids algorithm described in Section 4. For a chosen cluster and two-week time period, we plot all of the customer trajectories associated with the cluster and allow for toggling of trajectories by cluster. The tool can be accessed at http://67.202.19.246/proj/finw.html by requesting login credentials from the corresponding author. Further details of the analytics infrastructure are included in Section B.

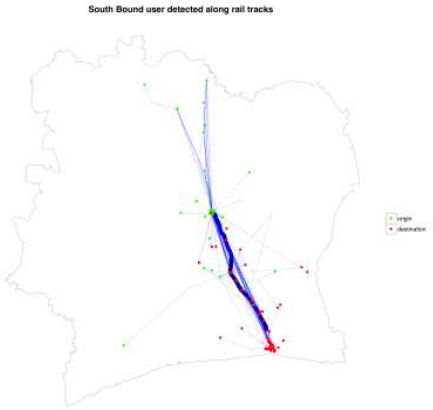


Fig. 10. This figure shows the mobility traces of all users who travel south on the segment of the railroad on a particular day.

7 Conclusion

This paper has identified a number of challenges in detecting mobility patterns from individual trajectories defined by antenna locations across the Ivory Coast. One such challenge is the highly variable antenna density which somewhat opaquely impairs the estimate of daily commute distances. Our approach partially addresses this issue by using a k-medoids algorithm to attempt to cluster the antennae into a small number of artificial geographic regions in which the antennae are more evenly distributed. By further associating user cohorts with each cluster, we are able to study the temporal and distributional characteristics of daily mobility traces per cohort.

We find within each cluster that average daily call volumes are anti-correlated with the number of detected commuters and, in turn, correlated with the average daily distance measured by mobility traces. Because of a suspected undesirable causational effect between call volume and distance measured, we measure the average distance travelled between calls and find evidence of changes in mobility patterns around national holidays and religious events. We also find possible evidence of power-outages through very large decreases in daily call volumes which yield spurious mobility traces on these days.

Another central challenge is how to detect whether a mobile phone user is travelling on a particular segment of transport infrastructure such as a road or railroad. Detection of routes travelled along transport infrastructure not only leads to improved trip distance estimates but ultimately serves to inform transport planners about the origin and destination of users who use such a segment. We demonstrate a simple methodology for transport segment detection which uses postGIS and postgres applied to the D4D datasets. We further introduce a preliminary cloud-based GIS tool for visualizing user trajectories and it is the subject of future work to enable the visualization tool to fully automate route detection.

Acknowledgements

We are grateful to Professor Terence Parr for setting up the Amazon EC2 account and Deron Aucoin for his contributions to the earlier stages of this project.

A Data Manipulation in Postgres

Data Aggregation: Data was aggregated over SET1 and SET 2 to provide daily summaries per user and per antenna respectively. An example of how the daily trip distances for each user were aggregated is provided below.

Distance Travelled Calculation: A daily user aggregate view ("mobility trace") of SET2 was created. This view has a column with the aggregated list of time and antenna position aggregated for each day.

Example

UserID: 6836

Connection Date: $2011-12-16$

Aggregated Antenna List: $06:51-619, 06:55-596, 08:45-596, 13:21-619, 16:54-619$

Computed Distance: $26.175km$

B Analytics infrastructure

Using an EC2 instance in the Amazon cloud, we built a server-side data analytics stack consisting of Apache 2.4.3, postGres 9.1 and PHP 5.4.11. A combination of html and d3.js is used to create a web-based GIS visualization tool, which is run and tested in the Chrome browser (version 24.0.1312.57 m).

The visualization tool provides user-selected parameters to a php script, which connects to and queries the postGres database. The query returns rows consisting of user IDs, the connection date-time, and the user's longitude-latitude pair. This php script packages the query results into an array to be exported for processing in javascript.

Figures and animations are created in D3js (data-driven documents), where geographic coordinates are projected into SVG data. D3 requires the data to be in geoJSON, which is a JSON-style formatting that allows for Point features (an array with a longitude-latitude pair) and LineString features (an array of longitude-latitude pairs) among others. We reformat the raw longiture-latitude pairs for each user ID into the geoJSON format and combine all of the users together into a single feature collection.

Within the javascript, we iterate through the user IDs and create a feature collection for each date requested by the user. We store each user ID in a javascript object, and we hash the user ID for later reference. Each feature collection consists of a feature with either a LineString or Point Geometry type, as specified by the geoJSON format (see http://www.geojson.org/geojson-spec.html). For each new user ID we encounter, we create the geoJSON point feature, add it to the feature collection and store its position in the feature collection in the hashmap. If we encounter the same user ID again, we access the feature collection using the ID as the

hash key, convert the point feature to a LineString feature and add the new longitude-latitude pairs in the LineString array.

References

[1] L. Akoglu and C. Faloutsos, *Event detection in time series of mobile communication graphs*, in Proc. of Army Science Conference, 2010, pp. 1–8.

[2] D. Berry, *The computational turn: Thinking about the digital humanities*, Culture Machine (12), 2011, pp. 1–22.

[3] V. Blondel, M. Esch, C. Chan, F. Clerot, P. Deville, E. Huens, F. Morlot, Z. Smoreda and C. Ziemlicki, *Data for Development: The D4D Orange Challenge on Mobile Phone Data*, 2012, pp. 1–10.

[4] J.E. Blumenstock, D. Gillick and N. Eagle. *Whose Calling? Demographics of Mobile Phone Use in Rwanda*, Association for the Advancement of Artificial Intelligence, 2010, pp. 1–2.

[5] A.L. Dabalen and S. Paul, *Estimating the Causal Effects of Conflict on Education in the Ivory Coast*, Policy Research Working Paper, 2012, pp. 1–31.

[6] Diva-GIS, *http://www.diva-gis.org/gdata*

[7] Global Administration Areas, *http://www.gadm.org/*.

[8] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, 3rd ed., 2006.

[9] P. Wang, T. Hunter, A. M. Bayen, K. Schechtner and M.C. Gonzalez, *Understanding Road Usage Patterns in Urban Areas*, Sci. Rep. (2), No. 1001, 2012.

Predicting Human Mobility Patterns in Cities

Xiao-Yong Yan, Chen Zhao, and Wenxu Wang*
Department of Systems Science, Beijing Normal University, Beijing 100875, P.R. China

Despite the importance of predicting population movements in cities, we continue to lack a high accurate and low data requirement approach to address this issue. Here, we present a conduction model without free parameter to capture the underlying mechanism that determines human mobility patterns at the city level. We use various mobile data collected from four cities with different sizes to validate the predicting ability of our model. We find that insofar as the spatial distribution of populations is available, the observed mobility patterns, including distance distribution, destination travel constraints and flux, can be reproduced precisely. Our model has potential applications to many fields relevant to mobile behavior in cities, especially in the absence of previous mobility measurements.

Predicting human mobility among different locations in a city or a country is important since it is not only a fundamental problem in geography and spatial economics science [1] but also has many practical applications in urban planning [2], traffic engineering [3, 4], epidemiology of infectious disease [5–7] and location-based service [8]. Since 1940s, many trip distribution models [9–15] have been proposed for solving this problem, among them the best-known is the gravity model [10]. Despite being widely used to predict mobility patterns at different spatial scales [16–19], the gravity model relies on special parameters estimated from systematic traffic data. Once lacking previous mobility measurements, the gravity model is no longer valid. Similar limitation exists in all trip distribution models containing context-specific parameters, such as intervening opportunity model [9], random utility model [11] and so on.

Recently, a parameter-free radiation model [12] has been proposed, providing a new insight into the long history of modeling population movements. The radiation model has a solid theoretical foundation and can successfully describe observed mobility patterns ranging from long-term migrations to inter-county commutes. Particularly, the model does not rely on adjustable parameters and only need as input the spatial distribution of the population. However, the radiation model has been validated being inaccurate in predicting human mobility at the city scale [20, 21]. Cities are the main sources of communicable disease, traffic congestion and environment pollution [6, 22] partly resulting from the population mobility. The important means to reduce these disadvantages are planning more efficient transportation systems and optimizing traffic management strategies, which all depend on our ability to predict the human travel patterns in cities [23]. Yet we continue to lack a high accurate but low data requirements approach for predicting city mobility patterns. We argue that this is mainly attributed to the underestimate of relatively high mobility of people in cities as compared to larger scales, such as travelling among counties. Within cities, especially metropolis, high development of traffic systems considerably facilitate travel of citizens to locations with more opportunities and attractiveness, regardless of time and economical costs. In this sense, the models that are quite successful in capturing mobile patterns at large spatial scales, e.g., the radiation model, are generally unsuitable to be used at the city scale. The key to solve this problem lies in uncovering underlying mechanisms that play dominant roles in intra-city travelling and differ from larger spatial scales.

In this paper, we present a conduction model in the absence of free-parameter to predict human mobility patterns in a variety of cities. We find that such mobile behavior has underlying similarity with heat conduction, so that can be modeled by a heat-conduction-like process. In particular, the model is based on the stochastic decision making process of the individual's destination selection. Before an individual selects a location to travel, s/he will weight the benefit of each location's opportunities, and the more opportunities a location has, the higher benefit it offers and the higher chance of it being chosen. The number of a location's opportunities is difficult to be straightforward measured but can be usually reflected by its population. Insofar as the population distribution is available, it is reasonable to assume that the number of opportunities of a location is proportional to its populations. Different from the radiation model's assumption that individuals tend to select the nearest locations with the largest benefit, we enlarge the possible chosen area of individuals to be the whole city, due to the relatively high mobility of people at the city scale. Despite the relatively high mobility, the possibility of travel still decays as the distance between origin and destination increases. This is due to the fact that on average there are more opportunities associated with populations along a longer travel trajectory, such that an individual is of higher probability being trapped before arriving at the presumed destination. In other words, the attraction of a location is naturally decreased along with the increment of distance. This scenario is somewhat analogous to the attenuation of temperature in heat conduction. In this point of view, we present a conduction model (see details in **Materials and Methods, Section A**) and calculate the attraction of a location i derived from the location j with actual opportunities o_j is $o_j^{'} = o_j(\frac{1}{S_{ji}} - \frac{1}{M})$, where S_{ji} is the total population in the circle of radius r_{ij} centred at location j (including the origin i and destination j), M is the total number of population in the city. Using a similar analytical method to

*wenxuwang@bnu.edu.cn

the one of the radiation model (see details in **Materials and Methods, Section A**), we get the travels from location i to location j is

$$T_{ij} = T_i \frac{m_j(\frac{1}{S_{ji}} - \frac{1}{M})}{\sum_{k\neq i}^{N} m_k(\frac{1}{S_{ki}} - \frac{1}{M})}, \quad (1)$$

where T_i is the travels departed from location i, m_j is the number of population of location j and N is the number of locations in the city. We will next use real travel data to validate the prediction ability of the model.

RESULTS

We used human daily travel data collected by mobile phone, GPS and traditional household surveys from four cities (see details in **Materials and Methods, Section B** and **C**) to test the prediction ability of the conduction model, as well as compare with the performances of two well-accepted mobility models. One is the radiation model [12] with a form

$$T_{ij} = T_i \frac{m_i m_j}{(m_i + s_{ij})(m_i + m_j + s_{ij})}, \quad (2)$$

where s_{ij} is the total population in the circle of radius r_{ij} centred at location i (excluding the origin i and destination j) and other variables have the same meaning as in Eq. (1). Another is the gravity model [10], which originates from an analogy with Newton's gravity law and has many modified versions so far. Here we employ the origin-constrained gravity model [3] for reference to predict the city mobility patterns, which has a form

$$T_{ij} = T_i \frac{m_j f(r_{ij})}{\sum_{k\neq i}^{N} m_k f(r_{ik})}, \quad (3)$$

where $f(r_{ij})$ is a function of the distance r_{ij} and other variables have same meaning as in Eq. (1). The distance function $f(r_{ij})$ can take different forms, such as power [21], exponential [15], truncated power [18] or Hill function [19]. By experiments we find the gravity model with power function $f(r_{ij}) = r_{ij}^{-\gamma}$ can describe the cities' mobility patterns well, so we use this function and estimate its parameter γ with real travel data of four cities (see details in **Materials and Methods, Section D**).

We first investigate the travel distance distribution produced by the models based on the real data. Travel distance distribution is an important statistical property to characterize the human mobility behavior [24–27] and reflect the economic efficiency of a city [1]. We find that, as shown in Fig. 1, the conduction model and the gravity model can both reproduce the observed distributions of travel distance well. Although in some cases (Fig. 1a and Fig. 1b) the performances of the gravity model are slightly better than that of the conduction model, consider the parameter-free nature of the conduction model, the superiority of the gravity model thus becomes negligible. We also find that the radiation model significantly underestimates the long-distance (longer than about 2 km) travel

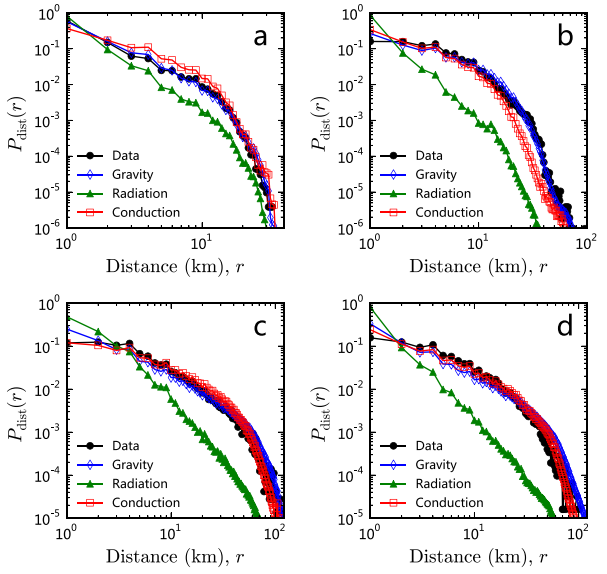


FIG. 1. **Comparing the travel distance distributions generated by different models**. **a**, Abidjan. **b**, Beijing. **c**, Chicago. **d**, Seattle. $P_{\rm dist}(r)$ is the probability of a travel between locations at distance r.

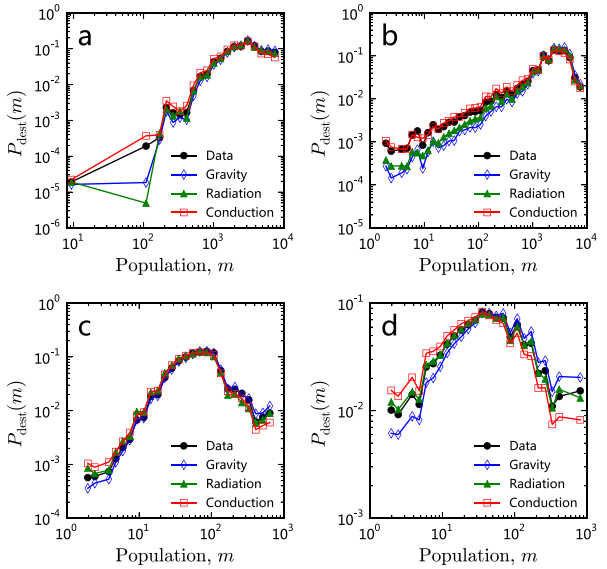


FIG. 2. **Comparing the destination travel constraints of different models**. **a**, Abidjan. **b**, Beijing. **c**, Chicago. **d**, Seattle. $P_{\rm dest}(m)$ is the probability of a travel towards a location with population m.

in all cases. This is due to the fact that the assumption of radiation model does not consider relatively higher mobility that allows for selecting better location, going beyond nearest location at the city level compared to country scale. The fact that the gravity model and the conduction model can both reproduce the real travel distance distributions suggests that the two

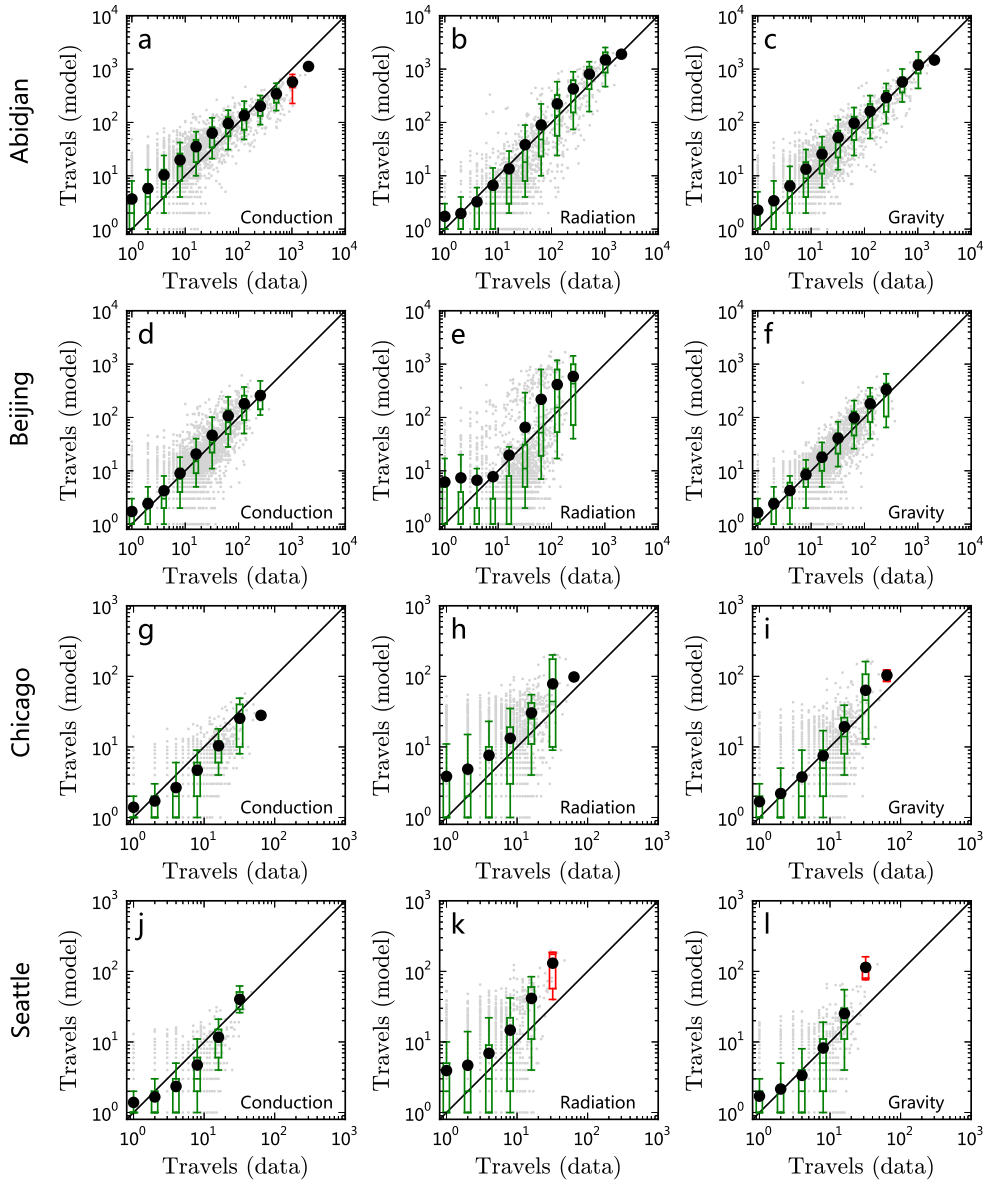


FIG. 3. **Comparing the observed fluxes with the predicted fluxes**. The grey points are scatter plot for each pair of locations. The black points represent the average number of predicted travels in the different bins. The boxplots (D1, Q1, Q2, Q3 and D9) represent the distribution of the number of predicted travels in different bins of number of observed travels. A box is colored green if the line $y = x$ lies between D1 and D9 in that bin and is red otherwise.

models' basic assumptions are reasonable under the scale of city. However, the gravity model uses a distance-decay function with parameters to match the real data and our model can generate appropriate results only using the population distribution data.

We further calculate the probability of a travel towards a location with population m, $P_{\rm dest}(m)$, from both observed data and the models. $P_{\rm dest}(m)$ is a key measure to check the accuracy of the origin-constrained mobility models (the radiation model, conduction model, and gravity model used here are all origin-constrained) because that the origin-constrained model can not guarantee the modeling travels to a location equal to real travels arrived at that location [3]. From Fig. 2 we can see that our model agree with the real data better than or equal to the gravity model and in some cases (Fig. 2a and Fig. 2b) it outperforms the radiation model.

A more straightforward measure of models' prediction ability is to compare the travel fluxes for all pairs of locations generated by the models with the fluxes observed in real data [12], as shown in Fig. 3. We find that, except in the case of

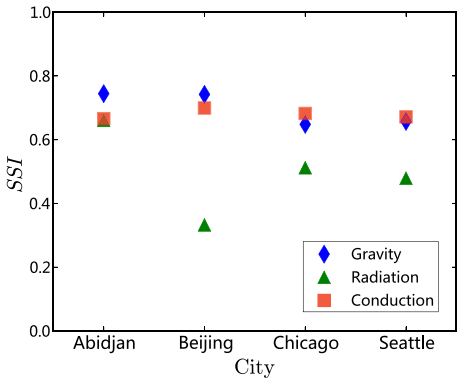


FIG. 4. **Comparing the prediction ability of deferent models based on Sørensen similarity index (SSI)**.

Abidjan, the average fluxes predicted by the radiation model significantly deviate from the real fluxes; in most cases, the prediction results of the gravity model agree with the real fluxes well; the conduction model's predictions are slightly worse than the gravity model's, but not far away from the real data. It is noticed that the boxplot used in Fig. 3 is not, more or less, an appropriate approach to distinguish the difference of the predictions between the models. For example, in Fig. 3e, the radiation model's results significantly deviate from the real data but the boxes still be colored with green. To make a more clear comparison between the models' predictions, we further use the Sørensen similarity index [28] to measure the degree of agreement between the travel matrices predicted and observed (see details in **Materials and Methods, Section E**). The result is plotted in Fig. 4, from which we can see that in most cases the conduction model outperforms the radiation model and has similar performances with the gravity model.

CONCLUSIONS AND DISCUSSION

We developed a parameter-free conduction model to reproduce and predict mobile behavior in cities with different sizes, economic levels and cultural backgrounds. Our results are in good agreement with real data with respect to travel distance distribution, destination travel constraints and flux, suggesting that the model captures the fundamental mechanisms governing the human daily travel behaviors at the city scale. Although the gravity model can yield similar prediction accuracy, it needs parameters estimated from previous mobility measurements. In contrast, the conduction model only require the population distribution as input data so can be used in the absence of systematic traffic data.

The radiation model, despite also being parameter-free and performing well at the large scale, underperforms at the city scale. The problem lies in the underestimation of relatively high mobility at the city scale due to the well-developed public traffic systems in cities. In particular, the radiation model assumes that due to the restrain of mobility, people tend to select nearest location rather than a further location with more opportunities to improve their circumstances. The conduction model can successfully overcome this problem by casting the mobile behavior in cities into the framework of heat-conduction-process. Insofar as only population distribution is available, our model can offer the best prediction of mobile patterns at the city scale, as an alternative to the radiation model that is available at the inter-city scale.

It is noteworthy that even though the conduction model can provide appropriate predictions for city mobility patterns, its accuracy still remain at a relatively low level. The travel matrices generated by the model have about 70% common part with the real data (see Fig. 4). Although such accuracy can suffice the requirements in many areas of application *e.g.* urban planing and epidemic modelling [29], it is far away from the reported average upper limit of predictability in human mobility [23]. We deem the main reason is that the conduction model is essentially a kind of aggregate travel models [3] which are based on the behaviors of the groups of travellers, or the "average individual". While the real individuals' behaviors are quite different [27] and difficult to be characterized by the aggregate models. We believe a mobility prediction model taking into account the diversity of individual travel behaviors would be worth of pursuing in the future research.

MATERIALS AND METHODS

A. Derivation of the conduction model

We use the heat conduction process to simulate the decay of the available opportunities of a location i derived from a location j with actual opportunities o_j. Because the number of a location's actual opportunities is proportional to its population, we assume at initial time the heat quantity Q_j (*i.e.* the actual opportunities o_j) of a location j is $m_j c\Theta_0$, where m_j is the number of population of location j and can be understood as the mass of the location, c is the specific heat capacity and Θ_0 is the initial temperature of j. Then we rank all other $N-1$ locations based on their distances to location j and number the closest location as 1, the second closest location as 2, *etc.* and assume at initial time the heat quantities of these locations are all zero. According to the heat conduction process, the heat quantity conducted from j to its closest location is

$$Q_1 = m_1 c\Theta_1 = m_j c(\Theta_0 - \Theta_1), \tag{4}$$

where Θ_1 is the temperature of location 1, which is equal to

$$\Theta_1 = \frac{m_j}{m_j + m_1}\Theta_0. \tag{5}$$

Similarly, the heat quantity conducted to the second closest location is

$$Q_2 = m_2 c\Theta_2 = (m_j + m_1)c(\Theta_1 - \Theta_2) \tag{6}$$

and the temperature is

$$\Theta_2 = \frac{m_j + m_1}{m_j + m_1 + m_2}\Theta_1 = \frac{m_j}{m_j + m_1 + m_2}\Theta_0. \tag{7}$$

Recursively, we can get heat quantity conducted to the ith location is

$$Q_i = \frac{m_i m_j c\Theta_0}{S_{ji}}, \tag{8}$$

where S_{ji} is the total mass (population) in the circle of radius r_{ij} centred at location j (including the location j and location i). For each individual of location i, the heat acquired from location j is

$$q_i = \frac{Q_i}{m_i} = \frac{m_j c\Theta_0}{S_{ji}}. \tag{9}$$

In addition, when reaching thermal equilibrium, the heat conducted to each individual have same quantity

$$q^* = \frac{m_j c\Theta_0}{M}. \tag{10}$$

Therefore, for each individual of location i the real difference of heat quantity conducted from location j, or the number of available opportunities of location j, is

$$o_j^{'} = q_i - q^* = m_j c\Theta_0(\frac{1}{S_{ji}} - \frac{1}{M}). \tag{11}$$

We next use a similar analytical framework presented in ref. [12] to derive the expression of our mobility model. We assume the benefit of an available opportunity with a single number, z, randomly chosen from distribution $p(z)$. Thus, each location with available opportunities $o^{'}$ is assigned $o^{'}$ random numbers. Because we assume an individual selects the location having the largest benefit to travel, the probability that an individual in location i selects location j to travel is

$$p_{ij} = \int_0^\infty \mathrm{d}z P_{o_j^{'}}(z) P_{\sum_{k\neq i,k\neq j}^N o_k^{'}}(<z), \tag{12}$$

where $P_{o_j^{'}(z)}$ is the probability that the maximum value extracted from $p(z)$ after $o_j^{'}$ trials is equal to z, $P_{\sum_{k\neq i,k\neq j}^N o_k^{'}}(<z)$ is the probability that $\sum_{k\neq i,k\neq j}^N o_k^{'}$ numbers extracted from the $p(z)$ distribution are all less than z. Because that

$$P_{o_j^{'}}(z) = \frac{\mathrm{d}P_{o_j^{'}}(<z)}{\mathrm{d}z} = o_j^{'} p(<z)^{o_j^{'}-1}\frac{\mathrm{d}p(<z)}{\mathrm{d}z} \tag{13}$$

and

$$P_{\sum_{k\neq i,k\neq j}^N o_k^{'}}(<z) = p(<z)^{\sum_{k\neq i,k\neq j}^N o_k^{'}}, \tag{14}$$

we can get

$$p_{ij} = o_j^{'}\int_0^\infty \mathrm{d}z \frac{\mathrm{d}p(<z)}{\mathrm{d}z} p(<z)^{\sum_{k\neq i}^N o_k^{'}-1} = \frac{o_j^{'}}{\sum_{k\neq i}^N o_k^{'}}.$$

Combine Eq. 15 and Eq. 11 and give the number of travels departed from location i, T_i, we can obtain the average number of travels from location i to j is

$$T_{ij} = T_i p_{ij} = T_i \frac{m_j(\frac{1}{S_{ji}} - \frac{1}{M})}{\sum_{k\neq i}^N m_k(\frac{1}{S_{ki}} - \frac{1}{M})}. \tag{15}$$

B. Data descriptions

1. *Abidjan mobile phone dataset*

The dataset contains 607,167 mobile phone uses' movements between 381 cell phone antennas in Abidjan, the biggest city of Ivory Coast, during a two-week observation period. Each movement record contains two endpoints' coordinates (longitude and latitude). The dataset is based on anonymized Call Detail Records (CDR) of phone calls and SMS exchanges between five million of Orange's customers in Ivory Coast. To protect the customers' privacy, the customer identifiers have been anonymized by Orange Company.

2. *Beijing taxi passengers dataset*

The dataset contains passengers tracks of more than 10,000 taxis in Beijing during an one-week observation period [21, 30]. When a passenger get on or get off a taxi, the coordinates and time are recorded automatically by a GPS-based device installed in the taxi. From the dataset we extract 1,070,198 taxi passengers travel records.

3. *Chicago travel tracker survey dataset*

Chicago travel tracker survey was conducted by Chicago Metropolitan Agency for Planning during 2007 and 2008, which provides a detailed travel inventory for each member of 10,552 household in the greater Chicago area. The survey data are available online at http://www.cmap.illinois.gov/travel-tracker-survey/. Because some participants provided one-day travel records but others provided two-days, to maintain consistency, we only extract the first-day travel records from the dataset. The extracted data include 87,041 trips, each of which includes coordinates of the trip's origin and destination.

4. *Seattle household activity survey dataset*

The dataset of Seattle household activity survey conducted by Puget Sound Regional Council in 2006 are available online at http://www.psrc.org/data/surveys/. This Dataset includes basic demographics, activities and travel characteristics collected from every member of 4,746 households in Seattle metropolitan area during a consecutive 48-hour travel period. We extract 64,570 trips with the coordinates of their origin and destination from the dataset.

C. Data preprocessing

The raw travel data of the four cities are all based on disperse endpoints with latitude and longitude coordinates. However, the mobility prediction models need inputting data based on zones (namely *locations* in the main text) [3]. Because the

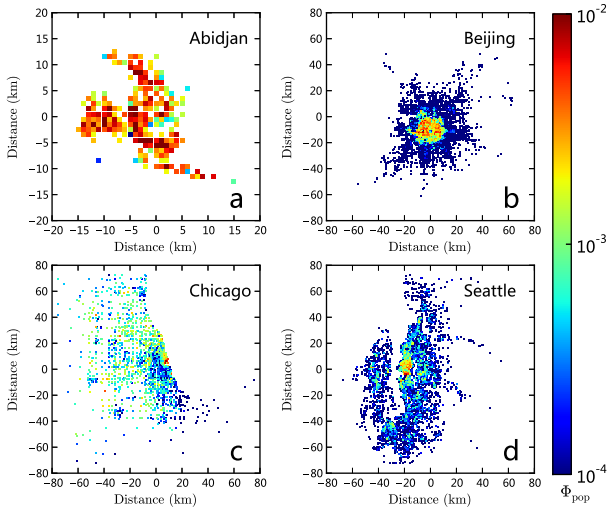


FIG. 5. **The zone division and population density distribution of four cities**. **a**, Abidjan, 219 zones. **b**, Beijing, 3,025 zones. **c**, Chicago, 1,406 zones. **d**, Seattle, 3,175 zones. The density function $\Phi_{pop}(i)$ represents the probability of finding a travel started from zone i.

absence of predefined zoning system (like states or counties in country-wide) in city, we must divide the city into several zones before using the data to feed mobility models. For simplicity, we divide each city into equal-area square zones with the area of 1 km^2. Fig. 5 shows the zone division results and the number of zones for four cities. We assign an origin (or destination) zone ID for each travel if the travel's departure (or arrival) endpoint falls within the range of that zone. Then we can accumulate the total number of travels departed from any zone i, T_i, and the total number of travels from zone i to zone j, T_{ij}. In general, the number of travels departed from a zone is proportional to the population of the zone [12]. Because the absence of detailed data on the population spatial distribution of the cities, we use the travellers data approximately represent the population density distribution, as shown in Fig. 5.

D. Estimating the Parameter of the gravity model

Before using the gravity model it is necessary to estimate their parameters. The goal of the parameter estimation is making the model reproduces the mobility patterns observed from real data as close as possible. Here we use Hyman method [31], an almost standard method for calibrating gravity model in transportation planning [3], to estimate the gravity model's parameter. Hyman method try to find an optimal parameter to minimize the the difference between the modelled average travel distance and the real average travel distance

$$E(\gamma) = \left| \frac{\sum_i \sum_j T_{ij}(\gamma) r_{ij}}{\sum_i \sum_j T_{ij}(\gamma)} - \frac{\sum_i \sum_j T_{ij} r_{ij}}{\sum_i \sum_j T_{ij}} \right|, \qquad (16)$$

where $T_{ij}(\gamma)$ is the number of travels from zone i to j generated by the gravity model with the parameter γ, T_{ij} is the real number of travels from zone i to j. It can be seen as a problem that finding a root of equation $E(\gamma) = 0$. This problem can be easily solved by many root-finding algorithms, such as secant method and bisection method [32]. The parameters having been estimated for four cities are respectively $\gamma = 2.47$ for Abidjan, $\gamma = 1.71$ for Beijing, $\gamma = 2.14$ for Chicago and $\gamma = 1.93$ for Seattle.

E. Sørensen similarity index

Sørensen similarity index is a statistic used for comparing the similarity of two samples and mainly useful for ecological community data [28]. Ref. [15] used a modified version of the index to measure if the real fluxes are correctly reproduced (on average) by the mobility prediction models, which has a form as

$$SSI = \frac{1}{N^2} \sum_i^N \sum_j^N \frac{2\min(T'_{ij}, T_{ij})}{T'_{ij} + T_{ij}}, \qquad (17)$$

where T'_{ij} is the travels from location i to j predicted by models and T_{ij} is the real travels. Obviously, if each T'_{ij} is equal to T_{ij}, the index is 1; if all T'_{ij} are far away from the real values, the index is close to 0.

[1] Barthélemy. Spatial networks. *Phys. Rep.* **499**, 1-101 (2011).
[2] Batty, M. The size, scale, and shape of cities. *Science* **319**, 769-771 (2008).
[3] Ortúzar, J. D. & Willumsen, L. G. *Modelling Transport (4th Ed.)* (Wiley, 2011).
[4] Helbing, D. Traffic and related self-driven many-particle systems. *Rev. Mod. Phys.* **73**, 1067-1141 (2001).
[5] Hufnagel, L., Brockmann, D. & Geisel, T. Forecast and control of epidemics in a globalized worlds. *Proc. Nat. Acad. Sci. USA* **101**, 15124-15129 (2004).
[6] Eubank, S. *et al*. Modelling disease outbreaks in realistic urban social networks. *Nature* **429**, 180-184 (2004).
[7] Balcan D., Colizza V., Goncalves B., Hu H., Ramasco J. R. & Vespignani A. Multiscale mobility networks and the large spreading of infectious diseases. *Proc. Nat. Acad. Sci. USA* **106**,

21484-21489 (2009).

[8] Scellato, S., Noulas, A. & Mascolo C. Exploiting place features in link prediction on location-based social networks. in. *Proc. of the ACM KDD11*, 1046-1054 (2011).

[9] Stouffer, S. A. Intervening opportunities: a theory relating mobility and distance. *Am. Sociol. Rev.* **5**, 845-867 (1940).

[10] Zipf, G. K. The P_2P_2/D hypothesis: on the intercity movement of persons. *Am. Sociol. Rev.* **11**, 677-686 (1946).

[11] Domencich, T. & McFadden, D. *Urban travel demand: a behavioural analysis* (North-Holland, Amsterdam, 1975).

[12] Simini, F., González, M. C., Maritan, A. & Barabási, A.-L. A universal model for mobility and migration patterns. *Nature* **484**, 96-100 (2012).

[13] Simini, F., Maritan, A. & Néda, Z. Continuum approach for a class of mobility models. *arXiv:1206.4350* (2012).

[14] Noulas, A., Scellato, S., Lambiotte, R., Pontil, M. & Mascolo, C. A tale of many cities: universal patterns in human urban mobility.*PLoS ONE* **7**, e37027 (2012).

[15] Lenormand, M., Huet, S., Gargiulo, F. & Deffuant, G. A universal model of commuting networks. *PLoS ONE* **7**, e45985 (2012).

[16] Jung, W. S., Wang, F. & Stanley, H. E. Gravity model in the Korean highway. *Europhys. Lett.* **81**, 48005 (2008).

[17] Krings, G., Calabrese, F., Ratti, C. & Blondel, V. D. Urban gravity: a model for inter-city telecommunication flows. *J. Stat. Mech.* **2009**, L07003 (2009).

[18] Kaluza P., Koelzsch A., Gastner M. T. & Blasius, B. The complex network of global cargo ship movements. *J. R. Soc. Interface* **7**, 1093-1103 (2010).

[19] Goh, S., Lee, K., Park, J. S. & Choi, M. Y. Modification of the gravity model and application to the metropolitan Seoul subway system. *Phys. Rev. E* **86**, 026102 (2012).

[20] Masucci, A. P., Serras, J., Johansson, A. & Batty, M. Gravity vs radiation model: on the importance of scale and heterogeneity in commuting flows. *arXiv:1206.5735*, (2012).

[21] Liang, X., Zhao, J., Li, D. & Xu, K. Modeling collective human mobility: understanding exponential law of intra-urban movement. *arXiv:1212.6331* (2012).

[22] Bettencourt, L., Lobo, J., Helbing, D., Kuhnert, C. & West, G. B. Growth, innovation, scaling, and the pace of life in cities. *Proc. Nat. Acad. Sci. USA* **104**, 7301-7306 (2007).

[23] Song, C., Qu, Z., Blumm, N. & Barabási, A.-L. Limits of predictability in human mobility. *Science* **327**, 1018-1021 (2010).

[24] Brockmann, D., Hufnagel, L. & Geisel, T. The scaling laws of human travel. *Nature* **439**, 462-465 (2006).

[25] González, M. C., Hidalgo, C. A. & Barabási, A.-L. Understanding individual human mobility patterns. *Nature* **453**, 779-782 (2008).

[26] Roth, C., Kang, S. M., Batty, M. & Barthélemy, M. Structure of urban movements: polycentric activity and entangled hierarchical flows. *PLoS ONE* **6**, e15923 (2011).

[27] Yan, X.-Y., Han, X.-P., Wang, B.-H. & Zhou, T. Diversity of individual mobility patterns. *axXiv:1211.2874* (2012).

[28] Søensen, T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on danish commons. *Biol. Skr.* **5**, 1-34 (1948).

[29] Truscott, J. & Ferguson, N. M. Evaluating the adequacy of gravity models as a description of human mobility for epidemic modelling. *PLoS Comput. Biol.* **8**, e1002699 (2012).

[30] Liang, X., Zheng, X., Lv, W., Zhu, T. & Xu, K. The scaling of human mobility by taxis is exponential. *Physica A* **391**, 2135-2144 (2012).

[31] Hyman, G. M. The calibration of trip distribution models. *Environment and Planning* **1**, 105-112 (1969).

[32] Ravindran, A., Ragsdell, K. M. & Reklaitis, G. V. *Engineering Optimization: Methods and Applications (2nd Ed.)* (Wiley, 2006).

Combining call records and road data for strategic disaster response planning

Zhuojie Huang
Department of Geography
University of Florida
seenhzj@gmail.com

Udayan Kumar*
Computer and Information Science and Engineering
University of Florida
ukumar@cise.ufl.edu

Abstract

Emergency response for disaster is critical for low-income countries. Due to its abundant water resources and a high population density situated in the coastal zone, Abidjan is subjected to high risk of torrential flooding from storms in rainy season that might impact a large number of populations. The aim of this research is to identify movement patterns of population for better disaster response planning. With the cell phone records, we identify a movement network that indicates population movement. A road network is constructed with road junctions and connections extracted by the OpenStreetMap. All these spatial information with population density are aggregated by the Thiessen polygon defined by the cell tower locations. Our analysis showed that the centre of Abidjan may suffer from congestion problem after a disaster (flooding) due to concentration of large amount of traffic from the risk area. Also, our initial analysis indicates that the movement network and the road network present significant similarities in terms of network modularity. This may help understand the population movement pattern from more readily-available road network data when movement data is not available.

1 Introduction

Sudden and unprepared nature disasters will cause displacement of large amount of population, which have profound economic and social impacts [1], [2]. It could be anticipated such disasters will be more often coupled with climate changes [3]. This is a challenge particularly for low income country with poor infrastructures and insufficient emergence response. Recent research have identified the severe population displacements after disaster in Haiti [4] and Bangladesh [2]. Cote D'ivoire is in the top 10 country list that at risk from intensification of storm surges and Abidjan is ranked twelfth in terms of the number of affected population [5]. In our research, we identified the areas in Abidjan which have a high risk subjected to flooding after storm. We studied the likelihood for displacement based on the predefined movement network before a disaster happens. Incorporated with road network, areas that in the surge zones will be identified and the spatial vulnerability of disease response will be presented.

*equal contribution

2 DATA

2.1 Spatial Information for Abidjan

Abidjan is the largest coastal city in Cote D'ivorie, currently around 7 million people live in the metropolitan area. Here we combine and integrate these background spatial information: the spatial boundary for global administrative area is obtained from the Global Administrative Areas Database (http://www.gadm.org/). The cell phone tower locations are extracted from the D4D challenge. The population information is extracted from the Afripop projects (www.afripop.org). The Cote D'ivoire digital elevation map (DEM) is obtained from the Shuttle Radar Topography Mission from NASA (http://www2.jpl.nasa.gov/srtm/) with a spatial resolution of 90 meter * 90 meters. We followed a similar approach in [6] to define a tower at risk when it has contained grid with elevation less than 10. All these information is aggregated under the Thiessen polygon created by the cell phone tower.

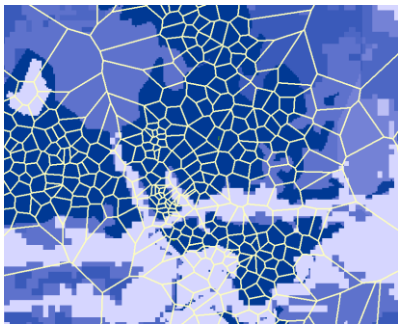
2.2 Connectivity Network based on social network and road network.

To understand connectivity both at social and infrastructure level, we utilize two data sets: 1. Movement data (Network 1) and 2. Road network data (Network 2).

Movement data is obtained from call records provided by Set 2 of D4D dataset [7]. This data set provides 10 sets of individual trajectories for 50000 random users for 2 weeks duration. Using this data we compute aggregate user movement between the towers; for e.g., if a user made/received a call at tower T1 and then made/received next call at tower T2, this user is assumed to move T1 to Tower T2. All these movement segments' counts are summed up in a connectivity matrix between towers.

The road network is obtained from a volunteer provided geographic data, OpenStreetMap (www.openstreetmap.org/). An ArcGIS extension (http://esriosmeditor.codeplex.com/) is utilized to extract the road network with junctions from OpenStreetMap. For each of Thessian polygon that a cell phone tower serves, information on junction counts and junction connectivity is aggregated for further inspection. Through these, a

weighted road network with connectivity between towers is constructed.



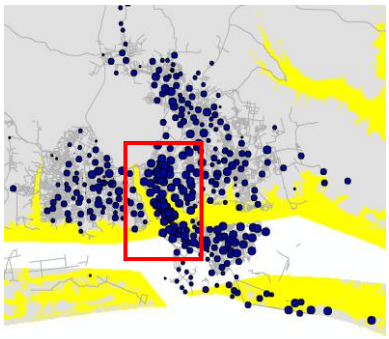
a)

b)

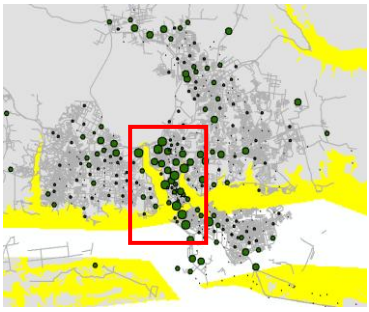
Figure 1.a) Thiessen polygon generated by the cell phone tower to extract the spatial information. b) Cell phone towers (highlighted) are at risk. Yellow represents area with an elevation less than 10 meters.

3 Result

With the integrated DEM, 292 towers (in Abidjan) are found in the risk zone with an estimated reachable population of 2.7 million. Areas where congestions may happen after flooding were inspected through two metrics in Network 1 and 2: First, from movement dataset population movements are calculated from risk areas (areas around and including Abidjan with elevation less than 10m) to other areas (shown the in-degree in Fig. 2a). A high in-degree indicates higher social connection and thus most probable places to go after a disaster. Second, betweenness centrality is calculated of the road network aggregated at tower level (Fig. 2b). Higher betweennes (denoted by a bigger circle) suggests that a larger number of shortest routes go through the roads near a tower thus higher traffic is more likely to go through these areas. Comparing these two graphs, places can be identified where congestion and road block is probable to happen thus obstruct aid delivery.



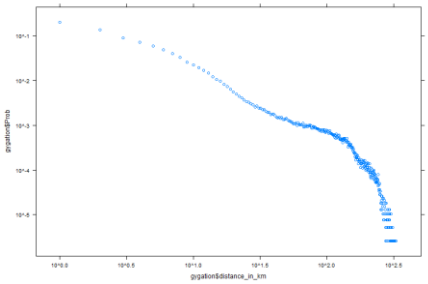
a)



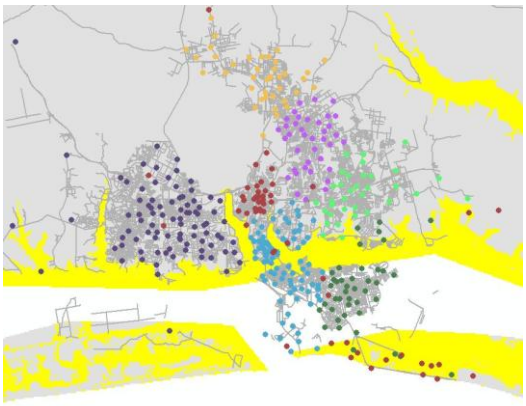
b)

Figure 2. a) Spatial distribution of "in degree" of the cell phone movement network (Network 1). b) Spatial distribution of betweenness centrality of the road network (Network 2). Larger dots represent larger value. The red rectangles represent the likely congestion areas

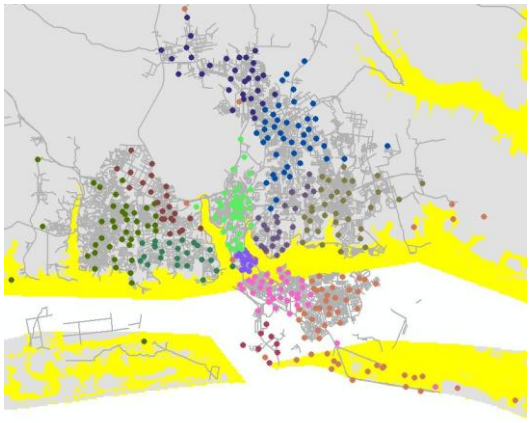
Further, we calculate the radius of gyration based on two weeks' window of the D4D data [7]. In this short period, 50% of the people in Abidjan can be observed moving in a radius of gyration of 5 kilometres (Fig. 3a). This may imply that their social connection may be limited to short distances. Next, we compare the community structure of movement network and road network utilizing a multilevel community detection method [8]. We found that community structure of both the networks is similar with the Rand index [9] 0.79, which indicates homogenous structure for two partition schema. Fig. 3b and 3c show the communities for both the network.



a)



b)



c)

Figure 3 a): Radius of gyration plots in the two weeks window. b): Movement Network generated by the cell phone data (Network 1). c): Road connectivity pattern (Network 2). Colours alike represent the same community (sub-graph). Yellow represents area with an elevation less than 10 meters.

4 Discussion

Nature disasters have adverse impacts on social and economic lifecycle. Population displacements, after a disaster, are more likely to affect people in lower income group [2]. To access the magnitude and distance population displacements, we construct complex networks with movement segments and road connectivity. Our major finding is the similarity observed between the actual movement patterns from cell phone data and the road connectivity pattern. Our results suggest that population movements are constrained but correlated with the topological structure of the road network. For area where cell phone data is not available, analysing the road connectivity network might reveal the local population movements and vice-versa. Further integration of other high risk areas as sources of population movement and the potential flows on the road network can inform the stakeholder about the areas where most likely to get affected when flooding occurs. Overall, our preliminary findings show promise in integrating spatial and network information to support decision makers in making better disaster evacuation plan.

5 Reference

[1] J. Levine, A. Esnard, and A. Sapat, “Population displacement and housing dilemmas due to catastrophic disasters,” *J Plan Lit*, 2007.

[2] C. L. Gray and V. Mueller, “Natural disasters and population mobility in Bangladesh.,” *Proc Natl Acad Sci*, vol. 109, no. 16, pp. 6000–5, Apr. 2012.

[3] M. Van Aalst, “The impacts of climate change on the risk of natural disasters,” *Disasters*, 2006.

[4] L. Bengtsson, X. Lu, A. Thorson, R. Garfield, and J. von Schreeb, “Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: a post-earthquake geospatial study in Haiti.,” *PLoS Med*, vol. 8, no. 8, p. e1001083, Aug. 2011.

[5] S. Dasgupta and B. Laplante, “Climate change and the future impacts of storm-surge disasters in developing countries,” *Center for Global Development*, vol. Working Pa, no. 182, 2009.

[6] S. Dasgupta, B. Laplante, S. Murray, and D. Wheeler, “Exposure of developing countries to sea-level rise and storm surges,” *Climatic Change*, vol. 106, no. 4, pp. 567–579, Dec. 2010.

[7] V. Blondel, M. Esch, C. Chan, and F. Clerot, “Data for Development: the D4D Challenge on Mobile Phone Data,” *CoRR*, pp. 1–10, 2012.

[8] V. Blondel and J. Guillaume, “Fast unfolding of communities in large networks,” *J Stat Mech Theor Exp*, pp. 1–12, 2008.

[9] L. Hubert and P. Arabie, “Comparing partitions,” *J Classif*, vol. 218, pp. 193–218, 1985.

Mitigating Epidemics through Mobile Micro-measures

Mohamed Kafsi, Ehsan Kazemi, Lucas Maystre, Lyudmila Yartseva,
Matthias Grossglauser, Patrick Thiran
School of Computer and Communication Sciences, EPFL
first.last@epfl.ch

Abstract. Epidemics of infectious diseases are among the largest threats to the quality of life and the economic and social well-being of developing countries. The arsenal of measures against such epidemics is well-established, but costly and insufficient to mitigate their impact. In this paper, we argue that mobile technology adds a powerful weapon to this arsenal, because (a) mobile devices endow us with the unprecedented ability to measure and model the detailed behavioral patterns of the affected population, and (b) they enable the delivery of personalized behavioral recommendations to individuals in real time. We combine these two ideas and propose several strategies to generate such recommendations from mobility patterns. The goal of each strategy is a large reduction in infections, with a small impact on the normal course of daily life. We evaluate these strategies over the Orange D4D dataset and show the benefit of mobile micro-measures, even if only a fraction of the population participates. These preliminary results demonstrate the potential of mobile technology to complement other measures like vaccination and quarantines against disease epidemics.

1 Introduction

Modeling and effectively mitigating the spread of infectious diseases has been a long-standing public health goal. The stakes are high: throughout human history, epidemics have had significant death tolls. In the mid-14$^{\text{th}}$century [15], between 30% and 50% of Europe's population died due to the Black Death. In 1918, the Spanish flu pandemic caused an estimated 50 million deaths worldwide [33]. More recently, the 2002–2003 SARS pandemic that originated in Hong-Kong and spread worldwide caused the death of 774 [36]. These events highlight not only the scale of the problem but also our vulnerability, past and present. The situation worsens in times of crises. A recent example is the ongoing cholera outbreak in Haiti: it started in 2010, a few months after a major earthquake. Cholera is a recurring issue in West African countries as well, with many deaths reported each time. Effective measures against an epidemic require an accurate and up-to-date assessment of the situation, a very fast response and a strong coordination, which are colossal organizational efforts under tight time constraints. To this day, there is no uncontested way of preventing epidemics in general. Traditionally, many methods that have been used involve top-down approaches such as vaccination campaigns, the set-up of medical shelters, travel restrictions or quarantines [19]. These methods have several drawbacks: they are difficult and slow to be put into place, they can be expensive and also freedom-restrictive. It is clear that any improvement could have a tremendous impact and translate into significant welfare gains.

In our work, we focus on human-mediated epidemics (transmitted by human contact, e.g., influenza). For these epidemics, human mobility clearly plays a crucial role in that it enables the epidemic to travel and spread geographically. We will explore new mitigation methods

and expand the solution space. In particular, we argue that taking advantage of mobile technology opens up many possibilities for mitigating the spread of an epidemic in original and distinctive ways. Importantly, mobile technology is unique in that it is allows the *personalization* of countermeasures through precise measurements at the individual level, as well as individualized recommendations. It is this combination of information extracted from mobile data and subsequent personalization of prevention advice that opens up novel ways of mitigating an epidemic. We envision a mobile service that sends recommendations that invite the individuals to adapt their behavior, for example by delaying or canceling a trip. More generally, we formulate subtle, precise and minimally restrictive personalized behavioral rules that, if followed even partially, will have a positive global effect on the epidemic.

1.1 Context and Contributions

Our work was spurred by the *Data for Development* challenge[1] organized by France Telecom-Orange, a global telecommunications operator. Participants in this challenge have access to data gathered from 2.5 billion calls made by 5 million users in Ivory Coast. The goal is to find an original and creative use of this data that contributes towards the social, economic and environmental development of Ivory Coast. Four different datasets were derived from call detail records (CDRs) recorded over a period of 5 months, from December 2011 to April 2012. Blondel et al. [8] provide a detailed description of the datasets. Among these, two are mobility traces containing the time and the location at which a sample of the users made their phone calls. In order to protect the users' privacy, the datasets reflect different trade-offs in terms of the location's accuracy and the time span over which the trace is provided. We use this data to build a home location and time-dependent model of human mobility in Ivory Coast, which allows us to accurately capture population movements across the country (Section 4). These mobility patterns then power the core of our epidemic model, which allows us to analyze epidemic outbreaks at the level of single individuals (Section 5).

Beyond these models, our main contribution is to foster the idea of a mobile service that sends personal recommendations to help mitigate an epidemic. The mobile service is an original idea that has several advantages over existing methods. In particular, we introduce and motivate the concept of *micro-measures*, individual countermeasures tailored to their recipients' specific behavior; this new approach is the opposite of the one-size-fits-all pattern that characterizes most traditional mitigation measures. We present several concrete such micro-measures and discuss their potential (Section 3). Finally, we empirically evaluate their effectiveness using our epidemic model and provide some insights into further research directions (Section 6).

2 Related Work

Infectious diseases, also known as transmissible diseases are one of the the major causes of deaths in human societies. An epidemic is a rapid and extensive spread of a transmissible disease that affects many individuals in an area, community, or population. In order to study epidemics, scientists need to describe them mathematically, which enables them to

[1]See: `http://www.d4d.orange.com/`. The challenge was launched in mid-2012 and ended on February 15th, 2013.

predict epidemic outbreaks and to find strategies for decreasing mortality rates, and hence the costs to the economy. In their seminal work, Kermack and McKendrick [23] introduce a SIR model with three distinct classes of populations: susceptibles, infectives and recovered. This simple, yet powerful, model is very popular for modeling the evolution of epidemics in populations. Hethcote [17] reviews different extensions of this model (SIS, SI, SEIS and etc.), as well as threshold theorems involving measures such as the reproduction number, which is the average number of secondary infections caused by an infected individual when in contact with a population of susceptibles.

Instead of modeling an epidemic for the population in a region, it is possible to increase geographical granularity by dividing the original region in sub-regions, and then study the SIR model for the population of each region [1, 5, 26, 34]. By assuming that human contacts are responsible for disease transmission, the disease spread among sub-regions is driven by the mobility of individuals. Sattenspiel and Dietz [26] take into account the home region of individuals in order to simulate their mobility. One of the simple approaches to modeling population mobility is the gravity model that is based on two assumptions: Mobility flux between two regions is proportional to the product of their population's size. It decays as the distance separating them increases [29]. For example, recently, Rinaldo et al. [25] study the Haiti cholera outbreak (2010–2011) and try to predict the next outbreaks of cholera, using the gravity model and rainfall as drivers of disease transmission. By using a stochastic computational framework, Colizza et al. [9] study the epidemic propagation on a larger scale: They analyse the effect of airline transportation (complete worldwide air travel infrastructure complemented with census population data) on global epidemics.

In order to improve the realism of epidemic models, we need to build more accurate and data-driven mobility models. CDRs collected by cellular services are used for studying human mobility, because they represent a rich source of information about mobility [2, 3, 4, 14, 20, 31]. For example, Gonzalez et al. [14] analyze the trajectory of 100,000 mobile phone users over a six-month period. They find that human trajectories exhibit very regular patterns, hence we can model each individual mobility with only a few parameters. Isaacman et al. [21] model how a large population move within different metropolitan areas. Because of the sporadic nature of CDRs, Ficek and Kencl [12] use a Gaussian mixture model to reproduce probabilistically location of users between two consecutive calls. Based on the number of unique antennas observed by each user, Halepovic and Williamson [16] assume that some proportion of the population are static and always stay in their home regions.

The development of strategies for controlling epidemics such as influenza is one of the high priorities of global public health policies [11, 13, 18, 19]. SIR models, which incorporate mobility between regions, represent powerful tools for designing and testing different strategies to control epidemics. The quarantine is one of the methods often used to limit the spread of infectious diseases within human populations. We lack information however about the effectiveness of quarantine on controlling epidemics. Sattenspiel and Herring [27] use records of the influenza epidemic, which took place in Canada at 1918-19, to investigate the effect of quarantine. They show that a quarantine is effective only when mobility is restricted, and that it depends on its application-time and duration. In addition to these issues about the effectiveness of quarantine, there are issues, that include implementation challenges, economic cost and the violation of civil rights, especially in the cases of long confinement or isolation from society. Another way to control epidemics is to vaccinate the susceptible

population in a series of pulses called pulse vaccination [24, 28, 30, 37, 38]. For example, Zaman et al. [37] define a control optimization problem based on the SIR model. They try to compute the optimum percentage of susceptible population to be vaccinated at each time. This method requires the vaccination of at least 10 percent of the susceptible population at each time step, in order to make a small change in the epidemic behaviour of the infectious disease.

3 Mobile Micro-Measures

Traditional epidemic mitigation methods consist of heavy, top-down approaches such as blockades, quarantines or large-scale vaccinations. As an alternative, we suggest that mobile technology could enable a much richer and sophisticated set of mitigation measures for human-mediated epidemics, which we name *micro-measures*. Let us illustrate our vision by describing a simple scenario.

> Jean, an 18 year old inhabitant of Ivory Coast living in Northeastern Bouaké, would like to play pickup football. He knows that a meningitis outbreak just surfaced in his district, and he does not want to take any risk. Bouaké happens to be part of a pilot program of a mobile service that helps mitigate the spread of meningitis. Using his mobile phone, he sends a short request to the service that instantly computes the following personalized recommendation for him: to minimize the risk, he should try the football field a few kilometers southwards, instead of going to the one he is used to. It would be best if he took the *gbaka* (small bus) in about 17 minutes, this way he would avoid contact with the kids coming back home from the school nearby.

Of course, this scenario presents an idealized and naive view of reality; Jean might not have a cell phone to begin with, the bus might not have such a precise schedule, and there might not be alternative locations where people are playing pickup football. It nevertheless gives an overview of the level of refinement that can be achieved through personal recommendations. The main properties of such a service are as follows:

Personalized. Recommendations are generated and communicated on an individual basis. Mobile technology enables this in two ways: first, it allows for a quantity of valuable behavioral information (such as location and activity) to be recorded and second, it provides a readily available unicast communication channel.

Adaptive. As the epidemic progresses and each individuals' intentions are discovered, the recommendations are instantly adapted. The personalization of mobile micro-recommendations ensures their effectiveness. Such recommendations, in contrast with most large-scale mitigation efforts, would typically require much less time to be set up and would always be in phase with the current state of the epidemic.

Microscopic. In contrast with a one-size-fits-all policy that typically considers an epidemic from a macroscopic perspective, micro-measures tend to focus on subtle and local changes. These changes, when looked at independently, are mostly insignificant; but taken together, they result in important global improvements.

State-independent. An additional property of the service is that it is epidemic-state independent: the recommendation should not depend on whether the individual is infective or not. First, it does not require prior knowledge about the state of an individual: it is often hard to determine precisely when he becomes infected. Second, it aligns the incentives: without additional knowledge, everyone can expect to benefit from following the recommendation—this might not necessarily be the case when the state is known.

This lays the foundation of our approach but does not yet suggest any concrete mitigation scheme. Still, there are fundamental questions related to the feasibility of micro-measures. Under which conditions do small, local changes (such as an individual agreeing to commute slightly earlier) have a global impact? How many individuals need to cooperate, and how does this, significantly alter the dynamics of the epidemic ? An epidemic can often be seen as being either supercritical (the epidemic grows) or subcritical (it declines). What microscopic changes are susceptible to cause a phase transition? Although a precise characterization of these changes and, by extension, rigorous answers to these questions are beyond the scope of our work, we intend to show initial evidence of the relevance of such a mobile service.

3.1 Concrete Micro-Measures

Beyond a theoretical argument, our contribution is the description and evaluation of three concrete strategies we can use to generate micro-measures. They represent initial baselines for further developments. Let us first note that contacts between individuals can broadly be categorized into two groups: the deliberate contacts are, for example, between family members or at work, whereas the accidental contacts are formed by random encounters, for instance, while shopping or commuting. At a high level, our approach is to maintain deliberate contacts and rewiring the accidental ones. The idea is to *weaken* the links in the contact network that form the path through which the epidemic spreads. By changing its structure, we seek to decelerate the dynamics and drive the epidemic down to a sub-critical level.

	CutCommunities	DecreaseMix	GoHome
Knowledge to maintain	List of communities of locations	Social communities of users	State of the epidemic across regions
Recommendation	Do not cross community boundaries	Stay with your social circle	Go/stay home
Intuition	Weakening the weak geographical links	Segmenting social communities	Home is a safe place

Table 1. We recapitulate the main characteristics of the three strategies we have implemented to mitigate the spread of epidemic.

3.1.1 CutCommunities strategy

It is clear that mobility drives the spread of an epidemic. A straightforward strategy would therefore be to reduce long-range contacts, be it at the expense of reinforcing local ones. Uniformly reducing mobility is, however, both expensive and inflexible. To overcome

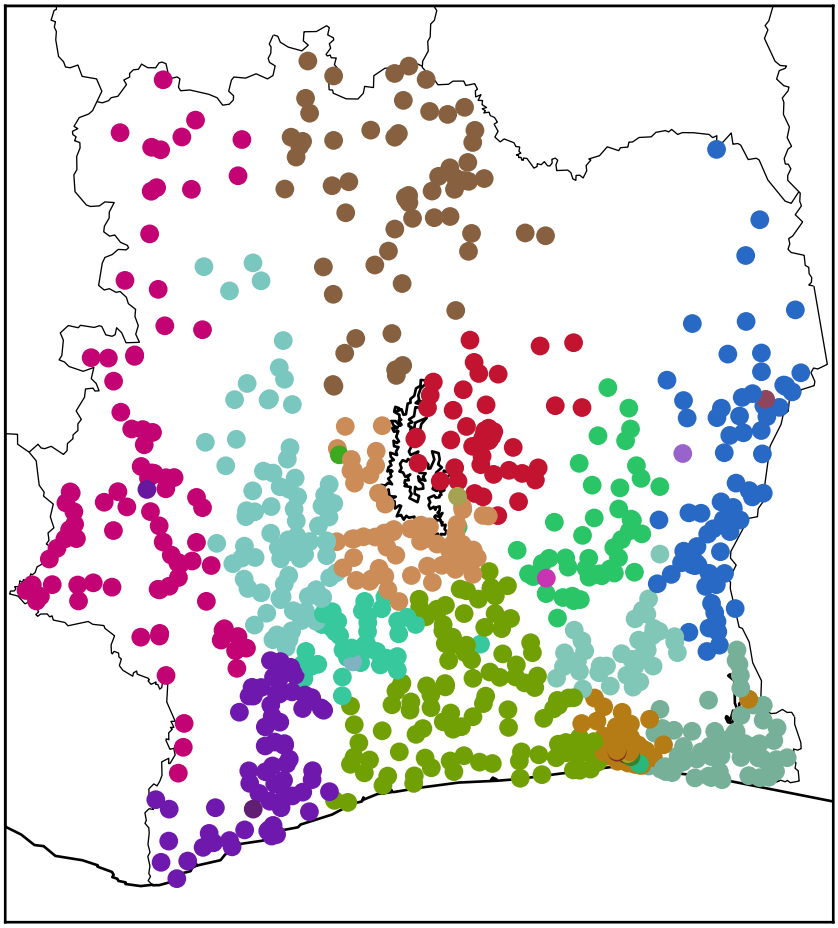


Figure 1. We find 30 communities in the mobility network (Section 6) when using the Louvain community detection algorithm. It is not surprising that these communities reflects the geographical proximity between nodes, as trips are more frequent between close antennas than between distant ones.

this, our first strategy, CutCommunities, takes into account *communities of locations* in the mobility network, and focuses on reducing human mobility over inter-community links—this is, in a sense, analogous to *weakening the weak links* in the network. The main difference with a simple blockade is that our strategy is able to adapt to changes in the network (mobility patterns vary over time, cf. Section 4). In practice, the service operator would maintain a list of location communities identified through the mobility patterns of its userbase; when an individual checks whether a trip is safe, the service would verify whether it crosses community boundaries and, if this is the case, discourage the individual from making this trip[2]. If additional per-location information is available about the current state of the epidemic, recommendations could be further corrected according to the strength

[2]As a relaxation of this counter-measure, we could consider *postponing* the trip instead. Simply by delaying certain trips, we could prevent harmful interactions between groups of individuals. This is analogous to *time-division multiplexing*; a slight change in the habits of a group of people might significantly change the contact surface.

of the epidemic at the individual's current and projected locations.

3.1.2 DecreaseMix strategy

Instead of acting on mobility to segment contacts across location communities, we also consider segmentation *social communities*. The aim is to separate individuals *inside* the same location, e.g., by making them visit different aisles of the same supermarket at different times. Putting in place such a segmentation is more sophisticated than in the case of mobility, but this strategy is the perfect example of another extremal point in the solution space. The service operator would keep a list of social communities and would communicate a distinctive tag (e.g., a color) to individuals according to their community. Individuals would access locations differently, depending on their tag; for example, seating in a theater would be organized in such a way that contacts between communities are minimized. We are aware that this strategy could raise many concerns, because it segregates people, therefore great care would be needed if it were to be implemented. Despite this, we retain it because it reflects a different trade-off with respect to CutCommunities: instead of discouraging individuals from going to certain locations where they can be in contact with everyone, we allow them them go everywhere, but restrict the contact network.

3.1.3 GoHome strategy

We consider a third case where the service recommends individuals to *go home*. The intuition behind this strategy is that we assume that when at home, the contact rate decreases. Whereas the previous strategies target the individuals' location or contact network, this one is distinctive in that it affects the the rate of contact. With information on the progress of the epidemic across locations, the operator could prioritize sending advice to those individuals whose cooperation would yield the greatest effect. In Section 6, we will provide a detailed evaluation of the three described strategies. Before doing so, we will introduce the mobility and epidemic models used for our assessments.

4 Mobility Model

Because the spread of epidemics depends greatly on the mobility of infected individuals, and on the locations where they interact with other individuals, a realistic, data-driven mobility model is a essential tool for simulating realistic epidemic propagation. It should therefore model population mobility, take into account certain microscopic aspects at the individual level, and still allow simulations of epidemic propagation to scale up to millions of individuals. Moreover, it should capture the main differences between the mobility of different groups of individuals, where a group is constituted of individuals exhibiting similar mobility profiles. To construct a mobility model that fulfils these requirements, our intuition is : The home location of individuals strongly shapes their mobility patterns because the places they visit regularly e.g., their workplaces, schools or the shopping centers, depend on the proximity to their home. Typically, we expect the most visited location (home) and the second most visited location (school, university or work) to be geographically close to each other. In addition to this geographical aspect, mobility is strongly time-dependent: Individuals commute between home and work during the weekdays, with a substantial change in their travel behavior during the weekends.

Definition	Domain	Explanation
$\mathcal{A} = \{1, \ldots, 1231\}$	-	Set of antennas
$\mathcal{SP} = \{1, \ldots, 255\}$	-	Set of sub-prefectures
k	$\mathbb{N}$	Time resolution
$sp_{\text{home}}(u)$	$\mathcal{SP}$	Home sub-prefecture
$a_{\text{home}}(u)$	$\mathcal{A}$	Home antenna
$X(n)$	$\mathcal{A}$	Antenna
$t(n)$	$\mathbb{N}$	Absolute time
$h^k(n)$	$\{1, \ldots, k\}$	Period of the day
$d(n) = \text{day}(t(n))$	$\{1, \ldots, 7\}$	Day of the week
$w(n) = \text{weekday}(t(n))$	$\{0, 1\}$	Day type: weekday or weekend

Table 2. List of the definition and domain of the variables relative to user u, as well as those describing his n^{th} visit.

Building on this, we make the assumption that the individuals that share the same home-location exhibit a similar mobility pattern. Therefore, we construct a location and time-based mobility model that depends on the variables presented in Table 2. The conditional distribution of the location $X(n)$ of user u depends on his home antenna $a_{\text{home}}(u)$, but also on the time of the visits $(h^k(n), w(n))$:

$$p\left(X(n)|u, t(n)\right) = p\left(X(n)|h^k(n), w(n), a_{\text{home}}(u)\right). \tag{1}$$

First, we choose the the time resolution $k = 3$ in order to divide the day in 3 distinct periods: Morning (6 am to 1 pm), afternoon (1 pm to 8 pm) and night (8 pm to 6 am). Second, conditioning on the parameter $w(n)$ allows us to distinguish between weekdays and weekends. Finally, the home antenna $a_{\text{home}}(u)$ of user u is defined as the most visited antenna during the night period. Consequently, given the period of the day, the day type and the home antenna of user u, the distribution of the location that he might visit (1) is a multinomial distribution with $|\mathcal{A}|$ categories.

4.1 Learning and Evaluating Mobility Models

In order to build our model from data, we analyse `SET2`, one of the datatsets provided by France Telecom-Orange [8]. It contains high-resolution trajectories of 500,000 users, observed over a two-week period. We focus on this datatset, as it offers the highest geographical resolution : Individuals' locations are observed across antennas. To avoid having to deal with users whose location samples are very sparse, we consider only the users who visited more than 1 antenna and made on average more than 1 call per day. In order to evaluate the realism of our mobility model, we separate the data into two parts: For each user, we put 90% of the calls in the training set and the remaining 10% in the test set. First, we build a mobility model by learning from the training set by using a maximum likelihood estimator . Then, we evaluate our mobility model by computing the average log-likelihood of the calls belonging to the unseen test set. The average log-likelihood reflects how well our model generalizes to unseen data. As the test set might contain some locations not visited by a given class of users in the training set, the maximum likelihood estimate of the distribution (1) assigns zero probability to these observations. We cope with this by assuming

Mobility model	Average log-likelihood
Our model	-1.07
SPM	-1.67
TM	-2.9
MC	-6.49

Table 3. Log-likelihood of the unseen data from the test set. Our mobility model significantly outperforms the baseline models since its predictive power, with respect to the test set, is higher.

that the distribution (1) is a multinomial distribution drawn from an exchangeable Dirichlet distribution, which implies that the inferred distribution (1) is a random variable drawn from a posterior distribution conditioned on the training data. A more detailed description of this smoothing procedure is given by Blei et al. [6].

We tested several variants of mobility models by varying their structure and parameters (time resolution, day of the week, etc). To have three representative baseline models for comparison, we choose three predictors out of the several variants we tested.

The first baseline model is a time-based mobility (TM) model

$$p\left(X(n)|u,t(n)\right) = p\left(X(n)|h^k(n),w(n)\right), \tag{2}$$

where all mobile-phone users exhibit the same time-dependant geographical distribution. The second baseline is a location-dependent first order Markov chain (MC)

$$p\left(X(n)|u,t(n),X(n-1),\ldots,X(0)\right) = p\left(X(n)|X(n-1)\right), \tag{3}$$

where the current location of a user depends only the location he visited just before. The third baseline is a time and sub-prefecture dependant mobility model (SPM)

$$p\left(X(n)|u,t(n)\right) = p\left(X(n)|h^k(n),w(n),sp_{\text{home}}(u)\right), \tag{4}$$

where the home of a user is represented by a sub-prefecture instead of an antenna. This implies a more important aggregation of users, where two users who share the same home sub-prefecture, have the same mobility pattern.

The experimental results are shown in Table 3. The first order Markov chain (MC) performs the worst. This is not surprising since the time difference between two call records varies greatly, ranging from a few minutes to a few days. The location associated with a call made in the past few hours or days does not necessarily affect the current location. As the location data is sporadic, it is not surprising than any model that learns from transitions performs poorly, and is outperformed by time-based models. Our model performs the best; and by comparing it to the time-based model (TM), we realise that knowing the home-locations of users enhances the predictive power of the mobility model. Moreover, the granularity of home locations is crucial: Our model significantly outperforms the sub-prefecture dependent mobility model because it has a finer granularity of the home-location.

A realistic mobility model is an essential building block of a realistic epidemic propagation model because mobility drives population flows between regions, and therefore the geographical proximity between individuals. In the next section, we introduce the epidemic model we use to simulate a local epidemic propagation.

5 Epidemic Model

Building up on the mobility, this section introduces our epidemic model. It is based on a discretized, stochastic version of the *SIR* model [23]; Tables 4 and 5 provide an overview of the different parameters and quantities used throughout the section. We assume that the size of the population (N individuals) remains constant—there are no births nor deaths, a reasonable assumption if the time horizon is limited to at most a few months. Under the *SIR* model, an individual can be either *susceptible* to the disease, *infective*, or *recovered* from the disease and immunized against further infections[3]. We assume that most of the population is initially susceptible, except for a small number of infective individuals that form the seed of the epidemic. Individuals successively go through the susceptible, infective and recovered states; a desirable outcome would have many individuals stay susceptible without ever becoming infective. The basic *SIR* model assumes *random mixing* of the whole population: any individual meets any other one with a uniform probability. In our model, we relax this strong assumption by taking into account the *mobility*. We spread the population across M regions; each region bears its own *SIR* process where the corresponding meta-population mixes at random. These regional processes are independent and isolated, and the only way the epidemic crosses regional boundaries is through human mobility [22]. In summary, regional interactions take place uniformly at random, whereas global interactions are shaped by the individuals' mobility.

N	total population
M	number of regions
N_i^*	initial population of region i, where $i \in \{1, \ldots, M\}$
L	number of different mobility classes
β	contact probability
g	recovery probability

Table 4. Parameters of the epidemic model.

5.1 Local Epidemic Dynamics

In order to work at the individual level, we adapt the classic deterministic *SIR* model in order to have a discrete-time stochastic variant. The contact probability β and recovery probability g are constant across all regions[4]. For a region $i \in \{1, \ldots, M\}$ we compute, at each time step, the force of infection $\lambda_i = \beta \frac{I_i}{N_i}$. This quantity represents the probability of making a contact that results in an infection. During a time step, every susceptible individual gets infected independently at random with probability λ_i, while every infective

[3] In the literature, this state is sometimes known as *removed*. The important point is that they do not participate in the epidemic anymore.

[4] These quantities are *rates* in the continuous time *SIR* model. In order to carry over the characteristics of the *SIR* model to our discretized version, we need to ensure that the sampling interval is short enough to ensure that $\beta, g < 1$.

c_l	mobility class l, where $l \in \{1, \ldots, L\}$
$\mathbf{S}_i$	distribution of the number of susceptible individuals in region i across classes. $\mathbf{S}_i = (S_{i,c_1}, \ldots, S_{i,c_L})$
$\mathbf{I}_i$	distribution of the number of infected individuals in region i across classes. $\mathbf{I}_i = (I_{i,c_1}, \ldots, I_{i,c_L})$
$\mathbf{R}_i$	distribution of the number of recovered individuals in region i across classes. $\mathbf{R}_i = (R_{i,c_1}, \ldots, R_{i,c_L})$
S_i	number of susceptible individuals in region i, equal to $\|\mathbf{S}_i\|_1$
I_i	number of infected individuals in region i, equal to $\|\mathbf{I}_i\|_1$
R_i	number of recovered individuals in region i, equal to $\|\mathbf{R}_i\|_1$
N_i	population of region i, where $i \in \{1, \ldots, M\}$
λ_i	infection probability for region i. $\lambda_i = \beta \frac{I_i}{N_i}$

Table 5. Notation for various quantities related to the epidemic.

individual recovers independently at random with probability g. If we denote by ΔX_i the variation of $X_i, X_i \in \{S, I, R\}$ after one time step, it is easy to see that

$$\begin{aligned} \mathbb{E}(\Delta S_i) &= -\lambda_i S_i \\ \mathbb{E}(\Delta I_i) &= \lambda_i S_i - g I_i \\ \mathbb{E}(\Delta R_i) &= g I_i, \end{aligned}$$

which are the expected difference equations for the *SIR* model under the random mixing assumption. We note that our model has many similarities with that of Colizza et al. [10], used to model the SARS pandemic.

5.2 Implementation

To allow for distinctive mobility patterns across the population, individuals belong to one out of L classes $\{c_1, \ldots, c_L\}$ that fully characterize their mobility patterns. In accordance with the mobility model (Section 4), the individuals' class is determined by their home antenna. The implementation is best understood when decomposed into two distinct, successive phases: a *mobility* phase where individuals can move between regions, and an *epidemic* phase where individuals get infected or recover.

Mobility phase We consider every individual. Suppose the individual is in region i; the mobility model assigns a new region j according to its mobility class. If $i \neq j$ we update the vectors $\mathbf{X}_i$ and $\mathbf{X}_j$ accordingly, where $\mathbf{X} \in \{\mathbf{S}, \mathbf{I}, \mathbf{R}\}$ depends on the current state of the individual.

Epidemic phase We consider every region $i \in \{1, \ldots, M\}$. We begin this phase by updating the infection rate λ_i given the current values of N_i and I_i. Every infected individual then recovers with probability g, while every susceptible individual gets infected with probability λ_i. $\mathbf{S}_i$, $\mathbf{I}_i$ and $\mathbf{R}_i$ are updated accordingly.

This process is repeated until the end of the period of interest.

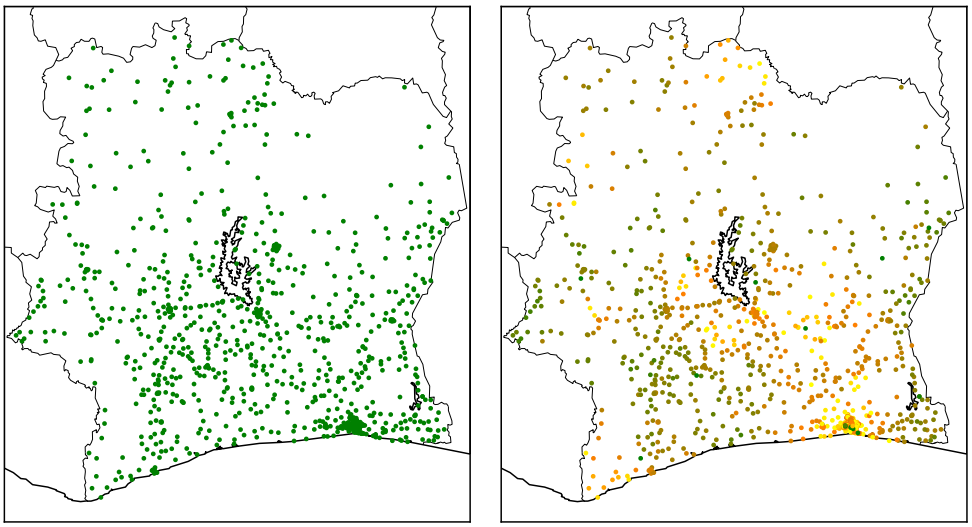


Figure 2. Snapshots of a sample epidemic process where each dot represents a region (here, the surroundings of an antenna). Colors indicate the relative proportion of infective individuals. Initially, just a few individuals form a seed of infectives (left). A little more that 9 days later, the epidemic has spread over most of the country (right).

6 Empirical Evaluation

Next, we use our models to test the strategies previously described in Section 3. Before evaluating our strategies, we first explain how the epidemic model is parameterized and how epidemic spreads are quantitatively characterized.

6.1 Model Parameters and Evaluation Metrics

In order to be consistent with our mobility model, the epidemic model defines regions to be the area surrounding the antennas ($M = 1231$). Hence, we will use the words *region* and *antenna* interchangeably. As an individual's mobility is tied to his home antenna, we distinguish among $L = 1231$ different classes. To initialize the population attached to each antenna, we use data from the AfriPop project [32] which provides us with Ivory Coast population figures at the hectare level; to account for the fact that not every individual is mobile, we allow only 55% of the population to move during the mobility phase[5], which roughly corresponds to the proportion of the population in the 15-to-64 age bracket [35]. Days are divided into three time steps in order to match the mobility model[6], and the typical time horizon is between 100 and 400 time steps (i.e. 1–4 months). Contact and recovery probability are usually set to $\beta = 1$, respectively $g = 0.5$; Although these synthetic values do not directly match any well-known disease, they are still qualitatively close to realistic cases, such as influenza. All our simulations start with a seed set of 23 infectives

[5]This distinction is rather crude and could certainly be further refined. However we deemed it to be sufficient for our purposes.

[6]Notice that this is not a formal requirement. We use this subdivision mainly for simplicity.

p	Affected movements	Maximum
0.90	10.91%	21.38% ($ts = 42$)
0.99	12.57%	22.91% ($ts = 51$)
1.00	5.32%	12.20% ($ts = 33$)

Table 6. Proportion of movements affected when using the CutCommunities strategy for three different values of the compliance probability p. We indicate the overall average over the 80 time steps, as well as the maximum value.

distributed across 5 antennas[7] in the Attécoubé district of Abidjan.

In order to quantify the difference between epidemic spreads, we propose three metrics to evaluate the effectiveness of our mitigation strategies. Figure 3 shows how these quantities are related to the epidemic's evolution over time. For notational clarity, let $X = \sum_{i=1}^{M} X_i, X \in \{S, I, R\}$ be the total number of individuals in each state over the country as a whole. As these quantities evolve over time, they are functions of the time step n. The first metric is the *size of the largest outbreak* or, equivalently, the maximal proportion of infective individuals,

$$I^* = \max_n \frac{I(n)}{N}.$$

The reasoning behind this metric is self-evident: in most cases, the larger the proportion of infective individuals, the more difficult the control of the epidemic. It is also, broadly speaking, a good indicator of the epidemic's strength. Our second metric is closely related to the first one, but considers the complementary dimension: it measures the *time of the largest outbreak*,

$$T^* = \arg\max_n I(n).$$

Delaying the moment at which the epidemic reaches its peak allows individuals and governments to have enough time to adapt their behavior, respectively, to deploy measures. Finally, our last metric captures the tail behavior of the epidemic: it measures the *final proportion of recovered users*,

$$Q^* = \lim_{n\to\infty} \frac{R(n)}{N}.$$

Note that we would like to minimize this metric. After the epidemic dies out, all individuals are either recovered or susceptible, and a low proportion of recovered individuals means that a high percentage of the population did not go through the infective state at all.

6.2 Results

We now take a closer look at our three proposed strategies. We will describe how we instantiate them and we provide qualitative and quantitative assessments with respect to their effectiveness.

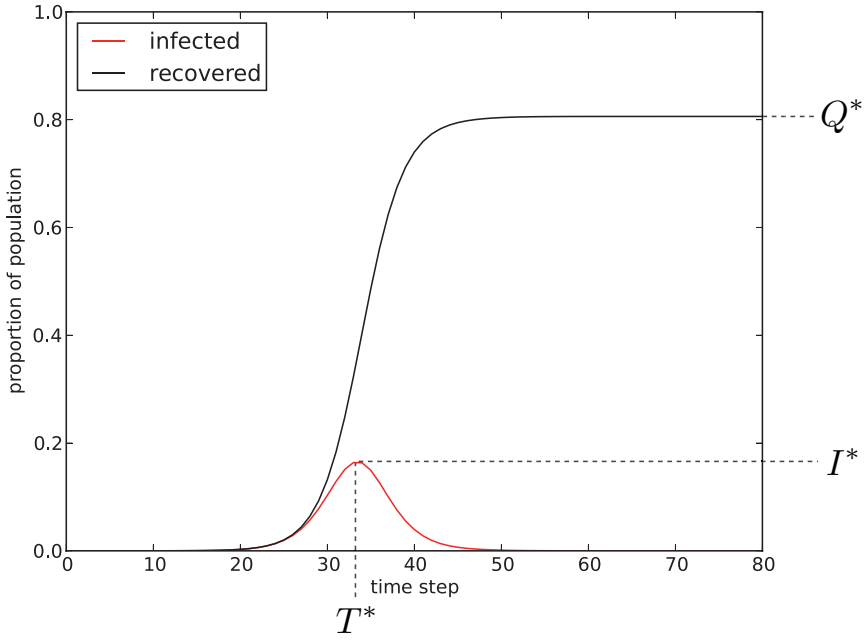


Figure 3. Metrics used to evaluate the effectiveness of mitigation strategies. I^* indicates the magnitude of the epidemic's peak, T^* the time at which the peak happens, and Q^* describes the asymptotic number of individuals that got infected and recovered.

6.2.1 CutCommunities strategy

The first strategy divides the country into location communities according to the network of mobility. We consider the weighted, undirected graph where nodes represent antennas, and edge weight is equal to the average number of trips between the two endpoints (regardless of direction). We use the Louvain community detection algorithm [7]; Figure 1 shows the 30 identified communities. It is interesting—but not surprising—to note that the communities are roughly geographicaly based[8]. This confirms our hypothesis stating that there are *geographical weak links*. Micro-measures are then generated as follows: when an individual checks whether a trip is safe, the service first verifies whether the trip crosses community boundaries and whether the current or projected locations are affected by the disease; if both of these conditions are met, the individual is discouraged from making the trip. The recipient then complies with probability p.

Figure 4 shows the effect of CutCommunities for different values of p. Compared to the baseline ($p = 0$), the strategy affects the size I^* and the time T^* of the epidemic's peak. However, it does not change much the tail behavior: Q^* stays constant at around 0.8, except for the degenerate case where $p = 1$, which represents a blockade around the community initially infected. We also observe that there seem to be two infection phases,

[7] In the datasets provided by France Telecom-Orange, these antennas have the following identifiers: 57, 146, 330, 836, 926.

[8] As a sidenote, we ran the Louvain method on a number of other graphs generated from the datasets provided for the D4D challenge, including one derived from `SET1` representing total antenna-to-antenna communications. The communities always displayed the same geographical clustering. Furthermore, we observed that mobility communities seem to be correlated to phone call communities.

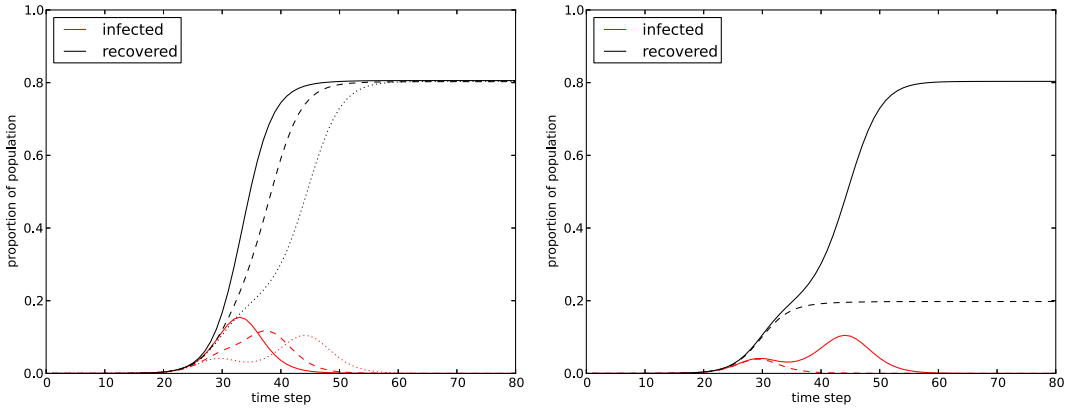


Figure 4. Shape of the epidemic under the CutCommunities strategy, $\beta = 1.0$, $g = 0.5$. On the left: solid lines represent the baseline ($p = 0$), dashed lines $p = 0.9$, dotted lines $p = 0.99$. On the right, we compare $p = 0.99$ (solid) to a complete blockade ($p = 1$, dashed).

made progressively more apparent as $p \to 1$, and that the blockade removes the second phase; these two phases correspond to infections happening inside, respectively, outside the initially infected community. Recall that this strategy only sends micro-measures to a fraction of the individuals, those who cross community boundaries—a case that by definition should not happen too often. It is therefore interesting to consider the number of trips actually *canceled* as a result: Table 6 lists the average and maximal proportion for different values of p. The numbers are quite low[9], suggesting that the communities form a natural partitioning of the regions. In conclusion, this strategy does not affect the asymptotic behavior of the epidemic but significantly shifts its peak. Altogether, it justifies the relevance of mobility-based geographical communities as a data source to generate micro-measures.

6.2.2 DecreaseMix strategy

Recall that this strategy assigns tags to individuals according to the social community to which they belong and segregates contacts across social communities. A service operator might use the call graph (i.e. the social network derived from who calls whom) to infer social communities in the population; unfortunately, we do have access to such data[10]. In order to quantify the effectiveness this strategy, we proceed as follows. Similarly to our mobility model, we make the assumption that the individual's community C is determined by his home antenna. The DecreaseMix strategy do not decrease the total number of contacts; instead it rewires contacts across communities to contacts inside the community. This is done by splitting the contact probability to into intra-community and inter-community

[9] That these proportions are lowest when $p = 1$ is due to the fact that the epidemic is local to the infective seeds' community

[10] The data provided for the Orange D4D challenge does include a dataset consisting of myopic views of the call graph. `SET4` is a sample of *egonets*, i.e. balls of radius two centred at a particular user. However, this dataset did not yield anything useful for our purposes.

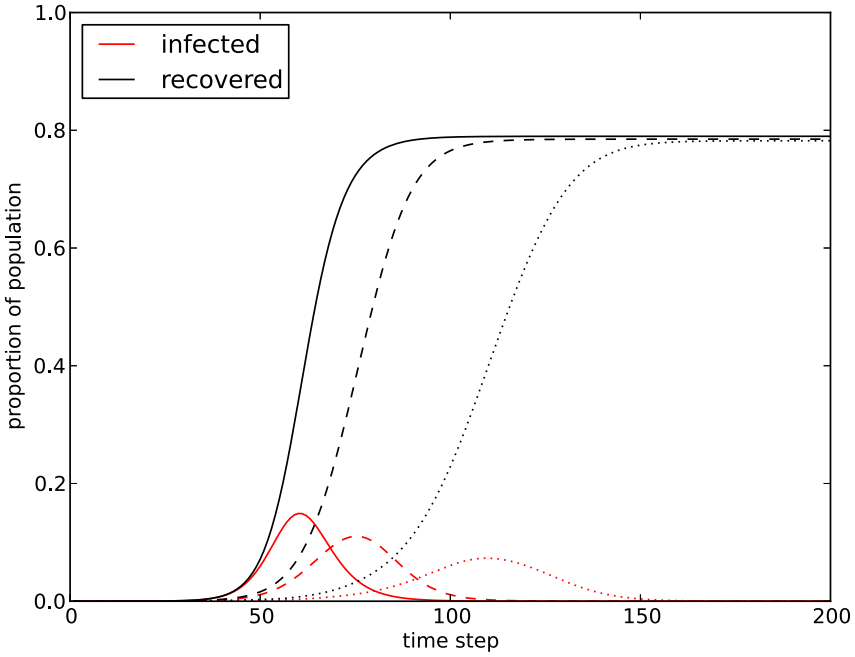


Figure 5. Shape of the epidemic under the DecreaseMix strategy averaged over 10 runs, $\beta = 1.0$, $g = 0.5$, for different values of the mixing parameter. Solid lines correspond to $q = 1.0$, dashed ones to $q = 0.1$, dotted ones to $q = 0.01$.

contact probabilities and introducing a mixing parameter q

$$\beta_{i,C} = \left(1 - q + q\frac{N_{i,C}}{N_i}\right)\beta$$
$$\beta_{i,\overline{C}} = \beta - \beta_{i,C}$$
$$\lambda_{i,C} = \beta_{i,C}\frac{I_{i,C}}{N_{i,C}} + \beta_{i,\overline{C}}\frac{I_{i,\overline{C}}}{N_{i,\overline{C}}},$$

where $N_{i,C}$ indicates the number of individuals of community C currently in region i, $N_{i,\overline{C}} = N_i - N_{i,C}$ and the other quantities follow the same convention of notation. The intuition is as follows: When $q = 1$, everyone mixes at random inside a region just as if no countermeasure was applied at all. At the other extreme, when $q = 0$, contacts happen only with individuals from the same community. Intermediary values of q allow us to play with the strength of the segregation.

We evaluate the effectiveness of DecreaseMix for different values of the mixing parameter q. Our simulations are parameterized with $\beta = 1.0$, $g = 0.5$ and $q \in \{1, 0.1, 0.01\}$; Figure 5 shows the average behavior of the epidemic over 10 runs. The main characteristic of this strategy is that it delays the epidemic outbreak. However, the slopes of the two curves at the strongest point of the epidemic are not very differentiated. As s result, the final proportion of recovered Q^* does not vary much. But by making it 10 or 100 times more likely to contact an individual of the same community, we delay T^* by approximately 5 and 16 days, respectively, on average. Our intuition about this phenomenon is that it takes more time for the epidemic to reach certain communities (as they are more segregated), but

p	Affected movements	Maximum
0.1	2.81%	5.21% ($ts = 190$)
0.5	15.80%	26.12% ($ts = 316$)

Table 7. Proportion of movements affected when using the GoHome strategy for two different values of the compliance probability p. We indicate the overall average over the 400 time steps, as well as the maximum value.

once a community sees its first case of infection, the spread is just as fast as before. We argue that one of the main limiting factors at play here is the random mixing assumption: if we were able to bring finer structural changes to the contact graph, the situation would look very different.

6.3 GoHome strategy

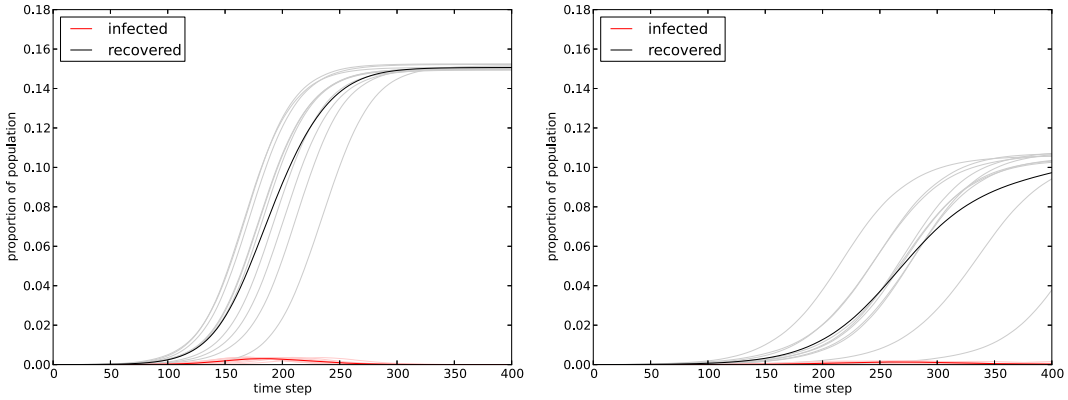


Figure 6. Shape of the epidemic under the GoHome strategy, $\beta = 1.0$, $g = 0.5$. Light curves indicate individual runs, dark curves indicate average. On the left: $p = 0.1$, on the right: $p = 0.5$.

Our last strategy advices individuals to go home or stay home. In order to focus the micro-measures on the most influential individuals, we assume that at each time step, the service operator knows the proportion of susceptible, infective and recovered individuals across locations. We suppose that before every trip, an individual sends a request to the service that compares the proportion of infectives in both source and destination, and recommends to go home if the destination has a proportion of infectives, *lower* than the source location. Individuals then comply with probability p. The main intuition behind this choice is to avoid sending infective individuals to highly susceptible locations. Note that we keep the state-independent assumption here: we do not know the state of the individual when sending out a recommendation. The second important assumption is that, when the individual is at at home, the contact probability is set to be equal to the recovery probability[11], i.e. $\beta_{home} := g$. This models the fact that there are less contacts at home, in term of accidental

[11] When contact and recovery probability are equal, the single-population *SIR* epidemic (under the random mixing assumption) does not develop anymore; setting $\beta_{home} := g$ can therefore be seen as the least change needed to stabilize the epidemic.

ones. Mixing is not exactly uniform anymore, and the infection probability is adapted as follows:

$$\lambda_{i,loc} = \beta_{home}\frac{I_i}{N_i}$$
$$\lambda_{i,vis} = \beta\frac{I_{vis}}{N_i} + \beta_{home}\frac{I_{loc}}{N_i}.$$

Quantities with *loc* and *vis* subscripts correspond to individuals whose home region is (respectively is not) i. Note that the contact probability of visitors can significantly decrease in a region where the proportion of visitors to locals is low.

This time, the effectiveness depends on the value of the compliance probability p. We use again $\beta = 1.0$, $g = 0.5$ and let $p \in \{0.0, 0.1, 0.5, 0.7\}$; Figure 6 shows the behavior of the epidemic over 10 runs. As opposed to the results obtained with the DecreaseMix strategy, we obtain significant improvements to Q^* as p increases[12]. This observation is not surprising because by suggesting to individuals to go home, we are directly reducing their contact probability, which is a determining factor of the epidemic's dynamics. It is also interesting to look at the actual number of trips that are affected (i.e., cancelled) because of the micro-measures; Table 7 shows that a relatively low number of trips have to be affected to noticeably impact the spread. In summary, this strategy has the potential to be quite effective, although the assumptions it makes deserve closer analysis.

7 Conclusion

In this paper, we explore the novel idea of using mobile technology in order to mitigate the spread of human-mediated infectious diseases. We motivate the concept of *mobile micro-measures* that consist of personalized behavioral recommendations given to individuals. By affecting, even partially, individual behaviors, we are able to globally impact the epidemic propagation. These mobile micro-measures have several original properties; they are adaptive, target individuals at the microscopic level and provide a rich set of mitigation methods. Using the data provided for the Orange D4D challenge [8], we first develop a realistic mobility model for the population of Ivory Coast. Then, we incorporate it into an epidemic model based on *SIR* in order to simulate the epidemic propagation, while taking into account population mobility. Taking advantage of this framework, we propose and evaluate three concrete strategies used to generate micro-measures. Our strategies weaken the epidemic's intensity, successfully delay its peak and, in one case, significantly lower the total number of infected individuals.

These preliminary results allow us to identify several research avenues. First, random mixing is the most limiting assumption. Being able to change the structure of human contacts at a finer level is a key component of more advanced micro-measures. The mobile call graph is an example of a source of information about social contacts, one that is readily available to mobile phone operators. Second, beyond our preliminary strategies, it is highly important to deepen our understanding of the key ingredients that make mobile micro-measures effective yet minimally restrictive. In parallel to mobile micro-measures, the

[12]Unfortunately, our simulation was limited to 400 time steps, which is not enough to clearly show the asymptotical behavior. The claim, however, is justified by looking at the *worst* runs whose slope quickly tends to zero.

availability of large-scale mobility data opens up new research directions in epidemiology: a more precise characterization of the relation between epidemic spread and human mobility patterns is an interesting topic we would also like to investigate in the future.

To conclude, we firmly believe that data-driven and personalized measures which take advantage of mobile technology are an important step towards effective epidemic mitigation.

Acknowledgements

We would like to thank Vincent Etter for his insightful comments and feedback about this paper.

References

[1] J. Arino and P. Van den Driessche. A multi-city epidemic model. *Mathematical Population Studies*, 10(3):175–193, 2003.

[2] A.-L. Barabasi. The origin of bursts and heavy tails in human dynamics. *Nature*, 435 (7039):207–211, 2005.

[3] M. A. Bayir, M. Demirbas, and N. Eagle. Discovering SpatioTemporal Mobility Profiles of Cellphone Users. In *WoWMoM 2009*, pages 1–9. IEEE, 2009.

[4] R. Becker, R. Cáceres, K. Hanson, S. Isaacman, J. M. Loh, M. Martonosi, J. Rowland, S. Urbanek, A. Varshavsky, and C. Volinsky. Human Mobility Characterization from Cellular Network Data. *Communications of the ACM*, 56(1):74–82, 2013.

[5] V. Belik, T. Geisel, and D. Brockmann. The impact of human mobility on spatial disease dynamics. In *CSE 2009*, volume 4, pages 932–935. IEEE, 2009.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[7] V. Blondel, J. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10): P10008, 2008.

[8] V. D. Blondel, M. Esch, C. Chan, F. Clerot, P. Deville, E. Huens, F. Morlot, Z. Smoreda, and C. Ziemlicki. Data for Development: the D4D Challenge on Mobile Phone Data. *arXiv preprint arXiv:1210.0137*, 2012.

[9] V. Colizza, A. Barrat, M. Barthélemy, and A. Vespignani. The role of the airline transportation network in the prediction and predictability of global epidemics. *PNAS*, 103(7):2015–2020, 2006.

[10] V. Colizza, A. Barrat, M. Barthélemy, and A. Vespignani. Predictability and epidemic pathways in global outbreaks of infectious diseases: the SARS case study. *BMC Medicine*, 5(34), 2007.

[11] N. M. Ferguson, D. A. Cummings, C. Fraser, J. C. Cajka, P. C. Cooley, and D. S. Burke. Strategies for mitigating an influenza pandemic. *Nature*, 442(7101):448–452, 2006.

[12] M. Ficek and L. Kencl. Inter-Call Mobility Model: A Spatio-temporal Refinement of Call Data Records Using a Gaussian Mixture Model. In *INFOCOM 2012*, pages 469–477. IEEE, 2012.

[13] T. C. Germann, K. Kadau, I. M. Longini Jr, and C. A. Macken. Mitigation strategies for pandemic influenza in the United States. *PNAS*, 103(15):5935–5940, 2006.

[14] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[15] R. Gottfried. *Black Death*. Simon and Schuster, 1985.

[16] E. Halepovic and C. Williamson. Characterizing and Modeling User Mobility in a Cellular Data Network. In *Proc. Workshop on PE-WASUN 20*, pages 71–78. ACM, 2005.

[17] H. W. Hethcote. The Mathematics of Infectious Diseases. *SIAM review*, 42(4):599–653, 2000.

[18] L. Hufnagel, D. Brockmann, and T. Geisel. Forecast and control of epidemics in a globalized world. *PNAS*, 101(42):15124–15129, 2004.

[19] T. Inglesby, J. Nuzzo, T. O'Toole, and D. Henderson. Disease Mitigation Measures in the Control of Pandemic Influenza. *Biosecurity and Bioterrorism: Biodefense Strategy, Practice, and Science*, 4(4):366–375, 2006.

[20] S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, M. Martonosi, J. Rowland, and A. Varshavsky. Ranges of Human Mobility in Los Angeles and New York. In *PERCOM 2011*, pages 88–93. IEEE, 2011.

[21] S. Isaacman, R. Becker, R. Cáceres, M. Martonosi, J. Rowland, A. Varshavsky, and W. Willinger. Human Mobility Modeling at Metropolitan Scales. In *MobiSys 2012*, pages 239–252. ACM, 2012.

[22] M. J. Keeling, L. Danon, M. C. Vernon, and T. A. House. Individual identity and movement networks for disease metapopulations. *PNAS*, 107(19):8866–8870, 2010.

[23] W. Kermack and A. McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society A*, 115:700–721, 1927.

[24] X. Meng and L. Chen. The dynamics of a new SIR epidemic model concerning pulse vaccination strategy. *Applied Mathematics and Computation*, 197(2):582–597, 2008.

[25] A. Rinaldo, E. Bertuzzo, L. Mari, L. Righetto, M. Blokesch, M. Gatto, R. Casagrandi, M. Murray, S. M. Vesenbeckh, and I. Rodriguez-Iturbe. Reassessment of the 2010–2011 Haiti cholera outbreak and rainfall-driven multiseason projections. *PNAS*, 109 (17):6602–6607, 2012.

[26] L. Sattenspiel and K. Dietz. A Structured Epidemic Model Incorporating Geographic Mobility Among Regions. *Mathematical Biosciences*, 128(1):71–91, 1995.

[27] L. Sattenspiel and D. A. Herring. Simulating the Effect of Quarantine on the Spread of the 1918–19 Flu in Central Canada. *Bulletin of Mathematical Biology*, 65(1):1–26, 2003.

[28] B. Shulgin, L. Stone, and Z. Agur. Pulse Vaccination Strategy in the SIR Epidemic Model. *Bulletin of Mathematical Biology*, 60(6):1123–1148, 1998.

[29] F. Simini, M. C. González, A. Maritan, and A.-L. Barabási. A universal model for mobility and migration patterns. *Nature*, 484(7392):96–100, 2012.

[30] L. Stone, B. Shulgin, and Z. Agur. Theoretical Examination of the Pulse Vaccination Policy in the SIR Epidemic Model. *Mathematical and Computer Modelling*, 31(4): 207–215, 2000.

[31] Y. Tanahashi, J. R. Rowland, S. North, and K.-L. Ma. Inferring Human Mobility Patterns from Anonymized Mobile Communication Usage. In *MoMM 2012*, pages 151–160. ACM, 2012.

[32] Tatem, A.J. Côte d'Ivoire AfriPop Data 2010 (alpha version). Emerging Pathogens Institute, University of Florida, 2010. URL `http://www.clas.ufl.edu/users/atatem/index_files/CIV.htm`.

[33] J. Taubenberger and D. Morens. 1918 Influenza: The mother of all pandemics. *Rev Biomed*, 17:69–79, 2006.

[34] J. Truscott and N. M. Ferguson. Evaluating the Adequacy of Gravity Models as a Description of Human Mobility for Epidemic Modelling. *PLOS Computational Biology*, 8(10):e1002699, 2012.

[35] United Nations, Department of Economic and Social Affairs. World Population Prospects, the 2010 Revision, 2010. URL `http://esa.un.org/unpd/wpp/index.htm`.

[36] World Health Organization. Summary of probable SARS cases with onset of illness from 1 November 2002 to 31 July 2003, 2004. URL `http://www.who.int/csr/sars/country/table2004_04_21/en/index.html`.

[37] G. Zaman, Y. Han Kang, and I. H. Jung. Stability analysis and optimal vaccination of an SIR epidemic model. *BioSystems*, 93(3):240–249, 2008.

[38] G. Zaman, Y. H. Kang, and I. H. Jung. Optimal treatment of an SIR epidemic model with time delay. *BioSystems*, 98(1):43–50, 2009.

Exploiting Cellular Data for Disease Containment and Information Campaigns Strategies in Country-Wide Epidemics

A. Lima, M. De Domenico, V. Pejovic, and M. Musolesi
School of Computer Science, University of Birmingham, United Kingdom

Human mobility is one of the key factors at the basis of the spreading of diseases in a population. Containment strategies are usually devised on movement scenarios based on coarse-grained assumptions. Mobility phone data provide a unique opportunity for building models and defining strategies based on very precise information about the movement of people in a region or in a country. Another very important aspect is the underlying social structure of a population, which might play a fundamental role in devising information campaigns to promote vaccination and preventive measures, especially in countries with a strong family (or tribal) structure.

In this paper we analyze a large-scale dataset describing the mobility and the call patterns of a large number of individuals in Ivory Coast. We present a model that describes how diseases spread across the country by exploiting mobility patterns of people extracted from the available data. Then, we simulate several epidemics scenarios and we evaluate mechanisms to contain the epidemic spreading of diseases, based on the information about people mobility and social ties, also gathered from the phone call data. More specifically, we find that restricting mobility does not delay the occurrence of an endemic state and that an information campaign based on one-to-one phone conversations among members of social groups might be an effective countermeasure.

I. INTRODUCTION

Health and well-being of populations are heavily influenced by their behaviour. The impact of the habits and local customs, including patterns of interactions and mobility at urban and regional scales, on health issues is remarkable [1]. The diffusion of mobile technology we are experiencing nowadays gives scholars an unprecedented opportunity to study massive data that describe human behavior [2]. An increasing number of people carries smart mobile phones, equipped with many sensors and connected to the Internet, for the whole day [3]. Data coming from a large number of people can describe trends in the macroscopic behavior of populations [4–6]. The results of the analysis of these trends can be directly applied to a number of real-world scenarios, and, more in general, to several applications where cultural and local differences play a central role. Analyzing this kind of data can provide invaluable help to support the decision-making process, especially in critical situations. For this reason, many public and private organizations are nowadays increasingly adopting a data-centric approach in their decisional process [7]. We believe that this strategy can be particularly useful in developing countries, which might have a lacking infrastructure[1].

Among the issues that developing countries are facing today, healthcare is probably the most urgent [9]. In these countries the effectiveness of campaigns is often reduced due to low availability of data, inherent limits in the infrastructure and difficult communication with the citizens, who might live in vast and remote rural areas. As a result, action plans are difficult to deliver. However, we believe that a data-centric approach can be an innovative and effective way to address these issues.

In this paper, we focus on containment of epidemics. We use movement data extracted from the registration patterns in a cellular network to evaluate the influence of human mobility on the spreading of diseases in a geographic area. In particular, we utilize this model to investigate how infectious agents might spread to distant locations because of human movement in order to identify optimal strategies that can be adopted to contrast the epidemics. We also evaluate how the collaborative effort of the population can be crucial in critical scenarios. For the reasons we mentioned before, in countries that are facing development challenges, vaccination campaigns are often hard to advertise to the population. Lack of communication and information is believed to be among the main causes of failure for immunization campaigns. The same applies to awareness campaigns that try to promote prophylaxis procedures that reduce the occurrence of contagion. However, in these cases, we argue that a collaborative effort leveraging individual social ties can be effective in propagating effective information (i.e., a sort of "immunizing information") to a widespread audience. Moreover, information received by people who are socially close can have a higher chance of leading to an actual action.

A large body of research has been conducted on models that describe the diffusion of diseases, with a particular recent interest on the role that human movement plays in spreading infections in large geographic areas [10–12], and also on the impact of human behavior on the spreading itself [12, 13]. With respect to the state of the art,

[1] We use the term "developing" to indicate countries that are assigned a low Human Development Index (HDI) by United Nations Statistics Division. We are aware of the limitations of this classification. As reported by UN, *the designations "developed" and "developing" are intended for statistical convenience and do not necessarily express a judgment about the stage reached by a particular country or area in the development process* [8].

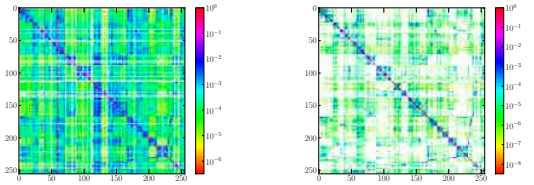
the main contributions of this paper can be summarized as follows:

- We propose an epidemic model based on a network of geographic metapopulations, which describes how people move between different geographic regions and spread the disease.
- We evaluate containment techniques based on the restriction of mobility of the most central areas. The centrality of the areas is extracted by building a movement network between all the geographic areas based on the mobility patterns of the individuals.
- We extend the model with a competing information spreading where *distance contagion* might take place. In other words, we study the dynamics of the system considering three characterizing aspects of the problem: the disease epidemics, human mobility and information spreading. This epidemics represents the diffusion of information related to measures to prevent or to combat the diseases, such as information about the ongoing vaccination and prevention campaigns in a certain area or actions that will help to limit spread of the infection, such as boiling water or avoiding contacts with people that are already ill.
- We evaluate the models by using the data provided by the Orange "Data for Development" [14]. We discuss the effectiveness of the containment strategies and, in particular, for the information dissemination strategy, we identify the degree of participation that is required to make it successful.
- We observe that restricting mobility by disallowing any movement from and to a limited set of subprefectures does not delay the occurrence of the endemic state in the rest of the country. We also find that a collaborative effort of prevention information spreading can be an effective countermeasure.

This paper is organized as follows. In Sec. II we briefly describe the four different datasets provided by Orange and we specify how they are used in the present study. In Sec. III we introduce our two models for epidemics and information spreading by taking into account human mobility and call patterns observed in Ivory Coast. In Sec. IV we present the results obtained by simulating several epidemics scenarios and evaluating mechanisms to contain the epidemic spreading of diseases. Finally, in Sec. V we summarize our main findings and we propose how the present study can be improved if more detailed data about mobility and calls will be available.

II. OVERVIEW OF THE DATASET

The data provided for the D4D challenge [14] consist of four datasets (identified by the labels SET1, SET2, SET3, SET4), containing information about user mobility and call patterns at various levels of granularity and time duration. We will now discuss how these datasets can be used to build a model which accounts for user mobility and information spreading.



(a) (b)

Figure 1: Logarithmic representation of the calls matrix (a) and the mobility matrix (b). Null values are indicated using the white color. For both matrices highest values are mostly concentrated along the diagonal, showing that communication and movement between sub-prefectures is highly uncommon. However, the calls matrix is visibly denser than the mobility matrix, confirming that phone contacts between different sub-prefectures are more usual than movement.

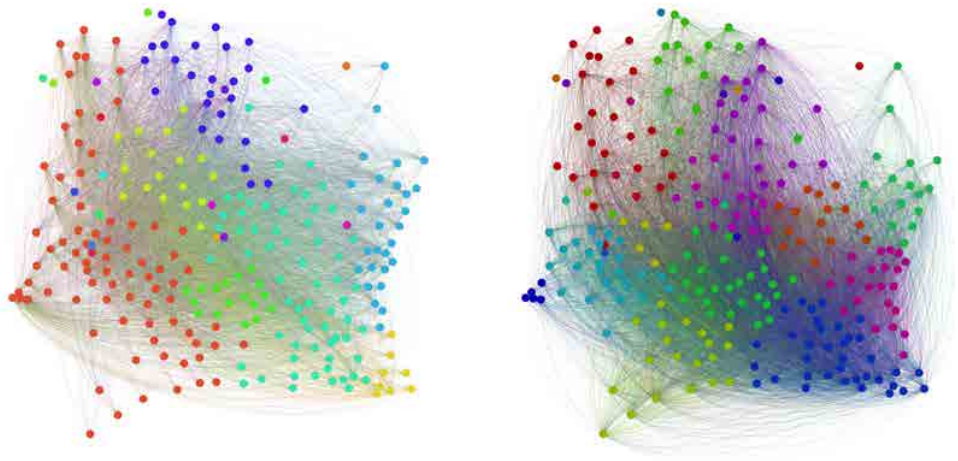
Two datasets contain information about mobility and communication patterns at macroscopic level. More precisely:

- The SET1 dataset contains the number and the duration of calls between pairs of cell phone towers, aggregated by hour. This dataset provides macroscopic information about communication in the country. We associate cell phone towers with the sub-prefecture they are located in, by using the supplied geographic position. Then, we evaluate the probability of a call being established between sub-prefectures i and j with:

$$c_{ij} = \frac{\mathcal{C}_{ij}}{\sum_k \mathcal{C}_{ik}}, \tag{1}$$

  where $\mathcal{C}_{ij}$ is the number of phone calls initiated from the sub-prefecture i and directed to the sub-prefecture j, during the entire period of observation. The term at denominator indicates the total communication flux between every pair of sub-prefectures and it is used to normalize the probability to check with Antonio. Using these values we build a calls matrix C, shown in Fig. 1(a). This matrix also shows high values along the diagonal, but it is distinctly denser, showing that calls between sub-prefectures are more common than movement. The vertical line at $x = 60$ identifies calls directed to the sub-prefecture that contains the capital.

- The SET3 dataset contains the trajectories of 50,000 randomly-selected individuals, at a sub-



(a) (b)

Figure 2: Geographic network obtained from call logs (a) and mobility traces (b). Color is used to indicate the community structure: nodes within the same community are represented with the same color.

prefecture level resolution, for five months.[2] This dataset can be used to estimate the probability that an individual moves from the sub-prefecture i to the sub-prefecture j:

$$m_{ij} = \frac{\sum_u \mathcal{M}_{ij}^u}{\sum_k \sum_u \mathcal{M}_{ik}^u}, \quad (2)$$

where $\mathcal{M}_{ij}^u$ is the number of times user u moves from the sub-prefecture i to j. The numerator counts how many times users who are in i move to j; the denominator normalizes this number by the total number of transitions from i to any sub-prefecture k. Using these values we build a mobility matrix M, shown in Fig. 1(b). By using this matrix, we model human mobility in the country as a Markov process [15]. We observe that the matrix is quite sparse and the highest values are concentrated along the diagonal. As the representation is in logarithmic scale, this demonstrates that the movement between sub-prefectures is present, but rather uncommon.

In Fig. 2(a) and Fig. 2(b) we show the geographic networks of calls and mobility, respectively. Nodes are positioned using the geographic locations of the sub-prefecture they represent, and their color indicates the community structure of the network based on [16].

The other two datasets provide microscopic information about mobility and communication patterns between individuals. Although we do not use them for the analysis in this paper, we now briefly outline how they could be used:

- The SET2 dataset contains fine-grained individual trajectories of 50,000 randomly sampled individuals over two-week periods. This dataset could be used to estimate the number of potential connections that an individual might have in a certain area, served by a cell phone tower.
- The SET4 dataset contains time-varying ego-networks of 5,000 users, describing the network of communication in time-slots of 2 weeks. If two users are connected by a link in a time-slot, it means that *at least* one call occurred during the two weeks under consideration[3]. The ego-network aggregated over the whole observation time, built considering every link that is present at least once,

[2] 17 sub-prefectures do not have any cell phone towers and for this reason do not appear in SET3. We discard these sub-prefectures from our analysis, since their users will be considered as belonging to nearby sub-prefectures.

[3] We have found that 1.31% of the total number of edges in ego-networks connect pairs of users who are neither egos nor first-level neighbors: therefore, we do not consider such edges in our analysis.

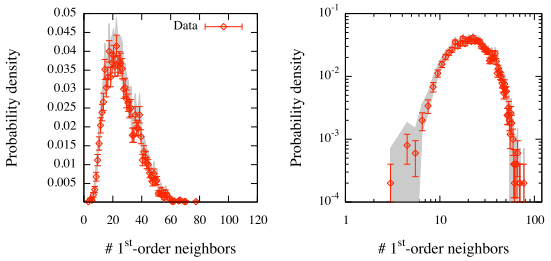


Figure 3: Distribution of friends for the ego-networks aggregated over time. Error bars indicate statistical uncertainty, while the shaded area represent 99% confidence intervals around the observed value.

describes the number of people contacted by an individual during the entire period. This dataset could be used to estimate the number of potential social connections that an individual might get in touch with. The degree distribution of the aggregate ego-network is shown in Fig. 3.

III. SPREADING MODELS

In this section we discuss two models: a model of disease spreading as a function of the mobility patterns of individuals between different geographic areas inferred from the cellular registration records and a model for information spreading among the same population, considering the social structure inferred from the call records. In the following section, we will evaluate the models using the data provided for the Orange Data for Development challenge.

A. Epidemic Spreading and Mobility

We will now present a model that represents the evolution of an epidemic taking place on a network of metapopulations. The aim of the model is to describe how the system evolves under the action of two processes, contagion and mobility. For this dataset, each metapopulation is composed by the individuals located in a particular subprefecture. Hence, the population is distributed in n different metapopulations, each having $N_i[t]$ individuals at time t. We make the simplifying assumption that there are no deaths and births in the considered time window, i.e., at each time $t = 1, 2, \dots T$ the total population is constant $\sum_{i=1}^{n} N_i[t] = N$.

We assume that contagion happens inside each metapopulation following a standard SIS model [17]. We indicate the number of infected and susceptible individuals at time t in a sub-prefecture i with $I_i[t]$ and $S_i[t]$, respectively. At each time t a person is either infected or susceptible, therefore $N_i[t] = I_i[t] + S_i[t]$.

Simultaneously, individuals move through the metapopulation network according to the *mobility matrix* M of dimension $n \times n$ extracted from the cellular traces. The generic element m_{ij} of the matrix represents the probability that a person moves from the metapopulation i to j, as described by Eq. 2[4]. This matrix describes how the state variables $N_i[t]$ evolve over time: $N_i[t+1] = \sum_{j=1}^{n} m_{ji} N_i[t]$. Under the assumption that individuals inside the classes I and S move consistently we can write the last relation also for the state variables $I_i[t]$ and $S_i[t]$[5]. The contagion-mobility combined system can then be described by the following set of equations:

$$
\begin{aligned}
I_i[t+1] &= \sum_{j=1}^{n} m_{ji} \left[I_j[t] + \lambda \frac{S_j[t]}{N_j[t]} I_j[t] - \gamma I_j[t] \right] \\
S_i[t+1] &= \sum_{j=1}^{n} m_{ji} \left[S_j[t] - \lambda \frac{S_j[t]}{N_j[t]} I_j[t] + \gamma I_j[t] \right],
\end{aligned}
$$

for each sub-prefecture $i = 1, 2, \dots, n$, with λ being the product of contact rate and contagion probability and γ being the recovery rate. The formulae inside the square brackets describe the evolution of n SIS models, one for each metapopulation. They are multiplied for the elements of the mobility matrix, which accounts for individuals moving between metapopulations.

This analytical model describes the expected outcome of a stochastic model where the following actions occur at each time step:

1. Each infected person in the sub-prefecture j causes the infection of new $\lambda \frac{S_j}{N_j}$ individuals inside j. This step is repeated for each sub-prefecture.

2. A new position i is assigned to each individual in the sub-prefecture j according to the probability density function $[m_{j1}, m_{j2}, \dots, m_{jn}]$. This step is repeated for each sub-prefecture.

B. Information Spreading

The model we presented in the last section tries to reproduce the spreading of a disease in a population where individuals change locations over time. The aim of this

[4] In general, this matrix can be time-varying, and it can be adjusted according to seasonal trends or real-time data at each step, for example following estimates based on historical data. In particular, this matrix can be used to study the impact of policies in real-time. However, in order to simplify the presentation, we use a matrix not changing over time. The treatment can be generalized, also applying the recent theoretical results related to time-varying networks [18, 19].

[5] This assumption can also be relaxed when data about the different classes of individuals is available, i.e., when a matrix for each class can be defined.

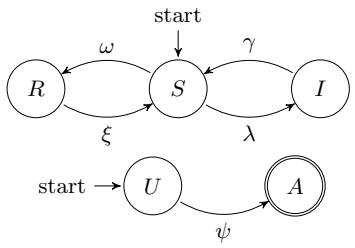


Figure 4: State machines describing the state transitions of a person with respect to the disease contagion (R=Resistant, S=Susceptible and I=Infected) and with respect to the information spreading (U=unaware, A=aware), respectively. A person starts in the susceptible and unaware states. We assume that aware individuals spread the information and cannot go back to the unaware state.

work is to analyze some scenarios and study the effectiveness of some containment techniques. In particular, as anticipated, we would like to investigate if a collaborative effort of the population is able, in theory, to reduce considerably the spread of the disease and what proportions should it have to be effective. More precisely, the population can disseminate information through personal social ties *immunizing*, such as information about prevention techniques, hygiene practises, advertisement of nearby vaccination campaigns and in general any information that can lead to a reduction of the number of contagion events.

In order to take into consideration these aspects, we now use a SIR model for each metapopulation, so that each person either belongs to the susceptible (S), infected (I) or resistant (R) category. At the same time, another simultaneous epidemic happens on the network of metapopulations, disseminating information that can make individuals resistant to the disease. In fact, a person also belongs to the category of unaware (U) or aware (A) individuals, with respect to the immunizing information. More formally, we have that $N_i[t] = I_i[t] + S_i[t] + R_i[t] = A_i[t] + U_i[t]$.

It is worth noting that this "immunizing epidemic" goes *beyond* the boundaries of metapopulations (subprefectures): in other words, it is a *distance contagion*. It is also important to remark that the states "aware" and "resistant" are substantially different. An unaware person that receives the information (i.e. has an "information contact") becomes aware with rate ψ; since the person is aware, he or she will start spreading the information as well. An infected person that receives the information becomes immune with rate ω. Additionally, individuals who have acquired immunity through information can lose it with rate ξ. The transition rates between states are summarized in Fig. 4. The model can be described by the following set of equations, specifying how state vectors evolve over time:

$$
\begin{aligned}
I_i[t+1] &= \sum_{j=1}^{n} m_{ji}\left[I_j[t] \ + \lambda\frac{S_j[t]}{N_j[t]}I_j[t] - \gamma I_j[t]\right] \\
S_i[t+1] &= \sum_{j=1}^{n} m_{ji}\Bigg[S_j[t] \ - \lambda\frac{S_j[t]}{N_j[t]}I_j[t] + \gamma I_j[t] + \xi R_j[t] + \\
&\qquad - \omega S_j[t]\frac{\sum_{k=1}^{n} c_{kj}A_k[t]}{\sum_{k=1}^{n} c_{kj}N_k[t]}\Bigg] \\
R_i[t+1] &= \sum_{j=1}^{n} m_{ji}\left[R_j[t] - \xi R_j[t] + \omega S_j[t]\frac{\sum_{k=1}^{n} c_{kj}A_k[t]}{\sum_{k=1}^{n} c_{kj}N_k[t]}\right] \\
A_i[t+1] &= \sum_{j=1}^{n} m_{ji}\left[A_j[t] + \psi U_j[t]\frac{\sum_{k=1}^{n} c_{kj}A_k[t]}{\sum_{k=1}^{n} c_{kj}N_k[t]}\right] \\
U_i[t+1] &= \sum_{j=1}^{n} m_{ji}\left[U_j[t] \ - \psi U_j[t]\frac{\sum_{k=1}^{n} c_{kj}A_k[t]}{\sum_{k=1}^{n} c_{kj}N_k[t]}\right] \qquad (3)
\end{aligned}
$$

for every $i = 1, 2, \ldots, n$. The fraction $\frac{\sum_{k=1}^{n} c_{kj}A_k[t]}{\sum_{k=1}^{n} c_{kj}N_k[t]}$ represents the probability that a call from an aware person occurs in the metapopulation j. It models the distance-contagion, and it is possible to verify that if the matrix is identical (absence of contacts between populations) it reduces to $A_k[t]/N_k[t]$, falling back to a model where contagion occurs only inside metapopulations.

This analytical model describes the expected value of a stochastic model where the following actions occur at each time step t:

1. Each infected person in the sub-prefecture j causes $\lambda\frac{S_j}{N_j}$ new individuals to get infected, inside j. This step is repeated for each sub-prefecture.
2. Each unaware person in the sub-prefecture j becomes aware with probability $\psi\frac{\sum_{k=1}^{n} c_{kj}A_k[t]}{\sum_{k=1}^{n} c_{kj}N_k[t]}$. This step is repeated for each sub-prefecture.
3. Each person in the sub-prefecture j who is susceptible, becomes resistant with probability $\omega\frac{\sum_{k=1}^{n} c_{kj}A_k[t]}{\sum_{k=1}^{n} c_{kj}N_k[t]}$. This step is reapeated for each sub-prefecture.
4. A new position i is assigned to each person in the sub-prefecture j according to the probability density function $[m_{j1}, m_{j2}, \ldots, m_{jn}]$. This step is repeated for each sub-prefecture.

IV. ANALYSIS

We initialize each scenario by allocating 22 million individuals (the estimated population size of Ivory Coast for July 2012 is 21,952,093 [20]) to different subprefectures across the country, according to the data in SET3. In each scenario we bootstrap the spreading process by infecting a fraction of the population (0.1%) distributed across metapopulations according to different criteria:

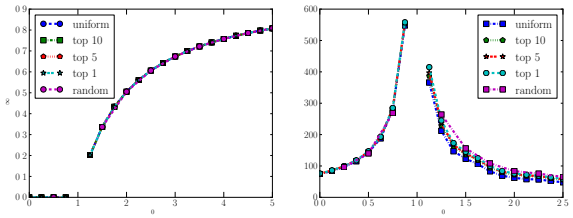


Figure 5: Fraction of infected population at the stationary state (left panel) and time required to reach the stationary state (right panel), for different values of r_0 and for different initial conditions. Missing values in the curves mean that, for the corresponding values, no stationary state is reached during the period of observation.

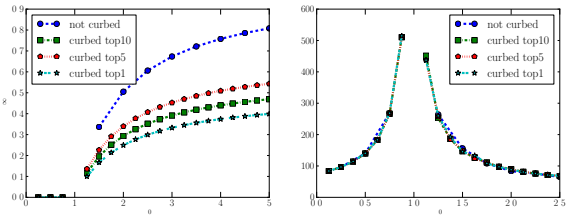


Figure 6: Fraction of infected population at the stationary state (left panel) and time required to reach the stationary state (right panel), for different values of r_0 when the epidemic starts from a random sub-prefecture, and different levels of geographic quarantine are applied. Missing values in the curves mean that, for the corresponding values, no stationary state is reached during the period of observation.

- Uniform distribution: every sub-prefecture gets a number of infected proportional to their population, i.e., every sub-prefecture has the same fraction of infected population.
- Random: a single sub-prefecture, chosen randomly, is the origin of the infection.
- Centrality based: the sub-prefectures are ordered by decreasing centrality values, then the first 1, 5 or 10 highest ranked sub-prefectures are chosen, as shown in Table I.

We study the evolution of the epidemics for a period of 6 months. We investigate multiple scenarios using the analytical model considering a large set of ranges for the key parameters. We conducted a series of Monte-Carlo simulations for multiple sets of parameters, confirming the validity of the analytical models presented in the previous section. In the following, we present results based on these models.

A. No Countermeasures

We will firstly explore the evolution of the epidemics in the case where no countermeasures are taken. In order to analyze the evolution of the system more clearly, we investigate two measures: the fraction of infected population i^∞ at the stationary state and the time required to reach the stationary state τ. In Fig. 5 we plot their values versus $r_0 = \frac{\lambda}{\gamma}$, which is the basic reproductive ratio of a classic SIS model [17]. As a future work, we plan to derive the analytical form of the basic reproductive ratio of our models, which take into account mobility and information spreading. We observe that for $r_0 = \frac{\lambda}{\gamma} < 1$ there is no endemic state (i.e., the final fraction of infected population is zero), whereas for $r_0 > 1$ a non-null fraction of population is infected. Values for $r_0 = 1$ are missing since no stationary state is reached within our observation window. In other words, for this particular scenario, experimental results show that the basic reproductive ratio of our model is very close to r_0; we expect this to be a consequence of the low inter-subprefectures mobility. We can also notice that the initial conditions do not affect i^∞ at all. Before the critical point (i.e., $r_0 = 1$) the choice of the initial conditions has also no impact on the delay time, whereas for $r_0 > 1$ it slightly affects the delay: epidemics that initially involves more sub-prefectures are slightly faster than the others.

Betweeness	Closeness	Degree	Eigenvalue
60	60	60	60
39	58	58	58
89	39	39	39
58	69	69	69
75	138	138	250
144	250	64	138
138	64	144	64
165	144	250	144
212	182	122	122
168	122	182	182

Table I: Highest ranked sub-prefectures, according to different definitions of centrality. We observe that the sets of the top 10 sub-prefectures ordered by centrality are very similar.

B. Geographic Quarantine

We now analyze the effects of curbing on the mobility between sub-prefectures, i.e., forbidding all the incoming and outgoing movement of a group of sub-prefectures. In order to do so, we calculate the centrality values of each sub-prefecture in the mobility matrix. We present the results for eigenvalues centrality. As it is possible to observe in Tab. I, the ranking based on other centralities is very similar. Then, for the quarantine operations, we select those with the highest centrality values. From a practical point of view, this is achieved by simply changing the $i-$th row and column in the mobility matrix, so that all the elements m_{ij} and m_{ji} are null, except for the elements $m_{ii} = 1$. For these scenarios, we randomly

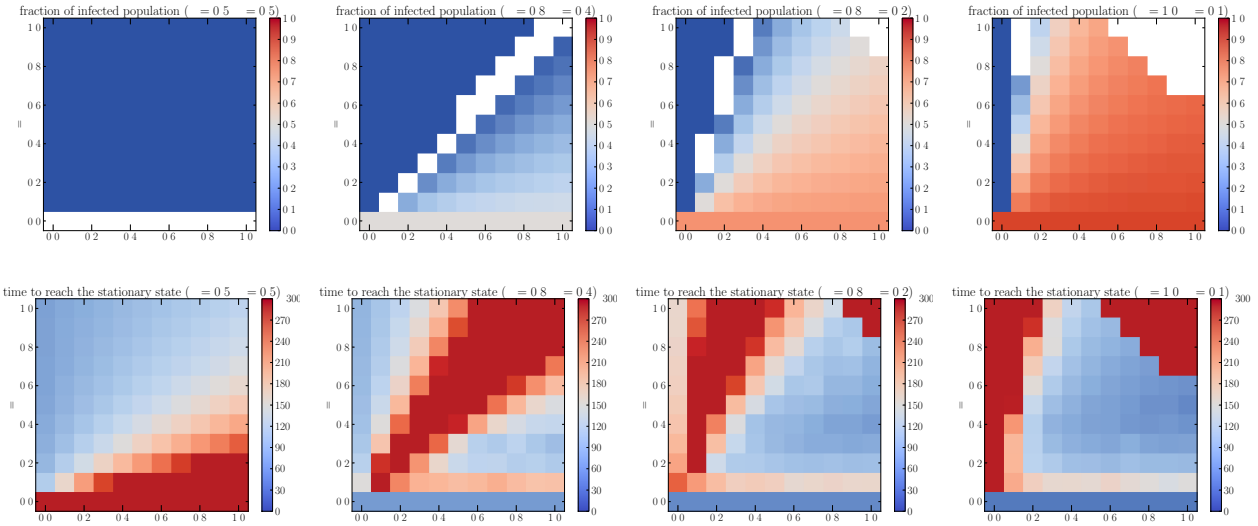


Figure 7: Fraction of infected population at the stationary state (top row) and time required to reach the stationary state (bottom row), for different values of $r_0 = \frac{\lambda}{\gamma}$ (from left to right 1, 2, 4, 10, respectively). White spaces show that no stationary state is reached during the period of observation.

choose a single sub-prefecture where the initial individuals are infected, and then we average i^∞ and τ over all runs. As shown in Fig. 6, the fraction of the infected population is sensibly affected by this measure, as the population inside the quarantined areas is protected from contagion. However, contrary to the intuition, the delay is not affected by the quarantine, even when the countermeasures involves 10 sub-prefectures, which account for almost half population. This suggests that such an invasive, expensive and hard to enforce measure reduces considerably the endemic size, but does not slow down the disease spreading in the rest of the country. For this reason, we now investigate a radically different approach to protect the population.

C. Information Campaign (Social Immunization)

We now show how a collaborative information campaign could help in contrasting the spread of the disease, following the model we presented in the last section. We initialize the scenario by distributing the immunizing information to 1% of the population, randomly chosen regardless of their location. These people will be informed and will be instructed to spread the information. In other words, we assume that they will contact their social connections, according to the call matrix.

In Fig. 7 we show the density plots describing i^∞ and τ for various values of r_0 (, for a subset of scenarios where $\omega = \psi$, i.e., when the information that spreads among the population has the same chance to immunize a person and to involve the person in the spreading process. This is consistent with a scenario where the same set of people who become aware also become immunized by the information they have received. Blank squares show that a stationary state was not reached for the corresponding set of parameters. The figure shows how contagious (ω=ψ) the immunizing information has to be with respect to how often people "forget" (ξ) in order to slow down the disease considerably and to reduce the endemic cases. When $\omega = \psi = 0$ we fall back to the model without information spreading, and the value of ξ does not affect i^∞ and τ. For $\omega = \psi > 0$ and $\xi = 0$ the fraction of infected population goes to zero in all cases, because the number of people aware of the information does not decrease, thus increasing the number of new immunized individuals at each step. We can notice that even for low values of participation ω and for information that gives temporary immunization ($\psi > 0$), the final fraction of infected individuals is considerably lower than in the case where no countermeasures are taken.

In Figs. 8 and 9 we show the density plots for ω and ψ when ξ is constant. In particular, we analyze the scenario for $\xi = 0$ (Fig. 8), which represents for example a scenario where the immunizing information is about vaccination campaigns (individuals who have been administered vaccination do not lose immunity). For every combination of parameters we have absence of endemic state even with the highest considered value of r_0. The two parameters that represent how individuals are likely to get involved both in the immunization and in the information spreading (ω and ψ) seem to have the same impact on the delay of the infection.

The value $\xi = 0.5$ (Fig. 9) describes the scenario when the information is about a good practice (e.g., boiling water, using mosquito nets, etc.), which loses its effec-

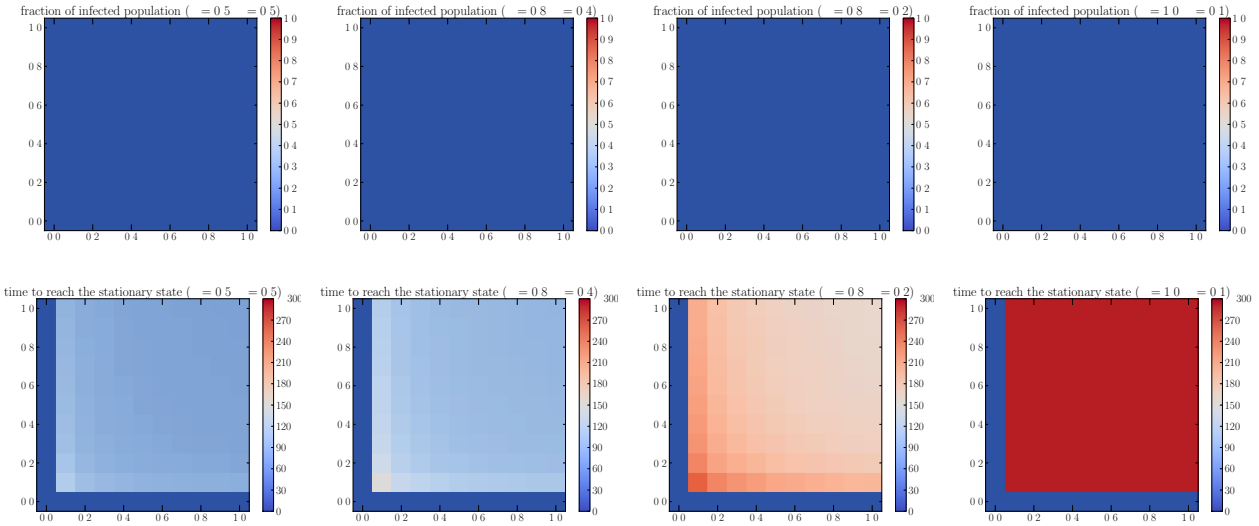


Figure 8: Fraction of infected population at the stationary state (top row) and time required to reach the stationary state (bottom row), for different combinations of $\frac{\lambda}{\gamma}$ (from left to right 1, 2, 4, 10, respectively) and $\xi = 0$. White spaces show that no stationary state is reached during the period of observation.

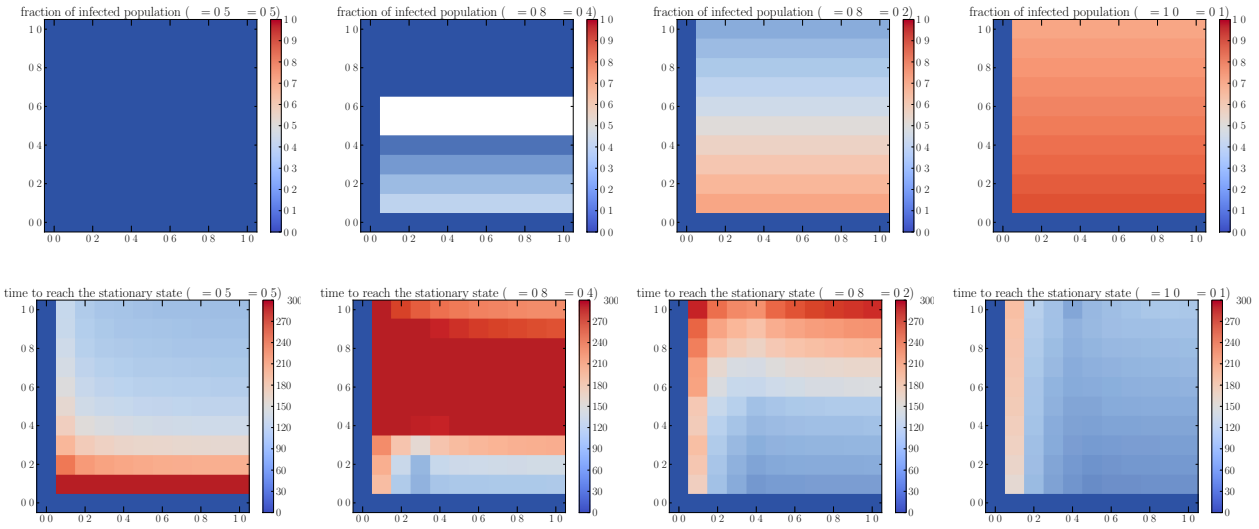


Figure 9: Fraction of infected population at the stationary state (top row) and time required to reach the stationary state (bottom row), for different combinations of $\frac{\lambda}{\gamma}$ (from left to right 1, 2, 4, 10, respectively) and $\xi = 0.5$. White spaces show that no stationary state is reached during the period of observation.

tiveness or it is stopped being used by a person with rate ξ. For this case we can notice that the fraction of infected population is independent from ψ, as rows in the density plot are of the same color. This suggests that, for this scenario, the rate at which people lose immunity does not affect the size of the endemic state.

V. CONCLUSIONS

In this paper we have presented a model that describes the spreading of disease in a population where individuals move between geographic areas, extracted from cellular network records. We have showed the evolution of the disease and we have evaluated two types of countermeasures, namely the quarantine of central geographic areas and a collaborative "viral" information campaign among

the population, by inferring the underlying social structure from the call records.

Our future research agenda includes the investigation of analytical aspects of the model, such as the derivation of the critical reproductive ratio R_0, i.e., the value that corresponds to the transition between an endemic and an endemic-free infection. Currently, the model is based on the assumption of a static mobility matrix: our goal is to refine the model by introducing time-dependent matrices, also exploring the application of the recent theoretical results related to temporal networks. We also plan to refine the model introducing specific contact rates for each metapopulation, potentially based on more fine-grained information about the number of encounters and the number of calls of each individual, if available. Finally, we plan to explore hybrid countermeasures, such as concurrent partial restrictions of mobility and targeted information campaigns.

Acknowledgments

The authors thank Charlotte Sophie Mayer for useful and fruitful discussions. This work was supported through the EPSRC Grant "The Uncertainty of Identity: Linking Spatiotemporal Information Between Virtual and Real Worlds" (EP/J005266/1).

[1] C.G. Helman et al., *Culture, health and illness.* (Arnold, Hodder Headline Group, London, United Kingdom, 2001), No. Ed. 4.

[2] D. Lazer, A.S. Pentland, L. Adamic, S. Aral, A.L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, et al., *Life in the network: the coming age of computational social science* Science **323**, 721 (2009).

[3] A. T. Campbell, S. B. Eisenman, N. D. Lane, E. Miluzzo, R. Peterson, H. Lu, X. Zheng, M. Musolesi, K. Fodor, and G.-S. Ahn, *The Rise of People-Centric Sensing* IEEE Internet Computing Special Issue on Mesh Networks (2008).

[4] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A.-L. Barabási, *Structure and tie strengths in mobile communication networks* Proceedings of the National Academy of Sciences **104**, 7332 (2007).

[5] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, *Understanding individual human mobility patterns* Nature **453**, 779 (2008).

[6] N. Eagle and A. Pentland, *Eigenbehaviors: Identifying structure in routine* Behavioral Ecology and Sociobiology **63**, 1057 (2009).

[7] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers, *Big data: The next frontier for innovation, competition, and productivity* McKinsey Global Institute (2011).

[8] United Nations Statistics Division, Standard Country and Area Codes for Statistical Use.

[9] *The health of the people the African regional health report* (World Health Organization, Regional Office for Africa, Brazzaville, Republic of Congo, 2013).

[10] V. Colizza, A. Barrat, M. Barthelemy, A.-J. Valleron, and A. Vespignani, *Modeling the Worldwide Spread of Pandemic Influenza: Baseline Case and Containment Interventions* PLoS Med **4**, e13 (2007).

[11] J. M. Epstein, D. M. Goedecke, F. Yu, R. J. Morris, D. K. Wagener, and G. V. Bobashev, in *Controlling Pandemic Flu: The Value of International Air Travel Restrictions*, PLoS ONE **2**, e401 (2007).

[12] S. Meloni, N. Perra, A. Arenas, S. Gómez, Y. Moreno, and A. Vespignani, *Modeling human mobility responses to the large-scale spreading of infectious diseases* Scientific Reports **1**, (2011).

[13] S. Funk, M. Salathé, and V. A. A. Jansen, *Modelling the influence of human behaviour on the spread of infectious diseases: a review* Journal of The Royal Society Interface **7**, 1257 (2010).

[14] V.D. Blondel, M. Esch, C. Chan, F. Clerot, P. Deville, E. Huens, F. Morlot, Z. Smoreda, and C. Ziemlicki, *Data for Development: the D4D Challenge on Mobile Phone Data* arXiv preprint arXiv:1210.0137 (2012).

[15] J. R. Norris, *Markov Chains* (Cambridge University Press, Cambridge, United Kingdom, 1998).

[16] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, *Fast unfolding of communities in large networks* Journal of Statistical Mechanics: Theory and Experiment **2008**, P10008 (2008).

[17] M. J. Keeling and P. Rohani, *Modeling infectious diseases in humans and animals* (Princeton University Press, Princeton, NJ, 2011).

[18] J. Tang, S. Scellato, M. Musolesi, C. Mascolo, and V. Latora, *Small-world Behavior in Time-varying Graphs* Physical Review E **81**, (2010), 055101(R).

[19] P. Holme and J. Saramäki, *Temporal networks* Physics Reports **519**, (2012).

[20] CIA, The World Factbook, 2012.

Linking the Human Mobility and Connectivity Patterns with Spatial HIV distribution

Katarina Gavrić, Sanja Brdar, Dubravko Ćulibrk, Vladimir Crnojević

Faculty of Technical Sciences
Trg Dositeja Obradovića 6
21000 Novi Sad, Serbia

{kgavric, brdars, dculibrk, crnojevic}@uns.ac.rs

ABSTRACT

An increasing amount of geo-referenced mobile phone data enables the identification of behavioral patterns, habits and movements of people. With this data, we can extract the knowledge potentially useful for many applications including the one that we tackled in this study - understanding spatial variation of epidemics. We explored the datasets collected by a cell phone service provider and linked them to regional HIV prevalence rates estimated from publicly available surveys. For that purpose numerous features were extracted from mobility and connectivity traces and related to the level of HIV epidemic in 19 Ivory Coast regions. By means of regularized regression model we identified key elements that impact the rate of HIV infections and by visualization of frequent trajectories, inter-region migrations and communications we strived to explain the spatial structure of epidemics.

Categories and Subject Descriptors

D.3.3 [**Database Management**]: Database Applications – *Data mining.*

General Terms

Algorithms, Experimentation

Keywords

Data Mining, Spatial Analysis, Human Mobility

1. INTRODUCTION

The HIV pandemic has devastating effects on entire human population in Africa. Ivory Coast has a generalized HIV epidemic with the highest prevalence rate in the West African region, 3 percent [1, 2]. Although the prevalence rate appears to have remained relatively stable for the past decade, nowadays there are several studies which declare that this number is even increasing, especially due to war conflicts [3]. Deeper understanding of the epidemics can help stop this trend and find ways to suppress it. Modern technologies that deal with human mobility phenomena may help respond to that challenge. Mobile phone communications engender the era of the big data by creating huge amounts of call detail records (CDR). Cell phone service providers collect these records whenever a phone is used for text messages or calls. The records contain time of action, identifies of sender and receiver, and cell tower used in communication. In this way, mobile phones uncover approximate spatiotemporal localization of users and provide immense resource for analysis of human mobility and behavioral patterns [4, 5, 6]. Recent studies report on use of mobile phone data in applications with great practical importance such are urban planning [7], disaster management [8], transportation mode inference [9], and traffic engineering [10]. Currently, there is a growing interest in the mining of mobile phone data for epidemiological purposes. This can advance research in epidemiology by shedding light on relationships between disease distribution, spread and incidence on one side and migrations, everyday movements and connectivity of people on the other side. Up to now, only a few studies have used mobile phone data to quantify those relationships. Wesolowski and coworkers explored the impact of the human mobility to the spread of malaria [11]. They analyzed CDR data collected by a mobile phone service provider in Kenya over one year period and discovered how mobility patterns contribute to the spread of the disease beyond what could be possible just by insects. The other study carried by Martinez and coworkers [12] investigated effect of government alerts during H1N1 flu outbreak in Mexico on the diameter of mobility of individuals. Bengtsson et al. [8] estimated population movements from a cholera outbreak area and suggested using obtained information for disease surveillance and resolving priority in relief assistance. Those pioneering works announce the emerging field of digital epidemiology [13].

The study we describe here is the first attempt to use mobile phone data for exploring complex structure of HIV epidemics. A lot of scientific effort is devoted to identifying the driving factors of HIV spread. Most frequently mentioned are poverty, social instability and violence, high mobility, rapid urbanization and modernization. The differences among these factors could explain spatial disparity in prevalence rates. In the study of Messina et al. geographic patterns of HIV prevalence in Democratic Republic of Congo were examined [14]. They showed that spatial factors: prevalence level in the range of 25 km and distance to the urban areas are strongly connected to the risk of HIV infection. Impact of the migration on the spread of

This material was prepared for the Data for Development Challenge (D4D) 2013 by Orange mobile network.

HIV in South Africa was studied in [15] where authors developed a mathematical model to compare the effects of migration and associated risk behavior. In the early stage of epidemics migration impacts HIV progression by linking geographical areas of low and high risk, while in the later stage by increasing high-risk sexual behavior. However, the migration was quantified through surveys where participants were questioned about movement history and the study included only two migration destinations. Nowadays, when we are overwhelmed with the mobile phone data that provides us with insight into movements and activity of millions of people over large areas we can try to utilize it for new studies of the epidemiology of HIV.

Therefore, we conducted a comprehensive analysis of three data sets offered within the Data for Development (D4D) Challenge [16]. Our research was guided by the following hypothesis: the risks for spreading HIV infection are associated with spatial and behavioral factors that may be detected from the available collection of data. We were particularly interested in tracking population movements and inferring the communication strength between regions of Ivory Coast with different prevalence rates.

2. DATA

2.1 HIV Spatial Distribution

In the beginning of this challenge, participants were encouraged to combine D4D mobile phone datasets with other datasets and use other sources of information. Therefore, we used the data set provided by United States Census Bureau, which contains the information about AIDS pandemic and HIV seroprevalence (infection) in population groups in developing countries. Using that dataset, we estimated the HIV prevalence rate between 2008-2010 in order to compare the results with DHS data (Demographic and Health Surveys), obtained by the Ministry of Fight against AIDS and the National Institute of Statistics [17]). Their survey presents the results for ten existing administrative regions of the country. Each administrative region consists of several regions (Figure 1).

Figure 1. The regions of Ivory Coast

The results of estimation are presented in Table 1. The prevalence rate increased during several years. Since the data on which we based our estimation are mostly sampled on pregnant women, we can expect that those numbers are slightly overestimated due to the higher risk of infection among them [18]. Nevertheless, that does not impact the results obtained, because pregnant women are equally represented in all samples and we were only interested in the rank order of regions according to HIV prevalence.

REGIONS	2005	2008-2010
Centre – East (Moyen – Comoé)	5.8	9.17
South (Lagunes, Agnéby, Sud Comoé, Sud Bandama)	5.5	8.91
Centre (Lacs, N'zi Comoé, part of Vallée du Bandama)	4.8	8.56
South-West (Bas Sassandra)	4.2	6.94
Centre-West (Fromager, Haut Sassandra, Marahoué)	3.7	4.85
Centre-North (part of Vallée du Bandama)	3.6	6.29
West (Dix-Huit Montagnes, Moyen Cavally)	3.5	4.76
North-East (Zanzan)	3.3	4.56
North (Savanes)	3.2	5.46
North-West (Bafing, Denguélé, Worodougou)	1.7	4

Table 1. HIV prevalence rate by administrative regions

2.2 D4D Data Sets

Mobile phone data sets originate from the Orange service provider in Ivory Coast and are further processed into four different D4D sets. Three of these were used in our study: SET1, SET2 and SET3. Although SET4 provides connectivity at user level and could be very informative for HIV epidemiology, it lacks spatial information. Users' ids cannot be related to the ids in the second or the third set and therefore were not able to approximate their home locations.

SET1 contains antenna -to- antenna communication traffic flow of five millions Orange customers aggregated in one hour time resolution. Each record has originating and terminating antennas of calls, number of calls and overall duration.

In SET2, individual movement trajectories were approximated. The original data has been split into consecutive two-week periods. In each time period, 50, 000 of the users were randomly selected and assigned anonymized identifiers.

Insight into the long term mobility (6-months long observation period) is possible through SET3. Spatial resolution is reduced from towers to sub-prefectures (255 administrative units). Records of this set contain user id, time stamp and sub-prefecture ids.

3. REGIONAL CONNECTIVITY

Available data on HIV prevalence rates in Ivory Coast limit the spatial resolution of our study. Therefore, we focus on regions as spatial units in order to be able to relate knowledge extracted from D4D sets to spatial prevalence distribution.

3.1 Graph Representation

A graph-based analysis was carried out for SET1 and SET3. We inferred pairwise connectivity of regions by measuring the flow

of communications and migrations between them. We were curious to investigate whether regions with higher HIV prevalence rates are more connected than those with lower.

To infer the inter-region communication graph we teased out the information from SET1. The first step was to assign each antenna to its region. Then, the communication flow is further aggregated at the region level by summing all antenna level flows between different regions. Nodes in the graph represent regions and edges measure the strength of human interaction expressed through the number of calls. Graph loops – edges that start and end in the same region were excluded from this part of the study. The next step was the normalization of edge weights. Since regions are unevenly populated we took that into account by dividing the edge weights w_{ij} (sum of all calls during the 6-month period between region i and j and vice versa) with a product of population numbers N_i and N_j (population estimates from 2010 [19] were used). The product $N_i * N_j$ is an approximation of all possible communication links between people in two regions. Finally, we filtered all obtained pairwise weights to create a 3NN graph. By adding only three strongest links for each region we can inspect the major directions and hubs of communication flow in Ivory Coast. The graph is presented at Figure 2.

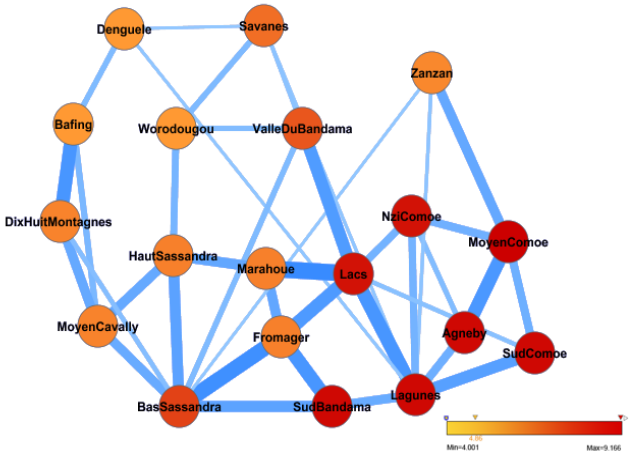


Figure 2. 3NN communication graph: Nodes represent Ivory Coast regions, arranged in geographical order and colored according to HIV prevalence rates. Links are inferred from inter –region communication flow during six months. Their color and width is proportional to the normalized flow between regions.

Nodes are geographically ordered and their colors indicate HIV infections rate: from red that denotes the regions severely affected by HIV to yellow for moderately affected regions. Added edges are presented with different widths and color intensities to highlight differences in their strength. We can notice that graph structure corresponds well with HIV spatial distribution; Southeast and South part of country that have higher prevalence rates, turn out to be the more connected part of the graph in terms of incidence edges, hubs and edge widths. The major hubs are Lagunes with 8 incidence edges, Bas Sassandra with 7 and Lacs with 6 are located in the area of the highest risk. Northwest part of the country is notably sparser in the graph, with no more than 3 thin incident edges. An interesting property revealed by this graph is that the gravitational law previously observed in inter-city commutations [20] is only partly supported. Although proximity of regions is correlated with the strength of the link between them, we have some exceptions as links from Zanzan and Denguélé to far away Lagunes rather than to some closer regions.

The inference of migration graph was done in way similar to that used for the communication graph. Here, the data source is SET3. First, we estimated for each user its home location in order to be able to assign him a home region. Then, we followed all his movements through time and detected transitions into other regions. Upon detection of a regional transition, the pairwise matrix of region to region migrations is updated - increased at position: home, detected host region. All users were processed in this manner and the final result is a pairwise matrix of overall migrations between regions, during the 6-month period. Before creating the 3NN graph the values in matrix were normalized. Here, the normalization was different than for SET1. From estimated home locations we can obtain the number of residents that were tracked in SET3 for each region. Edge weights w_{ij} (sum of all migrations between regions i and j and vice versa) with a sum of obtained resident numbers for region i and j: N_i and N_j. After this normalization step, 3NN graph was built. Its structure is presented in Figure 3. Normalized migration flows also support the spatial HIV distribution, since the graph is denser in the high risk area. The only link that significantly departs from the distribution and our expectations is the link between Worodougou and Sud Comoé region. A possible reason is that only 8653 residents of Worodougou were covered by SET3 and this number should be near 12 000 to be in accordance with the real population distribution. Nevertheless, even with correction that link would still exist. Rather, we can look at that link as a new insight. The strong connection may indicate the next hot spot of HIV epidemics and we can utilize that for prioritizing areas for intervention. Also, outliers can uncover strange occurrences in the field. In this particular case, the outlier could be related to mining of diamonds from Toubabouko field located in Worodougou and their export despite United Nations (UN) ban.

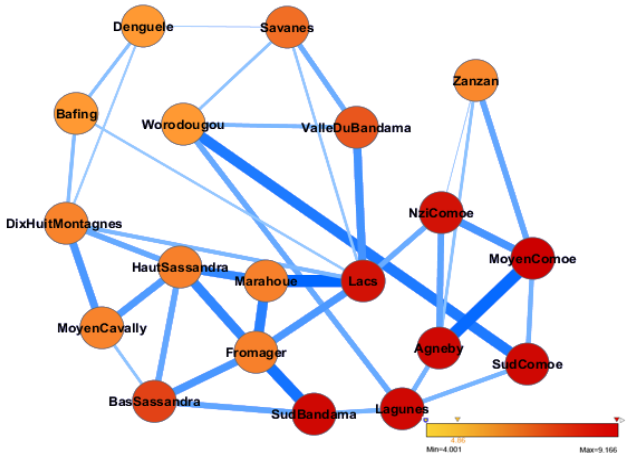


Figure 3. 3NN migration graph: The nodes represent Ivory Coast regions, arranged in geographical order and colored according to HIV prevalence rates. Links are inferred from inter–region migration flows during six months. Their color and width are proportional to normalized flow between regions.

3. 2 Frequent Trajectories

SET2 was analyzed using several geo-visualization techniques, with an emphasis on trajectory aggregation and clusterization. The initial idea was to determine the hubs with the highest level of connectivity and to identify the major routes taken among the hubs. Subsequently, we used route similarity clustering to identify standard paths taken among the hubs. The clustering was performed using the OPTICS (Ordering Points to Identify the Clustering Structure) algorithm that allows different distance functions to be applied. The idea is that two trajectories (P, Q) are repeatedly scanned in search for the closest pair of positions (D – distance threshold). In the course of scanning, two derivative distances are computed: the mean distance between the corresponding positions and a penalty distance. Skipping a position increases the penalty distance. Finding corresponding positions decreases the penalty distance. The final result is the sum of the two derivative distances. The size of the clusters obtained based on this distance measure, represents how frequently the route was used. We analyzed each sub-set separately and then combined them in order to get a representation for the entire set.

From Figure 4, it is obvious that main hubs are situated in the centre of each region and that communication is higher in the south part of the country. This confirms the hypothesis that the closer people are to the main routes and hubs, the higher is the chance to get infected with virus and also to transfer it.

Figure 4. Trajectory aggregation model based on OPTICS 'route similarity' clusterization

4. FEATURE EXTRACTION

Previous part of the study was more descriptive and focused on linking regions' connectivity inferred from D4D data with HIV spatial distribution. Here, we set up our work more explicitly. Numerous features were extracted in order to quantify behavioral and mobility patterns for each region. Than we built regression models and evaluate their performance in predicting regional prevalence rates. D4D SET1, 2 and 3 were analyzed separately.

From SET1 we extracted features related to intra-region communications. For each region we created average profiles of communication flow in one hour time resolution for weekdays and weekends. Profiles contain average number of calls and their average duration and they are normalized by the number of people in the region. Additionally we created aggregated features for night hours (22h-05h) during weekdays and weekends and also for whole days. In total, we derived 104 features from SET1.

Very often, the limited knowledge of an individual's trajectories can be significant for human mobility monitoring because individuals can be traced during certain period of time. SET2 contained high resolution trajectories of randomly sampled individuals over two-week periods. With features extraction, we gained the intervals when people are more active (working and non working days, working and non working hours, weekends, nights, etc.), as well as the home and visited regions per each user. We assumed that the time when people are more active increases the chance of their infection, as well as of virus transmission.

SET3 is firstly analyzed at user level, and then based on home location estimates, individual patterns of daily movements were aggregated into region - level features. We calculated various aspects of mobility such as gyration, radius, diameter and approximate sum of all distances that users travel [21, 22], counted numbers of distinct sub-prefectures that users visited within the home region and out of it during the 6-month period and under time constraints: only night hours or weekends. By tracking moving trajectories of users we determine in and out migrations for each region at different time scales: staying in host region more than 3, 5 or 10 days. Also, we measured how long on average, nonresidents stay in each region. Overall number of features from SET3 is 23.

5. REGRESSION MODELS

5.1 Elastic Net Predictive Model

The elastic net predictive model simultaneously does automatic variable selection and continuous shrinkage. It produces a sparse model with good prediction accuracy, while encouraging a grouping effect. The empirical results and simulations demonstrated good performance of the elastic net and its superiority over the lasso predictive model. The elastic net is particularly useful for problems where the number of features is higher that number of samples ($p >> n$). The prediction procedure can be divided into three steps: approximation of the unpenalized log-likelihood using iteratively reweighted least squares; application of soft-thresholding to take care of lasso contribution to the penalty; application of proportional shrinkage for the ridge penalty [23].

5.2 Ridge Regression

Ridge regression is a variant of ordinary multiple linear regression whose goal is to circumvent the problem of instability arising, amongst other, from colinearity of the predictor variables. It works with the original variables and tries to minimize penalized sum of squares. Like ordinary least squares, ridge regression includes all predictor variables, but typically with smaller coefficients, depending upon the value of the complexity parameter. The selection of ridge parameter plays an important role, because the adding of a small constant to the

diagonal elements of the matrix $X'X$ will improve the conditioning of a matrix [24].

6. RESULTS

Three types of experiments were performed using one of the regression models, for each set. For SET 1 and 3, we used elastic net predictive model and for SET 2 we used ridge regression.

Before learning the regression model we normalized the feature-space by dividing each feature with its mean value. Parameters of modes where estimated with leave-one-out cross-validation. The results of finally selected models are expressed through correlation coefficient and root mean square error (rmse) in Table 2.

SET	CORRELATION COEFFICIENT	RMSE
1	0.96	0.59
2	0.55	1.67
3	0.71	1.46

Table 2. Correlation coefficient and RMSE for models

Obtained models have different predictive powers. The model learned on SET1 performed the best probably due to high number of very detailed features that we extracted. The other reason could be in the fact that SET1 encompasses communications flow of 5 million people during the 6 months. Obviously this was enough to detect the regional patterns. Extracting the features from SET2 and SET3 is harder since complex dynamics of human movements is involved. Furthermore data sets covered fewer people and also shorter periods of time in case of SET2.

In Table 3 we report on a few interesting features for which we observed that in all validated models were stable – did not change the sign and they have high coefficients.

SET 1	DESCRIPTION	WEIGHT
Duration-w0h5	Average duration of calls during weekday in 05-06h time interval	0.88
NbVoice-w1h2	Average number of calls during weekend in 02-03h time interval	0.84
Duration-w1h22	Average duration of calls during weekend in 22-23h time interval	-1.00
SET 2	**DESCRIPTION**	**WEIGHT**
WeD	Time spent in each region during weekend days	0.69
NHo_We	Time spent in each region during weekend night hours (00 - 05h)	0.34
WoH	Time spent in each region during working hours (08 - 18h)	0.19
SET 3	**DESCRIPTION**	**WEIGHT**
RadiusNight	Maximum radius from home location during the night hours	-1.00
Gyration	Standard deviation from the average location of user	-0.77
InMigration	Counted movement of nonresident users from their home regions to observed region	0.46

Table 3. Features coefficient weights for 3 data sets

Features that stand out from SET1 are those related to the communication activities during night hours. The late night calls and their longer duration are positively associated with epidemic rate and they can be seen as indicators of risk behavior. On the other side duration of the early night calls is negatively associated.

After application of ridge regression prediction on features extracted from SET 2, the results showed that each of them have different informative weight. The most informative features were time spent in each region during weekend days, time spent in each region during weekend night hours and time spent in each region during working hours. Values across the regions indicate that higher activities in the sense of migration are related to the regions with higher risk. Also, the values of features across all regions showed that:

- Migrations of people are higher during the working hours, as well as one hour before and after work which can be explained with performing daily business duties and travel to and from work place.
- Migrations of people are higher during weekend due to the lack of specific contents in their own environment (malls, cinemas, sport contents etc.). Therefore, people are forced to travel to larger centers nearby in order to fulfill some of their secondary obligations.
- Migrations of people are higher during the weekend night hours (00-05h) which is the most important indicator. In this period, people usually go to larger centres looking for fun and entertainment, which significantly increases the risk of infection and transmission of infection.

In SET3, coefficients for regional average gyration and radius in the night hours have a negative sign. But this is not surprising; the studies have already shown that in the denser urban areas there is higher expectation of shorter movements [25]. The third feature listed in the table indicates that the more migration in the region the higher is the epidemic rate. This is something that we would intuitively expect and here the trained model learned it from our data.

7. CONCLUSION

This study showed how crude real world data can be used for significant knowledge extraction. We addressed the problems of HIV/AIDS spatial distribution prediction by analyzing human activity and mobility in the area of extracted features. However, the results leave a lot of room for improvement, especially in the field of defining the features which affect disease transmission the most. A detailed study on the feature selection methods is necessary and is a possible direction for further research. Also, the individual communication graphs which are geographically determined would be an immense source of information for uncovering the connectivity at a more detailed scale. They have significant potential to enable further progress in the domain of modeling communicable diseases [26].

8. ACKNOWLEDGMENTS

This work was partly supported by Serbian Ministry of Education and Science (Project III 43002) and COST Action IC0903-MOVE. Also, ours acknowledgements go to organizers of the Data for Development Challenge for sharing their data sets.

9. REFERENCES

[1] Kalipeni, E. and Zulu, L.C. HIV and AIDS in Africa: a geographic analysis at multiple spatial scales. *GeoJournal*, (04), 2012. doi: 10.1007/s10708-010-9358-6

[2] http://www.unaids.org/en/dataanalysis/datatools/aidsinfo

[3] Betsi, N.A. and Koudou, B.G. and Tschannen, A.B. and Pignol, A.M. and Ouattara, Y. and Madougou, Z. and Tanner, M. and Utzinger, J. Effect of an armed conflict on human resources and health systems in Côte d'Ivoire: prevention of and care for people with HIV/AIDS. *AIDS Care* 18: 356-365, 2006.

[4] Becker, R. and Cáceres, R. and Hanson, K. and Isaacman, S. and Loh, J.M. and Martonosi, M. and Rowland, J. and Urbanek, S. and Varshavsky, A. and Volinsky, C. Human mobility characterization from cellular network data. *Communications of the ACM*, 56(1): 74-82, 2013. doi: 10.1145/2398356.2398375

[5] Candia, J. and González, M.C. and Wang, P. and Schoenharl, T. and Madey, G. and Barabási, A.L. Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical*, 41(22): 224015, 2008. doi: 10.1088/1751-8113/41/22/224015

[6] Wesolowski, A. and Eagle, N. Parameterizing the dynamics of slums. *AAAI Symposium on Aritificial Intelligence and Development*, 2010.

[7] Becker, R. and Caceres, R. and Hanson, K. and Loh, J.M. and Urbanek, S. and Varshavsky, A. and Volinsky, C. A tale of one city: Using cellular network data for urban planning. *IEEE Pervasive Computing*, 10(4): 18-26, 2011. doi: 10.1109/MPRV.2011.44

[8] Bengtsson, L. and Lu, X. and Thorson, A. and Garfield, R. and von Schreeb, J. Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: a post-earthquake geospatial study in Haiti. *PLoS medicine*, 8(8): e1001083, 2011. doi:10.1371/journal.pmed.1001083

[9] Wang, H. and Calabrese, F. and Di Lorenzo, G. and Ratti, C. Transportation mode inference from anonymized and aggregated mobile phone call detail records. *Proceedings of the IEEE International Conference on Intelligent Transportation Systems*, 2010.

[10] Caceres, N. and Romero, L.M. and Benitez, F.G. and del Castillo, J.M. Traffic Flow Estimation Models Using Cellular Phone Data. *IEEE Transactions on Intelligent Transportation Systems*, 13(3): 1430 – 1441, 2012. doi:10.1109/TITS.2012.2189006

[11] Wesolowski, A. and Eagle, N. and Tatem, A.J. and Smith, D.L. and Noor, A.M. and Snow, R.W. and Buckee, C.O. Quantifying the impact of human mobility on malaria. *Science*, 338(6104): 267-270, 2012.

[12] Frias-Martinez, V. and Rubio, A. and Frias-Martinez, E. Measuring the impact of epidemic alerts on human mobility. PURBA2012 *Second Workshop on Pervasive Urban Applications*, 2012.

[13] Salathé, M. and Bengtsson, L. and Bodnar, T.J. and Brewer, D.D. and Brownstein, J.S. and Buckee, C. and Campbell, E.M. and Cattuto, C. and Khandelwal, S. and Mabry, P.L. and others. Digital Epidemiology. *PLoS Computational Biology*, 8(7): e1002616, 2012.

[14] Messina, J.P. and Emch, M. and Muwonga, J. and Mwandagalirwa, K. and Edidi, S.B. and Mama, N. and Okenge, A. and Meshnick, S.R. Spatial and socio-behavioral patterns of HIV prevalence in the Democratic Republic of Congo. *Social Science & Medicine*, 71(8): 1428-1435, 2010.

[15] Coffee, M. and Lurie, M.N. and Garnett, G.P. Modelling the impact of migration on the HIV epidemic in South Africa. *Aids,* 21(3): 343, 2007.

[16] Blondel, V.D. and Esch, M. and Chan, C. and Clerot, F. and Deville, P. and Huens, E. and Morlot, F. and Smoreda, Z. and Ziemlicki, C. Data for Development: the D4D Challenge on Mobile Phone Data. arXiv preprint *arXiv:1210.0137, 2013.*

[17] Enquête sur les Indicateurs du Sida, République de Côte d'Ivoire. *Ministère de la Lutte contre le Sida Institut National de la Statistique,* ORC Macro Calverton, Maryland, U.S.A., 2006.

[18] Gray, R.H. And Kigozi, G. and Serwadda, D. and Brahmbhatt, H. and Wabwire-Mangen, F. and Nalugoda, F. and Kiddugava, M. and Sewankambo, N. and Quinn, T.C. and Reynolds, S. J. And Wawer, M. J. Increased risk of incident HID during pregnancy in Rakai, Uganda: a prospective study. *Lancet*, 366. (9492): 1182-8, 2005.

[19] http://ivorycoast.humanitarianresponse.info

[20] Krings, G. and Calabrese, F. and Ratti, C. and Blondel, V.D. Urban gravity: a model for inter-city telecommunication flows. *Journal of Statistical Mechanics: Theory and Experiment*, (07): L07003, 2009.

[21] Gonzalez, Marta C and Hidalgo, Cesar A and Barabasi, Albert-Laszlo. Understanding individual human mobility patterns. *Nature*, 453(7196): 779-782, 2008.

[22] Csáji, Balázs Cs and Browet, Arnaud and Traag, VA and Delvenne, Jean-Charles and Huens, Etienne and Van Dooren, P and Smoreda, Zbigniew and Blondel, V.D. Exploring the mobility of mobile phone users. Physica A: Statistical Mechanics and its Applications, 2012.

[23] Zou, H. and Hastie, T. Regression Shrinkage and Selection via the Elastic Net, with Application to Microarrays. *Journal of the Royal Statistical Society*, Ser. B, 67, 301-320, 2006.

[24] El-Dereny, M. And Rashwan, N. I. Solving Multicollinearity Problem Using Ridge Regression Models. *International Journal of Contemporary Mathematical Sciences*, Vol. 6, no. 12, 585-600, 2011.

[25] Noulas, A. and Scellato, S. and Lambiotte, R. and Pontil, M. and Mascolo, C. A tale of many cities: universal patterns in human urban mobility. *PloS one,* 7(5), 2012.

[26] Bian, L. Spatial Approaches to Modeling Dispersion of Communicable Diseases - A Review. *Transactions in GIS*, 2012. doi: 10.1111/j.1467-9671.2012.01329.x

Using Mobile Phone Data to Supercharge Epidemic Models of Cholera Transmission in Africa: A Case Study of Côte d'Ivoir

Andrew S. Azman*[1], Erin A. Urquhart†[2], Benjamin Zaitchik‡[2] and Justin Lessler§[1]

[1]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA
[2]Department of Earth & Planetary Sciences, Johns Hopkins University , Baltimore, Maryland, USA

Abstract

Human movement is known to shape the transmission of infectious diseases however detailed data on human movement from West Africa are severely lacking. A cholera epidemic swept across West Africa this year, and computational models of transmission may be helpful to understand the dispersal pathway, and to help predict aspects of future epidemics. Here we use mobile phone records from 500,000 individuals in Côte d'Ivoir to parameterize a human mobility model and combine these results with detailed environmental data from the country to drive a stochastic cholera transmission model. We provide an example cholera transmission model and show how it can reproduce some general aspects of cholera epidemics seen in the region, including transmission hotspots. The results from this model (and extensions of it) can help improve our understanding of cholera transmission and guide targeted prevention and control efforts.

Introduction

Cholera is a waterborne bacterial disease responsible for an estimated 300,000 deaths, and millions of severe diarrhea cases each year[1]. The first cases of cholera from the current global pandemic appeared in Africa in the 1970's and large epidemics have occurred regularly throughout subsaharan and West Africa ever since. Côte d'Ivoire has suffered from multiple cholera outbreaks since the 1970s, and will likely remain at risk until the country's water and sanitation infrastructure is fundementally improved [2].

Attack rates in cholera epidemics throughout West Africa tend to vary by geographic area with heterogeneity seen at multiple spatial scales (e.g., heterogeneity in attack rate by district and by neighborhood within a city)[3]. The combination of how humans move between areas and how conducive environmental conditions are for cholera transmission in each of those areas plays a fundamental role in how cholera moves through a country. If we wish to understand the risk that different areas face from cholera emergence within a country, and perhaps target interventions accordingly, we must account for both of these factors.

Cholera is transmitted by ingestion of *Vibrio cholerae* bacteria, resulting from either direct transmission via fecal contamination of food, water, or fomites, without entry into a larger aquatic ecosystem, or

*aazman@jhsph.edu
†eurquhart@jhu.edu
‡bzaitchik@jhu.edu
§jlessler@jhsph.edu

through ingestion of seafood or water from aquatic reservoirs [4]. Environmental factors affect cholera transmission both through directly impacting the survival of the bacteria and through modifying individuals exposure to contaminated water. A large increase in the amount of water in an area may have two antagonistic effects. On the one hand, it may decrease cholera transmission potential by diluting the concentration of infectious *V. cholerae* in aquatic reservoirs. At the same time, heavy rain and flooding may increase exposure to contaminated drinking water by overwhelming the infrastructure and natural barriers that normally separated drinking water from waste water and feces.

While symptomatic cholera cases are likely to return home or visit a health care facility after infection, the likely location of their infection will depend on where they spend there day, and particularly where they consume food and water. Human movement is likely the primary driver of cholera spread throughout a country (and region) once an epidemic has begun [5], hence models of the spatial dispersal of cholera must account for these movement patterns. Models of infectious disease often rely on parametric models of human movement. In particular, gravity models are frequently used as they capture the ways in which people select destinations by both their size and distance for their current location [6]. The ways in which people move may vary largely between regions, an it is unlikely that models of movement in West Africa would be correctly parameterized using data from elsewhere in the world (e.g., North America, Europe).

Cell phone data provide a mechanism to develop and parameterize cholera transmission models that accurately reflect human movement in West Africa. Here we combine cell phone mobility trace data with high-resolution population density data to fit models of human movement within Côte dIvoire. Based upon the best fitting mobility model we develop a stochastic meta-population transmission model for cholera, incorporating fine-scale environmental data from climate models. We then use the transmission model to estimate the time-dependent risk of cholera throughout Côte dIvoire, estimate the timing of spread to different areas throughout the country under different scenarios, and finally explore simulations for the presence of transmission hotspots. We compare results broadly to existing cholera incidence data from Côte dIvoire, checking for similar seasonality and hotspots. The models presented here represent a first step in the creation of tools that can be used to understand disease transmission dynamics, predict the effect of interventions and help decision makers to more effectively plan and respond to cholera outbreaks.

Data

Four key data sources were used to construct detailed models of cholera transmission in Côte dIvoire: (1) human population data, (2) coarse cholera incidence reports from Côte d'Ivoir, (3) mobile phone records, and (4) meteorological data.

Population and Incidence Data
Human population density (Figure 2) estimates were derived from the Afripop database [7]. While few detailed data on cholera are available from past epidemics in Côte d'Ivoir, we used the World Health Organization Weekly Epidemiological Records (`http://www.who.int/wer/en/`) dating back to 1995 to provide some insight into the size and duration of epidemics in the country.

Mobile Phone Data
Mobile phone records and cell phone tower locations from Côte d'Ivoir were provided by a major cell phone network provider through the Data for Development (D4D) competition (`http://www.d4d.orange.com/`). The dataset used in this analysis consisted of records from 500,000 randomly selected mobile phone users each for a two-week period (ten two-week periods with 50,000 users each from Dec 1, 2011 to April 28, 2012). Each record (corresponding to a call or sms), consists of a unique identifier,

a time-stamp, and the unique id (n=1194[1]) for the tower handled their call. The call handling tower was assumed to be the closest tower to the individual although there are some cases (in dense urban areas) where a call is passed from one tower to another.

Meteorological Data

Meteorological and hydrological fields were then used to derive estimates of local inundation potential (referred to as flooding index below), size of the potential vibrio reservoir, and rate of vibrio decay due to die-off and filtration (Figure 1). Data for 851 days of data starting from January 1, 2010 (referred to as epidemic day 1) through May 1, 2012 were used in simulations. All meteorological data except for precipitation were drawn from the Global Data Assimilation System (GDAS; [8], 0.471 degree resolution) gridded analysis and were topographically downscaled to 5km resolution using standard lapse rate corrections (e.g., [9]). GDAS meteorological fields were supplemented with precipitation estimates from the Tropical Rainfall Measurement Mission (TRMM) Multisensor Precipitation Analysis (TMPA; Huffman, Adler et al. 2007). The three hourly, 25km gauge corrected estimates were used (product 3B42v7). GDAS meteorological fields and TRMM precipitation were applied as forcing data to offline simulations with the Noah Land Surface Model v3.2 ([10, 11]), implemented in the NASA Land Information System ([12]) at 5km spatial resolution and with a 15 minute time step. Noah simulations were used to generate estimates of surface hydrologic states, including near-surface and root zone soil moisture, on a daily basis over the period of analysis.

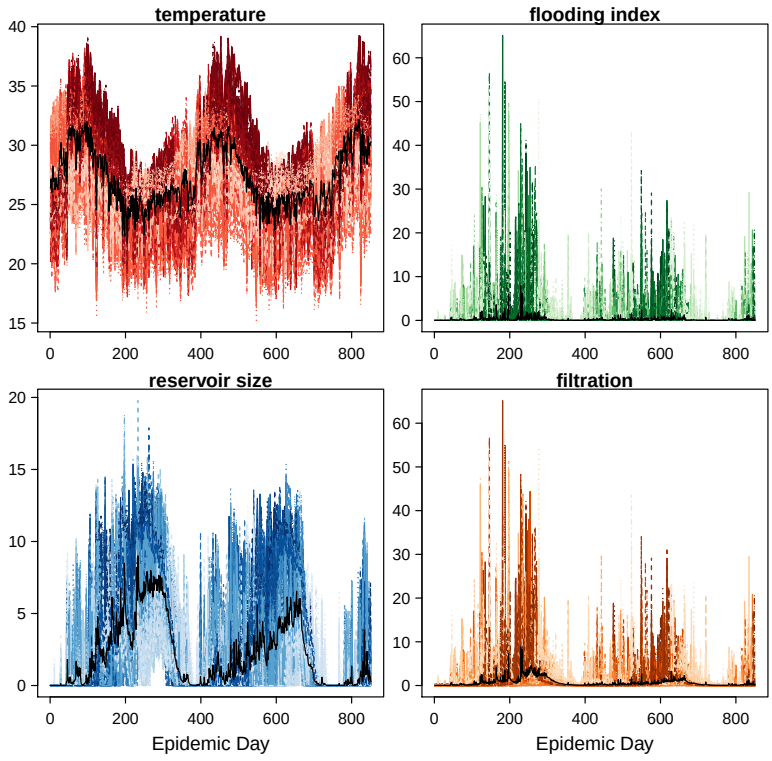


Figure 1: Key environmental data from 200 randomly selected locations with mean of all locations shown in black. Shading represents latitude going from light in the South to dark in the North. Day 1 represents January 1, 2010.

[1]towers with the same coordinates were collapsed into a single id

Methods

Mobility Model

High resolution mobile phone data such as those used in this analysis have information to help estimate how often people travel to different locations and on how long they stay there. Here we present a simplified parametric model of human movement, however, a separate model, utilizing more of the information contained in the data, will be incorporated the future.

We assigned home locations as the location with the most calls for each individual. In order to reduce the influence of multiple calls made within a short period of time from a single location, we down sampled the data and only included one call from a single location within 30 minutes of a previous call made from the same location. To discretize the country into home areas based on cell tower locations, we first created a Voronoi tessellation (Figure 2) of the country and then took the minimum of the Voronoi cell area or the area of a 10-kilometer circular buffer around the cell tower (to reflect the fact that cell towers have a finite range).

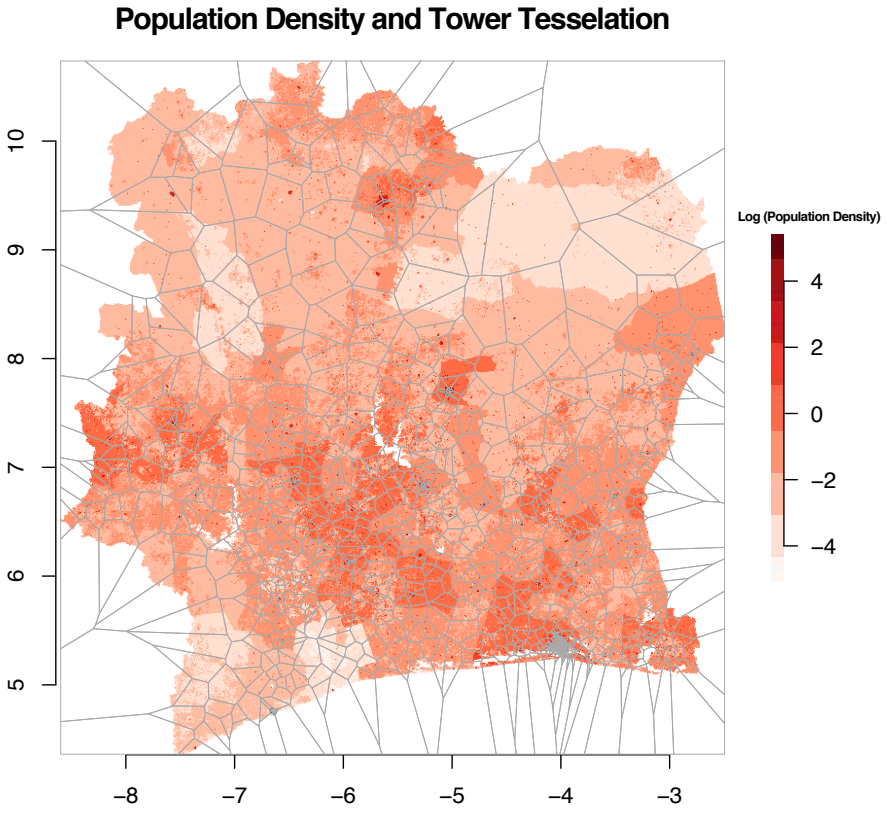


Figure 2: Log population density and Voronoi tessellation of country by cell phone towers.

In this simple gravity model, we model the probability that a particular person, k, (given their home location H_k) will be seen in any other location at any point in time ($\nu_{i,j}$). We define $\nu_{i,j}$ as,

$$logit(\nu_{i,j}) = \begin{cases} \alpha_0 + \alpha_1 log(P_i) & \text{if } i = j \\ \alpha_2 + \alpha_3 log(P_i) + \alpha_4 log(P_j) + \alpha_5 log(d_{ij}) & \text{if } i \neq j \end{cases} \quad (1)$$

where P_i is the population of the home location, P_j is the population of the area from which a call was made, and d_{ij} is the the distance between the two.

From these we can now write down the likelihood of a single person's call record from a day, $C_{k,t}$, as,

$$\Pr(C_{k,t}|N_{k,t}) = \prod_{j \in L} \nu_{i,j}^{n_{j,k,t}} (1-\nu_{i,j})^{n_{(-j),k,t}} \tag{2}$$

where, $N_{k,t}$ is the set of calls by k on day t to all locations, L is the set of all home location (i.e. towers), $n_{j,k,t}$ is the number of calls made by person k on day t from location j, and $n_{(-j),k,t}$ is the number of calls made by k on t from all locations except for j.

The likelihood of the full dataset, D, can be simplified to a product of data collapsed over all times as follows,

$$Pr(D|\vec{\alpha}, N_{\cdot,\cdot}) = \prod_{j \in L} \prod_{i \in L} \nu_{i,j}^{n^{\star}_{j,\cdot}} (1-\nu_{i,j})^{n^{\star}_{(-j),\cdot}} \tag{3}$$

where, $n^{\star}_{j,\cdot}$ is $\sum_{\{i:H_k=i\}} n_{j,k,t}$, and $n^{\star}_{(-j),\cdot}$ is $\sum_{\{i|H_k=i\}} n_{(-j),k,t}$, and $\vec{\alpha}$ is the vector of parameters.

We explored the likelihood surface of this model and estimated the parameters using MCMC techniques (i.e., the Metropolis-Hastings algorithm). After 500,000 iterations, convergence to the target distribution was visually assessed in parallel chains.

Transmission Model

We constructed a discrete-time stochastic susceptible-infectious-recovered meta-population model of cholera transmission in Côte d'Ivoir. In this model all transmission occurs through the environment by individuals shedding infectious bacteria in their home area - a transmission process that captures both direct-like and environmentally mediated transmission.

We divided the country into 5-km grid cells with each cell representing a fully-mixed population. At each daily time step, new infections in each location ($X_{i,t}$) are determined through a binomial process with the risk of infection ($\lambda_{i,t}$) for someone living in are i at t determined by the time-weighted probability of travel from an individual's home to another area ($\phi_{i,j}$) and the infection risk in each area ($\psi_{i,j,t}$). This infection process is specified as:

$$\phi i, j = \frac{\nu_{i,j}}{\sum_{\{i:H_k=i\}} \nu_{i,j}} \tag{4}$$

$$\lambda_{i,j,t} = \phi_{i,j} \psi_{i,j,t} \tag{5}$$

$$\lambda_{i,t} = 1 - \prod_{j \in L} (1 - \lambda_{i,j,t}) \tag{6}$$

$$X_{i,t} \sim Binom(\lambda_{i,t}, S_{i,t}) \tag{7}$$

The number of infectious ($I_{i,t+1}$) and newly recovered people ($\Delta R_{i,t+1}$) each day is updated according to:

$$I_{i,t+1} = I_{i,t} + X_{i,t} - \Delta R_{i,t} \tag{8}$$

$$\Delta R_{i,t+1} \sim Binom(I_{i,t}, \gamma) \tag{9}$$

where γ represents the mean infectious period.

Infection risk in each area is determined by environmental factors that affect the concentration of vibrios and the probability of ingesting contaminated water. We decompose infection risk in an area as:

$$\psi_{j,t} = \omega \cdot \chi_{j,t} \tag{10}$$

where, $\chi_{j,t}$ is the probability of ingesting contaminated water at t, and ω is the probability of infection given exposure to a specific dose ω. We used a beta-poisson dose-response model fit to data from human challenge experiments to estimate ω [2][13]. Infection risk is not determined solely by the number of vibrios, but by the concentration ($C_{j,t}$), a function of the number and the size of the reservoir ($W_{j,t}$) in each area. The reservoir size each day is estimated from meteorological data as a function of rain, evaporation and drainage each day.
We modeled $\chi_{j,t}$ as,

$$\chi_{j,t} = \mu_{ingest} \frac{e^{\chi_0 + \tau_{flood} \times flood_{j,t}}}{1 + e^{\chi_0 + \tau_{flood} \times flood_{j,t}}} \tag{11}$$

where, μ_{ingest} is the maximum ingestion risk, τ_{flood} is a scaling factor, and $flood_{j,t}$ is the flood index.

Each day the number of vibrios ($V_{j,t}$) in location j are updated according to:

$$V_{j,t+1} = V_{j,t} + I_{i,t}\xi - \delta_{j,t} V_{j,t} \tag{12}$$

The rate ($\delta_{i,t}$) at which vibrios decay each day (through die-off and physical processes) is also a function of environmental factors in each area and is modeled as:

$$\delta_{j,t} = \mu_{decay} \frac{e^{\delta_0 + \tau_{filtration} \times flltration_{j,t}}}{1 + e^{\delta_0 + \tau_{filtration} \times filtration_{j,t}}} \tag{13}$$

where, μ_{decay} is the maximum decay rate of vibrios, $\tau_{filtration}$ is a scaling factor, and $filtration_{j,t}$ is the filtration volume index representing water leaving the area through advection and groundwater infiltration.

Results

Human Movement
The median estimates for the parameters and 95% quantiles, are presented in Table 1. Through visual exploration of the empirical data, these parameters yield a reasonable fit to the data (e.g. Figure 3). As expected, the probability of being seen in a location decays with distance. For each additional 1 log change in distance (km) between locations, people have 0.32 times the odds of traveling to that location. For example, people have 0.32 times the odds of traveling to a place 100-km away compared to 10-km. Our estimates suggest that individuals who leave there home area are more likely to travel to an area of higher population density, regardless of the population density of their home area. For each additional 1 log people in an area, the odds of an individual from outside that area spending time in that area goes up by 34% percent. The odds that a person remains at home in a population of 100,000 is 0.69 times smaller than that of a person living in a population of 10,000.

Table 1: Gravity mobility model parameter estimates of key quantiles. Note: median values were used in all simulations presented in this manuscript.

quantile	α_0	α_1	α_2	α_3	α_4	α_5
2.5%	1.8704	-0.1603	-6.9926	0.0229	0.2897	-1.1270
50%	1.8791	-0.1594	-6.9815	0.0239	0.2905	-1.1266
97.5%	1.8876	-0.1584	-6.9707	0.0249	0.2913	-1.1262

[2]following the work from `http://wiki.camra.msu.edu/index.php?title=Dose_response_models_for_Vibrio_cholera`

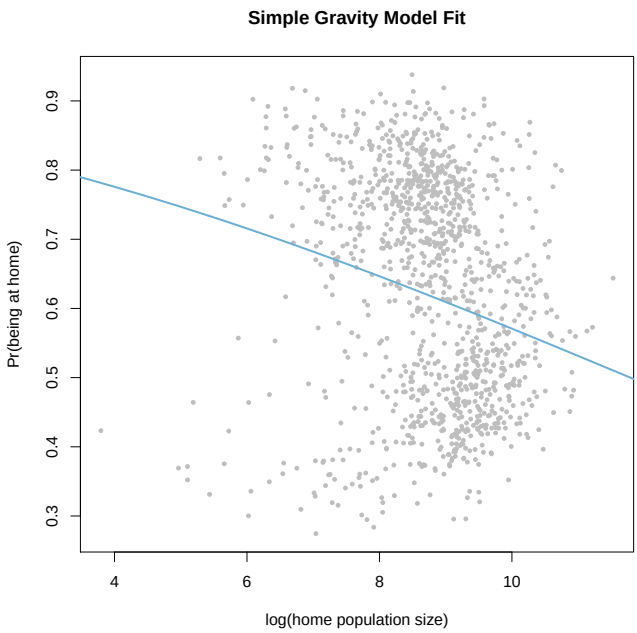


Figure 3: Fit of gravity model to probability of being seen at any point in home area. Grey points illustrate data, and blue line represent model fit.

Simulated Epidemics

Simulations of cholera transmission on a fully susceptible population (Figures 5,6 and `http://andrewazman.com/D4D/movies/`) for animations of model simulations) show rapid spread of the virus throughout the country, with new introductions leading to multi-year epidemics. Cholera incidence per 1,000 population is heterogeneous throughout the country with the areas of highest population density having the highest incidence rate. This latter observation is consistent with observations in previous cholera epidemics. In particular Abidjan, Bouak, and Daloa, show especially high incidence compared with the rest of the country, both in our model and in reports of previous cholera outbreaks. These location of hotspots appears to be independent of the location at which cholera outbreaks began.

In addition to epidemics size, the speed at which cholera cases first appear in an area is strongly correlated with the population size (Figure 4), with the densest areas seeing cases within days of introduction, while less dense areas may wait multiple epidemic years before seeing cholera cases.

Seasonality arose in our model solely from the use of climatic data (i.e., there was no independent seasonal forcing term). We found that cholera epidemics from new introduction tended to peak the months of October or November with smaller peaks around June. This is broadly consistent with the timing of epidemics seen in Côte d'Ivoir and elsewhere throughout the region. However, we did not see cholera clearly fall below detectable levels between epidemic years (see below for a discussion).

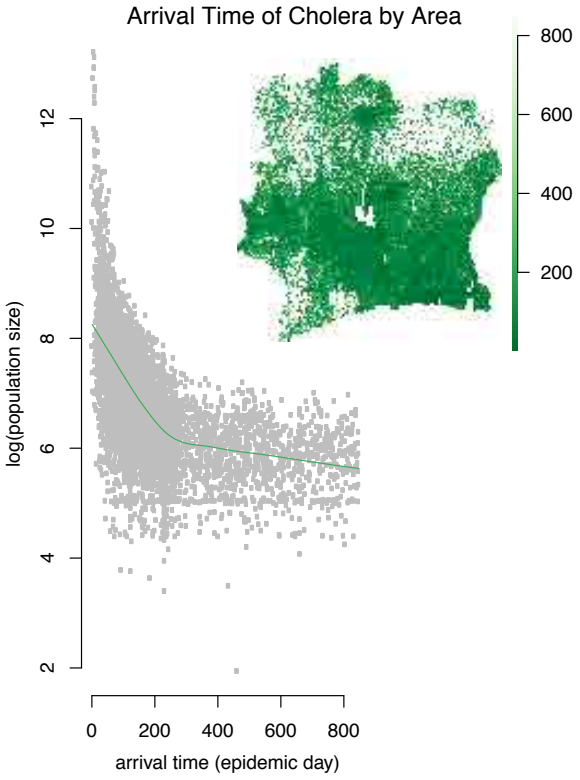


Figure 4: Arrival day of cholera from example run. Shown by population size (right) and geographically (upper-left)

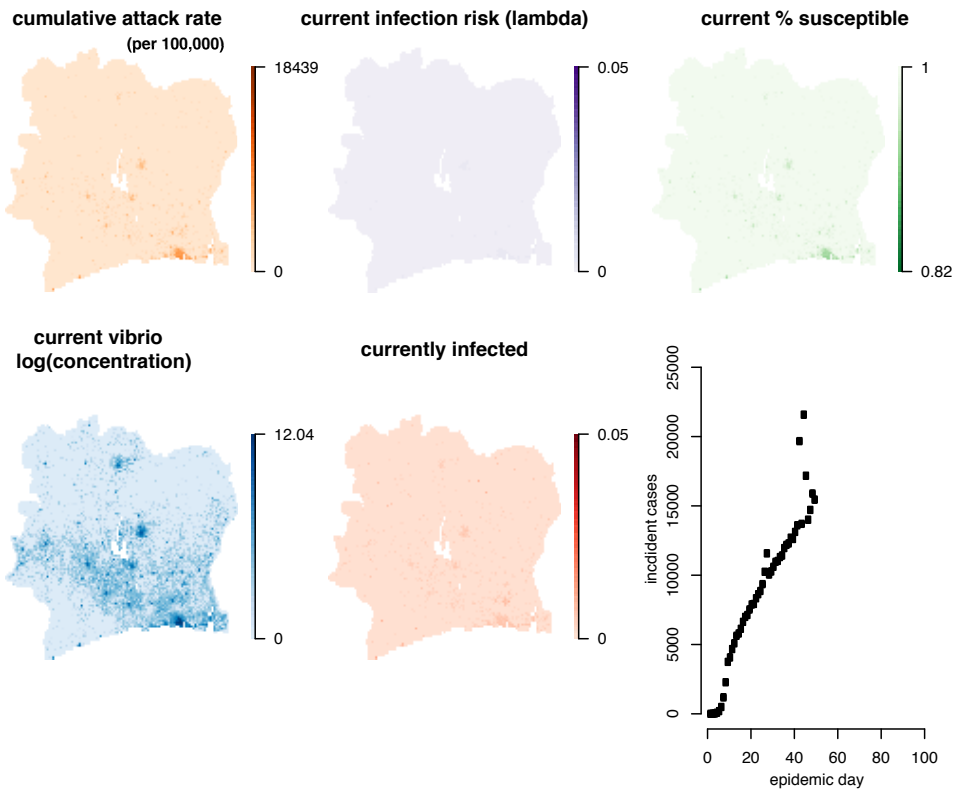


Figure 5: Example simulation at epidemic day 50. Very large epidemic.

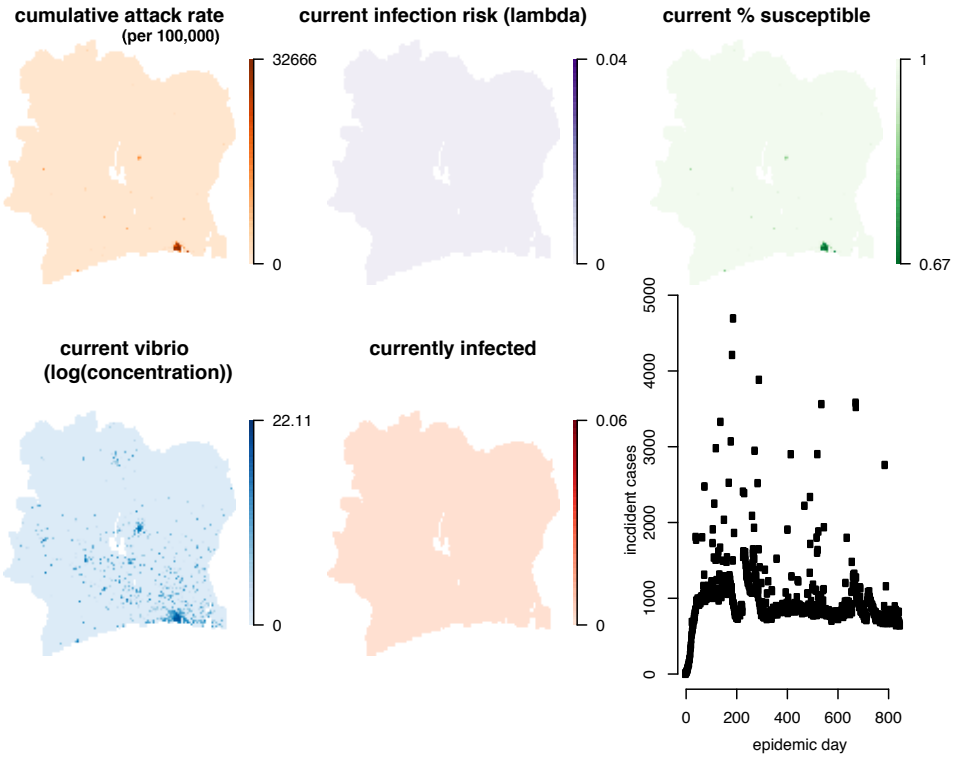


Figure 6: Example simulation at epidemic day 850. Smaller epidemic.

Discussion

This model represents an important first step in incorporating country-specific human mobility data and local environmental data into a cholera transmission model. While it is in its early stages, it serves as a proof of concept that models combining human mobility and environmental data can be used to create detailed models of cholera transmission at a country level. Cell phone use data plays a particularly important role in this process, as it provides a mechanism by which highly accurate human movement models can be constructed for a country. Cell phone use is high in developing countries, and mobile phones often play an even more important role in daily life African countries than they do in the United States and Europe. Here we have shown how this data can be leveraged to create high resolution models of disease spread.

While the current model serves as an important starting point, there is still much to be done before it becomes an practical public health tool. Fitting to incidence data from Côte d'Ivoir is necessary to tune model parameters and guarantee realistic dynamics. In addition, we simulated our results in a cholera naive population, whereas cholera has circulated periodically in Côte d'Ivoir since the 1970s, leading to heterogeneous immunity throughout the region which may highly impact dynamics. Furthermore, while our focus on home location based environmental transmission successfully captured key components of local dynamics, person to person spread and asymptomatic carries who continue with their daily activities still have a role to play. In particular, the periodic spikes seen in our simulations are consistent with jumps in transmission seem in empirical data, though the rapid decline in cases is not. These declines would likely be more gradual if direct transmission was also included in the model.

Large scale populations models are an important component of a suite of modeling and statistical techniques that help us to understand and predict disease spread. These models depend on understanding human movement in diverse settings. Cell phone data is an important tool in parameterizing these models that can be deployed in many contexts. As we develop highly predictive models we will be able to more effectively target interventions, both for immediate (e.g. oral cholera vaccination) and long term (e.g., infrastructure improvements) cholera control. More effective control of disease does not simply prevent the immediate human misery they cause. Healthy people have more opportunities to improve their conditions, develop vibrant economies, and live healthy, happy lives.

References

[1] Mohammad Ali, Anna Lena Lopez, Young Ae You, Young Eun Kim, Binod Sah, Brian Maskery, and John D Clemens. The global burden of cholera. *Bulletin of the World Health Organization*, 90(3):209–218A, March 2012.

[2] World Health Organization. Global Task Force on Cholera Control. Cholera Country Profile Cote D'Ivoir. Technical report, April 2011.

[3] Andrew S Azman, Francisco J Luquero, Amabelia Rodrigues, Pedro Pablo Palma, Rebecca F Grais, Cunhate Na Banga, Bryan T Grenfell, and Justin Lessler. Urban Cholera Transmission Hotspots and Their Implications for Reactive Vaccination: Evidence from Bissau City, Guinea Bissau. *PLoS Neglected Tropical Diseases*, 6(11):e1901, November 2012.

[4] J Glenn Morris. Cholera–modern pandemic disease of ancient lineage. *Emerging Infectious Diseases*, 17(11):2099–2104, November 2011.

[5] L Mari, E Bertuzzo, L Righetto, R Casagrandi, M Gatto, I Rodriguez-Iturbe, and A Rinaldo. Modelling cholera epidemics: the role of waterways, human mobility and sanitation. *Journal of The Royal Society Interface*, 9(67):376–388, February 2012.

[6] Duygu Balcan, Vittoria Colizza, Bruno Gonçalves, Hao Hu, José J Ramasco, and Alessandro Vespignani. Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences*, 106(51):21484–21489, December 2009.

[7] Andrew J AJ Tatem, Abdisalan M AM Noor, Craig C von Hagen, Antonio A Di Gregorio, and Simon I SI Hay. High resolution population maps for low income nations: combining land cover and census in East Africa. *PLoS ONE*, 2(12):e1298–e1298, December 2006.

[8] John C Derber, David F Parrish, and Stephen J Lord. The New Global Operational Analysis System at the National Meteorological Center. *Weather and Forecasting*, 6:538–548, December 1991.

[9] G E Liston and K Elder. A meteorological distribution system for high-resolution terrestrial modeling (MicroMet). *Journal of Hydrometeorology*, 2006.

[10] M B Ek, K E Mitchell, Y Lin, E Rogers, and P Grunmann. Implementation of Noah land surface model advances in the National Centers for Environmental Prediction operational mesoscale Eta model. *J Geophys Res*, 2003.

[11] F Chen, K Mitchell, J Schaake, and Y Xue. Modeling of land surface evaporation by four schemes and comparison with FIFE observations. *Journal of Geophysical Research*, 1996.

[12] S Kumar, C Peterslidard, Y Tian, P Houser, J Geiger, S Olden, L Lighty, J Eastman, B Doty, And P Dirmeyer. Land Information System: an Interoperable Framework for High Resolution Land Surface Modeling. *Environmental Modeling and Software*, (10):1402–1415, October 2006.

[13] R B Hornick, S I Music, R P Wenzel, R Cash, J P Libonati, M J Snyder, and T E Woodward. The Broad Street pump revisited: response of volunteers to ingested cholera vibrios. *Bulletin of the New York Academy of Medicine*, 47(10):1181–1191, October 1971.

Information Dissemination using Human Mobility in Realistic Environment- (E-Inspire)

Rachit Agarwal*, Vincent Gauthier*, Monique Becker*
*Lab. CNRS SAMOVAR UMR 5157, Telecom Sud Paris, Evry, France
Emails: {rachit.agarwal, vincent.gauthier, monique.becker}@telecom-sudparis.eu

***Abstract*—Dissemination of information in mobile adhoc networks has lately picked up lot of interest. Some studies argue that the dissemination in these networks should be contained while some argue that it should not. Research has found that it depends on the type of the application that is considered. For example, dissemination of mobile viruses should definitely be contained however, dissemination of emergency information should not. Moreover, in the regions where there is less connectivity and very few mobile devices, dissemination of packets is highly impacted. Towards this, we would like to propose a mechanism that could enhances dissemination of information in a sparsely populated mobile adhoc environment. We use the concept of metapopulation model and epidemic model and the results obtained after the analysis of the dataset provided by D4D Organizers. From the results we obtained, we could say that in our model we could reach the epidemic state in dissemination process using the movement pattern of the users (derived from the dataset provided by D4D organizers).**

***Index Terms*—Human Mobility, Information Dissemination, Variable Density.**

I. Introduction

Due to vast developments in wireless devices and mobile network, recently Pocket Switched Network (PSN) has been introduced [1]–[3]. A PSN is a mobile adhoc network formed when devices carried by humans interact with each other. Due to the human aided mobility, PSNs closely follow human mobility characteristics. Human mobility has vastly been studied and many spatio-temporal characteristic properties have been identified that define human mobility. Some of these properties include jump length, pause time, radius of gyration, frequency of visits, etc. Recently, studies have revealed that human mobility not only has spatio-temporal dimension but also has social dimension [4], [5]. It was also revealed that different characteristics of human mobility closely follow truncated power law. It was however also shown that the truncated power law was also due to the sampling of the data [6]. Due to the vast identified properties of human mobility, models usually use only some features of human mobility instead of incorporating all. Some models use temporal characteristics in form of periodic, aperiodic and sporadic nature while some use spatial like centric, orbital, random or social like group movement, etc. In [7] authors survey different mobility models using above mentioned features and clearly bring out the differences between the models.

Recently, epidemics across population have been the focus of lot of research and various models has been proposed. An epidemic model typically contains two states: Susceptible(S) and Infected(I). However, there are other states also used like Recovered(R), exposed or Latent(E) and Passively immune(M) by some models. A typical epidemic model consists of combinations of these states. A comprehensive survey about the epidemic model could be found in [8], [9]. In Communication networks, information dissemination has been closely related to epidemics across the population and consider SIS or SIR epidemic model, for example, [10]. Further, in communication networks, information dissemination has been shown to be influenced by many factors like bursty data [11], strength of the tie [12], source of the infection, number of infected devices, human mobility parameters [13], location preference [14], network structure [15], activity pattern [16], device characteristics [13], [17], altruism [18] etc. However, mostly the focus has been limited to the study of effects of human mobility on information dissemination. Recently, [19] showed that human mobility in some time can speed up the information dissemination rate while can also in some cases suppress the information dissemination rate. The speed up relates to higher probability of meeting susceptible population while reduction related to isolation of the infected device.

Moreover, mostly the models based on epidemics using human mobility considered homogenous population well spread across the area. However, Watts et al in [20] used the hierarchical metapopulation model for the dissemination process. In their model, Watts et al argued that clusters are evident in a large population and they affects the epidemic spread. They assumed SIR type epidemic model and allowed human mobility in terms of changing clusters with a probability related to levels of clusters jumped. Fig 1 shows clustering of humans into groups and possible transitions that could happen. Moreover, in Watts et al model, uniform distribution of population in the clusters was considered. However, in realistic case, there is a non-uniform distribution of population in the clusters and the overall population constantly changes with respect to time. This affects the dissemination process in terms of time taken to spread the epidemic in the area.

In a PSN, however, where the structure of the network is dependent on the humans and the characteristics of mobile device, we are interested to investigate how information dissemination takes place in the dynamic population in contrast to [13] where constant population size was used. Towards this, we use insights from the spreading in metapopulation model [21], data provided by D4D organizers and epidemic model to formulate our model. In our model, the population is non-uniformly divided into communities. These communities

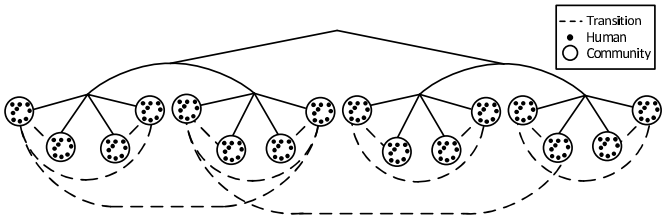


Fig. 1. Hierarchical Metapopulation Model [20].

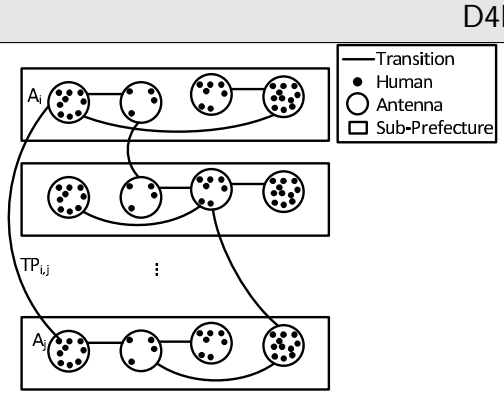


Fig. 2. Modified metapopulation Model for uneven population distribution and antenna-subprefecture hierarchy. Movement could occur between neighboring antenna. The neighboring antennas could be in different subprefecture. Here, a person from antenna A_i could move to antenna A_j with a probability $TP_{i,j}$. The antennas A_i and A_j belong to different subprefecture.

are the antennas to which the population is associated to in the country. A group of antennas formulate a bigger community known as sub-prefecture and further, collection of sub-prefectures form a country. According to the metapopulation model, a user transits from one community to anther using a transition probability, in our model this transition probability between two antennas is calculated through the analysis of the dataset provided by the D4D organizers and through the graph generated using voronoi tessellation. The transitions could happen between antennas of different subprefecture as two neighboring antennas could be in different subprefecture (Cf. fig. 2). More details on how we calculate transition probability can be found in section II-B. As a person can move within a community also, in our model, we assume that devices can also move within the community (antenna).

As we are concentrating on PSN, a device in a PSN can pair up with a device within its range and can transmit its data packets to the paired devices. A device having data packet to transmit is said to be in Infected(I) state while those not having the packet are considered to be in Susceptible(S) state. Once the device in S state has the information it changes its state to I. We use recovery rate also so as to model realistic scenario. The recovery rate would mean that the devices are only willing to transmit the information for certain period. As mentioned before, characteristics of a PSN is dependent on the humans. Human can switch off the device and can reopen them any time causing changes in the structure of the PSN due to change in the number of active devices. To capture this effect we use the Latent state (E) of the epidemic model. To distinguish between devices that are latent but Susceptible and are latent but Infected, we divide E into two states E_S and E_I. A device in E_S or E_I state does not participate in dissemination process. However, only devices in S and I participate in the dissemination process. Further, more comprehensive details about the model are mentioned in section sec. III.

Further, in this paper, we first analyze the dataset provided in section sec. II. We then provide detailed description of the model in section sec. III. We then provide results obtained in section sec. IV and finally conclude in section sec VI after providing future work in sec. V.

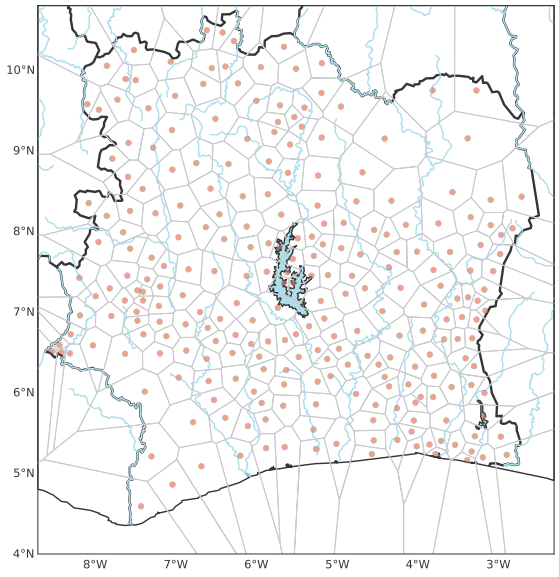
II. Data Analysis

In this section, we further analyze data collected by Orange for the region of Ivory Coast than what has been provided in [22]. The data is based on the calls made in the region of Ivory Coast and the mobility of the users. The region of Ivory Coast has been assigned number of antennas and is divided into sub-prefectures. The dataset contains the locations of these antennas and sub-prefectures in longitude and latitude format. The dataset is further divided into four sub-datasets out of which we are interested only in sub-dataset **SET2TSV** and sub-dataset **SET3TSV**. Sub-dataset **SET2TSV** and **SET3TSV** contains pruned mobility patterns of the users over 5 months. These sub-datasets has been formed by the logging of call information of the users in Ivory Coast. The sub-dataset **SET2TSV** relates users with antennas while sub-dataset **SET3TSV** relates users to sub-prefecture. Both these sub-datasets have information like, user, time, antenna or sub-prefecture. Moreover, another difference between the two sub-datasets is that sub-dataset **SET2TSV** has been sampled for 50,000 users while sub-dataset **SET3TSV** has been sampled for 500,000 users over 5 months. However, mobility from these two sub-datasets can only be inferred as the id of antennas and the sub-prefectures has been logged when the call was made and not the actual location of the user. Moreover, we are interested in the analysis datasets with antennas and sub-prefectures location and sub-dataset **SET2TSV** to get useful information that could be used in our proposed model.

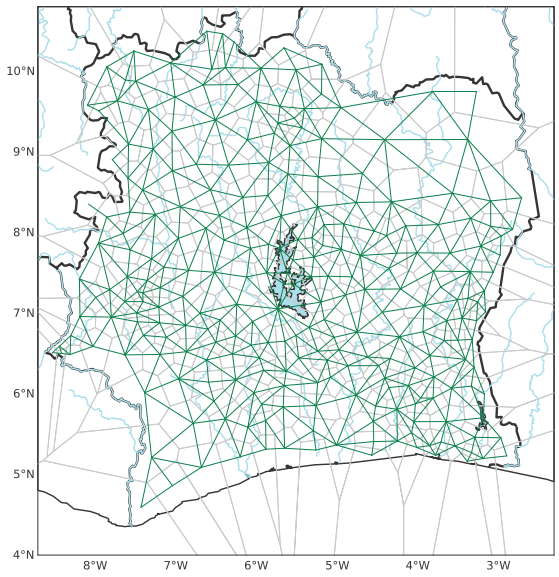
A. Analysis Dataset SUBPREF_POS_LONLAT.TSV and ANT_POS.TSV

We first analyze dataset **SUBPREF_POS_LONLAT.TSV**. We use the position information of the sub-prefectures to provide the visualization of the sub-prefectures in the region of Ivory Coast and visualize the Voronoi tessellation (Cf. fig. 3(a)). Using Voronoi tessellation we then generate a graphical structure that connects all sub-prefectures with common Voronoi edges. As an edge in the graph could lie well outside the country boundaries, we remove all those edges that bypass the country. We call the remaining graphical structure $G_{sub-pref}$, (Cf. fig. 3(b)).

Similar to sub-prefectures dataset, dataset **ANT_POS.TSV** has been provided for antenna locations. We perform similar procedure on this dataset and generate Voronoi tessellation, (Cf. fig. 4(a)), and the graphical structure $G_{antenna}$, (Cf. fig. 4(b)). Further, we assume that each antenna is assigned to a sub-prefecture. Depending on the Voronoi tessellation of the sub-prefectures we then provide an estimate of which antenna is assigned to which sub-prefecture (Cf. fig. 5(a)). This leads us to further visualize fig. 5(b) which is the frequency of number of antennas in the sub-prefectures. As

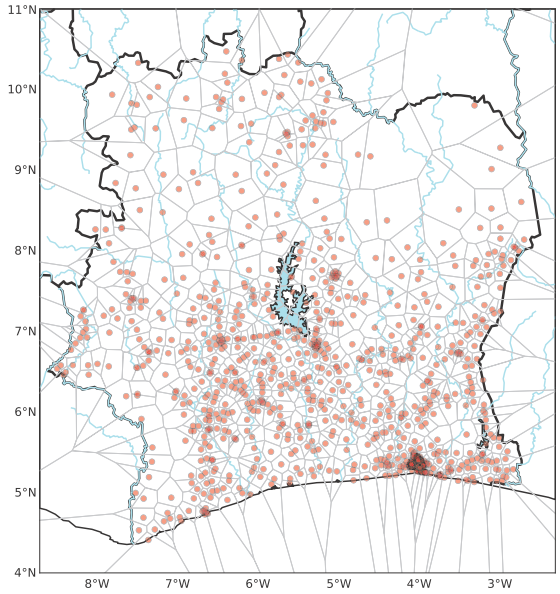


(a) Voronoi tessellation for the Sub Prefectures in Ivory Coast with Sub Prefecture locations.

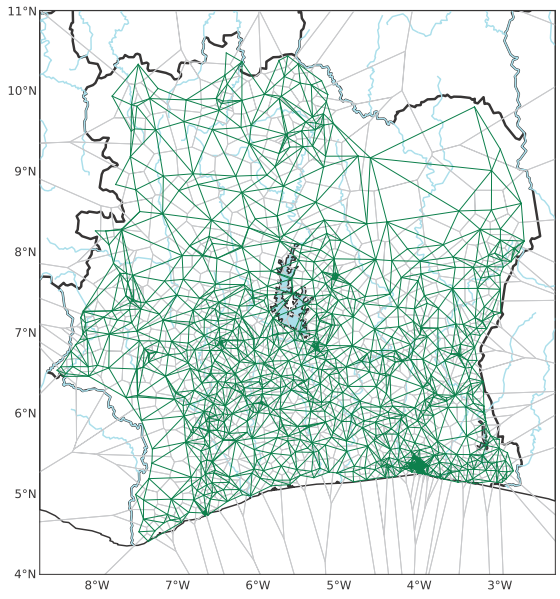


(b) $G_{sub-pref}$.

Fig. 3. Analysis of dataset **SUBPREF_POS_LONLAT.TSV**.



(a) Voronoi tessellation for the antennas in Ivory Coast with antenna locations.



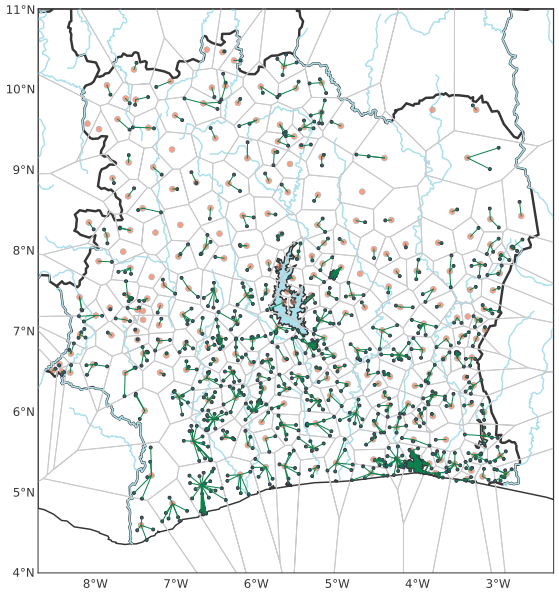
(b) $G_{antenna}$.

Fig. 4. Analysis of dataset **ANT_POS.TSV**.

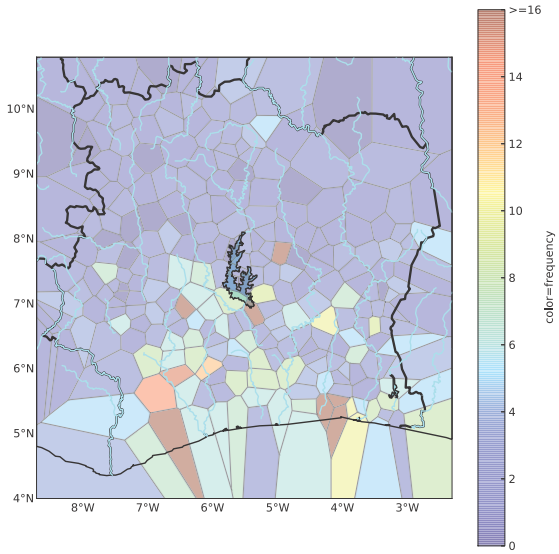
we have estimated the region of sub-prefecture using Voronoi tessellation, the results of the frequency of number of antennas in the sub-prefectures slightly vary from the frequency of antennas when actual sub-prefecture region in Ivory Coast are used, (Cf. fig. 6).

B. Analysis Sub-Dataset SET2TSV

As the sampling of the information in the datasets is based on the calls made, the data has very high percentage of users calling from same location. This sampling hampers the correct estimation of human mobility. We could only infer the mobility pattern of the user. Lack of actual user coordinates leads us to map the mobility of the user on $G_{antenna}$ to get an estimate of the user mobility. This would give us what antennas a user would have connected to while they were moving. This would give us valuable information like how many times a user stayed at a location and how many times the user took a certain path. Further, to estimate the mobility of the user we use shortest path between two antennas in the graph $G_{antenna}$. Thus, from the sub-dataset **SET2TSV** we map the mobility of one user using the $G_{antenna}$ and determine the antennas that the user might have been contacted to by the user while moving, (for user id 48930 Cf. fig. 7(a)). In the figure 7(a) the blue edges mark the edges that user traversed. The thickness of the edges determines the number of times the user traversed through that edge. On the other hand, the size of the node in the graph is determines the number of times the user has stayed at that antenna. We then perform this process for all the users in the sub-dataset **SET2TSV** for all the period and determine



(a) Antenna connected to the Sub Prefectures in Ivory Coast.



(b) Frequency of antennas in the Sub-Prefectures in Ivory Coast estimated using Voronoi tessellation.

Fig. 5. Antennas in the Sub-Prefectures.

a transition probability matrix ($Tm_{antenna}$). The $Tm_{antenna}$ contains the normalized weight of the edges accumulated over 5 month period, (Cf. fig. 7(b)).

We now describe our information dissemination model using the transition matrix formed in the section sec. II-B and metapopulation model in the next section.

III. Model

Consider N devices to be non-uniformly distributed in the region. The non-uniformity leads to a community structure in the region. We assume that the devices in a community (c) are associated to one and only one antenna in the region at any given time. As discussed in the introduction section, collection of these antennas form sub-prefecture and collection of sub-prefectures form a region. Further consider, the number

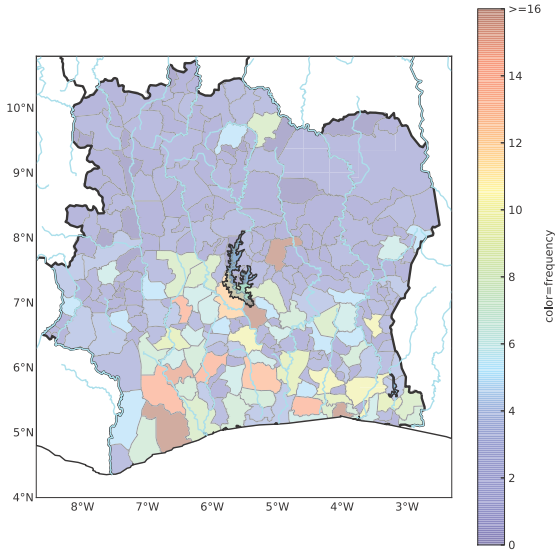
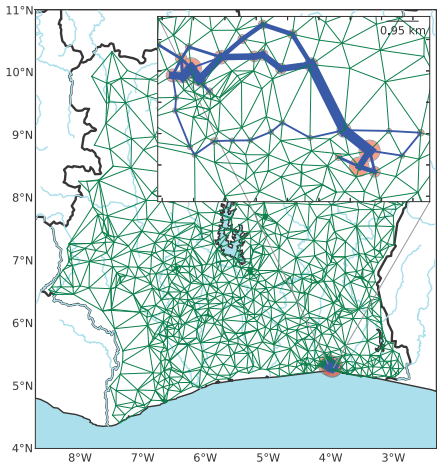


Fig. 6. Actual antenna frequency in Sub-Prefectures in Ivory Coast.

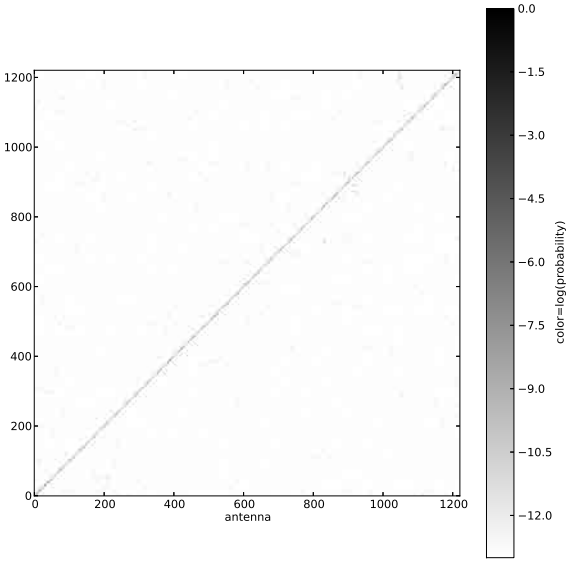
of antennas(communities) as N_c. As argued by Watts et al, community structure is evident in the population in a realistic scenario. A transition from one community to anther occurs with a probability [20] and the nature of the community [14]. This probability plays an important role in determining which community has to be joined. We use transition probability matrix ($Tm_{antenna}$) calculated after doing the data analysis in section sec. II-B to determine the jumps from one community to another. The transition probabilities also provide us the probability of staying in the same community. Staying in the same community would mean that the device has not moved out of the community. This would not restrict the movement of the device within the community bounds, i.e., the device would be free to move within the community bounds.

In-order to study the dissemination process, we assume the devices to be in any of the four states, S, I, E_S or E_I. State S would mean that a device is not having the information and is susceptible to receive it on the other hand state I would mean that the device has the information and will readily transmit it to other devices in its transmission range. Irrespective of whether a device is in state S or state I, the capability of the device to transmit or receive depends on the user. It could be possible that a device has the information but has been switched off by its user. This would hamper the transmission of the information from the device to other devices. Following the same argument, if a device does not have the information and it is switched off it would not be able to receive the information from other devices. We call such state of devices as latent state of a device and term them to be in state E_I and E_S respectively. At a later time, a device in a latent state could be switched on, this would mark the transition in the state of the device from either E_I and E_S to I and S respectively.

It has been argued in research that communities affect the dissemination of the information. The rate of the dissemination process within a community is more than the rate at which dissemination process takes place outside the community. This makes us to constrain the devices in state I to be able to



(a) Antennas reached by user 48930 while moving for first 2 weeks.



(b) Transition Probability Matrix.

Fig. 7. Sub-Dataset **SET2TSV**.

only transmit information to devices in the same community (having same antenna id), in state S and within its Tx. Further, in epidemic, each community has a different infection rate, β_c where c is a community. This is because of many factors like, density of the population and immunity strength of population in the community. Moreover, in a population an infected person has a recovery rate, δ_x where x is the person. In PSN infection rate and recovery rate would mean that devices in community c and in state I are willing to transmit the information with the rate β_c while a device x in state I is rejecting the information after some time with a rate δ_x. However, we assume that δ_x for all device in a community

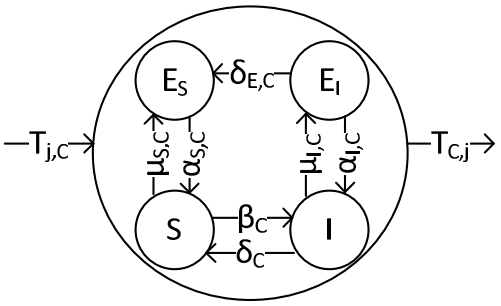


Fig. 8. State diagram with states S and I and their latent states E_S and E_I respectively with transition rates between states.

is same and is δ_c. In order to model β_c we use the area of the community as eq. 1. Thus, the type of epidemic model we consider is SIS with two additional states E_S and E_I.

$$\beta_c = 1 - \frac{A_c}{A_{max}} \tag{1}$$

where A_c is the area or the community c within a region and A_{max} is the area of the region under consideration.

The state diagram for a device to make a transition from one of the states to another can be given by fig. 8. Here, the transition from state S to state I at time $t + \Delta t$ depends on the infection rate β_c. In a community, devices can join as well as leave. The incoming rates and the outgoing rates are given by transition matrix found in sec. II-B. We call them as $T_{j,C}$ and $T_{C,j}$ where j is another community. The transition rates between S and E_S are given by $\mu_{S,C}$ and $\alpha_{S,C}$ while that between I and E_I are given by $\mu_{I,C}$ and $\alpha_{I,C}$.

In a community i, let S_i be the number of devices in state S at time t, I_i be number of devices in state I at time t, $E_{I,i}$ be number of devices in state $E_{I,i}$ at time t and $E_{S,i}$ be number of devices in state $E_{S,i}$ at time t. Considering initial conditions as $S(0) = N - \varepsilon$, $I = \varepsilon$ where $\varepsilon > 0$, from the model described above, we could formulate rate equations for one community i using mean field as follows:

$$\begin{aligned}\frac{\mathrm{d}S_i}{\mathrm{d}t} &= -\beta_i \frac{S_i I_i < k_{R,i} >}{N_i} \\ &+ \sum_{\forall j \in C; j \neq i} T_{j,i} S_j - \sum_{\forall j \in C; j \neq i} T_{i,j} S_i \\ &+ \delta_i I_i - \mu_{S,i} S_i + \alpha_{S,i} E_{S,i}\end{aligned} \tag{2}$$

$$\begin{aligned}\frac{\mathrm{d}I_i}{\mathrm{d}t} &= \frac{\beta_i S_i I_i < k_{R,i} >}{N_i} \\ &+ \sum_{\forall j \in C; j \neq i} T_{j,i} I_j - \sum_{\forall j \in C; j \neq i} T_{i,j} I_i \\ &- \delta_i I_i - \mu_{I,i} I_i + \alpha_{I,i} E_{I,i}\end{aligned} \tag{3}$$

$$\begin{aligned}\frac{\mathrm{d}E_{S,i}}{\mathrm{d}t} &= \mu_{S,i} S_i - \alpha_{S,i} E_{S,i} + \delta_{E,i} E_{I,i} \\ &+ \sum_{\forall j \in C; j \neq i} T_{j,i} E_{S,j} - \sum_{\forall j \in C; j \neq i} T_{i,j} E_{S,i}\end{aligned} \tag{4}$$

$$\begin{aligned}\frac{\mathrm{d}E_{I,i}}{\mathrm{d}t} &= \mu_{I,i} I_i - \alpha_{I,i} E_{I,i} - \delta_{E,i} E_{I,i} \\ &+ \sum_{\forall j \in C; j \neq i} T_{j,i} E_{I,j} - \sum_{\forall j \in C; j \neq i} T_{i,j} E_{I,i}\end{aligned} \tag{5}$$

$$\frac{\mathrm{d}N_i}{\mathrm{d}t} = \sum_{\forall j \in C; j \neq i} T_{j,i}S_j - \sum_{\forall j \in C; j \neq i} T_{i,j}S_i + \sum_{\forall j \in C; j \neq i} T_{j,i}I_j - \sum_{\forall j \in C; j \neq i} T_{i,j}I_i + \sum_{\forall j \in C; j \neq i} T_{j,i}E_{S,j} - \sum_{\forall j \in C; j \neq i} T_{i,j}E_{S,i} + \sum_{\forall j \in C; j \neq i} T_{j,i}E_{I,j} - \sum_{\forall j \in C; j \neq i} T_{i,j}E_{I,i} \quad (6)$$

where $N = \sum_{\forall i \in C} S + \sum_{\forall i \in C} I + \sum_{\forall i \in C} E_S + \sum_{\forall i \in C} E_I$, $< k_{R,i} >$ is the average degree of the devices in the network between devices in the community i with R being the area of the community i. For a network, $< k_{R,i} >$ could be modeled as eq. 7 [23]. Using the probability of connection in an area with a population and the attenuation factor (Communication network parameters), average degree could be defined as integral of probability of connection for a given density over the area, yielding eq. 7

$$< k_{R,i} > = a2\pi\overline{\rho}\frac{r_0\tau}{1-\tau} * \left[R\left(1 + \frac{R}{r_0\tau}\right)^{1-\tau} - \frac{r_0}{2-\tau}\left(1 + \frac{R}{r_0\tau}\right)^{2-\tau} \right]_{2l} \quad (7)$$

where $\overline{\rho}$ is the density of devices in the area, $r_0 = \sqrt{(A/(N\pi))}$ and a and τ are the communication network parameters > 0. The eq. 7 is modeled for static population, however, when mobility is introduced $< k_{R,i} >$ would change over time. This could be modeled as eq. 8 using p as the pause time of a device on a location [24].

$$< k_{R,i} > \approx \frac{Nr_0}{3}\left((4 - 2p + p^2) - \frac{4}{\pi}p^2 r_0 - 3(1-p)r_0^2\right) \quad (8)$$

Further, analyzing the eq. 3, we could say that there would be a growth in the population of devices in I when $\frac{\beta_i S_i < k_{R,i} >}{N_i} + \sum_{j=1}^{C} T_{j,i} - \sum_{j=1}^{C} T_{i,j} - \delta_i - \mu_{I,i} > 0$. This gives basic reproduction number as $\frac{\frac{\beta_i < k_{R,i} >}{N_i} + \sum_{j=1}^{C} T_{j,i} - \sum_{j=1}^{C} T_{i,j}}{\delta_i + \mu_{I,i}} > 1$ where epidemic would takeoff.

IV. Simulation and Results

We perform simulation in Python. Initially, each device operates in omnidirectional mode with the transmission range $Tx = 1km$. We consider $N = 5000$ in an area of $Area \approx 710 * 756km^2$ (Ivory Coast region). All the results use $Tm_{antenna}$ for mobility. We assume initially $N - 1$ devices are in susceptible state and only one device is in infected state. Out of these $N - 1$ susceptible devices some devices are in latent state.

As the preliminary results, we provide results for the information dissemination in the population using our model, (Cf. fig. 9). This result was obtained for the case when there is single infected device, $\delta_i = 0.975 \forall i \in C$, β_i as defined in eq. 1. The result shows the percentage of infected nodes over time

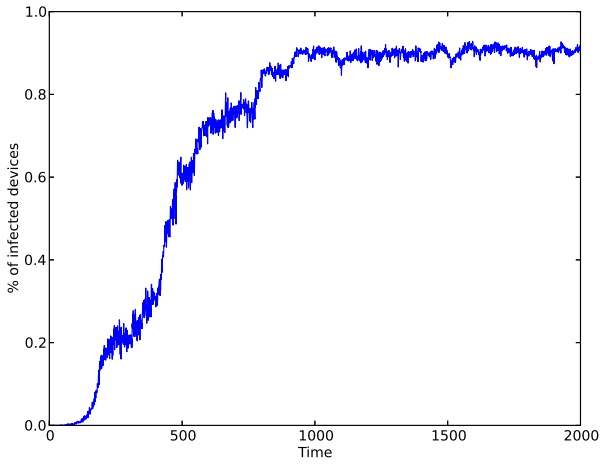


Fig. 9. The time taken to reach the epidemic state in the dissemination process.

normalized over active population in the area. Initially, due to more number of susceptible devices the rate of change in devices in I is more however, over time it reaches an epidemic state.

V. Future Work

Here we presented preliminary results. However, more comprehensive results are to be obtained and to be verified to see the effect of various parameters used in the model. Moreover, currently we have not used birth and death process, i.e., addition of new devices and removal of old devices. We also would like to incorporate heterogenous population density in our model. Addition of such concept could add more realism to the model. We would like to incorporate it as the future work. Further, a device in PSN can be equipped with multiple small antennas which could help in enhancing transmission radius for the device. Using multiple antennas gives rise to a beam of certain length and width. This technique is known as beamforming. Effects of beamforming have already been studied on information dissemination for both static and mobile networks with positive results [13], [25]–[28]. Incorporating beamforming in the model would definitively give an edge and help in enhancing information dissemination. A brief overview of how it could be done is explained in sec. V-A below.

A. Adding Beamforming to the model

Dissemination process could be enhanced using different ways. Some ways studied in literature are mobility and beamforming. Beamforming is a technique of using multiple device antennas in-order to get a long directional beam (long range link) with the same operational power as that of omnidirectional beam. Thus, in our model, we would also like to use beamforming. We assume that each device is equipped with m device antennas (DAs), where m could be different for different devices. Initially all devices use one DA for omnidirectional transmission with the omnidirectional

transmission range being Tx. The Tx could also be different for different devices. Beamforming is done by devices in state I. In the landmark paper by Watts and Strogatz, [29], the authors showed that using very few long range links network diameter can be considerably reduced while network clustering is maintained thereby escalating the dissemination phenomenon. We use this result to state that only 0.01% of the devices in state I are randomly chosen to beamform. The selected devices randomly choose m_x DAs from m available DAs to determine length and width of the beam. The best direction of the beam is chosen based on which direction has the maximum number of devices in S state. The beamforming device then beamforms in that direction, infects the susceptible devices and returns back to omnidirectional case. Further, beamforming is achieved by special arrangement of DAs. Some realistic ways include Uniform Linear Array Antenna model (ULA) [30]. Using m_x DAs according to ULA model would lead to a beam of length $m_x * Tx$ with different beamwidth for different angles $\in [0, 2\pi]$. We would use these unique directions and gain(beam) patterns of ULA model and determine number of devices in state S the beamforming device could connect using that direction. The direction having the maximum number of devices in S state is chosen for beamforming. let the chosen direction be θ_b. The width of the beam in the direction θ_b for the ULA model is given using eq. 9

$$g(\theta, \phi) = \frac{u(\theta, \phi)}{\frac{1}{4\pi} \int_0^{2\pi} \int_0^{\pi} u(\theta, \phi) \sin\theta d\theta d\phi} \tag{9}$$

where θ is angle with the z-axis, ϕ with the xy-plane, $u(\theta, \phi) \propto \left(\frac{\sin(m\psi)}{m*\sin(\psi)}\right)^2$, $\psi = \pi\Delta(\cos\theta - \cos\theta_b)/\lambda$ and Δ is the distance between 2 DA's.

Adding beamforming to the model would change the $< k_{R,i} >$ as 0.01% devices would beamform. We would like to use this concept and build our model towards enhancing information dissemination in realistic environment.

VI. Conclusion

In this paper we presented a model where information dissemination across the population is studied using movement probability from one community to another calculated using the dataset provided by the D4D organizers. We used concepts like epidemic model and metapopulation in our model. To realize the information dissemination process we have used SIS epidemic model with two additional states E_S and E_I. E_S and E_I states are the latent states where devices in these states are not involved in dissemination process. Our result shows that epidemic state could be reached in the our current setting.

VII. acknowledgement

The authors thank D4D organizers for providing the data.

References

[1] P. Hui, A. Chaintreau, J. Scott, R. Gass, J. Crowcroft, and C. Diot, "Pocket switched networks and human mobility in conference environments," in *Proceedings of the 2005 ACM SIGCOMM workshop on Delay-tolerant networking*, WDTN '05, (New York, NY, USA), pp. 244–251, ACM, 22-26 August 2005.

[2] P. Hui, A. Chaintreau, R. Gass, J. Scott, J. Crowcroft, and C. Diot, "Pocket Switched Networking: Challenges, Feasibility and Implementation Issues," in *Autonomic Communication* (I. Stavrakakis and M. Smirnov, eds.), vol. 3854 of *Lecture Notes in Computer Science*, pp. 1–12, Springer-Verlag,Berlin, Heidelberg, 03-05 October 2006.

[3] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott, "Pocket Switched Networks: Real-World Mobility and its Consequences for Opportunistic Forwarding," tech. rep., University of Cambridge, Cambridge, 2005.

[4] F. Ekman, A. Keränen, J. Karvo, and J. Ott, "Working Day Movement Model," in *Proceeding of the 1st ACM SIGMOBILE workshop on Mobility models - MobilityModels '08*, (Hong Kong), pp. 33–40, ACM New York, NY, 27-30 May 2008.

[5] D. Fischer, K. Herrmann, and K. Rothermel, "GeSoMo - A General Social Mobility Model for Delay Tolerant Networks," in *7th IEEE International Conference on Mobile Ad-hoc and Sensor Systems*, (San Francisco, CA), pp. 99–108, IEEE, 8-12 November 2010.

[6] S. Mossa, M. Barthélémy, E. Stanley, and L. N. Amaral, "Truncation of Power Law Behavior in Scale-Free Network Models due to Information Filtering," *Physical Review Letters*, vol. 88, pp. 138701(1–4), March 2002.

[7] D. Karamshuk, C. Boldrini, M. Conti, and A. Passarella, "Human Mobility Models for Opportunistic Networks," *IEEE Communications Magazine*, vol. 49, pp. 157–165, December 2011.

[8] H. W. Hethcote, "The Mathematics of Infectious Diseases," *SIAM Review*, vol. 42, pp. 599–653, Jan. 2000.

[9] T. Britton, "Stochastic Epidemic Models: A Survey," *Mathematical biosciences*, vol. 225, pp. 24–35, May 2010.

[10] N. Valler, B. A. Prakash, H. Tong, M. Faloutsos, and C. Faloutsos, "Epidemic Spread in Mobile Ad Hoc Networks: Determining the Tipping Point," in *Networking 2011* (J. Domingo-Pascual, P. Manzoni, S. Palazzo, A. Pont, and C. Scoglio, eds.), vol. 6640 of *Lecture Notes in Computer Science*, pp. 266–280, Valencia, Spain: Springer-Verlag, Berlin, 9-13 May 2011.

[11] M. Karsai, M. Kivelä, R. K. Pan, K. Kaski, J. Kertész, A. L. Barabási, and J. Saramäki, "Small but Slow World: How Network Topology and Burstiness Slow Down Spreading," *Physical Review E*, vol. 83, pp. 025102(1–4), February 2011.

[12] G. Miritello, E. Moro, and R. Lara, "Dynamical Strength of Social Ties in Information Spreading," *Physical Review E*, vol. 83, pp. 045102(1–4), April 2011.

[13] R. Agarwal, V. Gauthier, and M. Becker, "Enhancing Information Dissemination in Dynamic Wireless Network using Stability and Beamforming," *submitted*, pp. 1–16, October 2012.

[14] B. Wang, L. Cao, H. Suzuki, and K. Aihara, "Safety-Information-Driven Human Mobility Patterns with Metapopulation Epidemic Dynamics," *Scientific reports*, vol. 2, pp. 1–8, January 2012.

[15] V. Nicosia, F. Bagnoli, and V. Latora, "Impact of Network Structure on a Model of Diffusion and Competitive Interaction," *EPL (Europhysics Letters)*, vol. 94, pp. 68009(1–6), June 2011.

[16] A. Vazquez, B. Rácz, A. Lukács, and A. L. Barabási, "Impact of Non-Poissonian Activity Patterns on Spreading Processes," *Physical Review Letters*, vol. 98, pp. 158702(1–4), April 2007.

[17] D. Wang, Z. Wen, H. Tong, C. Y. Lin, C. Song, and A. L. Barabási, "Information Spreading in Context," in *Proceedings of the 20th international conference on World wide web - WWW '11*, (Hyderabad, India), pp. 735–744, ACM Press, New York, 28 March-1 April 2011.

[18] P. Hui, K. Xu, V. Li, J. Crowcroft, V. Latora, and P. Lio, "Selfishness, Altruism and Message Spreading in Mobile Social Networks," in *IEEE INFOCOM Workshops 2009*, (Rio de Janeiro), pp. 1–6, IEEE, New York, 19-25 April 2009.

[19] M. Kivelä, R. K. Pan, K. Kaski, J. Kertész, J. Saramäki, and M. Karsai, "Multiscale Analysis of Spreading in a Large Communication Network," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2012, pp. P03005(1–32), March 2012.

[20] D. Watts, R. Muhamad, D. Medina, and P. Dodds, "Multiscale, Resurgent Epidemics in a Hierarchical Metapopulation Model," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, pp. 11157–11162, August 2005.

[21] J. Arino and P. Van den Driessche, "Disease Spread in Metapopulation," *Fields Institute Communications*, vol. 48, pp. 1–13, 2006.
[22] V. Blondel, M. Esch, C. Chan, F. Clerot, P. Deville, E. Huens, F. Morlot, Z. Smoreda, and C. Ziemlicki, "Data for Development: The D4D Challenge on Mobile Phone Data." http://arxiv.org/pdf/1210.0137v1.pdf, 2012.
[23] G. Németh and G. Vattay, "Giant Clusters in Random Ad Hoc Networks," *Physical Review E*, vol. 67, pp. 036110(1–6), March 2003.
[24] C. Bettstetter, "On the Connectivity of Ad Hoc Networks," *The Computer Journal*, vol. 47, pp. 432–447, April 2004.
[25] F. Peruani, A. Maiti, S. Sadhu, H. Chate, R. Choudhury, and N. Ganguly, "Modeling Broadcasting using Omnidirectional and Directional Antenna in Delay Tolerant Networks as an Epidemic Dynamics," *IEEE Journal on Selected Areas in Communications*, vol. 28, pp. 524–531, May 2010.
[26] R. Agarwal, A. Banerjee, V. Gauthier, M. Becker, C. K. Yeo, and B. S. Lee, "Self-Organization of Nodes using Bio-Inspired Techniques for Achieving Small World Properties," in *2011 IEEE GLOBECOM Workshops (GC Wkshps)*, (Houston), pp. 89–94, IEEE, New York, 5-9 December 2011.
[27] Y. Li, Z. Wang, D. Jin, L. Zeng, and S. Chen, "Collaborative Vehicular Content Dissemination with Directional Antennas," *IEEE Transactions on Wireless Communications*, vol. 11, pp. 1301–1306, April 2012.
[28] R. Agarwal, A. Banerjee, V. Gauthier, M. Becker, C. K. Yeo, and B. S. Lee, "Achieving Small-World Properties using Bio-Inspired Techniques in Wireless Networks," *The Computer Journal*, vol. 55, pp. 909–931, March 2012.
[29] D. Watts and S. Strogatz, "Collective Dynamics of Small World Networks," *Nature*, vol. 393, pp. 440–442, June 1998.
[30] C. Balanis, *Antenna Theory-Analysis and Theory*. New York: Wiley, 1997.

Design and implementation of a tool for the Correlation between the rate of prevalence of a pathology and the flow of communication between diverse localities

Thomas Djotio Ndié
MASECNESS – LIRIMA
National Advanced School of Engineering
University of Yaoundé I
tdjotio@gmail.com,
djotio@ensp.uninet.cm
PoBox 8390 Yaounde Cameroon

Zephirin Nganmeni
MASECNESS– LIRIMA
National Advanced School of Engineering
University of Yaoundé I
nganmeni.zephirin@yahoo.fr
PoBox 8390 Yaounde Cameroon

Soulamite J. Nouho Noutat
MASECNES– LIRIMA
National Advanced School of Engineering
University of Yaoundé I
noutatsoulamite@gmail.com
PoBox 8390 Yaounde Cameroon

Abstract

We propose in this paper a tool to establish the correlation between communications from a telecommunication operator and the prevalence of any pathology of public health in a country. The sample data at our disposal, source of our analysis comes from the telecommunication service provider Orange Ivory Coast (IC), which will be our reference country. The pathology of public health on which we base our study is HIV AIDS whose prevalence rates for regions of IC were obtained from the Web. Our approach is based on the statistical method of principal component analysis that explores the links between variables, here the communication rate of Orange IC and AIDS prevalence. We initially calculated the correlation coefficient between the two variables in a region and afterwards, defined a linear regression model to estimate from the rate of communication in a Sub-division of the IC the prevalence of a pathology, in our case HIV AIDS.

1- Introduction

The mobile phone enables geographically remote people to stay connected through calls or SMS. It results from this interaction, an important dataset whose operation can be used to understand, characterize and anticipate phenomena on the state and/or the environment of a population. Indeed, the great challenge of such large amounts of data like the D4D, lies in the fact that not only their exploration and exploitation are difficult, but that they can also hide relevant information that can be useful in several contexts. The problem of our project revolves around the following question: what relevant information likely to contribute significantly to the development can hide in such a mass of data?

Areas of development are numerous like environment, health, economics, nutrition and safety. We are interested, in the following, in health and more specifically in health service. In 1952, the World Health Organization defines public health as the science and art of preventing disease, prolonging life and improving the physical and mental health at an individual and collective level. To achieve these objectives, the actors need reliable and meaningful data. The estimated prevalence of a pathology in a region can help to take preventive or curative measures against the pathology and protect the population. Thus, we propose in this paper a tool for correlation between the communication rate of a telecommunications operator such as a Orange Ivory Coast and the prevalence of a pathology in a region of IC. Our tool allows first, to calculate the correlation coefficient and then estimate from the rate of communication in a region, the prevalence of any

pathology. We have chosen as case study, the prevalence of HIV. The rate of communication results from the analysis of the D4D data made available.

Our approach is based on the statistical method of principal component analysis that explores links between variables (timestamp and location of calls and sms, ...) and similarities between individuals (mobile users). By applying mathematical and statistical theorems and methods, we define a linear regression model to estimate the prevalence rate knowing the rate of communication in a region. In Section 2 of this paper, we present the state of the art on technologies for storing computer data as well as the mathematical and statistical methods. In Section 3, we present the analysis and exploitation of the D4D's data. In Section 4 we present a case study on HIV and before concluding in Section 6, we present the interpretation of the results of the case study in Section 5.

2. State of the art

2.1. Presentation of the existing [VIN2012]

A database is a structured set of data stored in accessible media by the computer to simultaneously satisfy in a selectively and timely manner multiple users. The access to a database is managed by a Database Management System (DBMS) which is a consolidated software designed to provide a standardized interface between the database and applications. Before adopting a model of data management, it is important to understand the structure of the data available to us in this project.

The database consists of four (4) D4D subsets of data: (a) Communications between relay antennas, (b) Displacement trajectories: set of high-resolution data, (c) displacement trajectories: low resolution dataset, and (d) communication subgraphs.

2.1.1. Communications between relay antennas

This dataset provides the number of calls and the total duration of calls made between the relay antennas, aggregated per hour. It also informs which antenna initiated the call.

2.1.2. Trajectories: high resolution dataset

This dataset contains a sample of active users randomly drawn for each period of two-week. It provides the timestamp and the positions of the antennas for calls and SMS exchanged during this period of two week.

2.1.3. Trajectories: low resolution dataset

This dataset includes a sample of active users randomly selected. It provides the timestamp and location of calls and SMS messages exchanged by these users for a period of five months. Here, the location is not filled by the position of the antenna but the Sub-division.

2.1.4. Sub graphs communication

This dataset contains several thousand users (egos) randomly selected. From these users, it builds communication graphs based on all communications made between the egos and their contacts, second degree (neighbors of neighbors). Communications included in each sub graph are aggregated every two weeks over a period of five months.

2.2. Choice of the data management model

The question of information management raised from the systems operation is a real concern in Computer Science. The polling frequency, size, structure or the nature of the data are all factors that make the problem complex. The DBMS are widely adopted to store and use data from information systems. However, relational DBMS technologies nowadays have obvious limitations.

Indeed, aggregation of systems gave rise to mega-system, big size data sources on which operation requires new techniques and technologies. The data management solutions adopted without taking into account this expansion is then quickly outdated. This is the case of YouTube, which receives online every minute, the equivalent of twenty-four hours of video. Due to this rising power, they have been bought by Google in 2006, had developed new techniques for big size data warehouse uses in many of its services such as Google Reader, Google Maps, Google Earth, Google Code hosting, Gmail, etc.[OLBE2010].

2.2.1. The relational DBMS

The relational model is characterized by data storage techniques in two-dimensional tables, also called relationships. The data is manipulated by operators of relational algebra; the consistent state of the database is defined by a set of integrity constraints. The success of the relational model is due to the power and simplicity of its concepts. The volume of data already stored in relational model constitutes a strong inertia to move to other types of databases. Although very suitable in the context of business, relational databases have their limitations when used in a wider environment as a popular website in load balancing.

Indeed, web sites such as Google and Facebook, have several million entries in their databases and so for daily visits. Therefore, a single machine cannot manage the database. For reasons of reliability and efficiency, these databases are duplicated so that the service is not interrupted in case of failure. In addition, certain types of queries are not optimized in any relational database. Data in a relational database are stored in rows, therefore, to answer some queries, you should browse the contents of the entire table. Because of these limitations, new technologies have been developed to meet these needs, other types of architectures were then invented to meet these challenges. These are called as a whole "NoSQL databases" for "Not Only SQL».

2.2.2. NOSQL

NoSQL is a class of DBMS that does not take into account the classical architecture of relational databases. The logical unit is no longer the table, and the data are generally not handled with SQL. There are four basic categories of NoSQL [WAL]:

- **Columns oriented**

They are based on the concept of Google's BigTable. While the columns are static for a relational database, they are dynamic for a column-oriented database, it is possible to add columns dynamically and there is no cost of storage for null values. These databases are for use where there are needs to store per-user data and unique data; example: HBase[1], Hypertable [2] or Cassandra[3].

- **Based on Graph Theory**

They are based on graphs. The typical use case is the social networks where the aspect graph fully makes sense, but where complex relationships between the actors need to be described (machine parts, tree); example: Neo4J[4]

- **Key-value oriented**

Key-value databases are big hash-tables, its particularity lies in the fact that entities are in a majority of cases extracted from an identifier; example: DynamoDB[5].

- **Document Oriented**

[1]http://hbase.apache.org/
[2]http://hypertable.org/
[3] http://cassandra.apache.org/
[4]http://www.neo4j.org/
[5] http://aws.amazon.com/fr/dynamodb/

They are an evolution of key-value databases, where a key is associated with a document whose structure is free; example: CouchDB[6].

In this project, we need to handle datasets of approximately 2.5 billion records, making available to us in the context of D4D competition. The size of this database resulting from calls and SMS exchanged between about 5 million mobile phone users in Ivory Coast, is considerably high. We classified it in the category of big size data or simply big data, this is what will justify our operation's choice.

3. Exploitation of the D4D data

Despite the use of large data management techniques, the manipulation of datasets making available to us in the context of D4D competition requires a judicious storage choice.

3.1. Data storage

We have adopted a column-oriented NoSQL structure model of data storage. This is justified by the fact that the D4D data consist of information on users, relay antennas or Sub-divisions, which are autonomous entities. Thus, the column-oriented NoSQL presents itself as the most suitable structure to store the information for each entity. Cassandra is a tool that implements this type of data storage. The Cassandra project was initiated by Facebook in 2008, it is supported by the Apache Software Foundation. Its simplicity and open source distribution makes it an ideal research tool [DAT2012].

3.1.1. Structural organization of the D4D datasets in Cassandra

Cassandra data model is a dynamic schema of column-oriented data. This means that, unlike a relational model database, we do not need to model all the columns required by the application in advance and for each line, it is not obliged to have the same number of columns. Columns and their metadata can be added to the application as you need without stopping the application.

In Cassandra, a **KEYSPACE** is the data container of the application, similar to a database or a diagram in a relational database. Thus, we are working on a KEYSPACE called **D4D**. Inside the KEYSPACE, there are one or more **COLUMN FAMILY** objects, which are analogous to tables. Cassandra does not support relationships between COLUMN FAMILIES; there is no concept of foreign key in Cassandra and joins while querying is not supported. Each COLUMN FAMILY has an autonomous set of columns that are intended to be viewed together to meet the specific requirements of the application.

In our D4D KEYSPACE, we chose to store each dataset in its COLUMN FAMILY. Indeed, this choice is justified by the fact that each set of data can be operated independently. However, before storing the data in COLUMN FAMILIES, we performed optimizations on these data.

i. Communications between relay antennas

We stored the dataset in the COLUMN FAMILY named ***CF_SET1***. Using the ***ALGO_SET1*** storage algorithm (see Appendix), we have gone through the file ***SET1.TSV*** to reorganize information so as to have per hour and for each relay antenna, the number of outgoing calls, the number of incoming calls, the duration of outgoing calls and duration of incoming calls as shown in the following structure shown table 1:

Key storage (ID#DATE)	NBR_AP_SOR	DU_AP_SOR	NBR_AP_ENT	DU_AP_ENT
10#2012-04-28 23:00:00	**7**	**103**	**11**	**300**
…	**…**	**…**	**…**	**…**

Table 1: storage structure CF_SET1

[6] http://couchdb.apache.org/

Storage key: is the storage key composed of the identifier of the antenna concatenated with the date, the two elements being separated by the character '#'.

NBR_AP_SOR: is the number of outgoing calls from the antenna to the given date. It is obtained by summing all calls originating from the source antenna itself and whose date corresponds to a given date.

DU_AP_SOR: is the duration of outgoing calls from the antenna to the given date. It is calculated by taking the sum of all durations of calls originating from the source antenna itself and whose date corresponds to a given date.

In a dual manner, we define the number of incoming calls (**NBR_AP_ENT**) and duration of calls (**DU_AP_ENT**).

- **Computation of the size of the COLUMNFAMILY *CF_SET1***
  - The total number of antennae is: **1231**;
  - The total number of dates (number of hours over the period of 5 months or 150 days): **3600**.
  - The size of the COLUMN FAMILY CF_SET1 is: 3600 x 1231 = **4,431,600**.

On such a table, we can consider a linear browse. Indexing used, permits us to directly access the desired item (complexity: O (1)) and thereby facilitates treatment.

ii. Displacement trajectories: low resolution dataset

The dataset **SET2** (displacement trajectories: high resolution dataset) and the dataset **SET3** (displacement trajectories: low resolution dataset) have the same structure. The difference is that: in the first case, the location is indicated by the position of the relay antenna. In the second, it is indicated by the position of the Sub-division. For this purpose, we chose to base our study on the dataset SET3 which is larger. However, the theory can be applied to the dataset SET2.

We stored the dataset SET3 in the COLUMN FAMILY named ***CF_SET3***. Using the ***ALGO_SET3*** storage algorithm (see Appendix), we have gone through the **SET3.TSV** file and reorganized the information so as to have per hour and for each given pair of Sub-division (x, y), the number of users passing from Sub-division of x to the Sub-division y as shown in the following structure in table 2:

Key storage (ID_X# ID_Y# DATE)	**Nombre d'utilisateur qui sont allé de x vers y**
15#16#2012-04-28 23:00:00	19

Table 2: Storage Structure CF_SET3

The storage key is made by concatenating the identifier of the Sub-division of origin with that of the Sub-division of destination followed by the date, the elements being separated by the character "#".

- **Compuutation of the size of the COLUMNFAMILIES CF_SET3**
  - The total number of Sub-division is: **255**
  - The number of pairs of Sub-division is: C_{255}^{2} = **64,770**
  - The total number of dates (number of hours over the period of 5 months or 150 days): **3600**
  - The size of the COLUMN FAMILY CF_SET3 is: 3600 x 64770 = **233 172000**.

Remarks:

1. We have not worked with the **SET4** dataset (communication subgraphs) since in our study, mobility and the exchange rate between telephone users interests us more than the telephone interconnections between them.

2. For data concerning the coordinates of the relay antenna and those concerning the coordinates of the Sub-divisions, we have stored them as they were made available (by making the identity point to the coordinate pair longitude-latitude).

3. We also stored the table of dates. It consists of 3600 elements corresponding to 3600 hours or 5 months of communication (cf. ***ALGO_DATE*** appendix).

3.1.2. Physical architecture of data storage

We worked on a site consisting of three machines: Machine 1 Machine 2 and Machine 3. Machine 1 is the testbed, the machine 2 is the operational host and the Machine 3 is a real-time replication of the machine 2.

- **Characteristics of machines**
  - Machine 1: intel Core 2 (2 x 2 GHz), 3 Gb ofRAM
  - Machine 2: intel Core 2 (2 x 2 GHz), 4Gb ofRAM
  - Machine 3: intel Core 2 (2 x 2 GHz), 3 Gb of RAM

Our algorithms were implemented in Java and we worked in the NetBeans IDE version 6.9.

3.2. Querying Data and hypotheses

Thus stored, the data can be queried and observed on several ways.

3.2.1. Statistically

The exploitation of the SET1 dataset allows us to visualize the variation in time of the number of incoming and / or outgoing data from a given relay antenna. The goal is to correlate this variation with social phenomena. Data were collected over a period of 05 months, which corresponds to 3600 hours. We associate with each of these hours for each antenna, the number of incoming and / or outgoing calls.

In the example in Figure 1 below, we show the time variation of the number of outgoing calls from the antenna whose identifier is 1000. The x-axis corresponds to the time and the ordinate axis corresponds to the numbers of calls. Given the high number of graduation, we did not bother to write them on the axes. However, on the graphic, by placing the mouse pointer over a given point (in red) we have the corresponding date as well as the number of outgoing calls from the antenna on that date. To obtain the variation curve, we link each point to the next by a line segment.

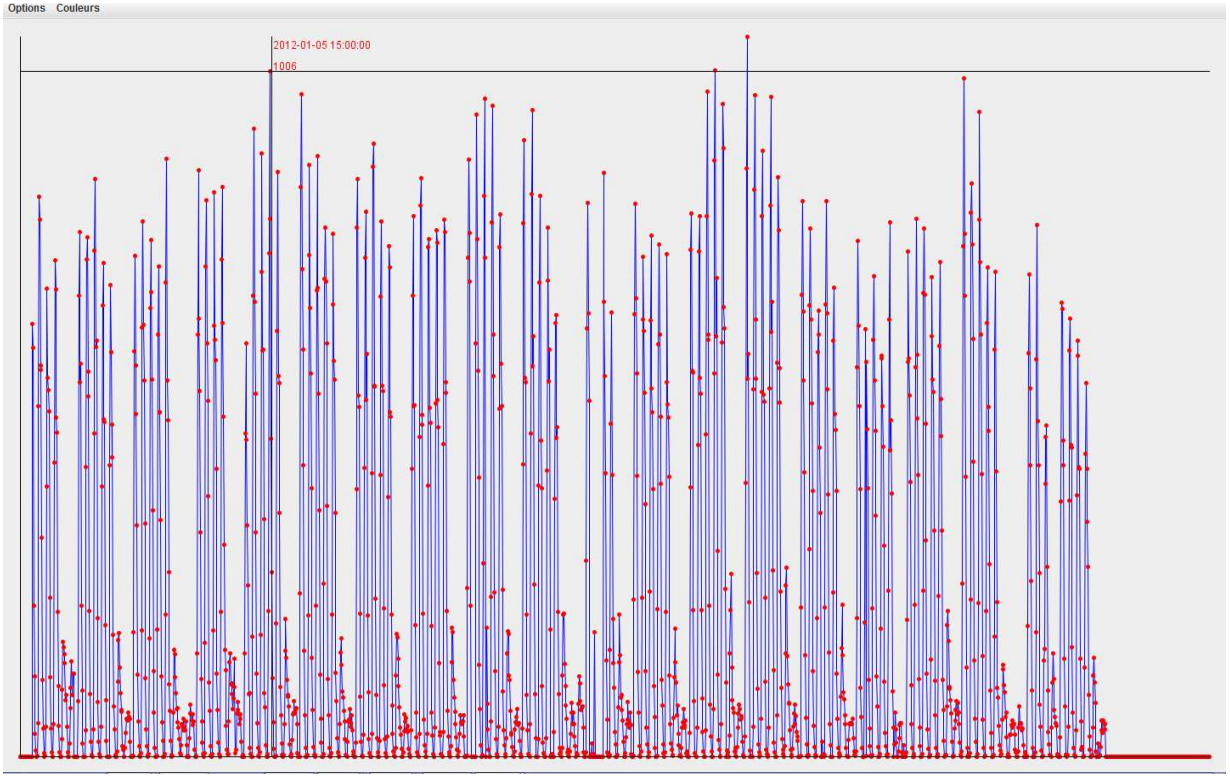


Figure 1: Graph showing the changes in the number of outgoing calls from the antenna identifier equal to 1000 as a function of times

For the same antenna, we present in Figure 2 below the graph representing the variation of the length of calls over time. It is obtained and read as in the case above.

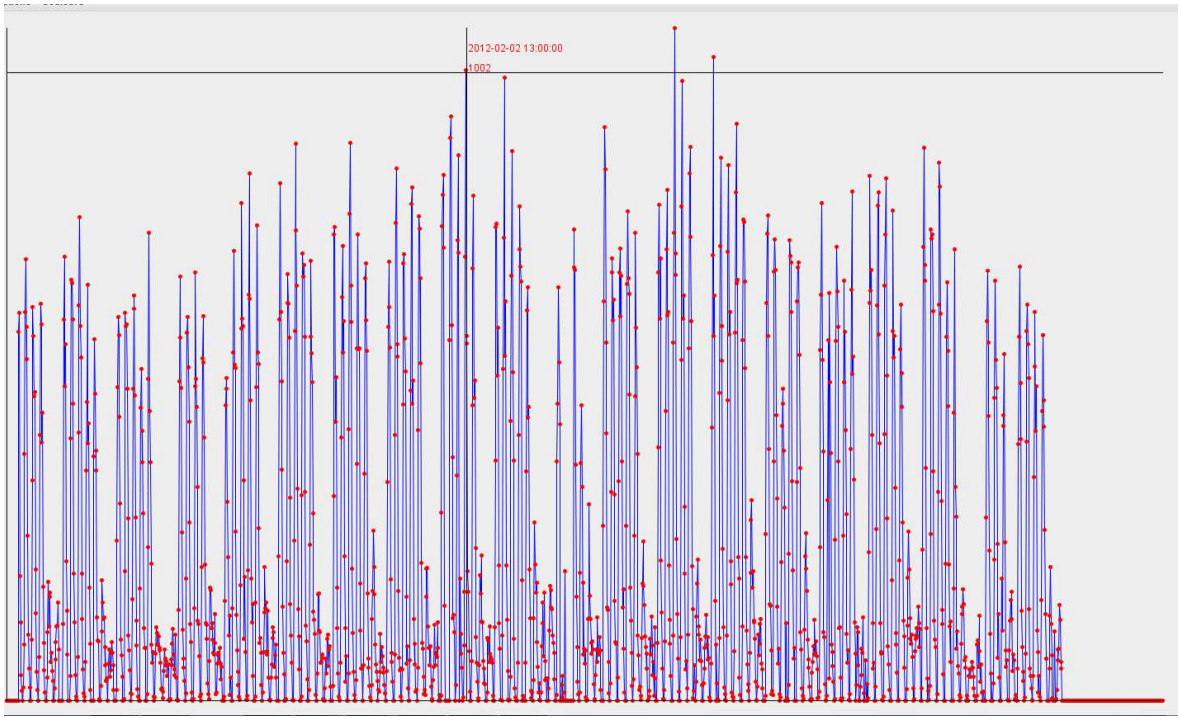


Figure 2: Graph showing the variations with time of the duration of outgoing calls from the antenna identifier equal to 1000

In the same logic, we represent the variations over time in the number of displacements between a Sub-division of origin A to a Sub-division B. In Figure 3 below, the Sub-division A of identifier equal to 60 and the Sub-division B of identifier equal to 61.

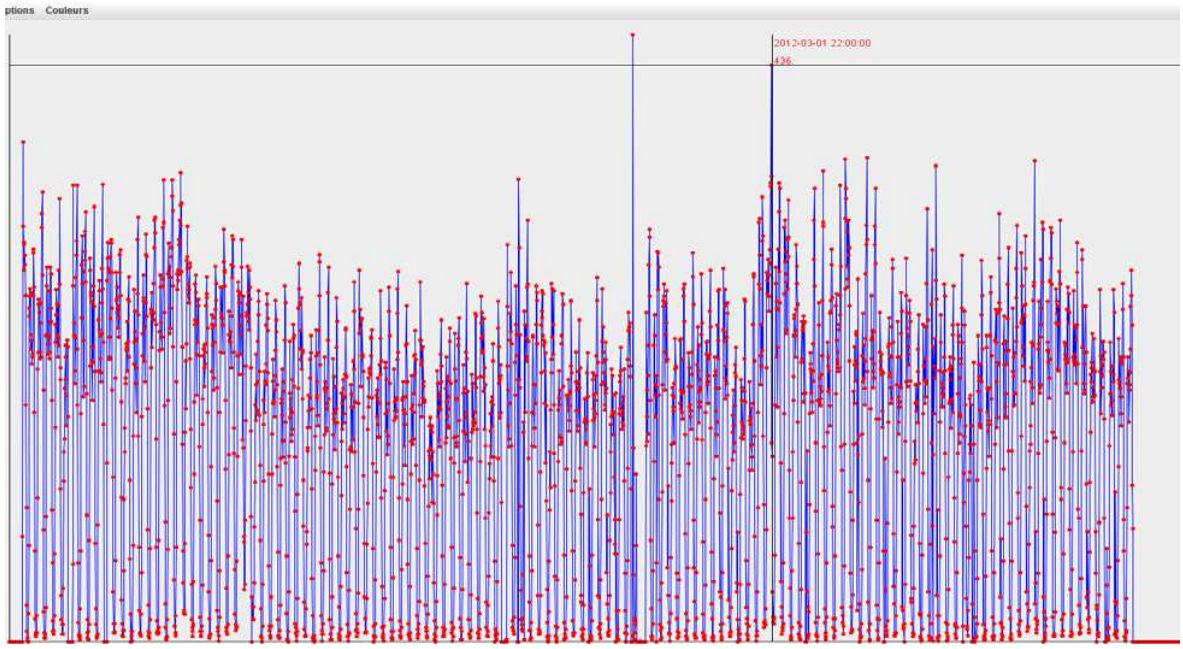


Figure 3: Graph showing changes over time in the number of displacements to the Sub-division of identifier equal to 60 to the Sub-division of identifier equal to 61

To better understand these changes, we considered a study of the geographical distribution of elements (relay antenna, Sub-divisions).

3.2.2. Geographical aspect

The study and the graphical representation of the distribution of antennas and Sub-divisions across the territory of IC resulted in two graphs: The first illustrated figure 4 concerning the position of the relay antennas, allows us to visualize their distribution but also have the identifier, coordinates of a given antenna simply by placing the mouse pointer. The second graph (figure 5) shows the same type of information for the Sub-divisions.

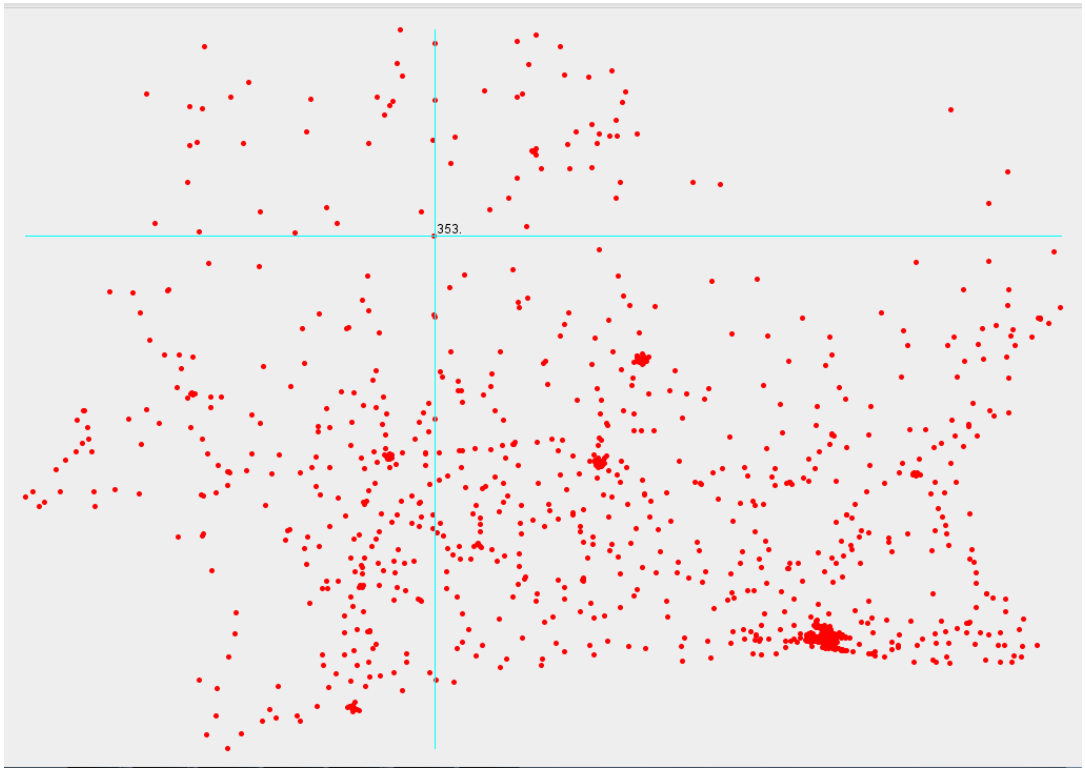


Figure 4: Graph showing the geographical distribution of relay antennas

The antenna located on Figure 5 above has for identifier 353, its coordinates in the file are: latitude: 8.740131 and Longitude: -6.193120.

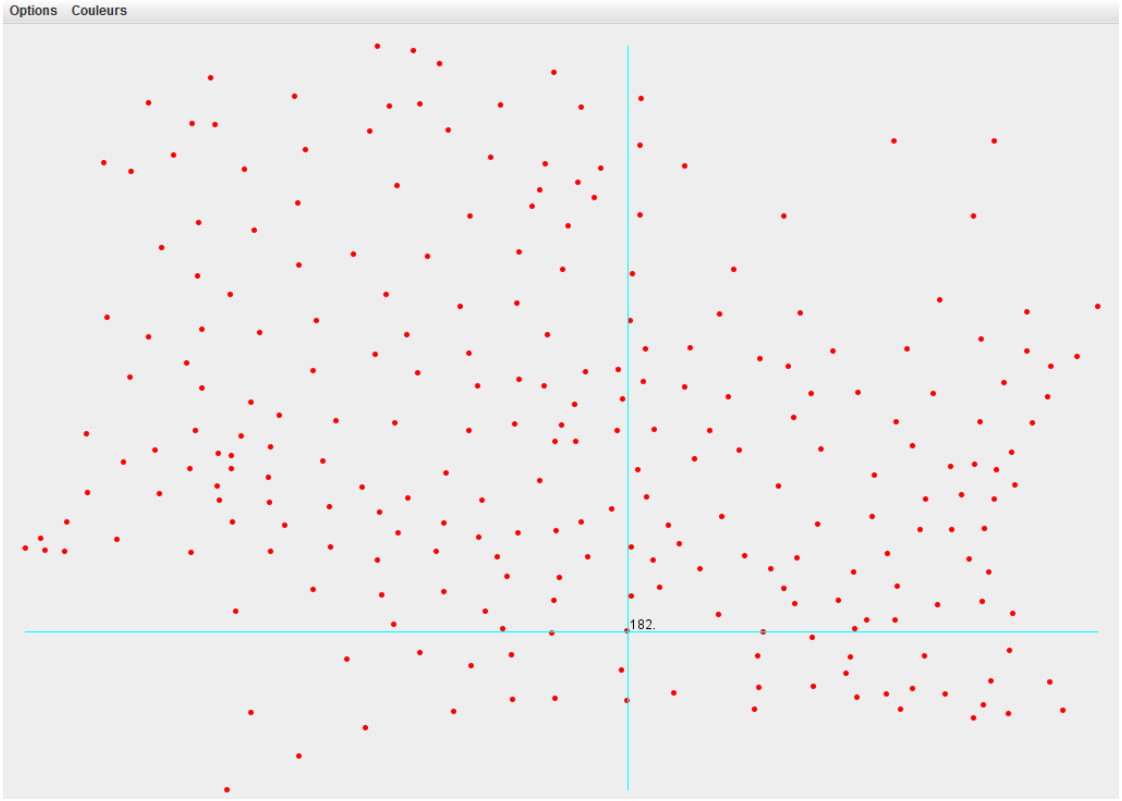


Figure 5: Graph showing the geographical distribution of Sub-divisions

The Sub-division located in Figure 5 has 182 identifying its coordinates are in the file are: latitude: 5.855107 and Longitude: -5.251963.

3.2.3. Interpretation of graphs

By varying the imaging parameters (identifier antennas, Sub-division of origin, destination Sub-division), we observed several graphs that have for certain parameters a uniform change, for others, an sawtooth evolution, etc. One can therefore wonder and make some assumptions: To what corresponds the extrema (maxima and minima) that characterize the variation of the flow communication generated by a given point (antenna) or by a geographical area (sub-division)?

On the graphs, we can clearly distinguish daytime periods characterized by spikes regular communication, nighttime periods marked by a low exchange rate. Beyond these trivial observations, we can intuitively consider that periods marked by extreme exchange rate (too high or too low), correspond to specific events. For example, a technical failure can paralyze telephone conversations over several days. In contrast, significant events such as New Year celebrations make grow the need for communication. These observations corroborated by the graphs confirm our idea that any event that may affect the telephone communication system (equipment-users) immediately affects the variation of the telephone exchange rate. Therefore, we must ask ourselves what kind of influence can exist between the events and the flow of communication. For example, how from the variation of the communication flow, we can identify a technical failure? It is clear that when the communication flow vanishes drastically and remains continuously over several hours or several days, you immediately think of a technical problem. However, we need well-developed models to identify a fault that moves slowly. Just like the study of antibody present in a human body attests or not to its infection to AIDS HIV, we wish to diagnose the communication state from the communication flow in a given area.

In this work, we wish to theoretically identify the type of relationship that exists between phenomena. For now, we simply study the relationship between the communication rate and the prevalence rate of HIV AIDS for the population in Ivory Coast regions. We suspect the existence of a linear correlation between these two variables; however, it is only after the mathematical approach that we can be sure.

4. Correlation Study: Application to the prevalence of HIV in Ivory Coast

4.1. Mathematical and statistical methods

4.1.1. Correlation calculation

Saying that there is a correlation between two variables means that the knowledge of the value taken by one provides information on the value taken by the other. The correlation coefficient measures the degree of linear association between two variables. It is calculated as follows:

$$\mathbf{r} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\,\mathbf{x}\,\sqrt{\sum(y_i - \bar{y})^2}} \qquad \textbf{avec } -1 < r < +1$$

Where:

- x and y respectively represents the explanatory variable and the dependent variable.
- The x_i represent the values taken by the explanatory variable.
- The y_i represent the values taken by the dependent variable.

The closer r is to -1 or 1, the more the correlation is stronger and the two variables are related. We consider that the correlation link is significant when r> 0.85 or r <-0.85. If r is positive, the two variables move in the same direction, if r is negative, the two variables move in opposite directions.

4.1.2. Calculation of variance and covariance

These calculations are given by the following expressions where $\bar{x}$ and $\bar{y}$ represents respectively the average x_i and y_i.

Var(X) =$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$

Var(Y) =$\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$

Cov(X, Y) = $\frac{1}{n}\sum_{i=n}^{n} x_i y_i - \bar{x}\bar{y}$

4.1.3. Linear regression model

Data: we have a sample of n pairs (xi, yi) for i = 1; ::: n independent of each other.

Linear regression: we seek a linear relationship between X = explanatory variable = regression variable and Y = dependent variable = response.

Linear model: Y = aX + b +ε where ε is a random variable called residual error satisfying E[ε] = 0 et Var(ε) = σ^2.

Regression line: Y = aX + b to adjust on data in the sense of least squares.

Estimated regression line (best fit line): Y = $\hat{a}$X + $\hat{b}$ with $\hat{a}$ =estimator of a and $\hat{b}$ = estimator of b. According to the least squares method:

$$\hat{a} = \frac{COV(X,Y)}{VAR(X)}, \quad \hat{b} = \bar{y} - \hat{a}\bar{x}$$

We call **predicted response** at the point x_i , the value $\hat{y}_i = \hat{a}x_i + \hat{b}$.

We call **residue** at the point x_i the value $e_i = y_i - \hat{y}_i$ representing the difference between the observed response y_i and the predicted response $\hat{y}_i$.

It is a question of calculating the adjusted coefficients $\hat{a}$ and $\hat{b}$ to determine the linear model between the explanatory variable X and the dependent variable Y.

4.2. Application to HIV

Ivory Coast is the first country affected by the HIV / AIDS in sub-Saharan Africa, with a prevalence rate of about 4.7% [EIS2005]. This rate is divided by region as follows shown in Table 3:

Region	Rate
Nord	3,2
Nord-Ouest	1,7
Nord-Est	3,3
Centre	4,8
Centre-Ouest	3,7
Centre-nord	3,6
Centre-Est	5,8
Sud	5,5
Sud-Ouest	4,2
Ouest	3,5
Ville d'Abidjan	6,1

Table 3: Prevalence of HIV by region in Ivory Coast [EIS2005]

Our objective in this section is to first calculate the correlation coefficient between the communication rate and the prevalence rate of HIV in one Sub-division, then define a linear regression between the two variables.

4.2.1. Approach for calculating the number of communication in a Sub-division

In our work, we have chosen one Sub-division by region: one with the greatest number of people. Our choice is motivated partly by the fact that, the more there is inhabitant in a Sub-division, the more it is possible to find the antennas, and secondly by the fact, that the most populated sub-division has a significant weight in the statistical calculations in the region.

To calculate the number of communication in a Sub-division, we first identified the geographic coordinates (latitude and longitude) of the selected Sub-divisions. Then we found identifiers (ID) in the corresponding archive **SET3.TSV** and finally we made the sum of all exchanges in the archive **SET3.TSV** identified in a Sub-division; this means the sum of all exchanges with the same Sub-division ID. This total is considered the number of communication in the Sub-division. We obtained the coordinates of the regions from GoogleMaps[7]. The table 4 below shows the Sub-divisions selected with longitude and latitude, population, and the ID number of communications.

Region	Sub-division	Population [OCHA2011]	Longitude	Latitude	ID	Number of communications
Savanes	Kanakono	309 882	-6,267314	10,360861	99	11 961
Bafing	Koro	28 929	-6,648476	9,828754	100	6 587
Worodougou	Seguela	82 378	-6,617047	8,052000	200	103 160
Denguele	Odienne	74 548	-7,713328	9,632629	240	52 587
Zanzan	Bondoukou	94 322	-2,814230	8,038299	212	230 799
lacs	Yamoussoukro	272 901	-5,148486	6,919306	56	1 371657
N'zi-comoé	Dimbokro	103 351	-4,740074	6,765964	173	157443
Haut-sassandra	Daloa	341 162	-6,439960	6,914402	144	915953
Marahoue	Bouafle	166 291	-5,723872	7,052564	22	479656
Fromager	Gagnoa	218 517	-6,020944	6,013225	138	550067
Vallée du Bandama	Bouake	727 674	-4,938431	7,791673	39	1 086475
Agneby	Agboville	126 241	-4,344698	6,077307	17	30876
Moyen-Comoe	Abengourou	137 298	-3,397551	6,426104	3	25544
Sud-Comoe	Grand-bassam	159 605	-3,768528	5,238152	198	683779
Lagunes	Anyama	196 572	-4,066367	5,517012	64	253447
Sud-Bandama	Lakota	88 368	-5,657858	5,837364	187	174191
Bas-Sassandra	San pedro	216 366	-6,672368	5,087328	122	1 534710
Dix-huits Montagnes	Man	208 211	-7,622609	7,142829	77	23997
Moyen-Cavally	Duekoue	135 896	-7,391132	6,723418	165	202747
The town of Abidjan	Abidjan	4 131 553	-4,038918	5,654023	21	20 670757

Table 4: Selected Sub-divisions with their characteristics

4.2.2. Calculation of the exchange rate and prevalence rate in Sub-division

The exchange rate in a region is calculated as follows:

$$\text{Exchange rate} = \frac{Number\ of\ communication\ in\ the\ region}{Number\ of\ inhabitants\ in\ the\ region.}$$

[7] https://maps.google.com

Assumptions: It is assumed that the prevalence of a Sub-division is equal to the prevalence of the administrative region in which it is located. The table 5 below shows the correspondence between the prevalence rate and the exchange rate in Sub-divisions.

administrative regions	Num	Sub-division	Exchange rate	Prevalence rates
Nord	1	Kanakono	3,86	3,2
Nord-Ouest	2	Koro	22,77	1,7
	3	Seguela	125,23	1,7
	4	Odienne	70,54	1,7
Nord-Est	5	Bondoukou	244,69	3,3
Centre	6	Yamoussoukro	5002,62	4,8
	7	Dimbokro	152,34	4,8
Centre-Ouest	8	Daloa	268,48	3,7
	9	Bouafle	288,44	3,7
	10	Gagnoa	251,73	3,7
Centre-nord	11	Bouake	149 ,31	3,6
Centre-Est	12	Agboville	24,46	5,8
	13	Abengourou	18,60	5,8
Sud	14	Grand-bassam	428,42	5,5
	15	Anyama	128,93	5,5
	16	Lakota	197,12	5,5
Sud-Ouest	17	San pedro	709,31	4,2
Ouest	18	Man	11,53	3,5
	19	Duekoue	149,19	3,5
Ville d'Abidjan	20	Abidjan	500,31	6,1

Table 5: Prevalence and exchange rates by Sub-division

4.2.3. Calculation of the coefficient of correlation, the variance and covariance

In continuation of our work, we denote by X (predictor) the statistical variable which represents the statistical exchange rate and Y (dependent variable), which represents the prevalence rate. Thus, x_i and y_i are respectively, the exchange rate and the prevalence rate in the Sub-division number I (see table 6 below).

Num	Exchange rate	Prevalence rates	Product		
i	x_i	y_i	$x_i y_i$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
1	3,86	3,2	12,352	187951,729	0,748225
2	22,77	1,7	38,709	171913,061	5,593225
3	125,23	1,7	212,891	97446,3629	5,593225
4	70,54	1,7	119,918	134581,857	5,593225
5	244,69	3,3	807,477	37134,8316	0,585225
6	5002,62	4,8	24012,576	20841288,4	0,540225
7	152,34	4,8	731,232	81255,7829	0,540225
8	268,48	3,7	993,376	28531,9394	0,133225
9	288,44	3,7	1067,228	22187,2941	0,133225
10	251,73	3,7	931,401	34471,1209	0,133225
11	149,31	3,6	537,516	82992,3911	0,216225
12	24,46	5,8	141,868	170514,488	3,010225
13	18,60	5,8	107,88	175388,414	3,010225
14	428,42	5,5	2356,31	80,532676	2,059225
15	128,93	5,5	709,115	95150,0393	2,059225
16	197,12	5,5	1084,16	57731,5951	2,059225

17	709,31	4,2	2979,102	73938,3111	0,018225
18	11,53	3,5	40,355	181360,146	0,319225
19	149,19	3,5	522,165	83061,5456	0,319225
20	500,31	6,1	3051,891	3958,42306	4,141225
Total	8747,88	81,3	40457,522	22560938,264736	36,8055
Average	**437,394**	**4,065**	**2022,8761**	**1128046,915**	**1,840275**

Table 6: Correspondence table

i. Point cloud

At this processing level, we have a table of correspondence between the values of the variable X and variable Y. We plotted the point cloud and the linear trend line. In this representation, the pair of (x=5002,62 and y= 4,8) does not appear. This point corresponds to the Yamoussoukro Sub-division, which presents a marginal remoteness region, compared to the overall distribution. To get a good view of the graph, we have ignored it.

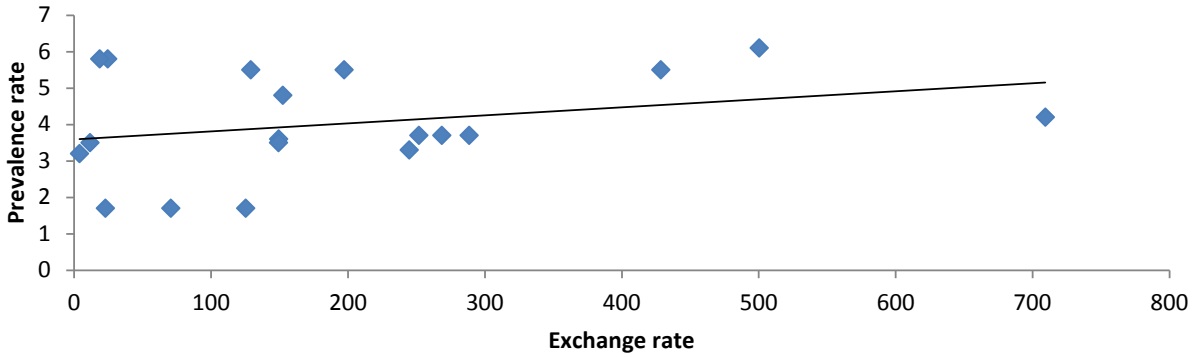


Figure 6: Point cloud and linear trend line

From the graph of Figure 6, there is a low linear relationship between the exchange rate and the prevalence rate. We will verify this statement using calculations.

ii. Correlation coefficient

$$\mathbf{r} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2}\,\mathbf{x}\,\sqrt{\sum(y_i-\bar{y})^2}} = \frac{\mathbf{15212252,17}}{\mathbf{4749,835604x6,066753662}} = \mathbf{0,11114245}$$

Remarks:

- r> 0 indicates that the two values change in the same direction. This explains the slight linear relationship observed graphically.
- 0.0 <r <0.5: this indicates that there is little correlation between the two values. So, we can conclude that there is little connection between the two variables. Therefore, it is possible to determine the model of linear relationship between en the two variables.

iii. Variance

Var(X) $= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ = 1128046,915

Var(Y) $= \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$ = 1,840275

iv. Covariance

Cov(X, Y) = $\frac{1}{n}\sum_{i=n}^{n} x_i y_i - \bar{x}\bar{y}$ = 244,86949

4.2.4. Estimated regression line

The estimated regression line is of the form Y = $\hat{a}$X + $\hat{b}$. We will calculate the coefficients $\hat{a}$ and $\hat{b}$.

$$\hat{a} = \frac{COV(X,Y)}{VAR(X)} = 0{,}00022$$
$$\hat{b} = \bar{y} - \hat{a}\bar{x} = 3{,}97$$

We obtain the estimated regression line (the line of least squares) following:

Y = 0,00022X + 3,97.

Nevertheless, the relationship between the exchange rate and the prevalence rate is not exactly linear because all points on the graph of Figure 6 are not aligned. Thus, there remains a residue calculation.

4.2.5. Simple linear regression model

The simple linear regression model allows to estimate the value of the prevalence rate (y_i) knowing the exchange rate (x_i) in a region. It is of the form: Y = aX + b + ε where ε is a random variable called residual error satisfying E[ε] = 0 and Var(ε) = σ^2 (E[ε] is the average of the estimated residuals). Using the estimated parameters $\hat{a}$ and $\hat{b}$, we can calculate for each pair (x_i, y_i) an estimated residue as $\hat{\varepsilon_i} = y_i - \hat{a}x_i - \hat{b}$. Therefore, this residue is specific to each couple.

We obtain the following simple regression model:

Y = 0,00022X + 3,97 + ε

5. Interpretation of results

After applying the mathematical model of linear correlation in the study of the relationship between the level of communication and the prevalence of HIV AIDS, we must identify some interpretations.

The value of the coefficient of linear correlation is egal to **0.11114245**, which proves that there is a weak linear correlation between the two variables and that they grow in the same direction. The fact that the correlation is low may mean that there are other factors that disrupt the relation between the two variables. Indeed, when the observations (x_i) deviate from the average value, the coefficient of linear correlation tends towards 0 (low linear correlation). Thus, the communication peaks observed during the events days and feasts as well as the low rate of communication (often zero over several weeks) probably related to technical failures are all disruptive whose consideration could significantly improve our model.

5.1. Social impact: the case of public health

The correlation that exist between the prevalence rate of HIV AIDS and the communication rate can allow the actors in charge of the fight against AIDS to better understand the target population and to deploy themselves more effectively. Indeed, if one knows how to evaluate after a specified period, the rate of the communications (calls and SMS) on a given geographic area, then using the correlation model proposed, we can estimate the prevalence of HIV of the geographical area. This study can be generalized to other types of pathology that affects public health such as malaria.

5.2. Business model

The exploitation of the model proposed in this paper in a real operating environment requires a minimum of strategy, resources and organization. As a first step, we must put in place a system for the collection, storage and manipulation of data generated by the telephone exchange. For this to be possible, there must be a dominant contribution of the telecommunication service providers or operators. Then it is necessary to analyze the mass of data to highlight the information one needs. The implementation of the graphical analysis techniques is necessary for a proper understanding of the distribution. Finally, we should highlight the results obtained in a view of decision-making.

6. Conclusion and future works

Based on the topic "*Design and implementation of a tool for the correlation between the prevalence rate of a pathology and the communication flow between different localities*," our participation in the D4D'2013 competition , was focused on the implementation of a data exploitation approach make available to us in order to establish a linear correlation between the communication exchange rate of inhabitants of a geographical area and the prevalence rate of a pathology. To achieve this, we have adopted an approach which consists of several steps. Firstly, we have elaborated from the Cassandra tool, an effective strategy of data management. Then, we have evaluated and graphically showed throughout the period of sample collection, the variations of communication flows. Finally, we have elaborated a statistics theoretical model whose application to the particular case of prevalence rate of HIV AIDS allowed us to have conclusive results.

However, the lack of up-to-date and specific data (based on the Sub-divisions and not on regions) on the prevalence rate of pathologies in Ivory Coast constrained us to make approximate choices such as using 2005 statistics on the prevalence rate of HIV AIDS in Ivory Coast. Beyond the results obtained at the end of this research exercise, we must emphasize on the main idea which consist of exploiting the communication flow of a telecommunications service provider to understand and explain the status of a population. It should probably pursue further study to unveil all its splendor. A track for further studies consists for example to remove from the sample the disturbance flow. This is data that intuitively have nothing to do with the pathology: the increased communications during feast days, the lack of communication in cases of failure, etc. Knowing that the more significant flow is that concerning the pathology (for example calls from families to hospitals to inquire on the status of their patients), we may think of a more effective use of data. To do this, you must have enough information about the study area: location of each hospital, its visitation rate, just to mention the few.

Bibliography

[DAT2012] DataStax, Apache Cassandra™ 1.1 Documentation, December 02, 2012

[EIS2005] *Enquête sur les Indicateurs du Sida, Ministère de la Lutte contre le Sida Institut National de la Statistique Project RETRO-CI Abidjan*, Ivory Coast , 2005

[MAL2012] Xavier MALETRAS, *LE NOSQL* – CASSANDRA, Thèse professionnelle, 27/05/2012

[OCHA2011] OCHA Cote d'Ivoire, *Estimation population 2010*, Septembre 2011

[OLBE2010] Olivier BENDAVID (18/06/10), *Biin The Cloud* – Mémoire de Master II – Académie de Montpellier –Université de Montpellier II – Sciences et Techniques du Languedoc.

[RAP2011] *RAPPORT ANNUEL 2011*, France Télécom-Orange, page 14 et page 74

[TIE] Dr. Tito Nestor TIEHI, *Activités des hôpitaux départementaux publics ivoiriens: une évaluation de l'efficacité technique par le bootstrap*

[VIN2012] Vincent D. BlondeL, Markus Esch, Connie Chan, FabriceCleroty, Ierre Deville, Etienne Huens, Fr Ed Eric Morloty, ZbigniewSmoreday, And CezaryZiemlickiy, *data for development: the d4d challenge on mobile phone data*, arXiv:1210.0137v1 [cs.CY] 29 Sep 2012

[WAL] Prof. Walter Kriha, *NOSQL DATABASES*, Selected Topics on Software-Technology Ultra-Large Scale Sites

Appendix

1. Algorithm for creating the *CF_SET1* column in Cassandra

Data of the algorithm

Input 1: file dataset SET1
Input 2: Reference to the keyspace COLUMNFAMILIES CF_SET1 D4D
Output data inserted into the COLUMNFAMILIES CF_SET1

Used data structures

ID_ANTENNE_COURANT: Integer initialized with the ID of the base station of the first communication
NOMBRE_APPELS_SORTANT: Integer initialized to 0
DUREE_APPELS_SORTANT: Integer initialized to 0

START OF THE ALGORITHM

While we have not reached the end of file,
Read a communication line in the file SET1
Identify the original antenna relay communication
Identify the number of calls related to the paper read
Identify the duration of calls to the paper read
Identify the date (day and timeslot) communication
If the ID of the antenna is equal to ID_ANTENNE_COURANT
Adds the number of calls to NOMBRE_APPELS_SORTANT
Adds the duration of calls DUREE_APPELS_SORTANT
Else
Insert (# ID_ANTENNE_COURANT date NOMBRE_APPELS_SORTANT) in CF_SET1
Insert (# ID_ANTENNE_COURANT date DUREE_APPELS_SORTANT) in CF_SET1
Replace ID_ANTENNE_COURANT value with the value read
NOMBRE_APPELS_SORTANT reset the number of outgoing calls read
Reset DUREE_APPELS_SORTANT with the duration of outgoing calls read
END_IF
END_WHILE
END_ALGORITHM

2. Algorithm for creating the *CF_KEYS* column in Cassandra

Data of the algorithm

Input 1: start_date, date (day - time) start observation set at 2012-12-05 5:00:00
Input 2: end_date, End Date 2012-04-28 11:00:00 p.m. fixed observation
Entry 3: Reference to the keyspace COLUMNFAMILIES CF_KEYS D4D
Output : data inserted into the COLUMNFAMILIES CF_KEYS

Used data structures

DAY TIME: Day - time set to the start date of observation
NUMBER: number of the current date, integer initialized to 1

START OF THE ALGORITHM

For DATE from start_date to end_date, **DO**
Insert (NUMBER, DAY-TIME) in CF_KEYS In NUMBER, put NUMBER + 1 to

END_For
END _ALGORITHM

3. Algorithm of the graphical representation of the number or the duration of outgoing calls of a relay antenna

Data of the algorithm

Input 1: ID_ANTENNE
Input 2: Reference to the keyspace COLUMNFAMILIES CF_SET1 D4D
Input 3: Reference to the keyspace COLUMNFAMILIES CF_KEYS D4D
Input 3: VALUE: Value to represent (call duration or number of calls)
Output: Curve representing the variations in the number or the duration of outgoing calls of the base station

Used data structures

COUNTER: Integer initialized to 1
ORDERED: Integer initialized to 0
ORDONNEE_MAX: Integer initialized to 0
KEY: Character set initialized with empty string
TABLE: Table memory of two dimensions

START OF THE ALGORITHM

For COUNTER from 1 to 3600, DO
In CLE, Put ID_ANTENNE + # + value corresponding to COUNTER in CF_KEYS
In ORDERED put in value corresponding to KEY and VALUE (Duration or NBREAPLS) in CF_SET1
Save the point with coordinates (COUNTER ORDERED) in TABLE
IF ORDERED greater than ORDONNEE_MAX Then
In ORDONNEE_MAX put ORDERED
END_IF
END_FOR

Reference

Calculate the scale using ORDONNEE_MAX and Window Size
Browse the table and plot the points
Join each point to the next by a segment
If modification of the size of the window, go to Reference.

END_ALGORITHM

4. Algorithm for creating of the *CF_SET3* column in Cassandra

Data of the algorithm

Input 1: file dataset SET3
Input 4: Reference to the keyspace COLUMNFAMILIES CF_SET3 D4D
Entry 3: Reference to the keyspace COLUMNFAMILIES CF_POSITION D4D
Output data inserted into the COLUMNFAMILIES CF_SET3

START OF THE ALGORITHM

WHILE has not reached the end of the file DO
Read a communication line in the file SET3
Identify the user id (ID_USER)
Identify the date of the communication (DAY TIME)
Identify the Sub-division that hosts the user (SUB-PREF_AP)
Watch the position of the user before the call (SUB-PREF_AV) in CF_POSITION
Put in CLE value: SUB-SUB-PREF_AV # # DAY TIME PREF_AP
IF its position does not exist **THEN**
initialize its position SUB-PREF_AP.
END_IF

IF position exists and is different from SUB-PREF_AP (current position) **THEN**

Update the record corresponding to the value in the table CF_SET3 CLE (number of people moved from the Sub-division SUB-SUB-PREF_AP PREF_AV to the date DAY-TIME) by adding 1 or set this value to 1 if n does not yet exist.

Set the position of the current user to SUB-PREF_AP in CF_POSITION

END_IF

END_WHILE

END_ALGORITHM

Human mobility and communication patterns in Côte d'Ivoire: A network perspective for malaria control

Eva A. Enns, PhD
John H. Amuasi, MPH, MBChB

University of Minnesota School of Public Health, Division of Health Policy and Management, Minneapolis, MN

Abstract

Malaria continue to be a great global health burden. Human mobility patterns play an important role in regional transmission by importing malaria to areas of low transmission following visits to high transmission regions. Using the data made available by France Télécom-Orange's "Data for Development" challenge, the objective of this work was to assess the importance of mobility and communication patterns for malaria control efforts. Based on mobile phone usage behavior, we constructed geographic networks reflecting patterns of human movement (based on where individuals made calls) and patterns of communication (based on the number of calls routed from one mobile phone tower to another). Our findings lead us to hypothesize that targeting vector control and behavior change campaigns to sub-prefectures of high mobility and communication may be a more cost-effective means of reducing national malaria prevalence. Targeting interventions to these regions may maximize the opportunity for downstream reductions in secondary malaria infections and for the diffusion of malaria prevention information.

Introduction

Malaria is a vector-borne protozoal disease caused by *Plasmodium* spp. and spread from human to human by the bite of the female *Anopheles* mosquito. Malaria continue to be a major global health, with the vast majority of the estimated 219 million cases and 660,000 malaria-related deaths occurring on the African continent [1]. Due to the limited flight range of the mosquito, human mobility is a key factor in the regional spread of malaria. The re-introduction of malaria to low transmission regions by individuals traveling from high transmission regions has been implicated as one of main factors contributing to the persistence of malaria, especially in areas where mosquito populations are small or have been reduced substantially, such as urban centers [2–6]. Mathematical modeling efforts have also shown that the re-introduction of disease through human mobility can be sufficient to maintain an endemic persistence of malaria in regions where, without these human movements, the disease could easily be eradicated [7]. Understanding human mobility patterns is thus critical to developing effective malaria control strategies as efforts towards elimination are intensified globally [8,9].

A wealth of information on human movement patterns has recently been recognized in the analysis of mobile phone usage data [10]. By tracking the location of the mobile phone antennas through which a given user's calls are routed over time, a temporal sequence of locations can be generated that map the individual's trajectory through space. Repeating this analysis over a large number of mobile phone users constructs a picture of population mobility trends, which may in turn have implications for disease transmission. Using mobile phone information to gain insight into human movement patterns and their implications for the control of infectious disease in Africa is promising: according to the World Bank, Africa is recognized as the region with the fastest rate of growth of mobile telephone subscriptions, growing at a rate of almost 20% each year with over 650 million subscribers at present [11,12]. Unthinkable just a decade ago, Africa now has more mobile subscribers than the United States or the European Union [13].

Mobile phone usage data has been analyzed in developing general models of human spatial-temporal behavior [14–16], inferring social networks [17,18], and understanding the implications of human movements in the spread of infectious diseases [19,20]. Only a handful of these studies have focused on mobile phone use in Africa: Tatem et al. and Wesolowski et al. used mobile phone data to characterize human movements in Zanzibar and Kenya, respectively, and assess their contribution to the regional spread of malaria [19,20]. These studies represent a growing interest in exploiting mobile phone data in the public health arena; however, the potential these applications have in informing health interventions and monitoring progress remains largely untapped.

The objective of this study is to assess the importance of mobility and communication patterns for malaria control efforts in Côte d'Ivoire. Malaria presents a particularly large burden to Côte d'Ivoire, where there were an estimated 2.5 million suspected malaria cases and 144,000 malaria infections requiring hospitalization in 2011 among its 21 million inhabitants [1]. Côte d'Ivoire has been characterized as having a low probability of success in malaria elimination efforts, largely on account of substantial incoming movement from other high transmission countries in West Africa, even though the country shares the same high endemic transmission intensities with other countries more likely to be successful at eradication efforts [9]. Côte d'Ivoire is also a country with a high mobile phone penetration. With an estimated 83% of its population using a mobile phone, the country is a well-suited setting for this kind of analysis. [21].

Methods

Mobile phone usage dataset

We based our analysis on a sample of annonymized mobile phone usage data from the France Télécom-Orange network over a five month period in Côte d'Ivoire. Orange is one of the major mobile phone service providers in Côte d'Ivoire: in 2011, Orange had over 5.8 million mobile phone subscribers in Côte d'Ivoire, representing 34% of the 17.3 million mobile phone users in the country [22,23]. These data were made available by Orange as a part of its "Data for Development" (D4D) challenge, with the goal of contributing to the socio-economic development and well-being of populations [24]. Two of the datasets made available as a part of the D4D initiative formed the basis of our analysis. The first dataset consisted of mobile trace data on a randomly-selected subset of 500,000 users, providing records of the location of every call made by that user over the entire 5-month period. Location information was recorded at the level of sub-prefectures, an administrative unit in Côte d'Ivoire, of which there are 255 in the country. From this dataset, we constructed a network representing human movement patterns from one sub-prefecture to another. The second dataset contained mobile phone antenna-to-antenna call records for all Orange mobile phone traffic during 5-month period, including the total number of calls made between two antennas and the total duration of those calls on an hourly basis. We used this dataset to construct a network representing communication (via mobile phone calls) from one sub-prefecture to another.

Mobility network

We constructed a network of human movements throughout Côte d'Ivoire based on the calling locations of a subset of users over the 5-month period. For each user, we looked at the series of sub-prefectures (in chronological order) from which they made calls. Whenever a call was not made in the same sub-prefecture as the previous call, we considered the user to have travelled from the former sub-prefecture to the latter one. We created a 255 x 255 adjacency matrix to represent these movements between sub-prefectures. For each pair of sub-prefectures, (p_1,p_2), we tracked both the total number of times users were observed to have traveled from p_1 to p_2 as well as the total number of unique individuals that that did so during the observed period. We considered movements to be directed (that is, moving from p_1 to p_2 is distinct from moving from p_2 to p_1), to account for potential differences in patterns of incoming and outgoing movements in certain locations.

Communication network

We constructed a network representing patterns of communication between sub-prefectures based on the number of calls routed through mobile phone antennas in one sub-prefecture to the mobile phone antennas in another. In the original dataset, the call traffic for each mobile phone antenna in the Orange network was recorded, including the antenna from which the call was initiated, the antenna of the call recipient, the total number of calls made between the two antennas and the total duration of those calls, reported on an hourly basis. For each pair of antennas, we aggregated the total number of calls and total duration of those calls over the 5-month period, ignoring all calls in any hour where the average call duration was less than one minute. We did not consider calls to be directed, since information could diffuse either to or from the initiator of the call. Using a geographic information system (GIS) analysis, we assigned each mobile antenna to a sub-prefecture based on its geographic coordinates provided as a part of the dataset, from which we constructed a between sub-prefecture communication network.

Contextual geographic information

To provide context to the observed mobility and communication patterns, we developed descriptions of the population density and prevalence of malaria by sub-prefecture. The most recent data available on population density (population per square kilometer) at the level of sub-prefectures in Côte d'Ivoire dates from the 1998 population census. At that time, there were 233 sub-prefectures in Côte d'Ivoire, as opposed to the current 255, which were created through the division of former sub-prefectures. Therefore, we assumed the same population density for all divisions of a given 1998 sub-prefecture and

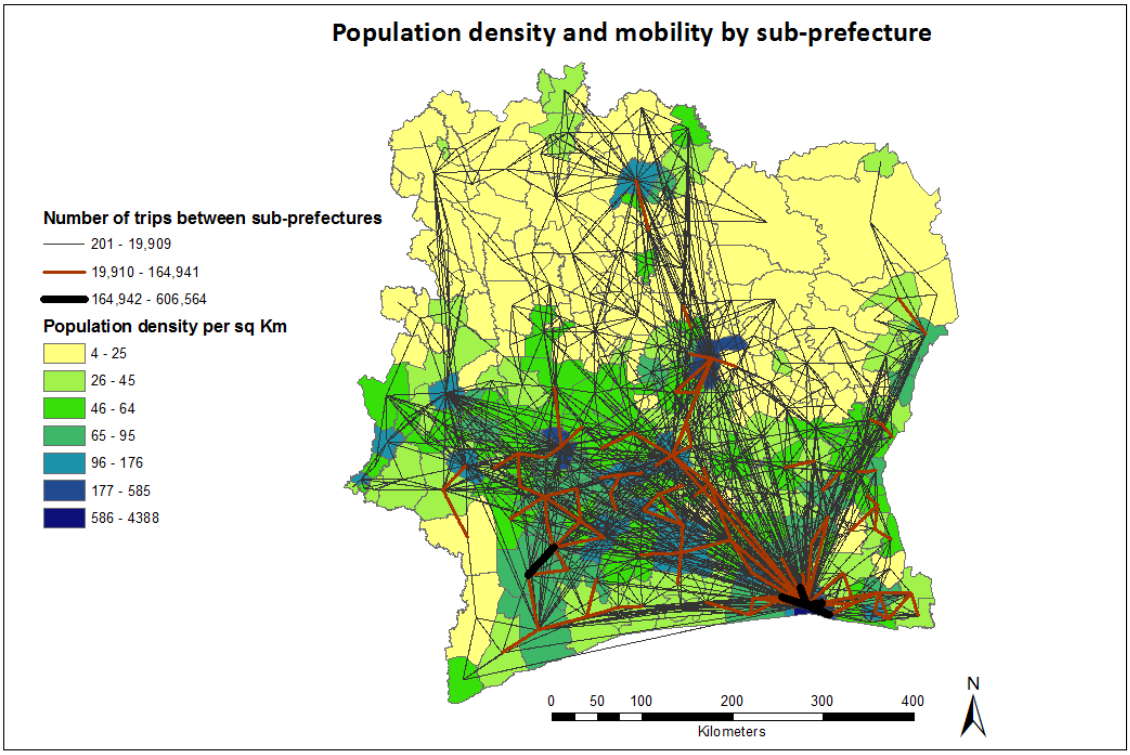
in this way created a population density map in terms of the current sub-prefecture borders. We based our estimates of sub-prefecture malaria prevalence on an analysis conducted by Raso et al. [25]. Using a Bayesian geo-statistical modelling approach, they predicted the geographic distribution of malaria prevalence for children under 16-years olds. Using these data, we calculated an average prevalence for each sub-prefecture.

Results

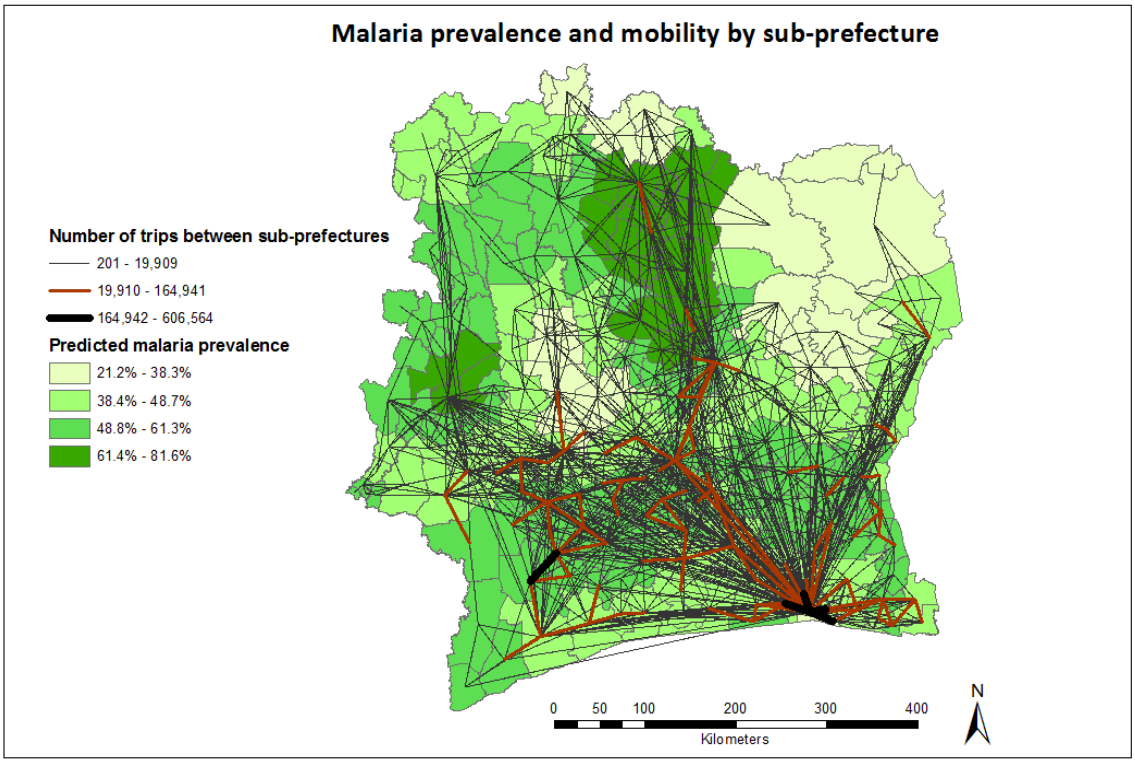
Over the 5-month period for which data was provided, nearly 419,000 users (84%) were observed to have made calls in at least two different sub-prefectures, although over 95% of users made at least half of their calls from a single sub-prefecture. A total of 16.9 million movements between sub-prefectures were observed, with each user making an average of 33 such movements over the observed period. Approximately 63% of movements were reciprocal (if movements were observed from sub-prefecture p_1 to sub-prefecture p_2, movements were also observed from p_2 to p_1). As is expected, a substantial fraction of observed movements were between urban areas, with 32% of movements occurring between the top 15% most densely populated sub-prefectures. The sub-prefecture of Abidjan, which contains the largest city in Côte d'Ivoire (also called Abidjan), alone accounted for 23% of all incoming and outgoing movements in the network. Not surprisingly, in the same way nearly 38% of the 500,000 mobile phone users made the majority of their calls in Abidjan. The relationship between mobility patterns and population density is shown in Figure 1(a).

A total of 282 million calls between antennas were observed with an average duration of 2.9 minutes per call. We noted that 43% of calls were made between antennas in different sub-prefectures and that these calls were on average twice as long as those made within a sub-prefecture (2.0 vs. 4.0 minutes). Similar to our analysis of mobile phone user movements, we find that densely populated areas dominate the communication patterns as well, though not to the same extent as the mobility patterns. Calls between the top 15% most densely populated sub-prefectures accounted for 27% of all calls made in the observed time period and a greater fraction of calls than of movements were made between the most densely populated and the least densely populated sub-prefectures (6.4% vs. 1.7%). The relationship between communication patterns and population density is shown in Figure 2(a).

We also compared the prevalence of malaria between sub-prefectures with strong connections, through either human movement or communication. Considering the strength of a connection between two sub-prefectures to be the number of trips that were made between them by the observed mobile phone users, the average difference in malarial prevalence among the top quartile of the most strongly

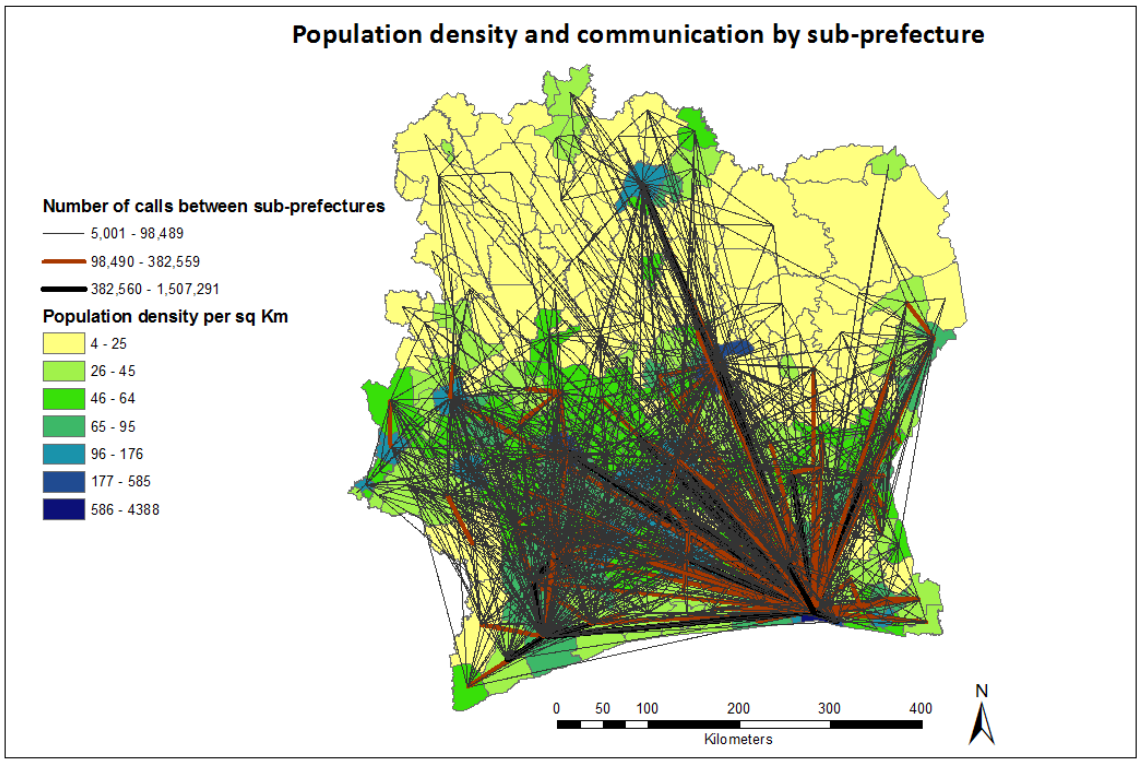


(a)

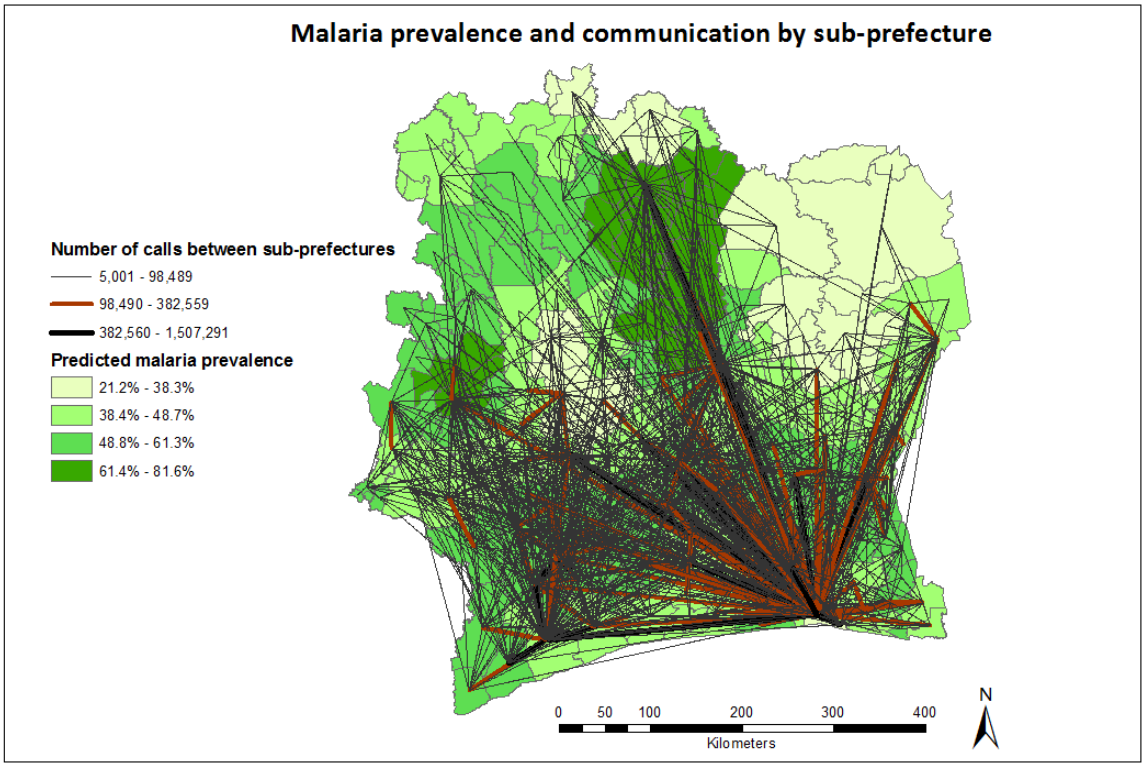


(b)

Figure 1: Movements of mobile phone users in Côte d'Ivoire between sub-prefectures are shown, weighted by the number of times those trips were made between December 2010 and April 2011, superimposed on (a) the 1998 population density and (b) the prevalence of malaria, as estimated by Raso et al. [25]. For clarity, only edges representing more than 200 trips over the 5-month observational period are shown.



(a)



(b)

Figure 2: Communication patterns of mobile phone users in Côte d'Ivoire between sub-prefectures are shown, weighted by the number of calls that were made between December 2010 and April 2011, superimposed on (a) the 1998 population density and (b) the prevalence of malaria, as estimated by Raso et al. [25]. For clarity, only edges representing more than 5,000 calls over the 5-month observational period are shown.

connected sub-prefectures was 8%, with 48% of those connections being individuals traveling from a region of higher prevalence to a region with lower prevalence. We find similar results when comparing the malarial prevalence of sub-prefecture pairs with strong communication connections (as measured by the number of calls made between them). Mobility and communication networks are shown in the context of malarial prevalence in Figures 1(b) and 2(b).

Discussion

Implications for malaria control

The role of movement in malaria eradication efforts was recognized as far back as in the 1950s by the World Health Organization [26]. Understanding movement was considered so cardinal to eradication efforts that recommendations were made to embed what was termed “geographical reconnaissance teams” with malariologists for co-operation and preparing of maps and obtaining data on population distribution and movement [2]. In fact, the role of movement has been described as a “major determinant of the required control strategy to achieve elimination and hold the line” [27].

From the findings presented, one is able to infer that within Côte d’Ivoire there much movement and communication between various sub-prefectures and that movement and communication appear to be correlated with population density. Although the official political and administrative capital city of Côte d'Ivoire is Yamoussoukro (the fourth most populous city after Abidjan, Bouaké, and Daloa), the very high level of movement and communication to and from Abidjan (the economic capital) suggests that movement and communication are largely on account of economic activity. A direct correlation between movement/communication and malaria prevalence is however not apparent. For example Abidjan is categorized within the lowest malaria prevalence group and has the highest level of movement and communication, whereas Yamoussoukro and Bouaké are categorized within the third highest malaria prevalence group out of four but still maintain a high level of movement and communication with other sub-prefectures.

A direct correlation between malaria prevalence and population density is also not apparent. For example Abidjan is the most densely populated sub-prefecture but is categorized within the lowest malaria prevalence group, whereas Bouaké is almost as densely populated as Abidjan and yet is categorized within the third highest malaria prevalence group. In addition, several of the very lowest population density sub-prefectures in the central, north-west, south-west and south-east parts of the country are categorized within the third highest malaria prevalence group, while Man, Korhogo and Bouaké, which are among the high population density sub-prefectures are categorized within the higher

malaria prevalence group. An absence of direct correlation between population density and malaria prevalence would however not be surprising, given that the prevalence of malaria is also determined by annualized rainfall and maximum land surface temperature [25].

Although malaria control interventions are generally carried out on a national scale, intuitively, strategic planning and resource allocation are often determined by population density [28,29] or malaria prevalence mapping [30–33]. We argue that the high level of mobility between sub-prefectures of varying prevalence allows for malaria to be "brought home" to a sub-prefecture of low transmission by infected individuals, following visits to areas of high transmission [6]. This phenomenon maintains prevalence levels so that intensified vector control efforts in low transmission sub-prefectures will not appear to sufficiently translate into reduced malaria prevalence. In the same way, intensified behavior change and communication (BCC) activities in high transmission sub-prefectures will not appear to translate into reduced malaria prevalence because "visitors" from low prevalence sub-prefectures will not know to adopt the preventive behaviors.

With increasing stability post the political and economic crises, levels of mobility are likely to increase along current patterns in Côte d'Ivoire, making it less likely to achieve a reduction in prevalence nation-wide if vector control is targeted with greater intensity towards areas of high transmission or high population density only. Levels of communication are also expected to increase along current patterns, suggesting that diffusion of information, which includes BCC, will also increase in the same pattern. Inhabitants of sub-prefectures with high levels of mobility and communication are more likely to diffuse malaria-related BCC information quickly to the rest of the country and influence behavior, both as they physically move and during communication. Inhabitants of high communication/mobility and low malaria transmission sub-prefectures (such as Abidjan) will also likely adopt safe practices which they will carry with them when they visit high transmission sub-prefectures, thereby reducing their risk of "binging malaria home". Inhabitants of low mobility/communication and high malaria transmission sub-prefectures will receive proportionate levels of vector control and BCC resources, but their low risk of serving as "malaria source sub-prefectures" and low probability of diffusing information and influencing behavior owing to low levels of mobility/communication will make them less likely to contribute to lowering the national malaria prevalence should resources be allocated mainly based on high transmission. Therefore we hypothesize that targeting vector control and BCC campaigns nation-wide, but with greater intensity in sub-prefectures of high mobility and communication, will be highly cost-effective by providing the greatest opportunity for a substantial downstream effect of reducing mean malaria prevalence nation-wide.

Data Limitations

Although a substantial amount of data was made available as a part of the D4D Challenge, there may be limitations to its representativeness. These data contain all call traffic over the five-month period on the Orange network, however it may over- or under-represent calls between certain geographic areas depending on the distribution of Orange subscribers and whether mobile phone users typically subscribe to multiple mobile phone plans, which may be used to call specific areas of the country based on perceived cost-effectiveness and other factors. We also cannot differentiate individual callers, so we do not know if any high-intensity calling patterns are due to widespread communication between two populations or a high level of contact between small numbers of individuals.

We based our analysis of human mobility in Côte d'Ivoire on the movement patterns of a random sample of 500,000 Orange subscribers, which may or may not be representative of the movement patterns of the general population. Furthermore, even for these 500,000 users, it is unlikely that all movements were captured in the mobile dataset. For example, if a user traveled from one sub-prefecture to another, but did not make a call, we will have no record of that movement. Thus, the completeness of this dataset depends on the intensity of mobile phone use in Côte d'Ivoire (how frequently calls are made) and how much of any individual's mobile phone usage is conducted over the Orange network, as some users may have multiple mobile phones with different carriers. The risk of contracting malaria increases with increasing length of time spent continuously in a high transmission area. Since this dataset does not allow us to adequately capture the length of time spent in a particular location, our findings might have variable implications for malaria control approaches.

Inferring a user's location based on the location of the antenna through which a mobile phone call was routed can be problematic due to mobile phone antenna oscillation, where a stationary user is assigned to any number of neighboring mobile phone antennas for to balance call traffic or due to changes in the radio frequency environment. We do not expect that this phenomenon significantly biased our findings because of the aggregate level at which we present our networks (at the sub-prefecture level). Antenna oscillation could therefore only affect users living near a border with an abutting sub-prefecture, which if present, we expect to be a small minority of users in the mobile phone dataset.

The mobile phone subscriber base itself may not be representative of the general population of Côte d'Ivoire. For example, a recent study in Kenya suggests significant biases in mobile phone ownership, and mobility in general, towards the more wealthy [34]. There is also evidence that mobile

phone owners in many African countries are more likely to be male and over 18 years old [21]. Also, lower socioeconomic status has been shown to be correlated with an increased risk of malaria infection [5]. Therefore, if mobile phone subscribers are not sufficiently distributed across the various socioeconomic segments of Côte d'Ivoire, then our findings may be less relevant in reflecting the behaviors of the populations who should be the focus of improved and expanded malaria control efforts.

Conclusions

Our hypothesis of a highly cost-effective reduction in mean malaria prevalence nation-wide by targeting sub-prefectures of high mobility/communication can be tested using simulation models of both disease transmission and information dissemination, and the costs and benefits of different intervention strategies for the control of malaria under a variety of scenarios can also be predicted. Further analyses will provide insight into the potential influence of mobility and communication networks on disease prevention efforts and will be useful for informing policy interventions aimed at the control of malaria and other infectious diseases.

References

1. WHO (2012) World Malaria Report 2012. WHO.

2. Prothero RM (1961) Population movements and problems of malaria eradication in Africa. Bull World Health Organ 24: 405–425.

3. Osorio L, Todd J, Bradley DJ (2004) Travel histories as risk factors in the analysis of urban malaria in Colombia. Am J Trop Med Hyg 71: 380–386.

4. Domarle O, Razakandrainibe R, Rakotomalala E, Jolivet L, Randremanana RV, et al. (2006) Seroprevalence of malaria in inhabitants of the urban zone of Antananarivo, Madagascar. Malar J 5: 106. doi:10.1186/1475-2875-5-106.

5. Ronald L a, Kenny SL, Klinkenberg E, Akoto AO, Boakye I, et al. (2006) Malaria and anaemia among children in two communities of Kumasi, Ghana: a cross-sectional survey. Malar J 5: 105. doi:10.1186/1475-2875-5-105.

6. Moonen B, Cohen JM, Tatem AJ, Cohen J, Hay SI, et al. (2010) A framework for assessing the feasibility of malaria elimination. Malar J 9: 322. doi:10.1186/1475-2875-9-322.

7. Cosner C, Beier JC, Cantrell RS, Impoinvil D, Kapitanski L, et al. (2009) The effects of human movement on the persistence of vector-borne diseases. J Theor Biol 258: 550–560. doi:10.1016/j.jtbi.2009.02.016.

8. Cohen JM, Moonen B, Snow RW, Smith DL (2010) How absolute is zero? An evaluation of historical and current definitions of malaria elimination. Malaria Journal 9: 213. doi:10.1186/1475-2875-9-213.

9. Tatem AJ, Smith DL, Gething PW, Kabaria CW, Snow RW, et al. (2010) Ranking of elimination feasibility between malaria-endemic countries. Lancet 376: 1579–1591. doi:10.1016/S0140-6736(10)61301-3.

10. González MC, Barabási A (2007) From data to models. Nat Phys 3: 224–225.

11. BBC (2011) Africa's mobile industry "booms". BBC.

12. World Bank (2012) ICTs Delivering Home-Grown Development Solutions in Africa.

13. Quartz (2012) Africa now has more mobile subscribers than the US or EU. Quartz.

14. González MC, Hidalgo CA, Barabási A-L (2008) Understanding individual human mobility patterns. Nature 453: 779–782. doi:10.1038/nature06958.

15. Candia J, González MC, Wang P, Schoenharl T, Madey G, et al. (2008) Uncovering individual and collective human dynamics from mobile phone records. Journal of Physics A: Mathematical and Theoretical 41: 224015. doi:10.1088/1751-8113/41/22/224015.

16. Bayir MA, Demirbas M, Eagle N (2009) Discovering spatiotemporal mobility profiles of cellphone users. 2009 IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks & Workshops: 1–9. doi:10.1109/WOWMOM.2009.5282489.

17. Eagle N, Pentland AS, Lazer D (2009) Inferring friendship network structure by using mobile phone data. PNAS 106: 15274–15278. doi:10.1073/pnas.0900282106.

18. Cho E, Myers SA (2011) Friendship and Mobility : User Movement in Location-Based Social Networks. 17th ACM SIGKDD international conference on Knowledge discovery and data mining. San Diego. pp. 1082–1090.

19. Tatem AJ, Qiu Y, Smith DL, Sabot O, Ali AS, et al. (2009) The use of mobile phone data for the estimation of the travel patterns and imported Plasmodium falciparum rates among Zanzibar residents. Malar J 8: 287. doi:10.1186/1475-2875-8-287.

20. Wesolowski A, Eagle N, Tatem AJ, Smith DL, Noor AM, et al. (2012) Quantifying the impact of human mobility on malaria. Science 338: 267–270. doi:10.1126/science.1223467.

21. Gallup World (2011) Mobile Phone Access Varies Widely in Sub-Saharan Africa.

22. France Télécom - Orange (2012) Financial Indicators in Côte d'Ivoire. Available: http://www.orange.com/en/group/global-footprint/indicateurs-financiers/financial-indicators-in-Cote-d-Ivoire. Accessed 7 February 2013.

23. Agence des Télécommunications de Côte d'Ivoire (2012) Service Mobile: Evolution du marché ivoirien. Available: http://www.atci.ci/index.php/Service-mobile/evolution-annuelle-sevice-mobile.html. Accessed 7 February 2013.

24. France Télécom - Orange (2012) Data for Development Challenge. Available: http://www.d4d.orange.com. Accessed 7 February 2013.

25. Raso G, Schur N, Utzinger J, Koudou BG, Tchicaya ES, et al. (2012) Mapping malaria risk among children in Côte d'Ivoire using Bayesian geo-statistical models. Malar J 11: 160. doi:10.1186/1475-2875-11-160.

26. WHO (1957) Expert Committee on Malaria: sixth report. World Health Organ Tech Rep Ser 38: 3–84.

27. The malERA Consultative Group on Modeling (2011) A Research Agenda for Malaria Eradication: Modeling. PLoS Med 8: e1000403.

28. Smith DL, McKenzie FE, Snow RW, Hay SI (2007) Revisiting the Basic Reproductive Number for Malaria and Its Implications for Malaria Control. PLoS Biol 5: e42. doi:10.1371/journal.pbio.0050042.

29. Snow RW, Guerra CA, Mutheu JJ, Hay SI (2008) International Funding for Malaria Control in Relation to Populations at Risk of Stable Plasmodium falciparum Transmission. PLoS Med 5: e142. doi:10.1371/journal.pmed.0050142.

30. Elyazar IRF, Hay SI, Baird JK (2011) Malaria Distribution, Prevalence, Drug Resistance and Control in Indonesia. Advances in Parasitology. Elsevier, Vol. 74. pp. 41–175.

31. Guerra CA, Gikandi PW, Tatem AJ, Noor AM, Smith DL, et al. (2008) The limits and intensity of Plasmodium falciparum transmission: implications for malaria control and elimination worldwide. PLoS Med 5: e38. doi:10.1371/journal.pmed.0050038.

32. Jorgensen P, Nambanya S, Gopinath D, Hongvanthong B, Luangphengsouk K, et al. (2010) High heterogeneity in Plasmodium falciparum risk illustrates the need for detailed mapping to guide resource allocation: a new malaria risk map of the Lao People's Democratic Republic. Malaria Journal 9: 59. doi:10.1186/1475-2875-9-59.

33. Omumbo JA, Noor AM, Fall IS, Snow RW (2013) How Well Are Malaria Maps Used to Design and Finance Malaria Control in Africa? PLoS ONE 8: e53198. doi:10.1371/journal.pone.0053198.

34. Wesolowski A, Eagle N, Noor AM, Snow RW, Buckee CO (2013) The impact of biases in mobile phone ownership on estimates of human mobility. Journal of the Royal Society, Interface 81.

Exploring Community Structure to Understand Disease Spread and Control Using Mobile Call Detail Records

M. Saravanan[1], P. Karthikeyan[1], A. Aarthi[2], M. Kiruthika[2] and S. Suganya[2]
[1]Ericsson Research India
[2]Meenakshi Sundrarajan Engineering College, Chennai, India
{m.saravanan@ericsson.com, p.karthikeyan@ericsson.com, msecarti@gmail.com}

Abstract: A **Mobile Agent Based Model** (**MABM**) is proposed in this paper with an aim to specifically control the spread of communicable diseases by taking into advantage the high degree of spatial and temporal regularities in human mobility pattern identified from their Mobile Call Details Records. With the widespread of diseases causing major public health problem, we argue that human mobility patterns not only influence the spreading, but are also useful for preventing and creating awareness of the diseases. Traditional epidemiological models which split the population into compartments that do not consider the individual information, deal with random clustering of human communities to capture the disease spreading process. In this paper, we have also proposed the **Spread Discovery Control** (**SDC**) **Model** to comprehend the spread of diseases by extracting the community structures and the analysis of mobility pattern of each agent (users) within the mobile network. The understanding of spread details helps us to propose the control strategy to avoid the spread of the epidemic disease on the specific region. The subdivision (communities) of the Mobile Social Telecom Network is validated by the process of identifying tie strength and specific agents (Influential users), who are chosen based on their degree of influence in spreading the disease to other agents inside a community. To realize our proposed models in a better way, we have modelled the communicable disease like Meningococcal Meningitis which has been initially observed from *Savanes region* and spreads to other places in the year February 2012 at *Ivory Coast*.

Keywords: Mobile Agent Based model, Spread Discovery Control model, Call Detail records, Community, Influential users, Spatio-temporal Behaviour

1 Introduction

Mobile phones call detail records has been extensively used in solving problems faced by operators to understand their customers and also for the varied areas of public interest, including education, public health and weather forecasting [16]. Today mitigating the effect of an epidemic has become the major public health concern for all countries. The outbreak of an epidemic is triggered due to person-to-person contact while the human contact networks in turn, exhibits a strong community structure. The call detail record is very useful in depicting the human contact and mobility traces and hence to find the social communities that are available in specific time period on particular location. Here we were exploring the possibilities for understanding the principles for telecom community structures that can lead to the prediction of the disease spread.

The use of traditional epidemiological models in predicting the spread of infectious diseases has been proved useful. They used the socio-economic and demographic characteristics to classify the communities [1]. The independencies in the human behavioural and mobility patterns can be noticed increasingly, in the recent past. And the traditional models fail to grab these inherent features. The spread of an epidemic can be modelled using Ordinary Differential Equations (ODE), Cellular Automata (CA) or Agent Based Modelling (ABM) [1]. The difference between ODE and other two lies in the fact that ODE does a very casual analysis at a larger level, whereas CA and ABM consider the smallest independent entities, agents as prime actors in influencing the community [3].

An agent based model is more flexible in cases where the community considered is complex and each individual is said to have their characteristic traits that distinguish them from others. The ABM defines each independent entity, with unique characteristic as an agent. They are free to move in space and with a unique set of agents to interact with, while each is flexible and autonomous with a set of rules to govern their action on the environment. Agent based models were developed to

effectively simulate the spread pattern of diseases by considering the uniqueness in each agent's behaviour [1]. However, most agent based models have not combined the real geographic environment with the agent environment [2]. In this paper, we propose a Mobile ABM that uses the Voice calls, SMS and MMS, etc related to each individual CDRs, and derives the social interaction from the individual's mobility patterns. This pattern identification has lead to realisation that each individual is being characterised by a particular travel distance and they travel to the set of few places frequently, which means they show temporal and spatial regularity [4].

The subdivision of the social network by considering the spatial and temporal aspects provides an accurate partitioning of entire network into smaller communities, in the best possible way. After the extraction of the communities, the structural properties of each community is analysed to find the related role of *strong and weak ties* in information spread among the members [5, 15]. Analysis of stronger ties allows in identifying unique users called *influencers*. They are the agents playing key role in spread. In this study we would like to focus them to control the spread of communicable diseases.

To initiate the control strategy, we have proposed Spread Discovery Control (SDC) model in this study that can provide insights into preventing the spread of disease between the communities by informing the highly influential persons who have high probability of spreading the alert to a large number of persons in his or her community. The identification of the influential users can help us to compute the various infectious and control rates possible among communities, based on whether the carrier is an ordinary agent or an influencer. We have analyzing the probability with which an influential user can spread the disease in the predicted number of rounds is very high compared to a normal user. From the early findings, we have understood that predicted infectious rates for a specific region customers can be used to determine the way of control activities in the communicable diseases spread.

Epidemic respiratory disease like Meningiococal Meningitis has been considered to be serious threat to human society in the African region was chosen as the case study. This epidemic reported the largest death toll in Africa, since 1996 so read. It forced 14 countries to implement surveillance. These countries reported 88199 suspected cases. Meningiococal Meningitis is a bacterial form of meningitis and affects the brain membrane [7]. It is transmitted from person-to-person through droplets of throat and respiratory secretions of carriers [8]. The SDC model was evaluated by studying the spread of this disease in Ivory Coast, Africa, during the 2012 epidemic. The disease originated in a place called Savanes region in Ivory Coast. The call detail records from December 2011 to April 2012 of this region was analysed and used to predict the pattern of spread of the disease.

2 Related Studies

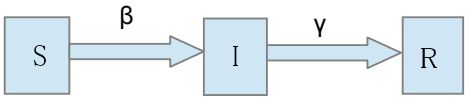
2.1 Epidemic Models

An Epidemic is said to have occurred when its spread exceeds a certain expected limit which is predicted from recent experience. Their outbreak can occur in different ways. The most common being (i) Common Source Outbreak- The affected persons were exposed to a common agent. This is said to be point source out break if all persons are affected from a single exposure. If the exposure is continuous or variable then it is termed as continuous outbreak. (ii) The second type is the propagated outbreak in which the disease spreads person-to-person (i.e.) the affected persons can further cause exposure [10]. The phenomenon of disease-spread can be explained on the basis the Diffusion principle. Diffusion, in general is one of several transport phenomena that occur in nature. A distinguishing feature of diffusion is that it results in mixing or mass transport, without requiring bulk motion. The disease diffusion process, is one in which an infectious disease spreads through

regions over time. We would like to understand the disease diffusion process which has spatial and temporal components in it. Temporal components of disease diffusion are time and duration while the spatial components of diffusion are the dispersal and structure aspects [10]. "The routes of commuting" are considered in dispersal aspect while the structure aspect “refers to the reciprocal relationships between the locations along the route". Contagious or hierarchical diffusions are described in dispersal component while structural component describes expansion or relocation processes. Relocation diffusion refers to movement of infected individuals from area of origin of disease to new areas. But expansion refers to the process in which the disease remains and intensifies in the original place and expands to the nearby places over time. Expansion diffusion occurs in contagious or hierarchical ways [10]. Hierarchical expansion diffusion involves transmission through an ordered sequence of places, in a socially structured population. Contagious expansion diffusion depends on direct contact between people, and the phenomenon spreads through a uniform medium. It is subject to distance decay which means nearby individuals or places have higher chances of contact. The diffusion of a disease is often analysed in mathematical epidemiology.

2.1.1 Compartmental Epidemic disease models

The SIR model [1] was proposed for studying the disease spread (disease diffusion) and it involves three distinct agents- Susceptible, Infected and Recovered agents. This model divides the population into three states: S, I, R referring to the susceptible, infective and recovered groups of people, respectively. R is referred to as, removed or refractory as well. Given a set of N individuals, infective persons have an average rate of contact of β per unit time, with individuals from all three groups and they die or become resistant (immune) at an average rate of γ per unit time.



The basic reproduction number (disease spread to N persons by a single carrier) is $R_o = \beta N/\gamma$. Here βN is the number persons contacted by the infective(carrier) per unit time and $1/\gamma$ is the infectious period. If $R_o > 1$ the disease spreads otherwise it dies out. The SEIR model is an extension of the SIR model. In this, a latent or exposed state is added. It represents the group of individuals who are exposed but not yet infective (have the virus but cannot infect others).

2.1.2 Agent Based Epidemic disease models

The Compartmental models don't consider the individualistic nature of humans. They assume that all persons in a particular compartment are identical. The compartmental models suffer from behavioural generalizations. The agent based models consider the granularity and capacity of each individual to interact independently. Agent based models are diverse in their behavioural rules. An Agent is autonomous, adaptive and self-directed.

Many interesting research studies exists. Episimdemics algorithm [10] simulates the spread of diseases in a realistic social contact network. The role of social networks epidemic simulations is twofold. First, the structure of the social contact networks greatly influences how disease spreads on these networks. Second, the network structure plays an important role in determining the efficiency of the parallel simulations by affecting the inter-processor communication patterns of a parallel algorithm and its implementation. This work develops a simulator for agent-based modelling and it uses a DES (Discrete Event Simulation) algorithm which changes its state based on the occurrence of event. There are two types of events Arrive and Depart [1], which trigger system change. The between-host transmission is used to specify the effect of interaction between individuals. The effect is given by a probabilistic function. The within-host progression is modelled as PPTS, an

extension of Finite State Machine.

Yanzhi Ren et. al [1] proposes that disease spread can be controlled by identifying the users in the infected person's contact graphs and kernel structure. The contact trace of each person represents the contact details of the users and it's used to determine the set of users who are likely to be infected by the particular user. People within the same kernel structure are considered to have more chances of encountering the disease and hence the persons within the infected person's kernel structure are alerted first.

2.2 Gravity Based Models

Gravity based model [18] is generally used to find the distance between any two locations. With the help of the basic Newtonian law of gravity, we derive the gravity model given by the formula

$$TV_{ij} = G\frac{Y_i Y_j}{D_{ij}} \quad (1)$$

where TV_{ij} is the total volume of trade between two trading partners i and j, Y_i, Y_j is the GDP of masses and D_{ij} is the distance between i and j. This is a model considering the interaction between two population centres based on Newton's Law of Universal Gravitation: two bodies in the universe attract each other in proportion to the product of their masses and inversely as the square distance between them.

2.3 Tie Strength and Influence

Tie strength refers to the links that connect individuals with others through the frequency and types of communications like voice calls, SMS, MMS among them. The strength of a tie is "a combination of the amount of time and the reciprocal services which characterize the tie" [17]. Based on the strength of the tie the entire network is sub divided into various communities. We can use the total time that two people spend talking to each other as a proxy for the strength of the tie – the more time spent communicating during the course of an observation period, the stronger we declare the tie to be [17].

2.3.1 Community Identification Methods

Identifying community structure is an important issue in network society and has attracted the attention of researchers in many fields [9]. Social Network analysis for telecom network is not much explored from a structural point of view. Community detection involves finding nodes that are closely and sparsely knitted. The closely knit nodes fall in a single community with more similar characteristics. There are several interesting techniques of community detection available to segment the total network. One such algorithm is the CNM algorithm [5] that helps in finding the communities using a hierarchical agglomeration algorithm for detecting community structure which is faster than many competing algorithms. Its running time on a network with n vertices and m edges is O (m d log n) where d is the depth of the dendrogram describing the community structure. This community identification algorithms optimizing the modularity of the partition in a greedy way [5], a method that even improved, does not allow to analyze more than a few millions nodes. Recently another new algorithm has been proposed to find communities based on modularity factor where the modularity is found and then the communities are further split using simulated annealing (SA) [5]. SA is a stochastic optimization technique that enables one to find 'low cost' configurations without getting trapped in 'high cost' local minima. This method is more 'transparent' than those relying on heuristic procedures. Furthermore, SA enables to carry out an exhaustive search and to minimize the problem of finding sub-optimal partitions. In this method one does not need to specify a priori of the number of modules; rather, the number of modules is an outcome of the algorithm. Moreover, it is

claimed that this algorithm works extremely fast, which allows to analyze systems of unprecedented sizes [12]. Another new approach based on self-organizing map has been proposed for community detection in both weighted and bipartite networks [13].

In the recent times, there is a lot of interest in studying the community structure in graphs for obtaining finer details. Most of the studies address properties of the graph including its size, density, distance, in and out degree distributions, small-world phenomenon, clustering coefficient, connectivity status, etc. Email graphs and online interactions have been studied in the context of explaining and analyzing friendship and demonstrating the small-world and navigability properties of these graphs [14, 15]. In telecom domain, Abello et. al [5] first experimented with the call graph of landline phones made on 1-day consisting of approximately 53 million nodes and 170 million edges. The Treasure-Hunt model [7] is an attempt to study a broad set of parameters that reveal various structural properties of mobile call graphs. It determines how the structure of these call graphs evolve over time. Uncovering such community structure not only helps in understanding the topological structure of large-scale networks, but also helps revealing the functionality of each component [9]. We also applied fast unfolding algorithm in our data set to generate location specific communities.

2.3.2 Identification of Influential Members

In this study, the prediction of disease spread by using the concept of heat dissipation is understand with an application of Game Theory. It finds its application in churn prediction as well. In [6], the author proposes that each churner is associated with a community formed from his contacts. A method based on Game theory is given to predict the community effect of churners using the concept of centrality obtained from Shapley value. With regard to network analysis, existing models which take into account the neighbouring nodes effects are bound by traditional centrality metrics which only consider the relationship between individual nodes and the rest of the network [21]. It is a limitation in some applications because the dynamics of combinations of nodes are completely ignored. Game theoretic centrality [21] has been proposed as a part of the framework in this study that would address the above limitation. That is, one might want to define centrality of a node n in terms of "the average extent by which the addition of n to an arbitrary set of nodes S reduces the distance between S and other nodes on the network". This is done by calculating the Shapley Value [21] of each individual node in an operator network. The added advantage is that it is computationally efficient. We have modified the algorithm mentioned in the paper with Dijkstra' technique to scale for larger data set and improves its efficiency.

Shapley Value (SV): The Shapley value represents the influential score for a given node in the network [23]. Influential nodes are the one who are not only active in participation but also holds strong influence among their neighbouring nodes. It is represented as:

$$SV_i = \sum_{v_j \in v_i \in N(v_i,d)} \frac{1}{1+\deg(v_j)} \quad (2)$$

where $N(v_i, d)$ represents nodes with d degree of separation from node v_j. There exists several other graph parameters for graph analysis but we have restricted our analysis to specific parameters that is applicable to node level analysis which have closed association to the events happening in telecom domain.

2.3.2 Identification of Susceptible Members

The Galton-Watson model [11] describes the branching processes to explain the concept of population growth. There is a single person in the first generation and as the number of generations increase, the population increases. And pk is the probability of persons contributing k descendants

to the next generation. The number of descendants per individual is given by

$$f = \sum_{k \geqslant 1} k p_k \qquad (3)$$

The Galton-Watson model exhibits phase transition : continuously varying parameter f results in radically different behaviour – namely, survival of population or, in the epidemic context, ongoing propagation of the disease in the infected individual.

2.4 Identification of Tie Strength

To discover the distribution pattern on a given dataset, cluster analysis has been used. Cluster analysis [13] groups objects (observations, events) based on the information found in the data describing the objects or their relationships. The goal is that the objects in a group will be similar (or related) to one other and different from (or unrelated to) the objects in other groups. There are different clustering techniques.

3 Uncovering Relevant Algorithms in SDC Model

In order to determine the influential agent, their mobility pattern and the infectious rate of each agent, SDC model implements the following algorithms.

3.1 Gravity model algorithm

We have adopted the following expression as a model for the number of consumers C_{ij} travelling between two nearby locations i and j :

$$C_{ij} = P \frac{N_i^{\alpha} N_j^{\gamma}}{\exp(\beta d_{ij})} \qquad (4)$$

where d_{ij} is the distance between the two locations in kilometers and N_i and N_j are the mobile phones users of the locations i and j, respectively. The above expression has four free parameters: the exponents α and γ, the inverse characteristic distance β and the proportionality constant P. A multivariate regression analysis is applied to obtain the values of the parameters that better fit our data as well as an estimation of the statistical significance. By applying the logarithmic transformations we get the expression as follows

$$\log(C_{ij}) = \alpha \log N_i + \gamma \log N_j - \beta d_{ij} + \log P \qquad (5)$$

Hence higher the value of C_{ij}, we obtain the maximum movement of the people between the two locations.

3.2 Community Identification Algorithms

We have applied fast unfolding algorithm for the community identification [5]. Fast unfolding algorithm is an alternate way of finding communities using modularity. The community formation algorithm is divided into two phases. In the first phase, each node has its individual community. Then the modularity is found with all its neighbours and the change in modularity value is evaluated. If there is a positive gain, the communities of nodes are merged into one. This procedure is applied to all nodes in the network. This first phase stops when a local maxima of the modularity is attained, i.e. when no individual move can improve the modularity. Part of the algorithm's efficiency results from the fact that the gain in modularity ΔQ obtained by moving an isolated node i into a community (C) can be easily calculated. If is the sum of the weights of the links inside C, is the sum of the weights

of the links incident to nodes in C, k_i is the sum of the weights of the links incident to node i, $k_{i,in}$ is the sum of the weights of the links from i to nodes in C and m is the sum of the weights of all the links in the network then, ΔQ is calculated as,

$$\Delta Q = \left[\frac{\Sigma_{in} + 2k_{i,in}}{2m} - \left(\frac{\Sigma_{tot} + k_i}{2m}\right)^2\right] - \left[\frac{\Sigma_{in}}{2m} - \left(\frac{\Sigma_{tot}}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)^2\right] \quad (6)$$

In second phase, each community is taken as node and the process is repeated, until all the weights are summed. Links between nodes of the same community lead to self-loops for this community in the new network. Once this second phase is completed, it is then possible to re-apply the first phase of the algorithm to the resulting weighted network and to iterate. These two phases are iteratively performed unless stabilized value is reached.

3.3 Applying Shapley value in identifying Influential Spreaders

The game theoretic network centrality assists in finding out the importance of each node in terms of its utility when combined with the other nodes [21]. In telecom network it wouldn't suffice to find the importance of a node as a mere standalone entity as in other centrality measures. Other works related to the finding of influential nodes in the network didn't address the influence of a node as a combination of several nodes in a network. Given a telecom network, the game theoretic network centrality indicates the coalition value of every combination of nodes in the network. We have used the *Dijkstra's algorithm* to efficiently track the shortest distance between a given node and its neighbor's in calculating Shapley value [23] for each node. The adjacent diagram shows the existence of influential agents in a single community

Program: Computing SV by running a game.
Input: Graph Network ingested from CDR.
Output: SVs of all nodes in network

```
foreach node v in Network do
   DistanceVector D = Dijkstra(v,Network);
   kNeighbours(v) = null;
   kDegrees(v) = 0;
   foreach node u <= v in Network do

      if D(u) <= k then
         kNeighbours(v).push(u);
         kDegrees(v)++;
      end
end
foreach node v in Network do
   ShapleyValue[v] = 1
      kDegrees(v)++;
      foreach node u in kNeighbours(v) do
         ShapleyValue[v]+= 1
         kDegrees(u)++;
      end
   end
   return ShapleyValue;
end
```

3.4 Clustering using K-Means algorithm

The K-Means [22] clustering algorithm is also called as the filtering algorithm. We have used K-Means clustering to understand the tie strength among the members in the community and to propose a strategy to simulate spread and control environments. In k-means clustering, we are given

a set of n data points in d dimensional space R^d and an integer k and the problem is to determine a set of k points in R^d, called centers, so as to minimize the mean squared distance from each data point to its nearest center.

3.5 Calculating Infectious Rates

Once the influential agent is identified based on the above algorithm, we determine the community in which the influential agent is present. We identify the infectious rate of the influential agent based on the susceptible people for the influential agent in each level. The infectious rate is determined based on the formula

$$P_i = 1 - \exp(\tau \sum_{r \epsilon R} N_r \ln(1 - rs_i \rho)) \quad (7)$$

where P_i is the probability that an infection is triggered in a susceptible agent i, τ is the duration of exposure; R is the number of levels that the agent covers the network; N_r is the number of susceptible agent in each level r ; s_i is susceptibility of individual i and ρ is the basic transmissibility of the disease. The value of ρ is identified as follows:

$$\rho = R_o/\eta \quad (8)$$

where R_o is the average number of calls done by the agent and η is the infectious period.

4. Observing specific regions of interest from the given dataset

The goal of the challenge is to address society development in a novel way by contributing to the socio-economic factors that for the well being of the population. We are provided with the datasets based on anonymized Call Detail Records of voice calls and SMS exchanges between four million of Orange operator customers in Ivory Coast between December 1, 2011 and April 28, 2012. The datasets contained the following:

(a) Antenna-to-antenna traffic on an hourly basis

(b) Individual trajectories for 50,000 customers for two week time windows with antenna location information

(c) Individual trajectories for 50,000 customers over the entire observation period with sub-prefecture location information

Fig. 1 Map of Areas with frequent epidemics of Meningococcal Meningitis

The spread of Meningitis disease is significant in Africa. It is claimed that climatic change contribute significantly the spread of the disease in Benin, Burkina Faso, Cameroon, The central African Republic, Chad, Cote d Ivoire which was shown in Fig 1, The democratic republic of Congo, Ethiopia, Ghana, Mali, Niger, Nigeria and Togo. This is an area of Africa where the diseases endemic and there are always a few cases. Ivory Coast located on the south coast of West Africa has about 18 million population and is subdivided into 19 regions with 255 administrative regions. In this paper we consider the case study of the communicable disease Meningitis spread, which originated in the region of Savanes and reach to different regions in Ivory Coast. Table 1 listed in WHO 2009 report shows the case of fatality rate during the outbreak of the disease Meningitis in the following regions. We have generated a Voronoi tessellation for entire Ivory Coast (Fig 2) based on the antenna id's. This helped us to identify the antennas that are closer and thus helps in grouping the locations that are near by.

Table 1: Fatality rate of Meningitis in African countries

Country	Cases of suspect meningitis	Deaths	Case fatality rate (%)	Predominant pathogen	Number of district in epidemic
Benin	381	38	10	NmW135	3
Côte d'Ivoire	281	39	13.9	NmW135	1
Ghana	369	37	10	NmW135	3

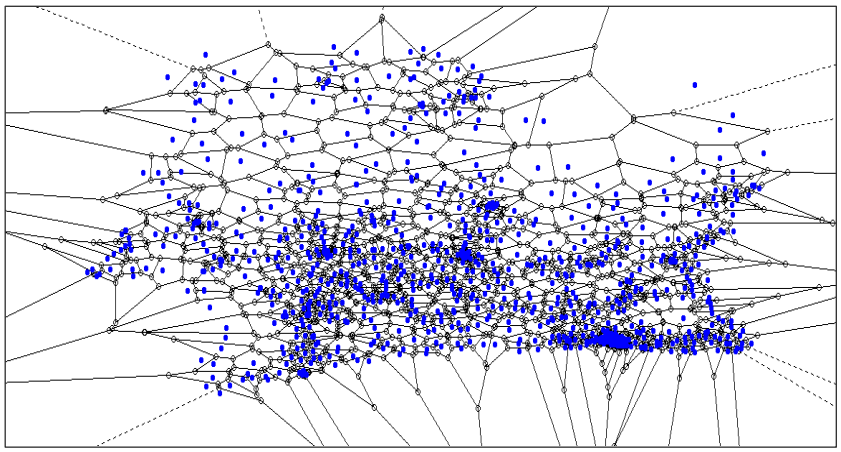


Fig 2 Voronoi Map of Ivory Coast

4.1 Users distribution in Savanes Region

We have the querying tool TORA [19], to generate mobile phone users dataset of Savanes region by using the antenna id coverage in the region. The dataset consisted of the people whose home location is Savanes and also the people who have travelled to Savanes as there is a high probability that the individual would contract the disease from the infected persons in Savanes. Savanes region consist of 46 antenna id in total. Totally it is found that 15,042 users have visited Savanes. Out of 15,042 users 2,029 of them have called the other users within the same region. From the calling and the called party details of 2,029 distinct users 323 communities were identified inside Savanes region using the community detection algorithm. The high dense communities (strong ties) were

identified by the modularity score determined by the Fast unfolding algorithm [5]. These calculated modularity score helps in identifying the size of each community depending on the modularity class as shown in Fig 3. Fig 4 shows the distribution of Antenna IDs of Savanes region formed using Voronoi Tessellation.

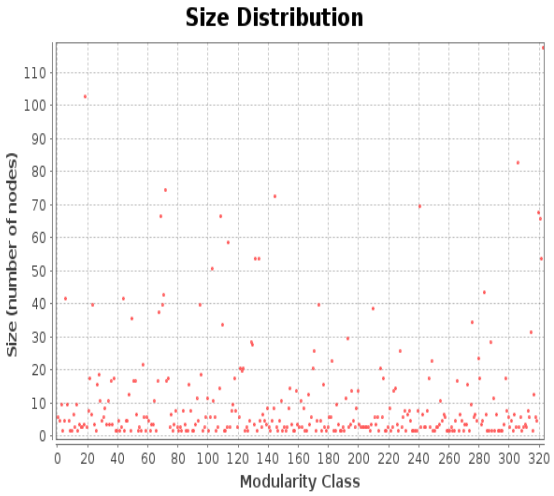


Fig. 3 Size distribution of Community

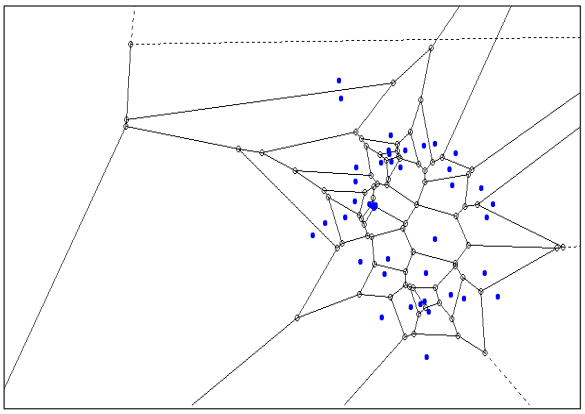


Fig. 4 Savanes Antenna distribution

4.2 Tracking user movements

Now we simulate the disease spread in Savanes region to understand the use of exploring mobile call details records. After discovering the disease spread pattern in Savanes, our next goal is to determine the spread of disease across the country so that the measures to mitigate the disease spread could be made easy. The movement of every user is traced based on the individual mobility pattern using the gravity model [18]. This mobility pattern of mobile phone users is used to identify the various other localities to which these users have travelled and thus estimated nearly to 23 regions. Fig 5 shows the details of mobile user's movement from Savanes to other regions inside Ivory Coast. It helps us to simulate the spread of disease pattern to other regions.

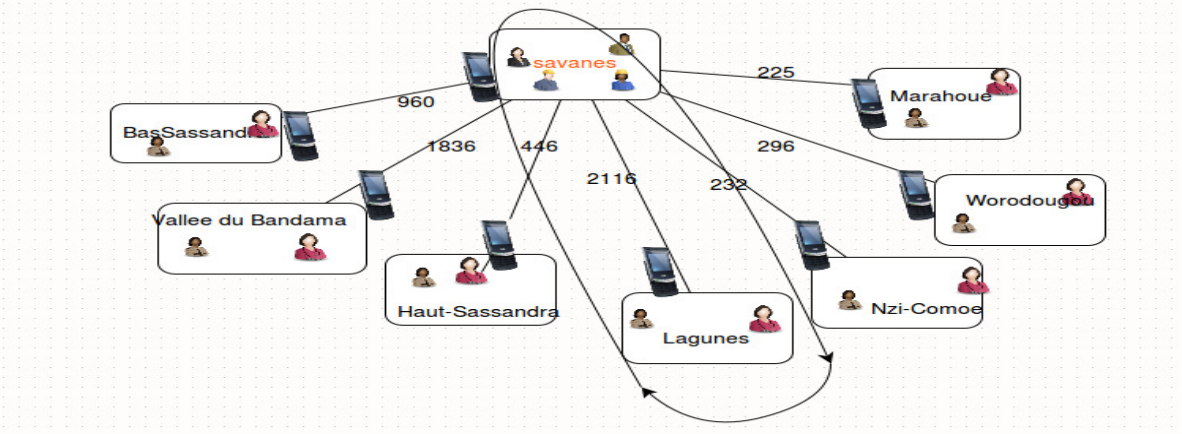


Fig 5. Tracking the user movements from CDR's

According to the data provided by WHO, it was found that the disease Meningitis spreads maximum from Savanes to Lagunes. While studying the pattern of movements in Table 2, we also

found that greater number of people moved from Savanes to Lagunes and hence we identified these two regions for applying our SDC model to figure out the controlling strategy in discussion section.

4.3 Selection of user distribution in Lagunes region

Out of 2589 users in Lagunes region, 618 of them have called the other users in Lagunes. Hence we have selected 2116 users directly. It clearly visible from WHO report that our findings are closely matches and depicts the spread of disease from Savanes to Lagunes. Thus applying the community extraction methods on both regions helps in understanding of the disease spread in an efficient manner. We use a Voronoi tessellation to define its coverage area. Fig 5 shows set of antenna ids with the original coverage area of each cell in the regions of Savanes and Lagunes.

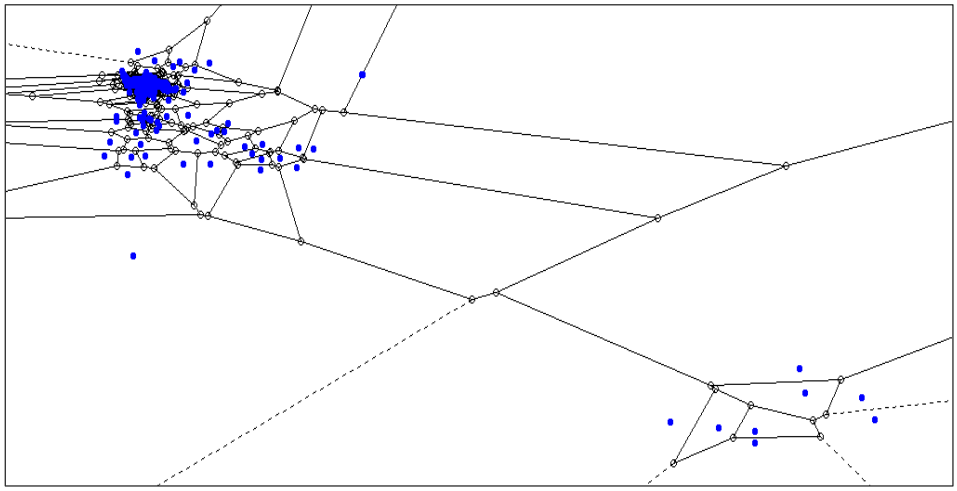


Fig 6. Voronoi map of Savanes to Lagunes regions

5 Conceptual View of our proposed Models

The daily activities of the people and their role in spreading of disease in the real world are simulated by creating a spatial and temporal model of disease transmission. The spreading pattern can be determined by extracting the dataset using the movement of people. The call detail records of every people act as a weapon for examine the spread of disease. It helps serving our purpose and identity the control measure to mitigate the rate of spreading.

5.1 Mobile Agent Based Model (MABM)

MABM is a new approach to the modelling system containing the Mobile Agents. A mobile agent is the agent who is autonomous, self directed and acting with different behavioural characteristics in his environment. An agent can function independently in its environment. We refer the agent's behaviour as the representation of a process that links the agents sensing its environment to its actions. Hence the behaviour of the mobile agent is thus the interactions with the surrounding communities forming the network of connections using the mobility pattern traced with the help of call detail records (CDR).

The CDR is comprehensive enough to be mined not only for understanding the human movement but also for supporting various applications from mobile tracking, monitoring to medical emergencies. We have understood that Meningitis disease spread has proved vitally fatal in the

recent years in the regions of Ivory Coast. Hence we investigate the feasibility and effectiveness of the collected CDR details of mobile phone to derive the temporal spatial distribution of the social community to make decision on the Spread, Discover and Control (SDC) model of our study.

5.2 Spread Discovery Control (SDC) Model

We have framed our model with following distinct entities as shown in Fig 7. There are three different categories of epidemiological modelling were considered in our model

1. **Spreading** of disease by the set of agents that are modelled using the information contained in the call detail records
2. **Discovering** the regions of the spread that are modelled by tracking the geographic network of the agents.
3. **Controlling** the spread of disease.

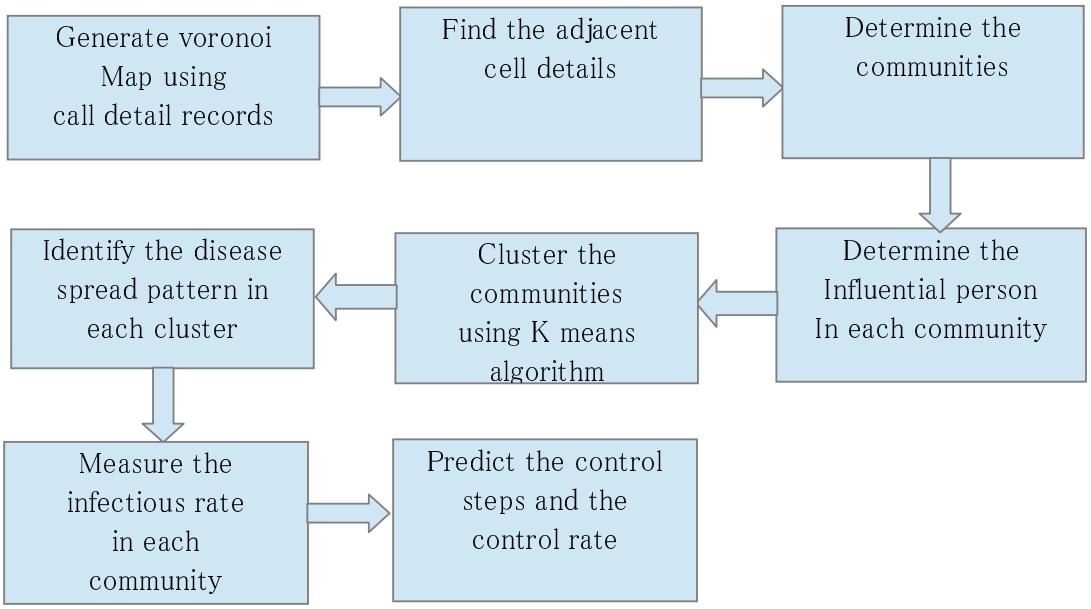


Fig. 7. General Structure of SDC Model

5.2.1 Spread Model

The Spread model illustrates the way to simulate the disease spreading with mobile agents. They are characterised by their individual mobility patterns and social networks. It focuses on fine-grained modelling of the spread of disease (e.g. Meningococcal Meningitis) in a large real-world social network. The spread has been initiated from Savanes region in Ivory Coast. We carry out the simulations and measures with the real time data of the call detail records of 4 million people of Ivory Coast. Based on their call patterns we have identified the users who are located in Savanes regions. The influential agent of the spread model is capable of spreading the disease faster in short span of time. In this study, we have determine the influential agent based on two criteria: i) the **total time** that two people spend talking to each other as one of the measure and the **number of times** the person has contacted, as greater the number of calls greater the probability that the agent might meet the people he has contacted; ii) the **number of people** the agent has called, as greater the number of people contacted wider the disease spread.

5.2.2 Discover Model

Discovery model is characterised by two main components of a) Mobility identification pattern b) Social network.

a) Mobility Identification

Every time the call is made, the details of calls like calling party, called party, BTS id, sub-prefecture id, connection date time etc are stored. Each time a CDR is generated we get a sample, where the user is at the time of call. Using this statistics we identify the users who move form Savanes to Lagunes. With this we get an idea of how the disease is spread from Savanes to Lagunes and also the persons who are responsible for the disease spread. The general structure of our model is as shown in Fig 7.

b) Social network formation

Of the records of 4 million users, the users in the region of Savanes were considered from where the disease Meningitis_originated. The disease spread across the locality occurs when the agent moves to the new locality physically which is determined from the CDR details. When the agent uses the antenna id on the new locality for his calling it implies that there is a physical movement of agent. Hence this helps us to track the mobility pattern of the agent from his base location. It is identified that nearly 2116 people have travelled from Savanes to Lagunes. Out of which 1763 have contacted the people in Lagunes region. Hence we identify that there are more chance that the disease spread across Savanes and Lagunes as shown in Fig 5 in Section 4.2.

5.2.3 Control Model

The persons causing the diseases in a home would be the person returning home after work or studies. They could be students or office-goers. The ultimate goal of any research work is to create mass awareness. Hence we consider the CDR's to identify the influential person in the given network. This person is the agent-influencer who has access to relatively large number of mobile users and hence can help in spreading awareness faster to control the disease. The geographical distribution of any place along with the cell towers located in the regions bordering the place can help tourist and other persons entering the disease prone areas, to gain access to preventive measures, symptoms, treatments and medical facilities available. This study thus uses the mobile records, data mining algorithms and statistics to process the mobile phone records and details of cell towers in addressing the root causes of diseases and sets up a priority to target the customers based on the importance with proper campaigns.

6 Experiments and Results

We analysed the data set given to us in all the possible ways by conducting various experiments and thereby deriving different attributes. These attributes were derived from the data set using the querying tool, Tora [19]. To get a clear picture, these data were visualised in a network graph using the visualisation tool Gephi [20]. The input file given to Gephi is extracted from Tora in the .graphml format. Various attributes of the graph including the details of the nodes, edges, connected components, modularity among the various others are obtained from Gephi. Numerous algorithms like fast unfolding algorithm, centrality algorithm are also implemented in Gephi for community detection and further processing of the network.

6.1 Detecting Communities in different regions

Provided with the details of 4 million dataset of Ivory Coast, we have identified the disease prone region as Savanes which matches with the information provided by WHO. After processing the call detail records, we have identified that most of the people travelled from Savanes to Lagunes. Hence we approach this problem in three different directions as follows:

1. Analysis of the spread pattern within the localities of Savanes
2. Analysis of the spread pattern for the people travelling from Savanes to Lagunes
3. Analysis of the spread pattern within the localities of Lagunes

Fig 8 depicts three different pie charts which show the distribution of the given dataset in three stages of our study with respect to the emergence of communities. The communities are classified based on the presence of number of nodes.

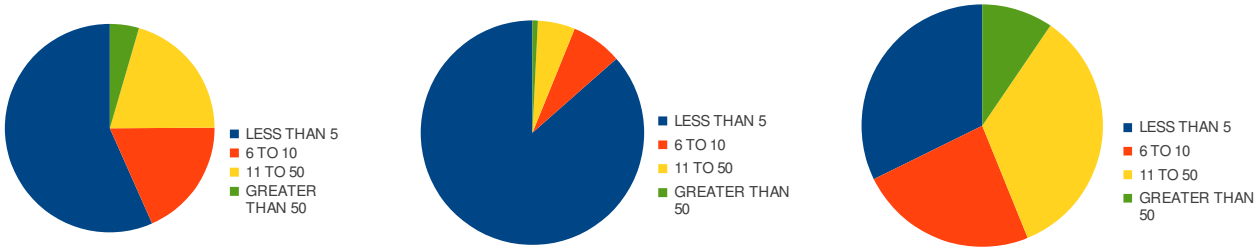


Fig 8. Distribution of communities in different regions like Savanes to Savanes, Savanes to Lagunes and Lagunes to Lagunes

As you can see, in the second pie chart in Fig 8, maximum number of communities contains lesser than 5 nodes. This is because, the second pie chart depicts the network of users who move from Savanes to Lagunes and have contact with the people in Lagunes. Hence as this set of users is moving from Savanes to Lagunes, the size of their network is considerably smaller when compared to the rest. Comparatively, the first and the third pie chart consist of greater number of nodes as they have a well established network in Savanes and Lagunes respectively. The total number of the communities and the modularity score identified for the respective stage is shown in Table 3.

Table 3 Communities modularity Score of different regions

Movement of people from	No of communities	Modularity score
Savanes to Savanes	323	0.986
Savanes to Lagunes	503	0.987
Lagunes to Lagunes	357	0.988

Modularity is a measure of the graph depicting the strength of connections within the community. Greater the modularity score, the denser the community is. With a higher modularity score, the nodes within that community have a stronger bonding rather than the nodes in different communities. As in the table 3, nodes within Lagunes have the highest modularity (0.988) as it has the most established network so as to enable easier disease spread. Hence even a single node can very easily spread the disease in Lagunes in a shorter span of time.

6.2 Impact of influential users

In each of the above said stages, we identified the influential users based on their respective Shapley values. Table 4 illustrates the obtained Shapley value scores for one single community.

Table 4. Sample Shapley values for the community given in fig 9

Node	Shapley Value
29630	0.5256410256410257
29631	0.5256410256410257
29610	0.5256410256410257
29611	0.5256410256410257
29592	3.442307692307693
29594	0.4423076923076923
29593	16.942307692307693
29596	0.4423076923076923
29752	0.4423076923076923
29595	0.5256410256410257
29655	0.5256410256410257
29598	0.5256410256410257
29654	0.5256410256410257
29732	0.4423076923076923

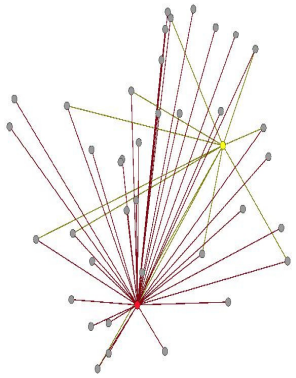


Fig 9. Identify influential user using Shapley Value

The community is shown in the above figure. The red colour node (29593) depicts the most influential node in this community. The yellow colour node (29592) depicts the second most influential node. Node 29593 has connections with almost all the other nodes in that community. The respective Shapley value of 29593 is given in the table 4. It has the highest Shapley value in that community. Or conversely, greater the Shapley value, the more influential the node is. This way, the Shapley value algorithm is used to identify the influential user(s) in the given community. If in case the influential node is also the infected node, then there is high probability that the entire community will be infected in a shorter span of time. Hence more the number of influential users, greater the disease spread. Table 5 depicts the total number of influential users in each stage.

Table 5. Identification of influential users in uncovered communities

	Total number of users	Total number of influential user	Percentage
Savanes to Savanes	2762	179	6.4808
Savanes to Lagunes	2147	81	3.7727
Lagunes to Lagunes	7526	329	4.3715

The percentage of influential users is higher within the Savanes and Lagunes region when compared to the set of users who are moving(Savanes to Lagunes).

6.3 Finding Infectious Rates

After generating the influential users on the emerged communities with different hop of nodes, we derive the infectious rate for each node. The value of expected fraction of infected members after r rounds is computed, and the number of persons infected in each round increases exponentially. The infectious rates of the influencers is calculated for each hop and compared with that of ordinary users. This is done cluster-wise. Three clusters are generated i.e, high, medium, low. This clustering is done using the K-Means clustering algorithm [22]. This has been done so as to trace the disease spread in various clusters. In the low cluster, the nodes are sparse, whereas in the high cluster the nodes are dense. The following tables (6-8) show the infectious rate calculation for the influential user in comparison with ordinary users.

Table 6. Average Infectious rates calculated for low cluster community nodes

	Single influencer	Single normal user	Average of normal users
Savanes to Savanes	0.9965	0.0726	0.0717
Savanes to Lagunes	0.9993	0.0720	0.0557
Lagunes to Lagunes	0.9992	0.0378	0.0296

Table 7. Average Infectious rates calculated for moderate cluster community nodes

	Single influencer	Single normal user	Average of normal users
Savanes to Savanes	0.9990	0.0777	0.0771
Savanes to Lagunes	0.9996	0.1222	0.1222
Lagunes to Lagunes	0.9999	0.0755	0.0603

Table 8. Average Infectious rates calculated for high cluster community nodes

	Single influencer	Single normal user	Average of normal users
Savanes to Savanes	0.9999	0.1585	0.1543
Savanes to Lagunes	1.0000	0.1976	0.1976
Lagunes to Lagunes	1.0000	0.1429	0.1364

Here, based on the number of nodes, the communities are all divided into clusters. As the size of the cluster increases, the proximity between the nodes in that particular community increases; thereby increasing the infectious rate. Within these three clusters we identify that the influential user has the maximum infectious rate. It is easily seen from the above tables that the influential user has contacts with almost all the other nodes in the given communities. From this we infer that the influential user has the maximum possibility of spreading the disease through out the community in very short span of time compare to normal user.

7 Discussion

The experiments we have conducted and discussed in section 6 clearly highlight the benefit of our models usage for understanding the disease spread in different regions. MABM is applied to identifies the behaviour of the mobile agent, based on its interactions with the surrounding communities, that forms the network of connections using the mobility pattern obtained from CDR. Our proposed SDC model is implemented to determine the areas in which the disease spreads, by dividing the larger network into smaller communities, which on analysing provides a wider

perspective of determining the disease spread pattern and help the concerned authorities to implement the control strategies to avoid the happening of major catastrophe. To prove the significance of our approach of identifying influential users in comparison with normal users, we have performed paired t-statistics [24] on the calculated infectious rates as part of SDC model implementation. T-statistics is used to determine the level of mean score difference between influential user spread rate with one single normal user and five normal users in different regions. Table 9 show the results of t-values with significant level to strengthen our proposal of start discussing the control measures from Influential users rather than the normal users.

Table 9 . T-statistics for mean score difference between influential user with normal user

No: of communities	Regions	Influential Vs Single normal user	Influential Vs Average of five normal users
71	Savanes to Savanes	132.93**	129.22**
35	Savanes to Lagunes	92.76**	90.53**
159	Lagunes to Lagunes	163.72**	163.12**

** - Significant at p<0.1 level

Now we can discuss our idea of applying control strategy in our SDC model. Fig 10 shows the community structure of the closely connected network of the users. The green colour highlights the infected influential user to which many other nodes are connected. The disease spread can be controlled by generating proper campaigns to connected neighbouring nodes with infected influential user. Measures can be applied before the spread of a communicable disease to protect a community from getting infected and to reduce the number of cases locally in the future. Once a communicable disease occurs and is identified in an individual, steps can be applied to reduce the severity of the disease spread by that person to other members of the community.

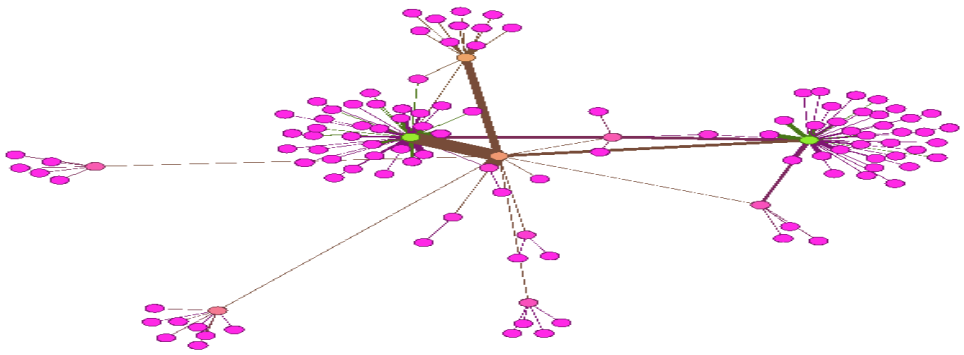


Fig. 10 A single community with Influential users

An application can be developed so as to curb the disease spread. Using this application the users who are connected to the infected persons can be notified of the disease. The infected user informs the operator about the status of his health condition. The operator checks whether the user is an influential user or whether he is connected to an influential user based on his presence in a community. This process of notifying the sub network of the infected user is done without affecting the privacy of the former. This way the disease spread can also be controlled in an effective way. Here the tourists also been properly guided if we understand their contacts and location.

8 Conclusion

This project thereby addresses the organized research in the area of multifarious systems models design, with a specific aim to improve the accuracy of predicting the disease prone areas and alert the customers and tourists with appropriate recommendations by incorporating real-world human mobility patterns from call detail records. The models proposed in this paper also help us to capture the complexity of people movements and the way of spread of disease to others. This way we can not only trace the disease spread to different geographical regions, but also provide a control measure for it. Again we consider the mobile phone records to identify the influential person in the given network. This person is the agent-influencer who has access to relatively large number of mobile users and hence can help in spreading awareness faster to control the disease. Finally, the geographical distribution of any place along with the cell towers located in the regions bordering the place can help tourist and other persons entering the disease prone areas, to gain access to preventive measures, symptoms, treatments and medical facilities available. The proposed models could impact all phenomena driven by human mobility, from epidemic prevention to emergency response, urban planning and Mega city development projects.

9 References

[1] Enrique Frıas-Martınez, Graham Williamson, Vanessa Frıas-Martinez, "An Agent-Based Model of Epidemic Spread using Human Mobility and Social Network Information", Telefonica Research, Madrid – Spain,School of Computer Science , University College, Dublin – Ireland

[2] Jiasheng Wang, Jianhong Xiong, Kun Yang, Shuangyun Peng, Quanli Xu, "Use of GIS and Agent-based Modelling to Simulate the Spread of Influenza", In Proceedings of 18th International Conference on Geoinformatics, 2010

[3] Charles M. Macal,Michael J. North, "Toward teaching Agent Based Simulation", Charles M.Macal,Michael J. North ,Argonne National Laboratory ,Center for Complex Adaptive Agent Systems Simulation (CAS2) ,9700 S. Cass Ave. ,Argonne, IL 60439, USA

[4] Marta C.Gonzalez , Ceasar A. Hildalgo and Albert-Laszlo Barabasi, "Understanding individual human mobility patterns". Nature, 453, 779-782 (5 June 2008) | doi:10.1038/nature06958; Received 19 December 2007; Accepted 27 March 2008.

[5] Saravanan M and Prasad, G "Labelling Communities using structural properties", In Proceedings of DBKDA, 2010.

[6] Saravanan, M and Yeshwanth V. "Evolutionary Churn Prediction in Mobile networks using Hybrid Learning", In proceeding of FLAIR conference, 2011. In Proceedings of the Twenty-Fourth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2011.

[7] http://www.who.int/csr/don/2009_03_25/en/index.html

[8] http://en.wikipedia.org/wiki/Epidemic

[9] http://www.openabm.org/book/export/html/2099

[10] Christopher L. Barrett,Keith R. Bisset, Stephen G.Eubank,Xizhou Feng, Madhav V.Marathe, "EpiSimdemics: an Efficient Algorithm for Simulating the Spread of Infectious Disease over Large Realistic Social Networks", In Proceedings of the 2008 ACM/IEEE conference on supercomputing, 2008.

[11] Patrick T.Eugster Rachid Guerraoui, Anne-Marie Kermarrec Laurent Massoulié, "Epidemic Information Dissemination in Distributed Systems", Computer Journal, Volume 37 Issue 5, May 2004, Pages 60-67

[12] Yanzhi Ren, Jie Yang, Mooi Choo Chuah , Yingying Chen, "Mobile Phone Enabled Social Community Extraction for controlling of Disease Propagation in healthcare" In proceedings of International conference on Mobile Adhoc and sensor systems, (MASS) 2011

[13]Narendra Sharma 1, Aman Bajpai 2, Mr. Ratnesh Litoriya , "Comparison the various clustering algorithms of weka tools", International Journal of Emerging Technology and Advanced Engineering, Volume 2, Issue 5, May 2012

[14]http://en.wikipedia.org/wiki/Cluster_analysis

[15]Bakshy E., Rosenn I., Marlow C., Adamic, " The role of social networks in information diffusion", WWW 2012, April 16-20, 2012, Lyon, France.

[16] Santi Phithakkitnukoon,Tuck W. Leong, Zbigniew Smoreda,and Patrick Olivier, "Weather Effects on Mobile Social Interactions:, A Case Study of Mobile Phone Users in Lisbon, Portugal. PLoS ONE 7(10): e45745. doi:10.1371/journal.pone.0045745

[17] David Easley, Jon Kleinberg , "Networks, Crowds, and Markets: Reasoning About a Highly Connected World", Cornell University, New York .

[18]Duygu Balcan, Vittoria Colizza,Bruno Goncalves,Hao Hu,Jose J.RAmasco and Alessandro Vespignani, "Multiscale mobility networks and the large scale spreading of infectious diseases", PNAS 106, 21484 (2009)

[19]http://torasql.com/

[20] Mathieu Bastian and Sebastien Heymann Mathieu Jacomy , “Gephi : An Open Source Software for Exploring and Manipulating Networks ”,Gephi, WebAtlas,Paris, France
[21] Aadithya, K. V. and Ravindran, B. (2010) "Game Theoretic Network Centrality: Exact Formulas and Efficient Algorithms". In the Proceedings of the Ninth International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2010), pp.1459-1460.
[22] Ian Witten and Eibe Frank, Data Mining: Practical MachineLearning Tools and Techniques, 2nd Edition, Morgan Kaufmann, ISBN 0120884070, 2005.
[23] Saravanan, M., and Vijay Raajaa G.S. “A Graph-based Churn Prediction Model for Mobile Telecom Networks”, In Proceedings of the 2012 International Conference on Advanced Data Mining and Applications (ADMA 2012), Nanjing, China, 15-18 December 2012
[24] Siegal S, Castellan NJ (1988) Nonparametric statistics for the behavioral sciences. McGraw Hill, Berkeley.

Large-scale Measurements of Network Topology and Disease Spread: A Pilot Evaluation Using Mobile Phone Data in Côte d'Ivoire

Rumi Chunara[1,2,§] and Elaine O. Nsoesie[1,2,3]

[1] Department of Pediatrics, Harvard Medical School, Boston, Massachusetts, USA

[2] Children's Hospital Informatics Program, Division of Emergency Medicine, Boston Children's Hospital, Massachusetts, USA

[3] Network Dynamics and Simulation Science Laboratory, Virginia Bioinformatics Institute, Virginia Tech, Virginia, USA

[§] Corresponding author (email: rumi@alum.mit.edu, telephone: 857.218.3607, fax: 617.730.0267, address: 1 Autumn St. Fourth Floor, Suite 433, Boston, MA 02215)

Abstract

Mobile phone calling records can provide information on people's behaviors and relationships with detailed resolution. This data offers a novel opportunity to augment spatio-temporal models of infectious disease spread. Mobile phone call records have been explored as a proxy for evaluating human mobility patterns in coordination with the spread of a pathogenic infectious disease, malaria. Alongside mobility, network topology also has been shown to play a role in the spread of infectious diseases. Here we investigate how network topologies measured through mobile phone call records relate to the evolution of a meningitis outbreak at the national scale. We simulate a meningitis outbreak on empirically derived network topologies via mobile phone records in Côte d'Ivoire from Dec. 2011 – April 2012. Also, we consider how the results compare to meningitis case data for the same time period.

Introduction

Political instability and environmental factors, amongst other drivers, have resulted in a high burden of infectious diseases in the sub-Saharan African country Côte d'Ivoire [1,2]. Disease burden in the country includes several preventable and treatable diseases: malaria, cholera and dysentery, tuberculosis, HIV/AIDS, measles, meningitis, yellow fever, polio and diarrheal illness. In the study of infectious diseases, mobile phone calling records have been examined as a proxy for movement. Spatio-temporal call record patterns have been used to identify sources and sinks of imported in the spread of malaria [3,4]. At the same time, studies using mobile phone data in other domains such as communication dynamics have extensively described further properties of mobile phone calling patterns including network topology, clustering coefficients and inter-call duration distributions [5]. These works have all guided this study, which uses empirical call record and disease case data to examine if network topology measured through mobile phone calling records can give insight into spread of a meningitis outbreak.

Contact networks naturally play a role in the spread of infectious diseases, and overall network topology has been shown to affect the rate and pattern of disease spread [6]. Numerous computational models have examined the spread of disease on different network structures [6], and it has been found that infectious diseases can spread more easily in scale-free and small-world networks than in regular lattices and random networks [7]. To validate and augment these theories, empirical network structure and corresponding disease data has been mapped using personal wireless sensors or surveys [8,9]. However, these methods are laborious and costly, and thus data have been difficult to obtain in scale. Mobile phone calling records offer opportunity for detailed information about real-life network topologies at larger geographic and temporal expanses while not requiring any added equipment or data gathering methods.

In parallel to the network information from mobile phone calling records, we also investigated disease event reports in Côte d'Ivoire during the same time period. HealthMap is a source of real-time information about infectious disease events around the

world [10]. We used HealthMap to understand what infectious disease outbreaks and events occurred in Côte d'Ivoire during the time period of the available data (Figure 1, Table 1). During this period, HealthMap shows ongoing or new disease events including HIV/AIDS, meningitis, malaria, cholera and dysentery, diarrheal illness and tuberculosis (Figure 1, Table 1). Generally these diseases have been found in Côte d'Ivoire year-round without significant seasonal effects measured [2]. Along with neighboring countries, though, Côte d'Ivoire is part of the "African meningitis belt", which includes areas from Senegal to Ethiopia and experiences seasonal large epidemics of meningitis annually between December and June [11]. The World Health Organization (WHO) reports weekly meningitis case information from the meningitis belt [11]. Meningococcal meningitis is a bacterial form of meningitis, caused by *Neisseria meningitidis*, and is a serious infection of the thin lining that surrounds the brain and spinal cord. It can cause severe brain damage and is fatal in 50% of cases if untreated. The bacteria are transmitted from person-to-person through droplets of respiratory or throat secretions from carriers. Increases in meningitis incidence levels have been attributed to environmental factors and transmission of *N. meningitidis* facilitated by overcrowding and large population displacements at the regional level [12]. Vaccines have been developed for some serogroups of *N. meningitidis*. However, groups of *N. meningitides* for which there is currently no vaccine have also been detected in parts of West Africa [13]. Simultaneously, multiple factors limit the widespread distribution and efficacy of vaccination [14]. Further, the drivers of the seasonal fluctuations are poorly understood, which limits the predictability of outbreaks and the dynamic response to immunization. Thus better surveillance of meningitis is required to understand the mechanisms of disease spread and for implementation of focused public health intervention and control measures.

Using the dataset of mobile phone calling records and patterns for Cote d'Ivoire made available through the D4D Orange challenge [15], we evaluated the use of mobile phone calling records as they relate to population networks and evolution of a meningitis outbreak. As far as we can tell, study of meningitis outbreaks in the meningitis-belt incorporating network information has been limited. In this paper, we present a contact

network model for meningitis in Côte d'Ivoire. We examine how a meningitis outbreak would evolve over networks described through mobile phone calling data, and compare this to information about a contemporaneous outbreak via weekly meningitis case data from the WHO. Finally we also compare how the calling data serves as a proxy for movement at the same geographic scale (national).

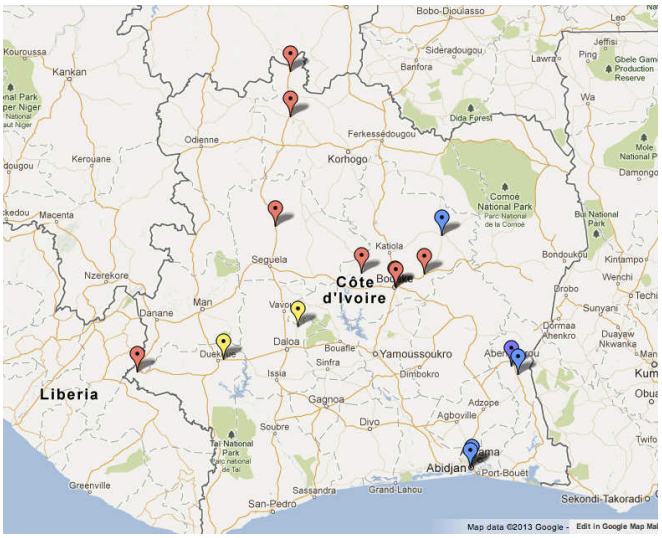


Figure 1. Location of HealthMap Cote d'Ivoire disease alerts. Disease event or related alerts from HealthMap for December 5, 2011 to April 22, 2012. Red markers indicate meningitis related reports, blue markers show cholera or water-related events, purple is for an area endemic with HIV/AIDS and yellow are other events (return of internally displaced peoples and Buruli Ulcer).

Location	Latitude, Longitude	Disease or disease-related event	Relevant dates from HealthMap articles
Dabakala	8.367, 4.433	No water	Jan. 7 – Jan. 19 2012
Tengrela	10.481, -6.409	Meningitis	Epidemic since mid-January, vaccination campaign Feb 16 – 19 2012
Kouto	9.894, -6.410	Meningitis	Epidemic since mid-January, vaccination campaign Feb 16 – 19 2012
Adjame	5.365, -4.023	Cholera	Month of January 2012
Bouake	7.683, -5.0167	Meningitis	January 30 2012 (report date)
Kani	8.477, -6.605	Meningitis	January 30 2012 (report)
Koko	7.690, -5.033	Meningitis vaccination campaign	March 10 2012
Bodokro	7.858, -5.476	Meningitis	January 30 2012 (report)
Locodjro naval base	5.327, -4.040	Diarrhea	January 25 2012
Toulépleu	6.577, -8.418	Meningitis	February 25, March 22 2012 (reports)
Brobo	7.717, -4.700	Meningitis	March 10 2012 (report)
Bediala	7.167, -6.300	Buruli Ulcer	April 10 2012 (report)
Adiake	6.467, -3.500	Cholera	Endemic
Niambly	6.744, -7.289	Return of IDPs	April 26 2012
Moyen-Comoe	(region)	HIV/AIDS	Endemic

Table 1 HealthMap Côte d'Ivoire disease alerts. Disease event or related alerts from HealthMap for the time period December 5, 2011 to April 22, 2012.

Methods and Analysis

Network Descriptions

A contact network can be used to describe disease transmission between individuals in a population. A node or vertex in the network typically represents an individual, and an edge is used to represent contacts between individuals. Weights on edges can represent the probability of contact, or duration of contact. Most nodes in a network are connected to at least one other node. The number of edges originating from a node is defined as the degree of the node. High degree nodes are typically more likely to become infected and to spread infection. A node's degree is therefore an indicator of the range of possible contacts that can lead to disease transmission to and from the node [16]. The degree distribution is therefore important in disease propagation.

For evaluation of the network behavior in this study, since longitudinal communication graphs (or spatial references for the graphs) were not made available as part of the data, we used an antenna-to-antenna dataset. In this data, the number of calls as well as the duration of calls between any pair of antennas was provided on an hourly basis. The antennas are uniquely identified by an antenna identification number and a geographic location for the entire observation period [15]. We assumed that connection to a certain antenna is by callers located nearest to that antenna, although specifics of this were not made available. In order to simulate spatial spread of disease across this region, we assumed the following: (i) if two antennas share a high number of calls, this could be an indication that individuals residing close to those antennas are likely to come in contact and (ii) the number of calls or call durations can be used as indicator of contact strength between individuals in these regions. We chose to use the duration in weighing the graphs since the number and duration of calls are significantly correlated ($\rho \geq 0.94$). We also assumed that there are no births in the population. The data was provided in 10 sets of 14 days each, thus representing aggregate network structure for each of these 10 time periods. We performed analysis on the aggregated number of calls between antennas for

each of the 10 data sets. These assumptions can present some limitations (discussed later), however we find them sufficient for illustrating the model.

Network Parameters

Generally, mobile phone networks have demonstrated small world properties including relatively high clustering coefficient (which measures the ratio of triangles to connected triplets in the network, and means within the network there is a high probability that two neighbors of one node are connected themselves), scale-free degree distribution (degree-distribution of the network decays as a power law, has a power-law regime followed by a sharp cutoff, or has a fast decaying tail such as an exponential or Gaussian) and short average path length between nodes [17,18]. We examined these properties to enable comparison of our findings to what has been theoretically determined about spread of disease on networks. As well, these properties enable comparison to other mobile phone calling networks that will be or have been examined. Although not exhaustive, this could indicate how reproducible the results are, and illustrates variation in the 10 network types.

We calculated the average path length and average clustering coefficient for each of the 10 networks [5]. Also, we confirmed that the networks show scale-free behavior [19]. We used a simple approach based on maximum likelihood estimation to examine the power-law behavior [20]. All of the data sets also show a clear deviation from power-law behavior (sharp cutoff), which is indicative of broad-scale or truncated scale-free networks; a type of small world network.

Population Movement

Mobile phone calling records have also been used to understand human movement in relation to disease spread [3,4]. Precise geo-located incidence or prevalence information about the ongoing meningitis outbreak was not available. The WHO reported endemic regions of the country were in two districts (Tengrela and Koutu) [11], while

simultaneously we found news reports of meningitis occurring in other parts of the country (Figure 1). Thus because of evidence for meningitis in various parts of the country, we examined any overall national changes in position from the antenna call data. We measured the average latitude of incoming and outgoing calls from the antenna network data over the entire network for each of the 10 data sets to examine if there was any substantial average movement of the population over the selected time frame.

Disease Model and Dynamics.

Most models for the study of infectious diseases can be classified as compartmental (deterministic or stochastic), meta-population, individual-based (also agent-based) or contact network models. Contact network models have been widely used in the study of infectious diseases such as influenza [16], but scarcely in the study of the spread of bacterial meningitis. In this study we present a Susceptible-Carrier-Infectious-Recovered (SCIR) model (Figure 2) [21], which is an extension of the Suspected-Infected-Recovered (SIR) model typically used to model the progress of an epidemic. Network information on which the epidemic is progressing can be used to more accurately understand this spread by incorporating information about the strength of connections between any two nodes in the model [8].

The construction of the contact network model entailed two steps. First, we defined a contact graph representing contact patterns between individuals in Côte d'Ivoire. Second, we defined a mathematical model describing meningitis spread through a human population. We discuss each of these steps in proceeding sections.

SCIR Modeling

As stated, we use a Susceptible-Carrier-Infectious-Recovered (SCIR) model (Figure 2) to describe disease transmission and progression. Typically each node represents an individual, but here we modeled antennas as the nodes (*Network Descriptions* section). Nodes in the network move through four disease stages. Initially a proportion of nodes in

the population are assigned to each of the disease states. A susceptible node becomes infected based on contact with a carrier. The probability that a carrier node v will infect a susceptible node u is given as [6]:

$$T(u,v) = 1 - (1 - r_{u,v})^{\tau} \quad (1)$$

This assumes discrete time steps. Here, $r_{u,v}$ represents the probability of transmission per unit time, and τ is the number of time steps u is infectious. Carriers can infect susceptible nodes without exhibiting symptoms of the disease. Carriers can either become infectious or return to the susceptible class. After the disease has run its course, infected nodes move to the removed state. The removed class therefore consists of nodes that were infected and then recovered or died from the disease and also nodes that are vaccinated at the start of the epidemic as discussed in the next section.

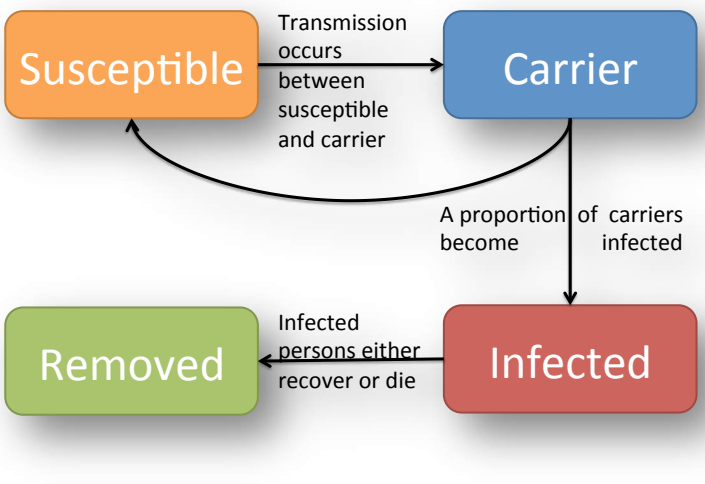


Figure 2: SCIR model of Meningitis transmission and progression. Individuals move through four disease states.

The model parameters were obtained from published reports. *N. meningitidis* carrier rates are approximately 10% in the meningitis belt [22]. However, this is expected to be higher in an epidemic. Thus we initialized the model by depicting 20% of the nodes as carriers. Mean carriage time is about 30 days [23], nodes are infectious for 1, 2 or 3 days [24], and death due to infection varies from year to year, however we set this to 10% based on the 1995- 1997 epidemic in which 25,000 of the 250,000 cases died [25]. The transmission rate between carriers and susceptible is unknown. We therefore ran a

sensitivity analysis on the transmission rate parameter to find values that result in an epidemic. The transmission rate was set at 20% for all simulations.

A proportion of individuals (nodes) can be vaccinated at the start or during the course of a simulation. Vaccinated individuals cannot contribute to the spread of the disease. Since there is meningitis vaccination available in Côte d'Ivoire [2], simulations in this study were run with a 20% vaccinated proportion at the start of the epidemic. We compared the resulting epidemics across the different networks.

Results

Disease Progression:

WHO reports the number of meningitis cases and deaths weekly in 9 countries in the meningitis belt [11]. The case data for Côte d'Ivoire are illustrated in Figure 3. Two districts reached the epidemic threshold during the period of this study (10 cases/100 000 inhabitants/week if no epidemic for 3 years and vaccination coverage < 80% or alert threshold crossed early in the season, else 15 cases per 100 000 inhabitants/week) [26]. HealthMap alerts regarding meningitis occurred throughout the period (Table 1).

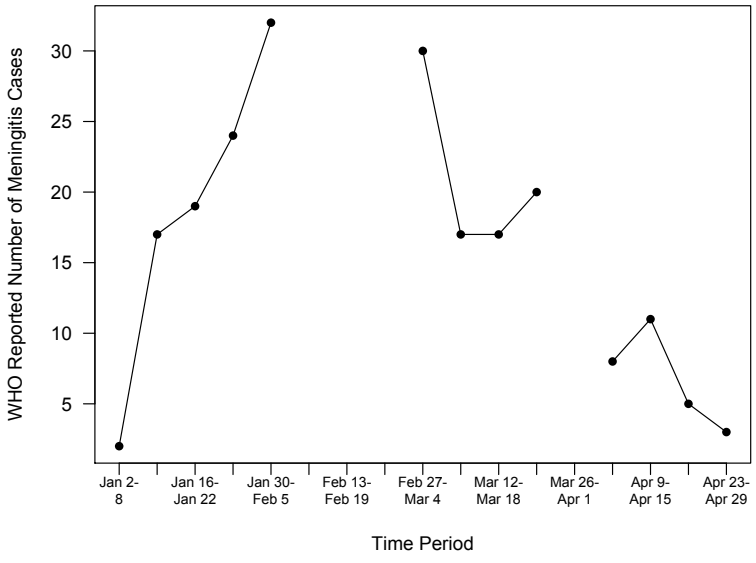


Figure 3. WHO meningitis case data by week. Weekly feedback bulletin on cerebrospinal meningitis on total number of cases for the country by week. Weeks with missing data have no marker in the plot.

Network Parameters:

The number of nodes, clustering coefficient and scaling exponent from the power-law fit are reported in Table 2. The degree distribution for the networks considered showed generally most nodes (antennas) having lower numbers of connections, and few nodes with a large number of connections. The scaling exponents from the power-law assessments were similar to that observed in previous studies of phone calling and e-mail networks; scale-free behavior of the degree distribution has been described with an exponent of -1.81 to -2.1 [5]. The number of nodes and scaling exponents remained relatively constant (don't significantly change from week to week). The clustering coefficient however, significantly changed each week ($p < 0.0001$), meaning that the network became more and more random (clustering coefficient decreases) from set 0 to set 7, then more organized again (more small-world in structure).

Network Number	Degree Distribution	Average Clustering Coefficient	Number of Nodes	Scaling Exponent (power-law regime)
0	Dec 5 - 18	0.849	1109	-1.30
1	Dec 19 - Jan 1	0.773	1100	-1.27
2	Jan 2 - Jan 15	0.723	1115	-1.29
3	Jan 16 - Jan 29	0.756	1093	-1.30
4	Jan 30 - Feb 12	0.752	1081	-1.30
5	Feb 13 - Feb 26	0.734	1030	-1.26
6	Feb 27 - Mar 11	0.724	999	-1.26
7	Mar 12 - Mar 25	0.702	929	-1.30
8	Mar 26 - Apr 8	0.744	1205	-1.25
9	Apr 9 - Apr 22	0.728	1211	-1.28

Table 2. Network properties of the antenna-to-antenna networks by 2-week periods.

Simulated Epidemics

As stated, a Meningitis epidemic was simulated over each of the ten networks. The same parameters were used in each of the simulations. Each simulation was run for 20 time steps, with each time step representing one week. The twenty weeks represent the time

period for which the mobile phone calling data was available. In addition, each simulation was replicated 6 times to capture the variability inherent in the system. The mean epidemic curve illustrating the number of infected individuals at each time step is illustrated in Figure 4. Although not explicitly visible in Figure 4, several of the simulated epidemics have multiple peaks. First peaks are likely due to depletion of the susceptible population, while peaks after are probably due to repopulation of the susceptible population by individuals moving from the carrier population. As shown in Figure 2, carriers can either become infected or susceptible again. The epidemic curves also have similar peak infected proportions. In all instances, the infected proportion is between 20% and 26% of the population. Carriers appear to drive the epidemic as expected, since they have the ability of infecting susceptibles and either progressing to the infectious class or returning to the susceptible class. As for peak time, Figure 5 shows that the epidemics simulated on each of the different network types peak at different time points. The results suggest that slight differences in the networks structure influence epidemics simulated over them. Networks 6 and 7 (described in Table 2) have the highest mean peak times compared to the other networks. These networks also have the lowest clustering coefficients (Table 2) and occur after the number of meningitis cases started to decrease (Figure 3).

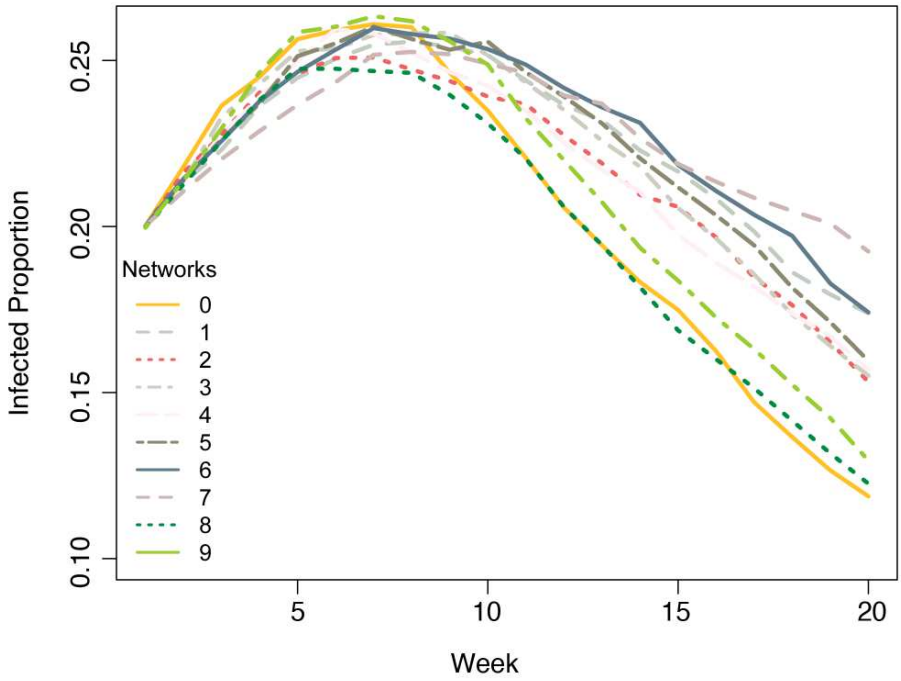


Figure 4. Proportion of infected individuals over time. The proportion of infected individuals over time, generated from simulating the epidemic on each network data set.

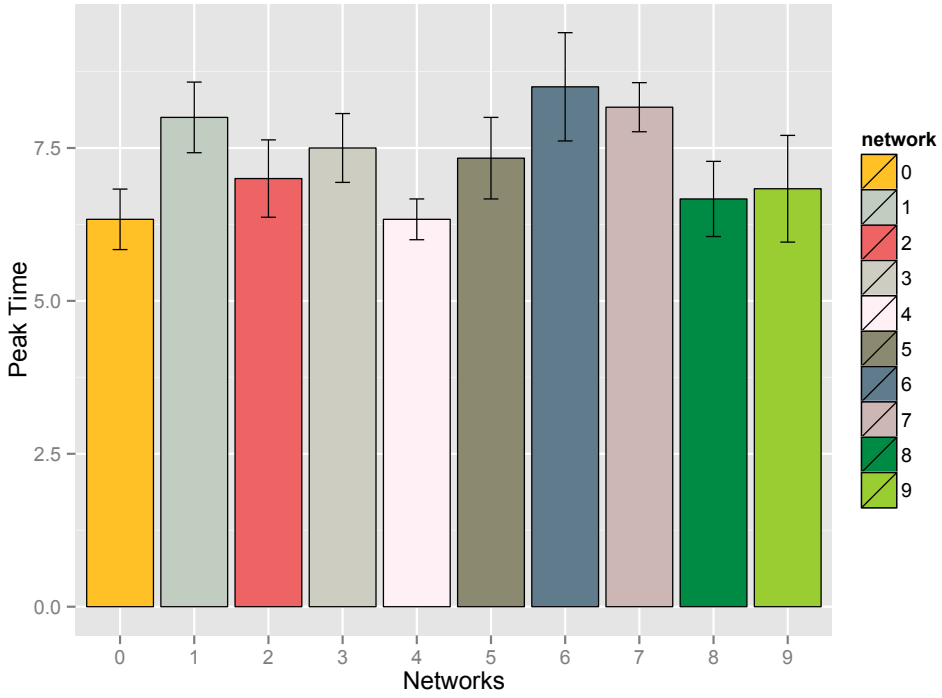


Figure 5. Peak infection week by network. Peak week of infection by network data set. Mean (height of the bars) and variance (error bars) from 6 simulations is illustrated.

Population Movement:

Figure 6 shows the average latitude of incoming and outgoing calls (via the antennas), for the entire nation. The average latitude of incoming calls both increase from the period of December 5 – 18 to March 12 – March 25, after which it decreases again. The maximum change in latitude is 0.285 degrees, which corresponds to about 33 km. In reference, the entire country is approximately 600 km from north to south, thus this represents a small change in latitude of the average calls.

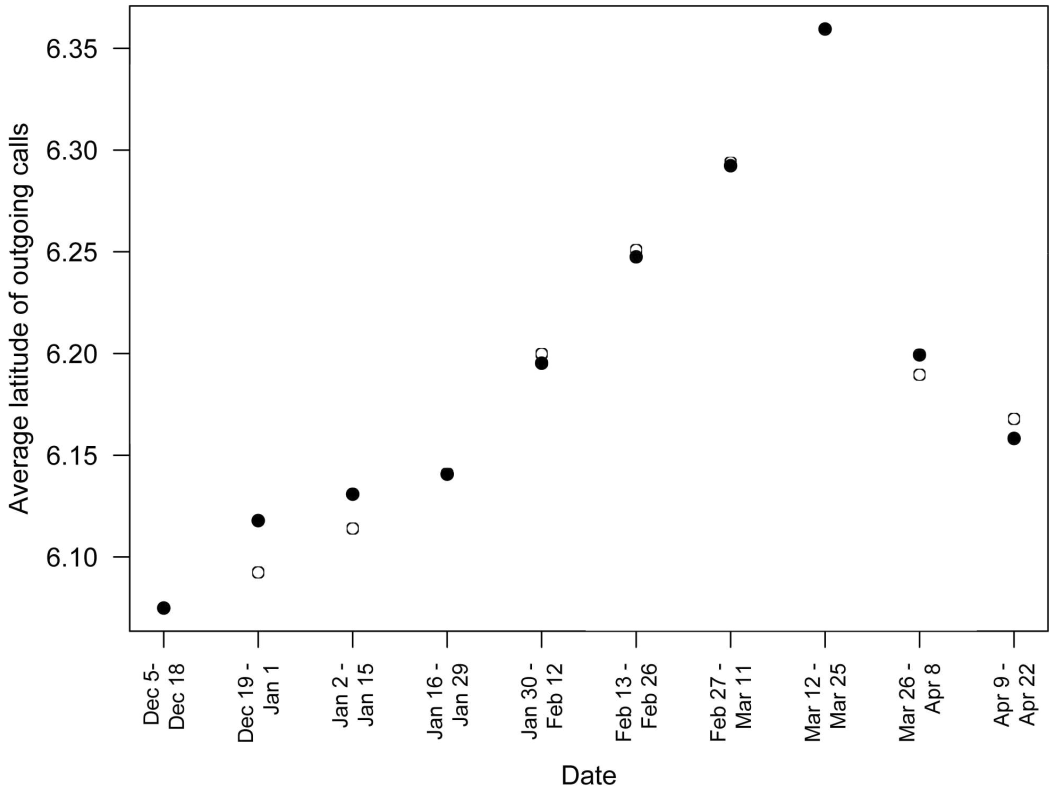


Figure 6. Average latitude of incoming and outgoing calls. Average latitude for all calls made in the country, by 2-week period. Incoming (open circles) and outgoing (closed circles) calls are tallied separately.

Discussion

This study combined an emerging area of investigation, the use of mobile phone call records in development, with another pertinent area in disease epidemiology and public health: disease spread on networks. Further analysis must be done to examine time periods outside of the data presented here to understand the relationships between network structures and disease burden more robustly. Through simulation, our initial results suggest that lower levels of clustering in the mobile phone networks are associated with later peak infection time periods. In turn, we also found that clustering coefficient of the networks decreased as the meningitis disease burden in Côte d'Ivoire empirically increased, reaching a minimum a few weeks after the cases peaked. A slight increase in overall incoming and outgoing call latitude was noted over the same time period, and peaked at the same time as the minimum clustering coefficient.

A main strength of this study is that it adds to a growing body of literature on the use of mobile phone calling records in development and more specifically infectious disease spread applications. As well, we are able to contrast and compare our simulated results regarding speed of disease spread given network structure with contemporary disease outbreak information from the WHO. This area is understudied at this scale because of a dearth of both large data about network structure and parallel disease information. To our knowledge, prior to this, compartmental models incorporating network, or strength of tie, information have not been studied for meningitis.

There are many limitations to this study. First, our results are only specifically applicable to this one disease outbreak and location (Côte d'Ivoire). In terms of the data itself, there were some shortcomings based on the data that was made available as part of the competition (missing records and no longitudinal and location information for individual callers). For the analysis, our use of the antenna nodes does not translate in a precise way to the individual or population of callers making the calls, and is an aggregate measure of clustering. Also the relationship between calling networks and real-life networks needs to be studied.

We could have analyzed the data at a finer time resolution instead of at the bi-week level (i.e. contact network structure at the weekly, daily or even hourly level would have been possible). We also do not evaluate the results by geography even though the precise geo-location of each antenna was provided. Instead, our results regarding disease outbreak evolution and infection rates are overall for the nation. If more detail about meningitis incidence, prevalence or vaccination rates in the country were available, the model could be augmented by more informed initial conditions and transmission parameters. Finally, although we show the common occurrence of changes in the calling network patterns, and disease spread in the country (via simulation and case data), we have not established any causal relationships.

Our results indicate the possibility that lower clustering coefficients in networks can correspond to slower spread of disease and potentially lowering of disease burden. Without any evidence of causality, these results are still interesting from the perspective of disease surveillance, for giving indication of when disease spread might be heightened. This can be useful to policy and public health practitioners in order to target appropriate control measures, such as vaccination. Even though meningitis epidemics are strongly seasonal, the drivers of these fluctuations are poorly understood, which limits the predictability of outbreaks and the dynamic response to immunization. Specifically for meningitis in Côte d'Ivoire, news articles reveal that immunization clinics were held in March [27], even though cases had already peaked well before this (Figure 2). Additionally, monitoring calling patterns could also provide a way to estimate and optimize, in real-time, the effect of vaccination campaigns. Thus, near real-time information that can be passively accrued offers a rapid, low-cost way to learn about disease dynamics. Higher levels of clustering in the mobile phone network may represent more people getting in touch to check on ill family or friends, or conversely may result in higher population density which has been shown to play a role in spread of infectious diseases [28]. The lag between peak case burden and minimum clustering coefficient may be indicative of the time needed for these effects to occur. The slight increase in call latitude could represent individuals travelling to areas in the north part of the country (or

returning). However more study, including analysis of the calling patterns outside of the time period provided, is required to strongly understand the mechanisms and time scales of these effects.

The study here proposes a new way by which mobile phone calling record data can be used to study disease dynamics and inform disease control measures; via calling network topologies. We examine these topologies on an aggregate scale via antenna-to-antenna calling record data. At this level, we detected significant changes in properties of the national calling network topology, which by simulation also reflected in the time course of meningitis spread. Further, these changes occurred during an ongoing meningitis outbreak in Côte d'Ivoire, and occurred simultaneously to evolution of the ongoing outbreak. Future research should examine these parameters in higher spatio-temporal resolution; both the disease burden and calling patterns on an individual level. For example, is the clustering occurring in particular geographic areas, and can that be related to disease incidence? Most importantly, before these results can be used practically, the same analysis of network characteristics must be evaluated for other outbreak situations and over longer time periods to investigate if similar patterns occur. Particularly for this outbreak, mobile phone calling patterns prior to December should be examined to understand the changes in network topologies prior to peak disease incidence. In conclusion, this study adds to the growing body of research on how mobile phone calling record data provides an opportunity to augment our understanding of disease dynamics. This can be useful for filling gaps in traditional surveillance systems in underdeveloped and developed areas around the world.

References

1. Bogich TL, Chunara R, Scales D, Chan E, Pinheiro LC, et al. (2012) Preventing pandemics via international development: a systems approach. PLoS Medicine 9: e1001354.
2. Organization WH (2010) Communicable Disease Epidemiological Profile - Cote d/Ivoire.
3. Tatem AJ, Qiu Y, Smith DL, Sabot O, Ali AS, et al. (2009) The use of mobile phone data for the estimation of the travel patterns and imported Plasmodium falciparum rates among Zanzibar residents. Malar J 8:287.: 10.1186/1475-2875-1188-1287.
4. Wesolowski A, Eagle N, Tatem AJ, Smith DL, Noor AM, et al. (2012) Quantifying the impact of human mobility on malaria. Science 338: 267-270. doi: 210.1126/science.1223467.
5. Ebel H, Mielsch L-I, Bornholdt S (2002) Scale-free topology of e-mail networks. Phys Rev E 66.
6. Newman MEJ (2002) Spread of epidemic disease on networks. Physical Review E 66: 016128.
7. Jeger MJ, Pautasso M, Holdenrieder O, Shaw MW (2007) Modelling disease spread and control in networks: implications for plant sciences. New Phytologist 174: 279-297.
8. Salathé M, Kazandjieva M, Lee JW, Levis P, Feldman MW, et al. (2010) A high-resolution human contact network for infectious disease transmission. Proc Natl Acad Sci U S A 107: 22020-22025.
9. Christakis NA, Fowler JH (2007) The spread of obesity in a large social network over 32 years. N Engl J Med 357: 370-379.
10. Freifeld CC, Mandl KD, Reis BY, Brownstein JS (2008) HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports. Journal of American Medical Informatics Association 15: 150-157.
11. The World Health Organization (Ongoing) Global Alert and Response (GAR) Epidemiological Information.
12. Organization WH (2012) Meningococcal meningitis Fact sheet N°141.
13. Boisier P, Nicolas P, Djibo S, Taha M-K, Jeanne I, et al. (2007) Meningococcal Meningitis: Unprecedented Incidence of Serogroup X—Related Cases in 2006 in Niger. Clinical Infectious Diseases 44: 657-663.
14. Miller MA, Wenger J, Rosenstein N, Perkins B (1999) Evaluation of meningococcal meningitis vaccination strategies for the meningitis belt in Africa. The Pediatric Infectious Disease Journal 18: 1051-1059.
15. Blondel VD, Esch M, Chan C, Clerot F, Deville P, et al. (2012) Data for Development: the D4D Challenge on Mobile Phone Data. arXiv preprint arXiv:12100137.
16. Meyers LA (2007) Contact network epidemiology: Bond percolation applied to infectious disease prediction and control. Bulletin-American Mathematical Society 44: 63.

17. Karsai M, Kivelä M, Pan R, Kaski K, Kertész J, et al. (2011) Small but slow world: how network topology and burstiness slow down spreading. Physical Review E 83: 025102.
18. Dong ZB, Song GJ, Xie KQ, Wang JY. An experimental study of large-scale mobile social network; 2009. ACM. pp. 1175-1176.
19. Amaral LAN, Scala A, Barthélémy M, Stanley HE (2000) Classes of small-world networks. Proceedings of the National Academy of Sciences 97: 11149-11152.
20. Jiang Z-Q, Xie W-J, Li M-X, Podobnik B, Zhou W-X, et al. (2013) Calling patterns in human communication dynamics. Proceedings of the National Academy of Sciences.
21. Vereen K (2008) An SCIR Model of Meningococcal Meningitis.
22. Martcheva M, Crispino-O'Connell G (2003) The transmission of meningococcal infection: a mathematical study. Journal of mathematical analysis and applications 283: 251-275.
23. Mueller JE, Sangaré L, Njanpop-Lafourcade BM, Tarnagda Z, Traoré Y, et al. (2007) Molecular characteristics and epidemiology of meningococcal carriage, Burkina Faso, 2003. Emerging infectious diseases 13: 847.
24. Vedros NA (1987) Evolution of meningococcal disease: CRC Press.
25. Immunization VaBDotWHO (2007) The evolving vaccine pipeline.
26. Organization TWH Humanitarian Health Action.
27. Kazony C (2012) Bouaké : 1200 enfants vaccinés contre la méningite.
28. Morse SS (1995) Factors in the emergence of infectious diseases. Emerging infectious diseases 1: 7.

Are gravity models appropriate for estimating the spatial spread of malaria?

Amy Wesolowski[1] and Caroline O. Buckee[2,3,*]

[1] Department of Engineering and Public Policy, Carnegie Mellon University, Pittsburgh, PA, USA.
[2] Department of Epidemiology, Harvard School of Public Health, Boston, MA, USA.
[3] Center for Communicable Disease Dynamics, Harvard School of Public Health, Boston, MA, USA.

Abstract:

Human mobility underlies the spatial transmission of malaria; infected and often asymptomatic individuals can carry malaria parasites further than the limits of mosquito dispersal, and contribute to malaria anywhere that is receptive to transmission. Refined spatial and temporal data are thus needed to accurately model how the dynamics of human populations contribute to epidemiological patterns of malaria, but human mobility data is often difficult to obtain on a scale relevant for malaria modeling. Here we use mobility data provided by Orange Telecom as part of the Data for Development (D4D) project to understand how human mobility contributes to malaria transmission in Cote d'Ivoire. We estimate malaria importation between sub-prefectures in Cote d'Ivoire (CIV), which is highly endemic throughout the country, as well as identifying sources and sinks of parasite importation. We compare these estimates to importation estimates derived from a gravity model, a common mathematical formulation for understanding population dynamics, to test the general utility of gravity models for malaria importation estimates. We find that although gravity models generally over-estimate parasite importation and are sensitive to assumptions about the duration of malaria infections, they can identify important source and sink locations.

*correspondence to: cbuckee@hsph.harvard.edu

Introduction

Human mobility can affect the geographic spread of infectious diseases worldwide [1-5]. Malaria is one of the most deadly infections, causing nearly 1 million deaths each year, primarily in young children and infants [6]. Sub-Saharan Africa burdens most of these deaths, and recent calls to eliminate the parasite from many countries have reinvigorated efforts to understand factors underlying transmission rather than simply focus on preventing disease. Since immunity prevents disease but not infection, individuals continue to become infected with the parasite throughout their lives. Indeed in highly endemic regions like Cote d'Ivoire (CIV), most infected people have no symptoms, so the bulk of malaria cases remain "unseen" [7, 8]. These asymptomatic carriers transport the parasite between regions, creating a dynamic landscape of transmission and making it difficult to understand the impact of regional interventions on the overall burden of infection and disease [9-12].

To estimate the impact of travel on malaria transmission, detailed spatiotemporal travel data are necessary [4, 13-14]. In particular, since infections are finite in duration and an individual's probability of being infected is geographically variable, the number and duration of journeys between particular locations are both required. In the absence of explicit data of this kind about mobility in Africa, models of spatial spread of infectious diseases often rely on gravity models, which assume that the only factors required to estimate movement are the physical distance between locations and the population size at each location [1,15]. Although these assumptions are simple and explicit, it has not been possible to test their validity among African populations.

Here, we use human mobility data from Cote d'Ivoire, provided by Orange Telecom as part of Data for Development Challenge, to estimate malaria importation between regions. We then examine how well gravity models perform relative to these estimates, and consider how sensitive gravity models are to variations in some of the most uncertain aspects of malaria epidemiology. We show that travel and parasite movement are both strongly correlated with population, but heterogeneous population density and travel patterns mean that importation is not homogeneous despite a fairly constant parasite rate across the country. We have identified sources and sinks with the capital city, Abidjan, being the most stable source and sink of both travel and malaria parasites, regardless of the modeling assumptions. Using a gravity model overestimates travel and importation, although it does identify the relative sources and sinks of importation.

Methods and Materials

Measuring subscriber travel patterns

Mobile phone call data records (CDRs) for 500,000 subscribers in Cote d'Ivoire were obtained from the Data for Development Challenge provided by Orange Telecom. The location (up to sub-prefecture) and date of each call made by these subscribers between the dates December 1, 2011 - April 28, 2012. We aggregated sub-prefectures to Cote d'Ivoire administration unit 3 to utilize other available spatial data. We approximated a daily location for each subscriber based either on the location of the majority of his or her calls or the location of their most recent call if no call was made during the day as done in previous studies [4]. Subscribers were assigned a primary location based on their

most frequent daily location. We quantified the number and duration of each trip taken outside of the primary location.

Parasite prevalence and population data

The main components of a spatial malaria transmission model are i) malaria prevalence data, ii) population density estimates, and iii) mobility data. We used the 1 km by 1 km gridded parasite rate (*PfPR2-10)* estimates from the Malaria Atlas Project to assign each settlement a malaria endemicity class from 1 to 7, as previously described [16]. Overall, the mean prevalence across the country is high (minimum: 44%, mean: 56%, median: 57%, maximum: 64%) (see Figure 1A) [16-17]. The parasite rate is higher on the southern coast towards the border of Ghana as well as in the center of the country (Koudou). Population estimates were obtained using the AfriPop data set (www.afripop.org) (see Figure 1B) [18].

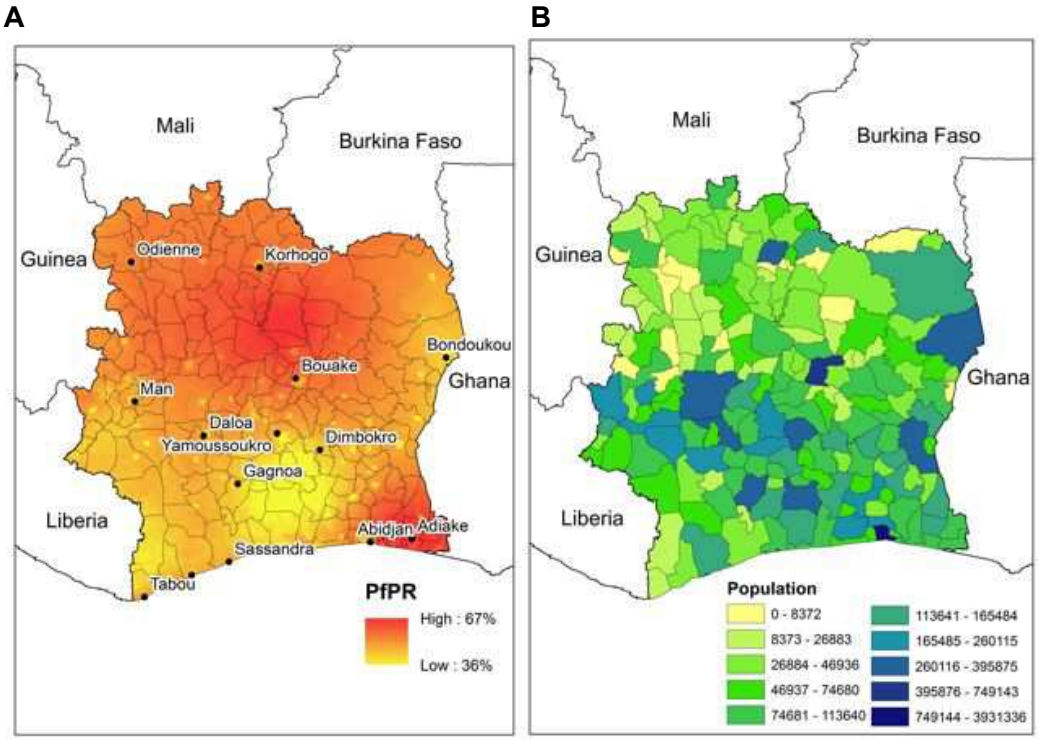


Figure 1: Parasite rate and population estimates for Cote d'Ivoire. A) Parasite prevalence ($PfPR_{2-10}$) is shown as a continuous surface for CIV. Major cities are also identified. B) The population for each location (sub-prefectures) is shown.

Quantifying importation by residents and visitors:

Travel by residents and visitors can contribute to malaria transmission in different ways. Residents traveling away from their home to a region with malaria have some probability of becoming infected during their trip and bringing parasites home with them (importation by residents). Visitors can bring parasites from their home location to the visited location if they stay overnight and are bitten by mosquitoes (importation by visitors). Since immunity is not sterilizing, we assume that naturally acquired immunity does not impact the probability of infection. We followed the same methodology as described in [4, 13].

We converted $PfPR_{2-10}$ to the entomological inoculation rate, EIR to calculate importation by returning residents. EIR is a measure of transmission intensity based on the number of bites by infectious mosquitoes per person per unit time. We used a modified version of the mathematical model from [4, 13-14] as follows: $EIR = -1.573 + 7.74 * PfPR_{2-10}$ setting $EIR = 0$ when $PfPR_{2-10} = 0$. We converted EIR values to $dEIR$, daily EIR values, as done previously [4, 13-14]. We then calculated the probability of each individual trip acquiring an infection P_r to each other location as:

$$P_r = \Pr\left(I((1 - (1+\propto b\, dEIR\, T_r)^{-\frac{1}{\alpha}}(\right.$$

where journeys last for T_r nights and b is the probability of infection given an infectious bite. We account for heterogeneous biting rates of mosquitoes on different individuals by an index α, in common with other malaria models [19]. We did not take into account individuals' immune status and thus these values should be compared relative to one another, not in terms of the impact on clinical cases. To include heterogeneity in infectiousness for individuals, we included the probability that each individual i will become infected ($\Pr(I()$. We choose a range of $\Pr(I)$ values from 10% to 100%. This discretization provides a more realistic approach to understanding individual contributions to transmission.

We calculated importation by visitors using the probability that the subscriber was infected (based on the $PfPR_{2-10}$ of their primary location) and initially assumed a duration of infection (DI) of 200 days [4, 13-14]. Each visitor can contribute at most

$$visitor(i,j) = \min\left(T_r, DI(* PfPR_{2-10}(i(\right.$$

per trip taken from his or her primary location ($i($ to each visited location (j). Here T_r is the trip duration and DI is the duration of infectiousness. We chose a range of DI values from 3 to 200 infectious days, to reflect the large uncertainty in the duration of infection with malaria in adults in particular. In general, the infectious period of a pathogen will be an important factor determining how mobility will impact its spatial spread.

To investigate the direction and total movements of parasites between locations, we analyzed 'source' and 'sink' locations. Each location as ranked based on their total emissions (sources) and receiving (sinks) of parasites.

Gravity model formulation

The most common model to quantify the flow of individuals between subpopulation is a gravity model [15]. The number of trips between locations i and j

$$N_{i,j} = \frac{population_i^{\alpha} population_j^{\beta}}{dist(i,j)^{\gamma}} + k$$

where $population_i, population_j$ are the populations of each location and $dist(i,j)$ is the distance between locations. Here we used Euclidean distance between the centroids of each location. The exponents, α, β, γ and intercept k were obtained from fitting the model to actual data using a generalized linear model with a Poisson specification [20-21]. It assumes that the only factors to estimate movement are locations, measured by the physical distance between locations, and importance, measured by the population size at each location. The simplicity of this model makes it a commonly used method to approximate movement between locations using empirical data. The exponents for separate gravity models describing each type of movement using population estimates from the census and Euclidean distance between the centroids of each county were estimated.

Duration of travel

A gravity model predicts the number of trips between locations. Modeling the importation of malaria parasites also requires the duration of each trip. To understand the importation of malaria parasites the duration of each journey is also necessary since the duration of each trip affects the likelihood of parasite movement. We modeled trip duration as an exponential decay fit to the actual distribution of trip lengths. All trips lasted between one and seventy-five days, with the majority of trips lasting a week or less (93%). We assumed an exponential decay for the length of each trip.

$$number\ of\ trips\ (l) = -a\exp(-l) + b$$

where l is the total duration (days) of each trip, $a = 1.14$, and $b = 0.0044$ (residual sum of squares = 0.0071).

Results

Travel patterns

The monthly average number of trips between all pairs of locations varied greatly with a mean of 34 trips (median: 2.5, 90% quantile interval 0.25 – 98). Figure 2 A,B show the monthly average number of trips originating and arriving (destination) at each location over the entire period of the data. The sub-prefecture containing the capital city, Abidjan, is the location with the most travel, by an order of magnitude. Unsurprisingly, less travel occurred in the northern, more rural parts of the country. More populated sub-prefectures and those containing major cities also experienced higher amounts of travel. The total amount of travel strongly correlated with population (correlation coefficient total origin: 0.967, total destination: 0.938). Figure 2 C shows the ratio of total origin to total destination travel, giving an indication of relative travel patterns. Sub-prefectures are colored if they are primarily a destination (blue) or origin (red), regardless of the total amount of travel. The areas that are primarily an origin of travel are not the most populous areas (see Figure 2C to compare) and in a number of instances, border neighboring countries. Although Abidjan is the most traveled to location, it is primarily an origin of travel and not highly skewed as either a source or sink of travel. The areas that are primarily destinations are often less populated, but this pattern is less clear.

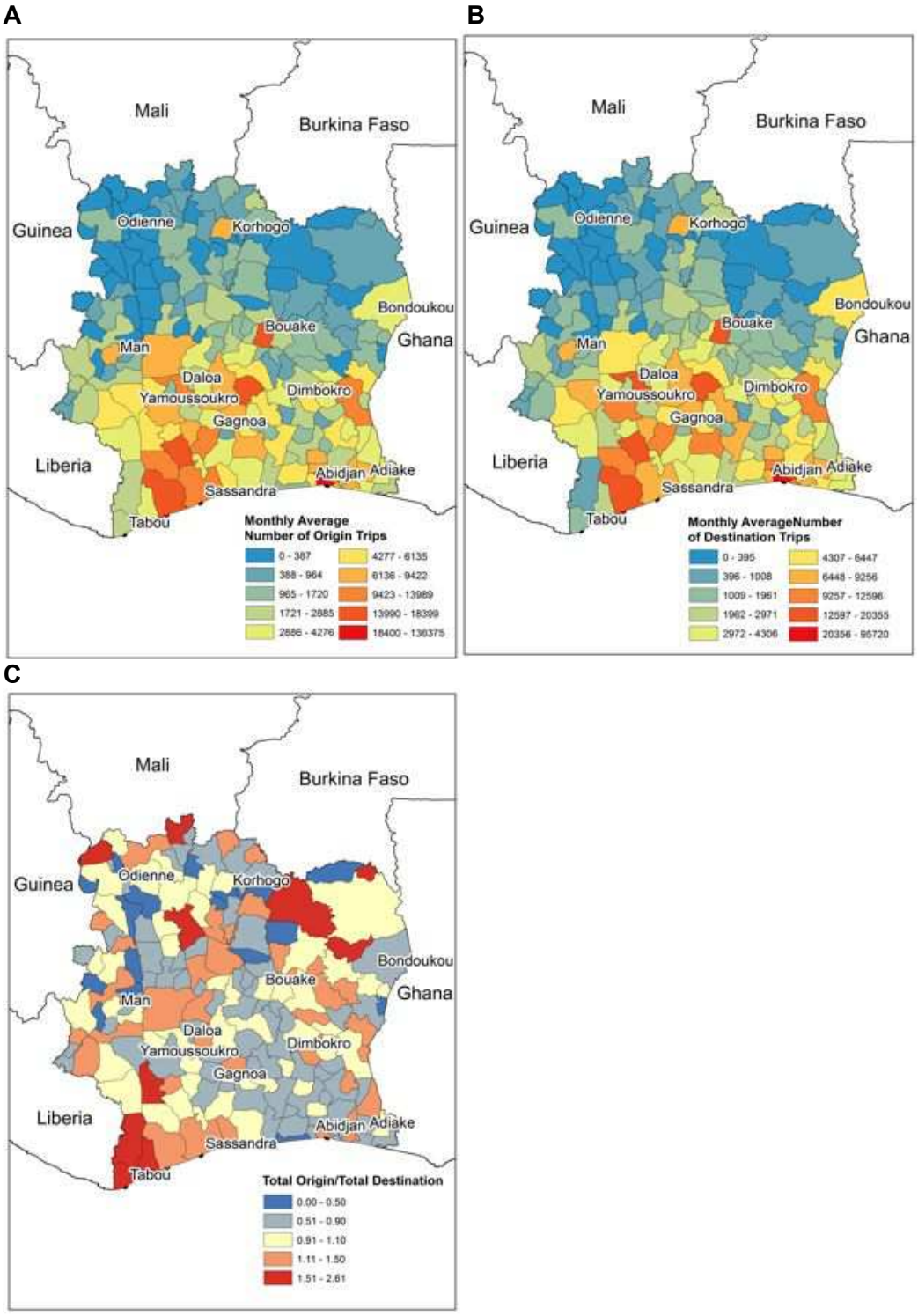


Figure 2: The monthly average amount of travel from (origin) and to (destination) each location. The monthly average number of trips quantified for each location is shown. The number of trips with each location as the origin (A) and destination (B) are highlighted. C) The ratio of total origin to total destination travel is shown. Sub-prefectures are colored based on if they are primarily a destination (blue) or an origin (red) of travel.

Sources and sinks of importation

To understand how these relatively complex patterns relate to malaria transmission, we apply our malaria data and model. In Cote d'Ivoire, malaria is endemic, making the impact of mobility patterns more difficult to disentangle than in areas that have a clear separation of malaria-endemic and malaria-free zones. We find that the impact of human movement on malaria in CIV varies spatially, although there exist consistent sources and sinks of importation. First, we identified sources and sinks of parasite importation using both resident (assuming a 100% chance of being infected) and visitors (assuming a 200 day duration of infection). Monthly average values for importation by residents varies greatly with a mean of 12 (median: 1, 90% quantile interval: 0 -98) (see Figure 3 A,B). Values for importation by visitors are substantially higher and vary over a larger range (mean: 58, median: 256 90% quantile interval: 0 - 174) (see Figure 3 C,D). This is unsurprising given that the maximum amount each trip can contribute to resident importation is one, whereas the value is technically unbounded for visitor contribution. The total importation source and sink values were strongly correlated with population (correlation coefficient resident source: 0.949, resident sink: 0.804, resident source: 0.948, resident sink: 0.89 all $p<0.001$).

Abidjan is the primary source and sink for both resident and visitor importation. For resident importation, southern sub-prefectures are the most dominant sources of importation (see Figure 3 A,C). The sub-prefecture containing Korhogo is the only dominant source of resident importation in the northern parts of the country. A number of more populated sub-prefectures (including the sub-prefectures contain Man, Bouake, Yamoussoukro, etc.) are primarily resident sinks (see Figure 3 B,D). The rank of visitor importation sources and sinks are nearly identical to the resident importation results (see Figure 4). In general, locations were not both majority sources and sinks (see Figure 4 A,B). In CIV, therefore, although parasite rates are high and relatively homogeneous across the country, heterogeneous population density and travel patterns means that the impact of mobility on importation is actually fairly heterogeneous. A correlate of this finding is that importation can impact the country-level impact of local malaria control programs. For instance, ignoring importation in Abidjan will likely reduce the efficacy of any control programs, whereas in the more rural northern parts of the country may not be greatly impacted by ignoring importation.

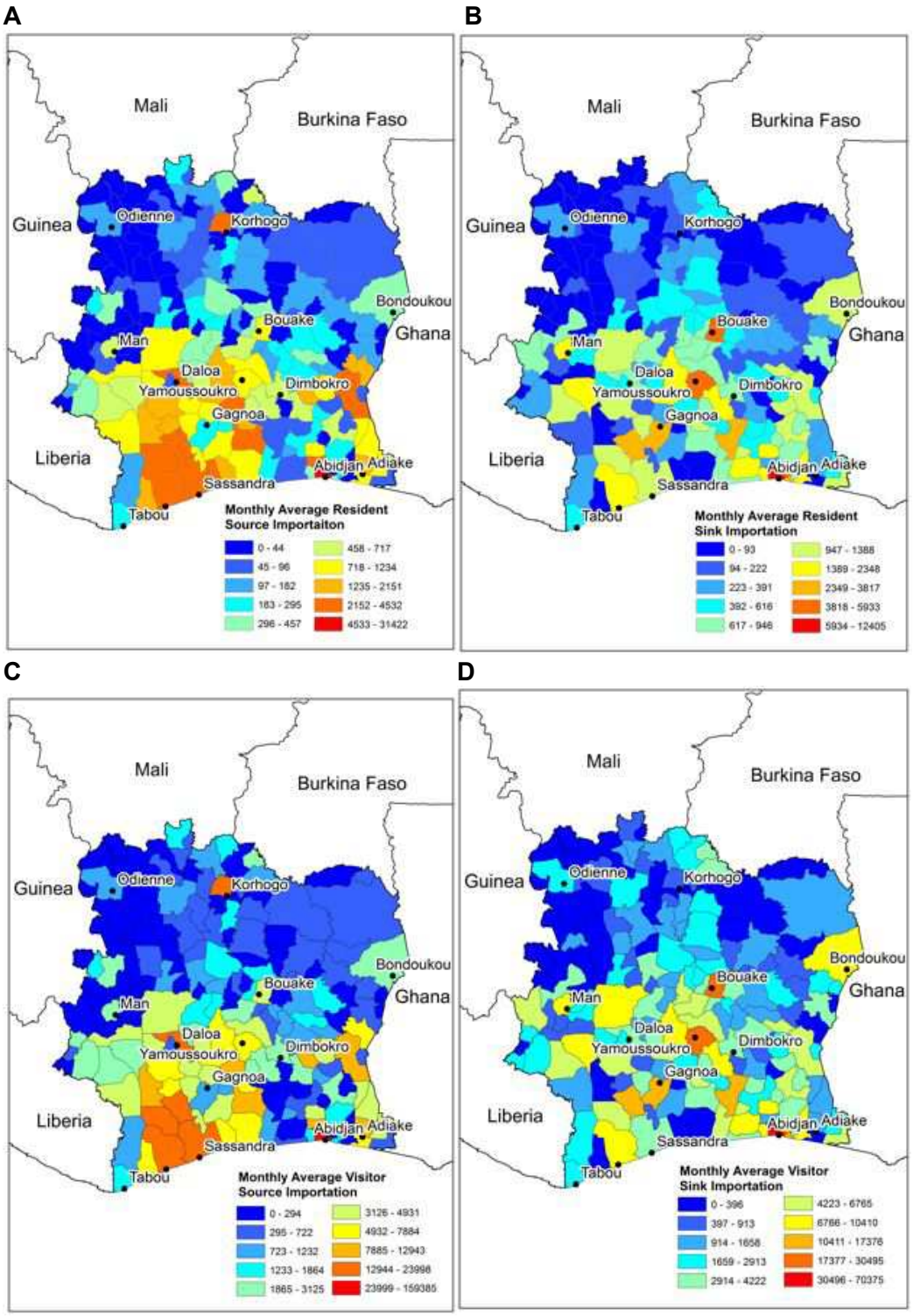


Figure 3: The monthly average amount of importation by residents and visitors. Each sub-prefecture is colored according to the total importation by residents A) as a source and a B) sink (for visitors C) source and D) sink). The top sources and sinks are shown in red.

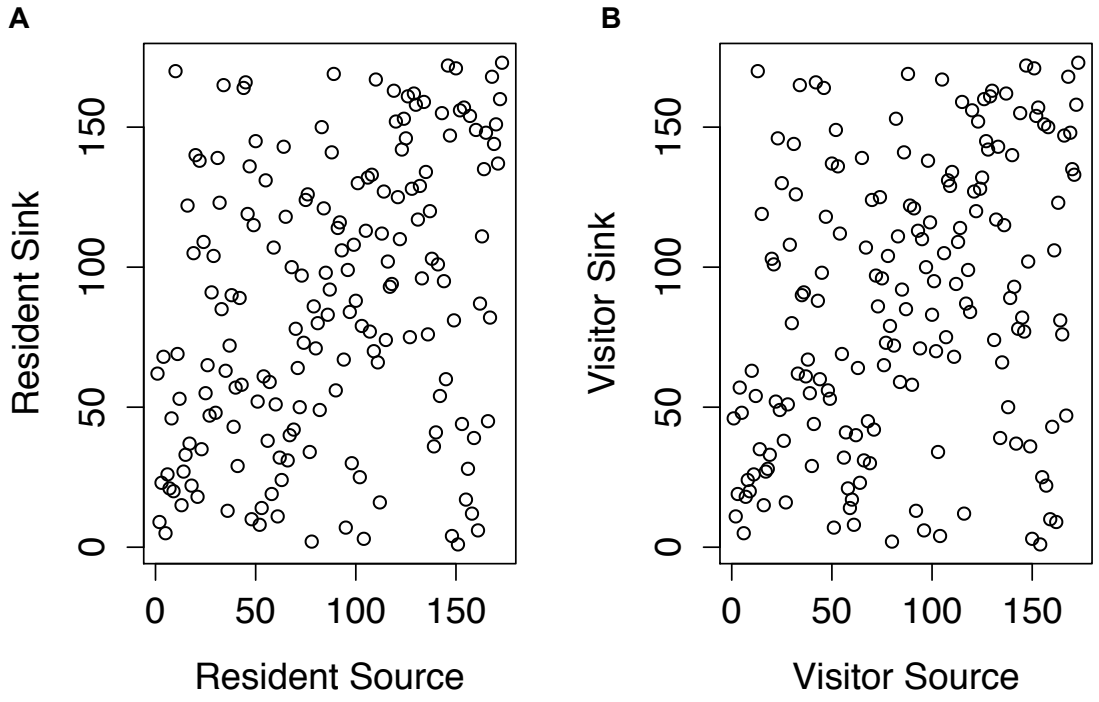


Figure 4: Sources and sinks of resident and visitor importation ranks. Each sub-prefecture was ranked according to its total importation value as a source or sink. For both resident and visitor importation values, major sources were not necessarily major sinks and vice a versa.

Modeling travel

Given the strong relationship between travel, importation, and population we modeled travel using a gravity model and compared estimates of importation (see Table 1). On average, the gravity model over estimated travel (39% greater) (see Table 2). Figure 5 shows the ratio of the data to predicted values from the model with the majority of values less than 1. In general, we identified few systematic biases where the model performed poorly. However, the model does the worst for trips over a short distance and when the true number of trips is greater than 10,000.

Table 1: Gravity model parameters. The estimated parameters from fitting a gravity model to the amount of travel between locations (reduction in deviance 56%).

Parameter	**Estimated Value**
α	0.8165
β	0.7814
γ	-1.518
k	-13.81

Table 2: A comparison of the gravity model results to data. The amount of travel estimated by the gravity model was compared to the data. In general, the gravity model over estimated travel.

	5%	50%	Mean	95%
Ratio	0.038	0.32	0.93	3.60

(data/model)				
Percent Increase (model>data)	-57%	52%	39%	93%

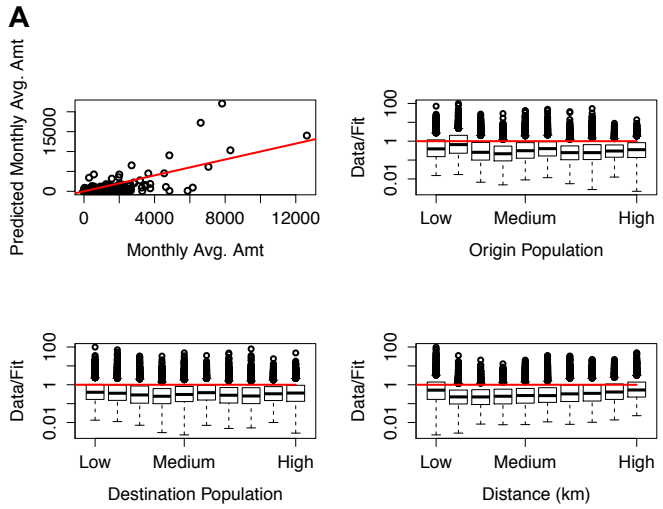


Figure 5: The predicted amount of travel from the gravity model. We fit a gravity model to the actual number of trips between locations from the mobile phone data. In general, the gravity model over predicts the amount of travel. There are no clear systematic biases in its estimated value for places of varying populations or over a range of distances.

The gravity model over estimates the amount of importation by residents and visitors (see Table 3, Figure 6 A, B) and can vary estimates greatly. The gravity model produced a wider range of results for visitor than resident importation values. However, as Figure 7 shows the model is still able to identify the same sources and sinks (adjusted R^2 total source rank resident importation: 0.78, total sink rank resident importation: 0.74, total source rank visitor importation: 0.77, total sink rank resident importation: 0.73, for all $p<0.001$).

Table 3: A comparison of the importation results. The amount of travel estimated by the gravity model was compared to the data. In general, the gravity model over estimated travel.

	5%	50%	Mean	95%
Ratio: Resident Importation (data/model)	0.039	0.36	1.12	4.42
Percent Increase of Model to Data: Resident Importation	-64%	46%	34%	93%
Ratio Visitor Importation (data/model)	0.13	0.65	1.6	6
Percent	-71%	21%	13%	78%

Increase of Model to Data: Visitor Importation				

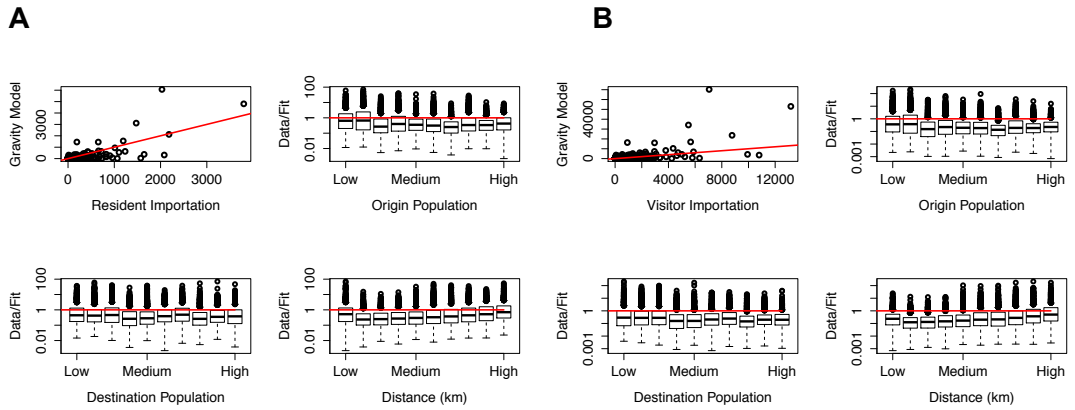


Figure 6: The monthly average amount of importation by residents and visitors using a gravity model. The estimated importation by residents and visitors compared to the predicted value using a gravity model is shown. In general, using a gravity model overestimated importation with no clear systematic biases based on the origin, destination population, or distance between the origin and destination.

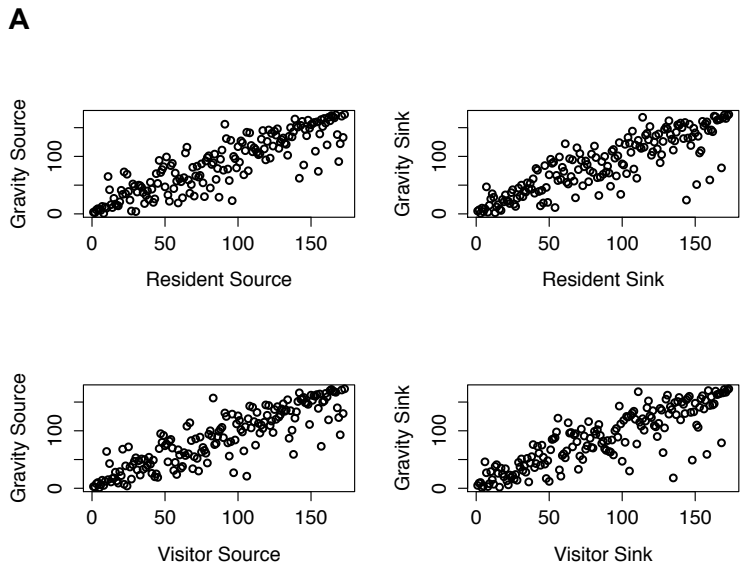


Figure 7: A comparison of major source and sinks identified using the data versus a gravity model. In general, a gravity model performed well identifying the source and sinks for both resident and visitor importation.

To further introduce heterogeneities in transmission modeling, we varied the percentage of individual trips infected for resident importation. Choosing a range of probabilities (between 0.1 and 1) produced results over an order of magnitude. Changing the percentage of individual trips infected does not change total importation estimates linearly since it is dependent upon the number of trips (see Table 4, 5).

Table 4: Summary statistics for the monthly average source and sink importation values by residents using data.				
	5%	50%	Mean	95%
Total Source: 10% infected	1	18	78	264
Total Source: 50% infected	6	79	337	1,151
Total Source: 100% infected	12	173	684	2,400
Total Sink: 10% infected	2	21	78	311
Total Sink: 50% infected	12	144	337	1,144
Total Sink: 100% infected	29	291	684	2,181

Table 5: Summary statistics for the monthly average source and sink importation values by residents using the gravity model.				
	5%	50%	Mean	95%
Total Source: 10% infected	2	24	68	183
Total Source: 50% infected	11	123	340	919
Total Source: 100% infected	22	245	680	1,814
Total Sink: 10% infected	4	29	68	271
Total Sink: 50% infected	20	141	340	1,357
Total Sink: 100% infected	40	284	680	2,730

We varied the duration of infectiousness (DI between 3 and 200 days) for visitor importation given uncertainties in infectiousness. Varying the duration of infection can change importation estimates by an order of magnitude (see Table 6, 7) for the highest estimates. However, it did not change the ability for a gravity model to identify major sources and sinks (adjusted R^2 between 0.74 – 0.77, $p<0.001$). Figure 8 shows the comparison of estimates using a shorter DI (<200 days) to our previously assumed DI (of 200 days). For all DIs greater than 100 days, the importation estimates are stable. Moreover, changing DI affects the range of importation estimates greater for a gravity model than the actual data.

Table 6: Summary statistics for the monthly average source and sink importation values by visitors using data.				
	5%	50%	Mean	95%
Total Source: 3 day DI	36	512	1,980	7,020
Total Source: 14 day DI	56	815	3,037	10,351
Total Source:	60	934	3,364	11,211

200 day DI				
Total Sink: 3 day DI	88	880	1,980	6,291
Total Sink: 14 day DI	158	1,508	3,037	9,426
Total Sink: 200 day DI	179	1,752	3,364	10,309

Table 7: Summary statistics for the total source and sink importation values by visitors using the gravity model.

	5%	50%	Mean	95%
Total Source: 3 day DI	60	660	1,836	4,931
Total Source: 14 day DI	120	1,250	3,502	9,401
Total Source: 200 day DI	264	3,664	7,631	20,626
Total Sink: 3 day DI	99	785	1,836	7,322
Total Sink: 14 day DI	184	1,502	3,502	13,875
Total Sink: 200 day DI	434	3,128	7,631	30,611

A

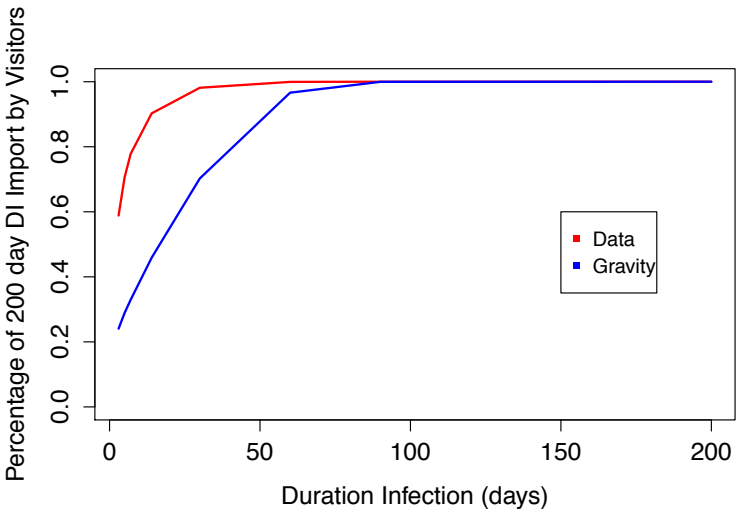


Figure 8: The impact of varying the duration of infection (DI). We varied the duration of infection and compared visitor importation estimates to the total visitor importation estimate (where we assumed a 200 day DI).

Discussion

Using mobile phone data and spatial malaria parasite prevalence estimates for CIV, we quantified parasite importation by residents and visitors due to travel. We identified sub-prefectures that were major sources and sinks of parasite movement. The sub-prefecture including the capital city was the clearest source and sink of parasite movement. However, top sources were not necessarily top sinks with the majority of major sources located in the southern, more populated portions of the country. Importation was strongly positively correlated with population. Modeling this type of heterogeneous importation for malaria transmission is particularly important because, unlike directly transmitted diseases, prevalence is driven by environmental determinants in addition to population density.

Obtaining detailed movement data is notoriously difficult and we used the most common spatial interaction model to investigate its impact on estimating importation, however. In general, we found that using a gravity model overestimated importation with no systematic bias based on population or distance between locations. However, a gravity model was able to identify similar sources and sinks as the data. We further introduced heterogeneities in our modeling assumptions by varying the percentage of individuals infected as well as the duration of infection. Varying the percentage of individuals infected non-linearly varied resident importation results due to heterogeneity in the number of trips between locations. Increasing the duration of infection increased visitor importation values with DI values greater than 100 producing the same importation estimates. Changing DI had a larger effect on importation estimates using a gravity model as opposed to data.

The data used in this analysis has inherent limitations. Mobile phone ownership bias is unclear in CIV and should be further studied to better understand the generalizability of these results [22]. The data are limited in the ability to estimate travel patterns, although we have previously found that substantial ownership bias did not greatly skew mobility estimates [23]. In addition, we analyzed movements on the sub-prefecture level due to data availability issues. This inherently misses smaller spatial scale movement patterns within a given sub-prefecture. In addition, we have used the Malaria Atlas Project estimates for $PfPR_{2-10}$, which were biased by poor data availability in CIV [17]. The data did not fall within the high season of malaria transmission in CIV and the variability in $PfPR_{2-10}$ as well as seasonal travel patterns is unknown and could skew our results.

The impact of human travel has a heterogeneous effect on malaria in CIV where prevalence is nearly highly endemic throughout the entire country. These results can be helpful in better understanding the malaria dynamics within the country and in particular the epidemiology in Abidjan [7]. As control programs within the country are improved, the impact of human travel may become more important making studies of these types more applicable for public health. In general, we propose that gravity models can be useful if the aim is to identify relative sources and sinks of parasite importation, but provide limited insights into the expected numbers of infections transported between regions.

Acknowledgements

A.W. was supported by the National Science Foundation Graduate Research Fellowship (0750271). C.O.B. was supported by the Models of Infectious Disease Agent Study program (cooperative agreement: 1U54GM088558). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of General Medicine Sciences or the NIH. The data was made available by French Telecom/Orange Cote d'Ivoire within the framework of the D4D Challenge.

References

1. D. Balcan *et al.*, Multiscale mobility networks and spatial spreading of infectious diseases. *Proc. Natl. Acad. Sci. USA* **106**, 21484-21489 (2009).
2. N. M. Ferguson *et al.*, Strategies for containing an emerging influenza pandemic in Southeast Asia. *Nature* **437,** 209-214 (2005).
3. I. R. Longini Jr. *et al.*, Containing pandemic influenza at the source. *Science* **309**, 1083-1087 (2005)
4. A. Wesolowski *et al.,* Quantifying the impact of human mobility on malaria. *Science* **338**, 267-270 (2012).
5. S. T. Stoddard *et al.*, The role of human movement in the transmission of vector-borne pathogens. *PLoS Negl. Trop. Dis.* **7,** (2009), doi:10.1371/journal/pntd.0000481.
6. S. I. Hay *et al.,* Estimating the global clinical burden of *Plasmodium falciparum* malaria in 2007. *PLoS Med.* **7**, (2010), doi:10.1371/journal.pmed.1000290.
7. S.-J. Wang *et al.,* Rapid urban malaria apprasial (RUMA) III: epidemiology of urban malaria in the municipality of Yopougon (Abidjan). *Malar. J.* **5**, (2006), doi:10.1186/1475-2875-5-29.
8. G. Raso *et al.,* Spatial risk profiling of *Plasmodium falciparum*parasitaemia in a high endemicity area in Cote d'Ivoire. *Malar. J.* **8**, (2009), doi:10.1186/1475-2875-8-252.
9. R. M. Prothero, Disease and mobility: a neglected factor in epidemiology. *Int. J. Epidemiol.* **6**, 259-267 (1977).
10. P. Bejon *et al.*, Stable and unstable malaria hotspots in longitudinal cohort studies in Kenya. *PLoS Med.* **7**, (2010), doi:10.1371/journal.pmed.1000304.
11. E. Dolgin, Targeting hotspots of transmission promises to reduce malaria. *Nat. Med.* **16**, 1055 (2010).
12. C. Lynch, C. Roper, The transit phase of migration: circulation of malaria and its multidrug-resistant forms in Africa. *PLoS Med.* **8,** (2011), doi:10.1371/journal.pmed.1001040.
13. A. Le Menach *et al.,* Travel risk, malaria importation and malaria transmisison in Zanzibar. *Nature* **93,** (2011), doi:10.1038/srep00093.
14. A. J. Tatem *et al.,* The use of mobile phone data for the estimation of the travel patterns and imported *Plasmodium falciparum* rates among Zanzibar residents. *Malar. J.* **8**, (2009), doi:10.1186/1475-2875-8-287.
15. A. Wesolowski *et al.*, The use of census migration data to approximate human movement patterns across temporal scales. *PLoS ONE*, **8**, (2013), doi:10.1371/journal.pone.0052971.
16. P. W. Gething *et al.,* A new world malaria map: *Plasmodium falciparum* endemicity in 2010. *Malar. J.* **10**, (2011), doi:10.1186/1475-2875-10-378.
17. G. Raso, *et al.,* Mapping malaria risk among children in Cote d'Ivoire using Bayesian geo-statistical models. *Malar. J.* **11**, (2012), doi:10.1186/1475-2875-11-160.
18. A. J. Tatem *et al.,* High resolution population maps for low income nations: combing land cover and census in East Africa. *PLoS ONE* **2**, (2007), doi:10.1371/journal.pone.0001298.
19. C. Dye, G. Hasibeder, Population dynamics of mosquito-borne disease: effects of slies which bite some people more frequently than others. *Trans. R. Soc. Trop. Med. Hyg.,* **80**, (1986).
20. A. Zeileis, C. Kleiber, S. Jackman, Regression models for count data in R. *J. Stat. Software,* **27**, (2008).

21. R. Flowerdew, M. Aitkin, A method for fitting the gravity model based on the Poisson distribution. *J. Reg. Sci.,* **22**, 191-202, (1982).
22. A. Wesolowski *et al.*, Heterogeneous mobile phone ownership and usage patterns in Kenya. *PLoS ONE,* **7**, (2012), doi:10.1371/journal.pone.0035319.
23. A. Wesolowsk *et al.,* The impact of biases in mobile phone ownership on estimates of human mobility. *J. R. Soc. Interface,* **10,** (2013).

On Models Characterizing Cellular Social Networks

Deepjyoti Deka and Sriram Vishwanath
Department of Electrical and Computer Engineering,
University of Texas at Austin, Austin, Texas 78712, USA

I. Introduction

Mobile phones have become an indispensable part of everyday life in today's world. From being simple mobile communication devices, cell-phones have graduated to become tools for accessing a host of web-based services like gathering news, posting on online social networks like Facebook, Twitter and for location-dependent services. Overall, cellular communication has become the predominant mode for information dissemination and connectivity between people. In this abstract, we focus on the network of connections between users/terminals and/or base stations of a cellular network induced by voice and text exchanges between them. This reflects an underlying social network which is dependent and in turn affects how close people are, both emotionally and geographically. A detailed understanding of this network can reveal important information about spread of information and viruses in the network. This can help in finding influential users and/or heavily used nodes and ways of optimally controlling of information flow in the network [7].

There has been substantial work on the relationship between the dynamics of a network and its vulnerability to its topological characteristics. For the physical network of buses representing the power grid, Albert et al. [1] and others [2] have studied the relation between grid failures and the topological structure of the power grid (as represented by the adjacency matrix and/or the network's degree distribution). There has been work along similar lines for social networks [4], the Internet [3] and population models for spread of diseases [5] and effective vaccination [6]. An accurate model of the degree distribution and characteristics of the adjacency matrix is a good starting point for the study of any network and is also the goal of this work.

For networks demonstrating a power-law degree distribution like the Internet and airline networks, Barabasi-Albert model [4] based on preferential attachment has been well studied. However, as seen in the cellular networks considered in this abstract, we find the underlying graph to be thin-tailed (exponential), and hence the principle of 'the rich gets richer' in the Barabasi-Albert model does not apply.

We provide a generative framework for understanding the cellular social networks by looking at the physical network evolution as a generative process and present a model that aims to explain the exponential degree distribution.We demonstrate the usefulness of our model by comparing the impact of infection propagation on both real networks as well as on our model.

The rest of this abstract is organized as follows. We present a description of the data-sets analyzed for this work in the next section. We then provide our generative model for the exponential degree distribution of cellular social networks in Section III. We analyze the degree distribution and provide numerical results in Section IV. We show the performance of our model in propagation of infection and compare the results with infection propagation in real networks in Section V. We present results, effects and future topics of study in Section VI.

II. Data-set on cellular social network

In this work, we consider the social network involving users of two different cellular networks. First, we investigate the social network of mobile phone calls made by users of an Asian telecom giant during the month of January 2009, and represent this by an undirected graph. In the graph generation, an undirected link in established between two nodes/users, if 4 calls or more are exchanged between them in the period. The giant component of this graph has 820000 nodes and 5 million links. Second, we study the data-set of phone-calls and SMS exchanges between 5 million Orange customers in Ivory Coast. This data set is provided through the D4D challenge [11]. In this case, we study antenna-to-antenna traffic on a hourly basis for a period of 150 days from December 2011 to April 2012. We consider the existence of an edge between two antennas if more than 100 exchanges have taken place in either direction between them. We take a threshold on the number of information exchanges needed for edge formation to exclude spurious communication in the data which do not reflect the usual patterns. A description of the data-set is given in [12]. Next, we present a generative model based on 2-D spatial Poisson point process theory.

III. Generative model for exponential degree distribution

We consider nodes to be randomly placed in space according to a Poisson point process P_λ with density λ. Every new node in the system connects to K of its nearest preexisting nodes in the network where K is a random variable. The nodes in the network are numbered same as the time step of generation. In a cellular social network, creating nodes within a given area represents formation of a plethora of ideas and interests among individuals and is followed by interaction (edge formation) with K others following similar interests. The formation of multiple connections by the nodes in the network can be seen as inducing increased information spread and a way to gather variety within the social network. The selection of 'nearest neighbors' is due to the tendency of forming links with similar users in cellular social network.

Similar users can be considered to be at a small emotional distance from each other which can further be a combination of social-economic considerations and physical distance between the communicating parties. Thus, in our generative model, each new node forms K connections into the existing network. We proceed to show that the network generated though our model has a degree distribution which is exponential in nature.

IV. Exponential fit for the generative model

We fit the degree distribution achieved from the generative model for different constant values of $K(= k)$ with exponential distributions as shown in Figures 1. We consider node creation over the specified area, simulate their positions randomly several times (500 iterations each), and then average our findings to obtain the degree distribution.

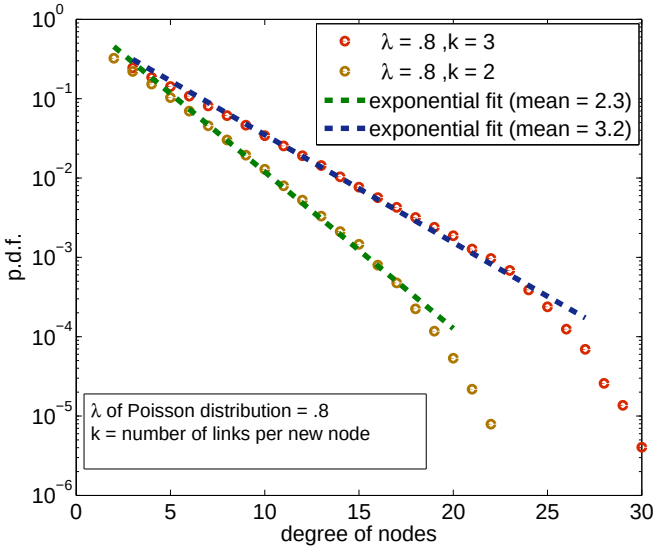


Fig. 1. Fitting exponential p.d.f. to the degree distribution for constant K

A. Fitting the degree distribution of social network of cellular users with exponential distribution

The degree distribution of the social network for the users in the Asian cellular network shows an exponential decay and a good exponential fit using our model is shown in Figure 2.

Next, we provide results from the D4D data-set. We plot the degree distribution of the antennas in the country which reflects the communication between the users they serve. From Figure 3, we see that the degree distribution has a good exponential fit validating our generative model. We measure the geographical distance between antennas sharing an edge. Here, we plot the distribution of the average distance that an antenna communicates over. Subsequently, we plot the distribution of the farthest distance that an antenna communicates with and finally show the distribution of the distance weighted by the number of calls that an antenna uses in communication. In Figure 4, we can see that the average distance weighted by the number of calls is shorter than the average distance reflecting the fact that most antenna-to-antenna communication happens

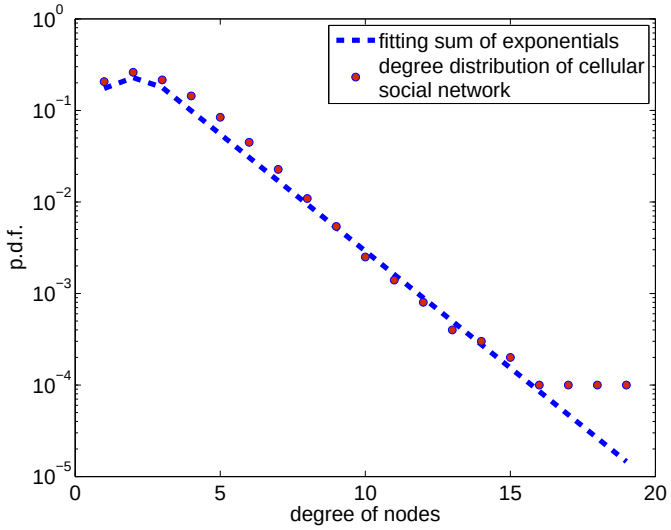


Fig. 2. Fitting exponential p.d.f. to the degree distribution of the Asian cellular social network

over short geographical distances. This is in agreement with our nearest neighbor approach in the generative model.

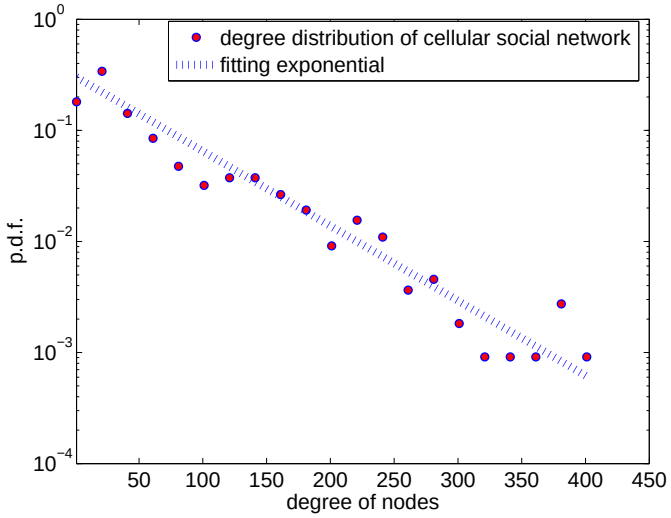


Fig. 3. Fitting exponential p.d.f. to the degree distribution of the D4D cellular social network

V. Vulnerability of the generative model

The users of the cellular network are at the risk of random and intentional attacks by disruptive agents that can propagate over the entire network. We consider the SIR (Susceptible-Infected-Removed) model of infection propagation in the network and analyze the network's vulnerability to it. The SIR model has nodes in one of three states: susceptible (S), infected (I) and removed (R). A node in the susceptible state gets to the infected state through any of its infected neighbors. Each infected neighbor can spread the infection to susceptible node through the shared edge with probability β. Each infected node itself gains resistance and enters the removed state with probability γ. Eventually, in a SIR model the infection dies out

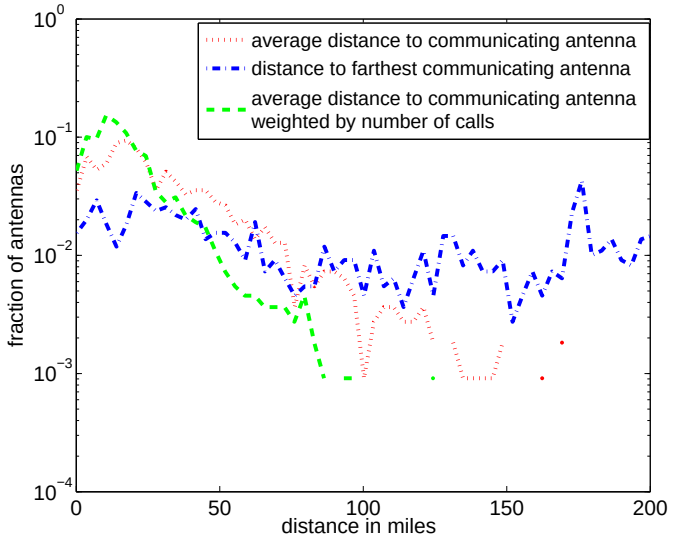


Fig. 4. Distribution of the distance over which antennas communicate

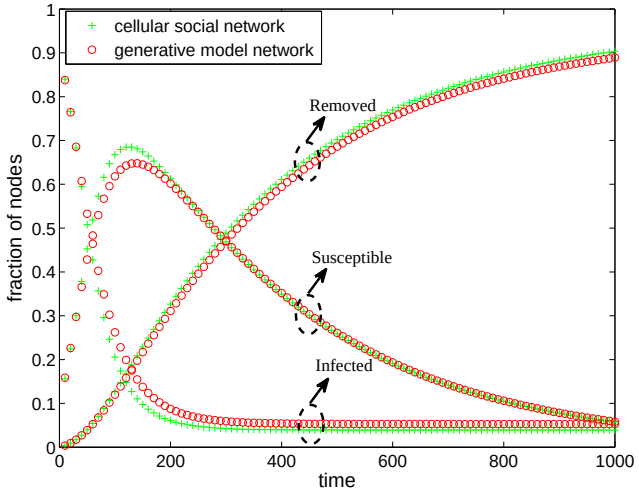


Fig. 5. SIR infection propagation among users of the Asian cellular network

with nodes being either in the removed state or the susceptance state.

A good review of the spread of infection under different network conditions and its mathematical formulation is found in [8] and [9]. These infection propagation models, originally meant for studying the spread of diseases also help in studying how malicious online viruses propagate and news/gossip dissipates in a information network like the cellular network. We simulate the SIR infection model on social network of users from the Asian cellular network and compare them with simulations on a similar sized model developed through our generative model. We show the results in Figure 5. The similarity in the trends of propagation of infection suggests that our model is a good representative of the real cellular social network.

VI. Results and Conclusion

In this abstract, we provide a generative model for social networks among cell-phone users. The models generate a graph representation for the network where new nodes are assumed to appear randomly at different locations within a given space and form links with K nearest neighbors. We demonstrate the efficacy of the models by analyzing their degree distribution, and also the propagation of infection on them. The degree distribution of the network obtained by the generative model has a good fit with the exponential degree distribution observed for real cellular networks. We look at one practical usage of our generative model, namely infection propagation. We test our model for SIR infection propagation and find that the propagation follows the same characteristics as seen in comparable networks. There are multiple uses for cellular usage patterns and the resulting social networks. They can be used to predict human activity, population density and ways to understand epidemic spread. We intend to focus on these issues as a part of our future research in this domain.

References

[1] R. Albert, H. Jeong, and A.-L. Barabasi, "Error and attack tolerance of complex networks", *Nature*, vol. 406, 2000.

[2] P. Erdos and A. Renyi, *On the evolution of random graphs*, Akad. Kiadó, 1960.

[3] L. Garber, "Denial-of-service attacks rip the Internet", *Computer*, vol. 33, no. 4, Apr 2000.

[4] A.-L. Barabasi and R. Albert, "Emergence of scaling in random networks", *Science*, vol. 286, 1999.

[5] D. J. Watts, R. Muhamad, D. Medina, and P. Dodds, "Multiscale,resurgent epidemics in a hierarchical metapopulation model", *Proc. of the National Academies of Science*, 2005.

[6] J. C. Miller and J. M. Hyman, "Effective vaccination strategies for realistic social networks", *Physica A*, vol. 386, 2007.

[7] D. Kempe, J. Kleinberg and E. Tardos, "Maximizing the spread of influence through a social network", *Proceedings of the 9th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, 2003.

[8] M. Boguna, R. Pastor-Satorras and A. Vespignan, "Epidemic spreading in complex networks with degree correlations", *arXiv:cond-mat/0301149v1 [cond-mat.stat-mech]*, 2003.

[9] D. Chakrabarti, Y. Wang, C. Wang, J. Leskovec and C. Faloutsos, "Epidemic Thresholds in Real Networks", *ACM Transactions on Information and System Security (TISSEC), 10(4)*, 2008.

[10] D. Deka and S. Vishwanath, "Network Models for Power Grids: A Generative Approach", *arXiv:1204.0165v1 [cs.SI]*, July 2012.

[11] http://www.d4d.orange.com/home.

[12] V. Blondel, M. Esch, C. Chan, F. Clerot, P. Deville, E. Huens, F. Morlot, Z. Smoreda and C. Ziemlicki, "Data for Development: the D4D Challenge on Mobile Phone Data", *http://arxiv.org/abs/1210.0137v2 [cs.CY]*, January 2013.

Neighborhood structures in socio-demographic and HIV infection conditions
Indication to the potential of mHealth for tackling HIV/AIDS in Ivory Coast

Ayumi Arai[a,*], Teerayut Horanont[b], Apichon Witayangkurn[c], Ryosuke Shibasaki[d]

[a] *Department of Socio-cultural Environmental Studies, University of Tokyo, Chiba 277-8568, Japan*
[b] *Institute of Industrial Science, University of Tokyo, Tokyo 153-8505, Japan*
[c] *Department of Civil Engineering, University of Tokyo, Tokyo 153-8505, Japan*
[d] *Center for Spatial Information Science, University of Tokyo, Chiba 277-8568, Japan*

Abstract

It is vital to know what factors relate HIV infection for effective intervention. In this paper, we employ distance metrics to quantify similarity in structural distribution of HIV infection conditions, socio-demographic attributes, and connectivity to information for all pairs of *departments* in Ivory Coast. Quantifying the similarity, we positively identify that the cluster-based approach, focusing on similar educational background distribution and income levels, would be a possible way of intervention. In addition, our estimation results indicate using the mobile phone for information provision (mHealth) would be an effective way although the results are not statistically significant. Considering possible differences in people's behavioral patterns based on their life-style or socio-demographic attributes, we expect the estimation results may be dramatically improved or even changed if socio-demographic attributes such as the age group were attached to the anonymized mobile phone data for further analyses.

1. Introduction

It has been increasingly discussed that people's activities and decision making processes are affected not only by neighbors whose distance is physically near but by neighbor who are linked through social networks (Topa 2001). Regarding HIV transmission dynamics, enormous efforts have been devoted focusing on socio-demographic homogeneity and heterogeneity between contract population groups (Anderson and May 1988; Anderson *et al*. 1988; Jacquez *et al*. 1988; Travis & Lenhart 1987). For instance, Anderson and May (1985) examine age-related heterogeneity between groups. Rothenberg *et al*. (1998) find social network structure is one of vital factors affecting

* Corresponding author
Email address: arai@csis.u-tokyo.ac.jp (A. Arai), teerayut@iis.u-tokyo.ac.jp (T. Horanont), apichon@iis.u-tokyo.ac.jp (A. Witayangkurn), shiba@csis.u-tokyo.ac.jp (R. Shibasaki)

the dynamics of HIV transmission. Keeling and Eames (2005) also suggest that it is important to understand how much characteristics of networks are mixed to predict epidemic patterns and intervention measures.

It is reported that levels of knowledge on the method to prevent HIV transmission remain fairly low particularly in West Africa including Ivory Coast, although comprehensive knowledge on HIV/AIDS has been increased in sub-Saharan Africa (Mishra 2009). Curtis and Hossain (1998) discuss the effect of settlement zones and economic diversity on the reproductive behavior of the West African women. They find positive correlation between economic growth rates and knowledge of modern methods of family planning. More interestingly, knowledge of modern methods is higher in zones where people's economic status is rather mixed.

Mobile phone is one of the fastest growing new technologies worldwide in information and technology. Studies exploring the potential of mobile phones to tackle health issues have been emerging (Noordam *et al.* 2011; Mechael 2009). The use of mobile phones in health systems and service provision is called mHealth, which is, for example, expected to reduce transaction costs in providing information on maternal health in timely manner in Bangladesh (Tamrat & Kachnowski 2012). In developing world 79% of the population is mobile-cellular subscribers whereas the global penetration reaches 87% (ITU 2011). Thus, there is considerable potential of mHealth for the global health in improving health and achieving equity in health particularly when it can be resolved by addressing lack of information in the developing world.

In this study, we examine relationships between socio-demographic neighborhood structures and HIV infection conditions using socio-demographic data and mobile phone data in Ivory Coast. We measure similarity in structural distribution of HIV infection conditions, socio-demographic attributes, and connectivity to information for all pairs of *departments* using normalized distance metrics.

2. Data

We use two different types of data for this study. One is socio-demographic statistical datasets including HIV blood test results from Demographic Health Survey (DHS) of Ivory Cost in 2004. Specific population is subsampled so that the population is common for both socio-economic survey and HIV blood tests in the DHS framework. The data consists of 5,087 females and 4,405 males, whose ages are between 15 and 49. We geo-referenced DHS data in 2004, which itself is not originally geo-referenced, using geo-referenced DHS data in 1995 and 1999. We link the cluster numbers, which are location based serial numbers commonly used in DHS surveys, by overlaying with GIS

map of Ivory Coast, and then aggregate at the *department* level. 333 individuals (3.5% of the original sample size) are dropped due to constraints in geo-referencing. We assumed that current geographical trends in HIV/AIDS infection rates do not differ markedly from those of 2004.

The other is the record of mobile phone calls during five month (from December 2011 to April 2012) in Ivory Cost provided by Orange. The original dataset contains 2.5 billion records, calls and text messages exchanged between five million anonymous users. We use the antenna-to-antenna data aggregated at the *department* level for our analysis as it represents the connectivity of people within and between *department*s. The number of calls as well as the duration of calls between any pair of antennas has been aggregated hourly basis. Our assumption is that the higher degree of information exchange, the more frequent mobile communication is. We use this information to characterize the social tight and knowledge exchange among people in individual *department*. We first process the data by calculating the average number of calls and the average duration of calls during five months and aggregated them using *department* administrative zone. This process helps to minimize failure interpretation due to the high-density antennas in the city area and to keep the consistency with other socio-demographic statistical datasets.

Movie 1a and Movie 1b[1] show intra-connectivity in Abidjan *department*. All nodes represent mobile antenna locations and the moving points explain the communication between pair of antenna. The links between nodes describe how strong the connectivity bridge between people in different areas. Movie 2[2] is the accumulate call density aggregated in five kilometers grid hourly. This implies the communication activity in each area and indirectly infers the population distribution in more timely, accurate and comparable manner.

3. Methodology

First, we measure intra-*department* structural distribution for six indicators: age group, the level of education attainment, level of income, proportion of HIV/AIDS infected population, level of knowledge on the technology to prevent HIV transmission, and ratio of average number of inter-*department* call to that of intra-*department* call (hereinafter connectivity). All indicators except for the connectivity, which derives from the mobile phone data, are computed gender wise. We use a common indicator for connectivity for male and female because demographic attribute are not attached to the mobile phone data. Then, inter-*department* structural similarity is computed using the

[1]Links to Movie 1a& Movie 1b:Mobile phone communications between antennas in Adbijan department on;
1a: 31st December, 2011 http://www.mobilesensing.org/d4d/movie1a.m4v and
1b: 1st January, 2012 http://www.mobilesensing.org/d4d/movie1b.m4v

[2]Link to Movie 2: Mobile phone communication density of a day on a weekday.
http://www.mobilesensing.org/d4d/movie2.avi

normalized distance metrics developed by Conley and Topa (2001)[3]. Six distance metrics, which are age distance, education distance, income distance, HIV infection condition distance, knowledge distance, and connectivity distance, are constructed for our estimation. The distance metric ranges from zero to 141. For instance, age distance for two *department*s, where each of them is dominated by a different age group, is close to zero or 141. Whether the value is close to zero or 141 depends whether the dominant age group is major or minor among given groups. Age distance between another set of *departments*, in which composition of age groups is fairly mixed, would value around median of the range. Computing the similarity as the distance for all pairs of *department*, six inter-*department* structural similarity matrices are formed.

Second, we visualize the inter-*department* similarity of each indicator by transforming into a set of 58 points in a Euclidian space using a method called Multidimensional Scaling (MDS). In the MDS space, inter-*department* similarity among 58 *departments* is shown as relative locations of 58 points. Besides visualizing the inter-*department* similarity, we visualize the similarity by clustering using Kernel density analysis. A continuous density surface is generated over all locations to let data pattern show itself. The density is a measurement of magnitude of similarity of features in the clusters.

Third, the following regression models are employed to estimate how inter-*department* structural similarity of HIV infection conditions correlate with that of socio-demographic attributes and connectivity by gender.

$$y_1 = \beta_{1,1i} x_{1,1i} + \varepsilon_1 \quad \text{... (1)}$$

$$y_2 = \beta_{1,2i} x_{1,2i} + \beta_{2,2} x_{2,2} + \varepsilon_2 \quad \text{... (2)}$$

$$y_3 = \beta_{1,3i} x_{1,3i} + \beta_{2,3} x_{2,3} + \beta_{3,3i} x_{1,3i}^2 + \varepsilon_3 \quad \text{... (3)}$$

Where y_k is the HIV distance vector, $x_{1,ki}$ is age distance vector where i=1, the education distance vector where i=2, income distance vector where i=3, knowledge distance where i=4, HIV distance where i=5, connectivity distance where i=6, $x_{2,k}$ is

[3] For instance, age distance metric, AD_{ij} is the Euclidian distance between the vector e_i of percentage of six age groups present in *department*$_i$ and the corresponding vector e_j in *department*$_i$ where n is number of variety of group or categories:

$$AD_{ij} = \sqrt{\sum_{k=1}^{n} e_{ik} - e_{jk}^2}$$

In our study, education attainment levels are classified into four groups: no education, primary, secondary, and higher. Wealth levels are classified into five groups: poor, poorer, middle, richer, rich. We follow the classification based on the DHS questionnaire. As the connectivity distance, we use percentage of inter-department calls and intra-department calls to average number of calls made in each department per day. HIV infection distance is calculated using percentage of HIV positive and negative population. Further detailed description of distance metrics and MDS is explained in Conley and Topa (2002).

Euclidian distance vector, ε_k is error term (*k=1, 2, and 3*).

In our estimation, we assume *x-vectors* from the centroid of the MDS space as the indicator of relative distance to measure relative location. Given time constraints, we use the *x-vector* instead of creating an indicator, capturing exact locations in the two-dimensional space. Though our estimation method is rough, results are still indicative.

4. Results

MDS locations for six metrics are shown in Fig. 1A to Fig. 1F. Fig. 1E(a) depicts relative locations of HIV distance of male. This configuration captures variations in HIV metric, having a goodness of fit of 99%. *Departments* where the proportion of HIV positive population ratio is high, such as *department* ID 5 and 33 are located in the right part of the figure. On the other hand, *departments* where proportion of HIV positive population is close to zero, such as *department* ID 46 and 37, concentrate in the left part of the figure. Incidentally, the area where several points overlap in the left part is the concentration of *department*s whose HIV positive ratio is zero.

Beside the distribution of relative locations, structural similarity is visually clustered in Fig. 2A to 2F. For instance, we can observe one distinct difference in clustering HIV distance metrics between male and female, Fig. 2E(a) and Fig. 2E(b). The clustering of the male is essentially dominated by one cluster whereas that of female has an additional cluster which locates upper side of the largest cluster.

Then, we run regression to estimate how the neighborhood structure of HIV infection conditions correlates with that of the age group, education level, income level, and knowledge level. Table 1 and Table 2 describe estimation results for male and female. For case (A) and (B) in Table 1 and Table 2, the results show the similar trend regarding neighborhood structure of the age group, education level, and income level for both male and female. Values of *t*-statistics are the largest for model (1) in all cases. The larger proportion of highly educated population, the smaller proportion of HIV infected populations. The higher ratio of the richer, the lower ratio of the HIV infected. However, for case (C), the neighborhood structure of the knowledge level, the results differ between male and female. Although both results are not statistically significant, we can observe simple linear trend for male and convex quadratic trend for female. That is, the more predominant population with correct knowledge on technology to avoid transmission of HIV, the lower HIV infection rates for male. More interestingly, for female, the more mixed population in knowledge level, the lower the HIV infection rates whereas the overall trend of female follows that of male. When the

socio-demographic structure matters to trends in HIV infection, which is affected by knowledge level, we could explore the potential of intervention through mHealth here.
Lastly, we also estimate how the neighborhood structure of the connectivity and some socio-demographic attributes correlate. The estimation results are shown in Table 3. Contrary to our expectations, the overall results are not statistically significant but the values of *t*-statistics for the age group and education levels are relatively larger. Value of *t*-statistics is relatively larger only for male. We consider one of most probable causes of such insignificant results is aggregation of mobile phone data where all individuals with various attributes are mixed. Because of the relatively larger value of *t*-statistics regarding the age group and education level, we expect the estimation results for the connectivity may be improved by attaching such attributes to the mobile phone record, which enables us to split the aggregated data into several groups.

5. Conclusion

Our estimation results suggest cluster-based approach, focusing on the education and income level structures, would be possible intervention for tackling HIV. In addition, the results have indication that the mobile phone could be one of tools for information provision to prevent HIV transmission. Considering possible differences in people's behavioral patterns based on their life-style or socio-demographic attributes, we expect the estimation results may be dramatically improved or even changed if socio-demographic attributes such as the age group were attached to the anonymized mobile phone data for further analyses. Further development of the method to estimate the relation between two sets of two-dimensional spatial distribution is necessary.

References:

Anderson, R. M., Medley, G. F., May, R. M., & Johnson, A. M. (1986). A preliminary study of the transmission dynamics of the human immunodeficiency virus (HIV), the causative agent of AIDS. *Mathematical Medicine and Biology*, 3(4), 229-263.

Andersen, R. M., & May, R. M. (1988). Epidemiological parameters of HIV transmission. *Nature*, 333(6173), 514-519.

Conley, T. G., & Topa, G. (2002). Socio-economic distance and spatial patterns in unemployment. *Journal of Applied Econometrics*, 17(4), 303-327.

Curtis, S. L., & Hossain, M. (1998). The effect of aridity zone on child nutritional status, West Africa spatial analysis prototype exploratory analysis. Calverton, Maryland: Macro International Inc.

ITU. 2011. ICT Facts and Figures. Retrived from http://www.itu.int/ITU-D/ict/facts/2011/material/ICTFactsFigures2011.pdf

Jacquez, J. A., Simon, C. P., Koopman, J., Sattenspiel, L., & Perry, T. (1988). Modeling and analyzing HIV transmission: the effect of contact patterns. *Mathematical Biosciences*, 92(2), 119-199.

Keeling, M. J., & Eames, K. T. (2005). Networks and epidemic models. *Journal of the Royal Society Interface*, 2(4), 295-307.

Rothenberg, R. B., Potterat, J. J., Woodhouse, D. E., Muth, S. Q., Darrow, W. W., & Klovdahl, A. S. (1998). Social network dynamics and HIV transmission. *Aids*, 12(12), 1529-1536.

Mechael, P. N. (2009). The case for mHealth in developing countries. *Innovations: Technology, Governance, Globalization*, *4*(1), 103-118.

Mishra, V. K. (2009). Changes in HIV-related knowledge and behaviors in sub-Saharan Africa. ICF Macro.

Noordam, A. C., Kuepper, B. M., Stekelenburg, J., & Milen, A. 2011. Improvement of maternal health services through the use of mobile phones. Tropical Medicine & International Health, 16(5), 622-626.

Topa, G. (2001). Social interactions, local spillovers and unemployment. *The Review of Economic Studies*, 68(2), 261-295.

Tamrat, T., & Kachnowski, S. (2012). Special delivery: an analysis of mHealth in maternal and newborn health programs and their outcomes around the world. *Maternal and child health journal*, 1-10.

Fig. 1A. MDS location of the age for (a) male and (b) female

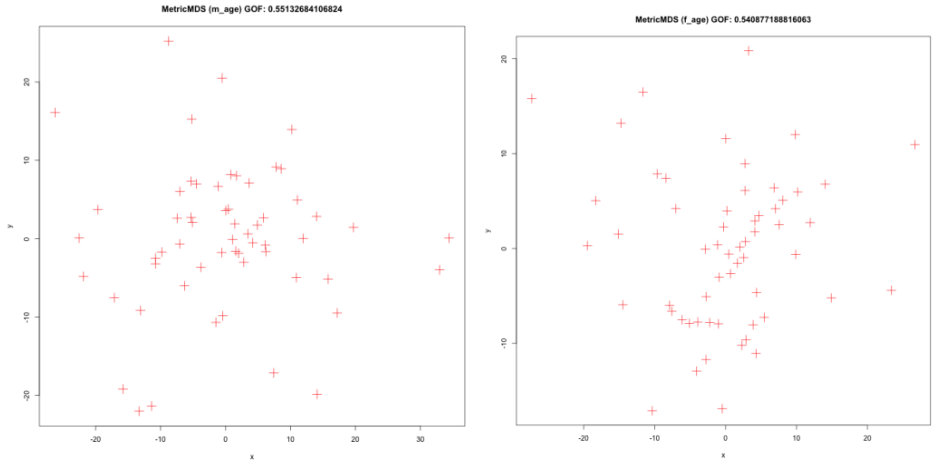


Fig. 1B. MDS location of the education attainment for (a) male and (b) female

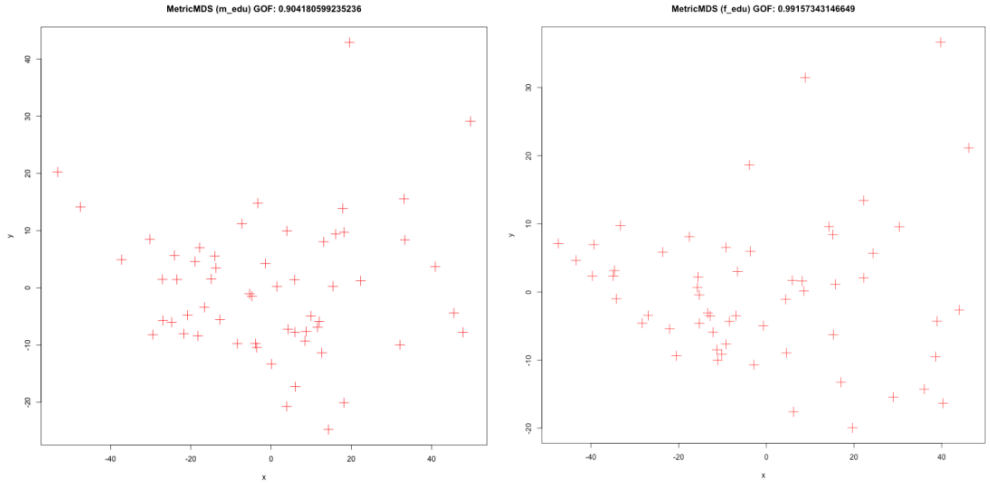


Fig. 1C. MDS location of the income level for (a) male and (b) female

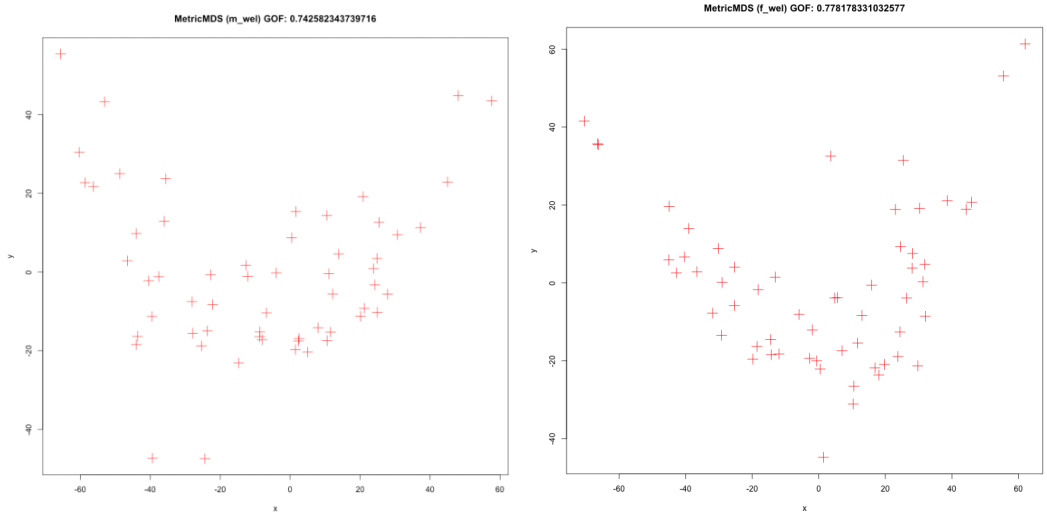


Fig. 1D. MDS location of the knowledge level on the technology to prevent transmission of HIV for (a) male and (b) female]

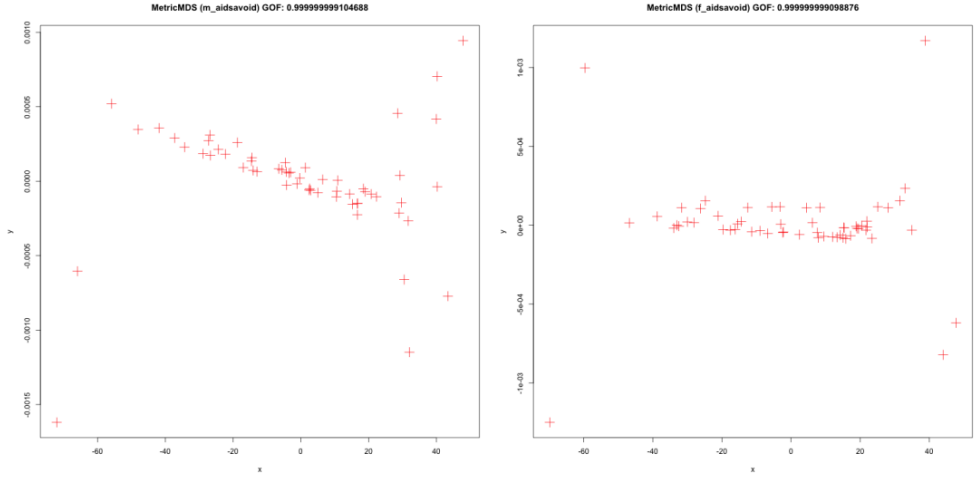


Fig. 1E. MDS location of proportion of HIV infection population for (a) male and (b) female

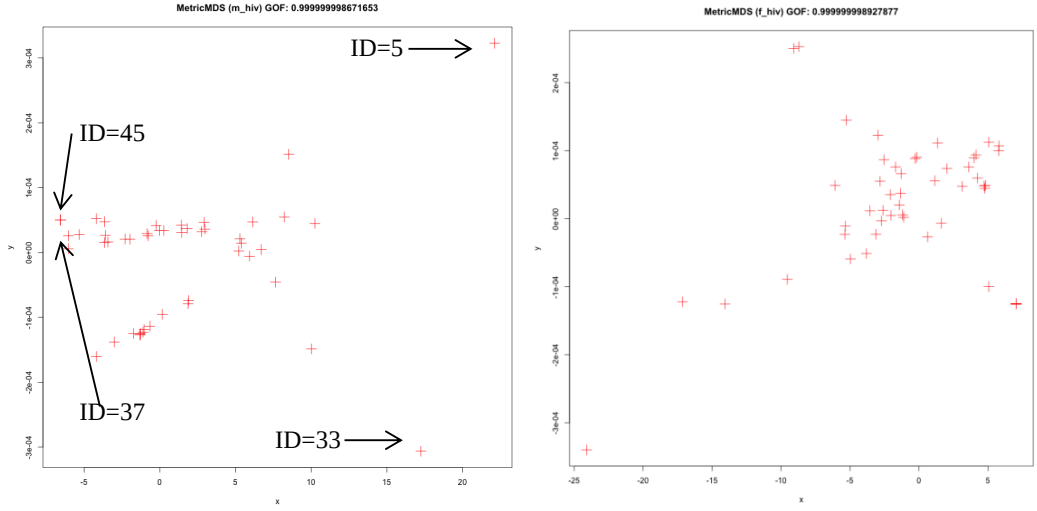


Fig. 1F. MDS location of connectivity

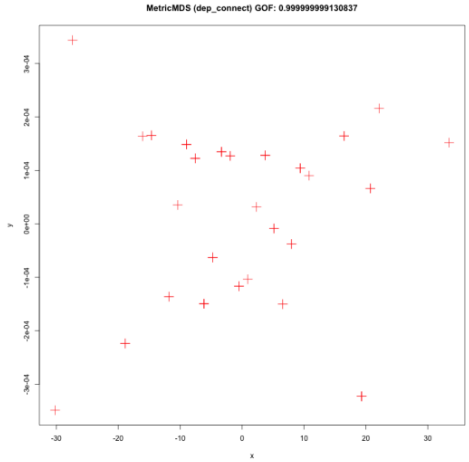


Fig. 2A. Kernel density clustering for the age group for (a) male and (b) female

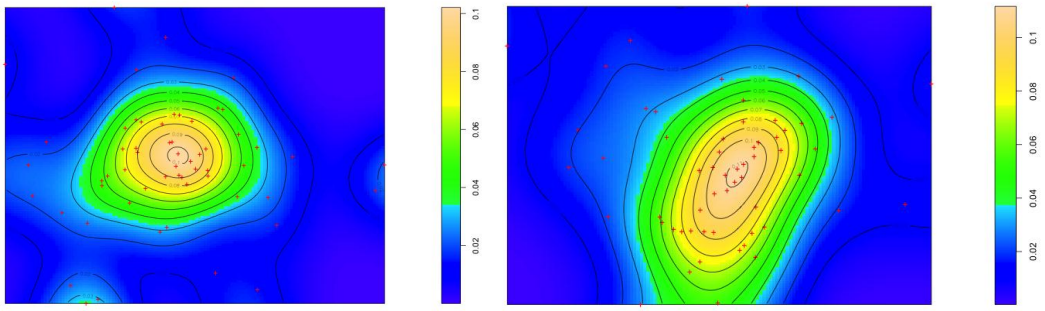


Fig. 2B. Kernel density clustering for the education attainment for (a) male and (b) female

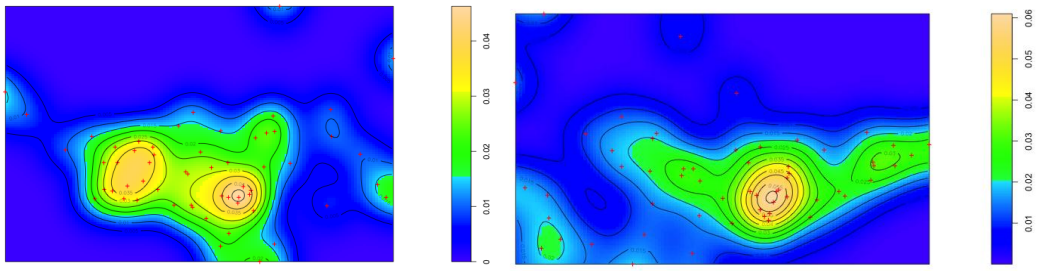


Fig. 2C. Kernel density clustering for the income level for (a) male and (b) female

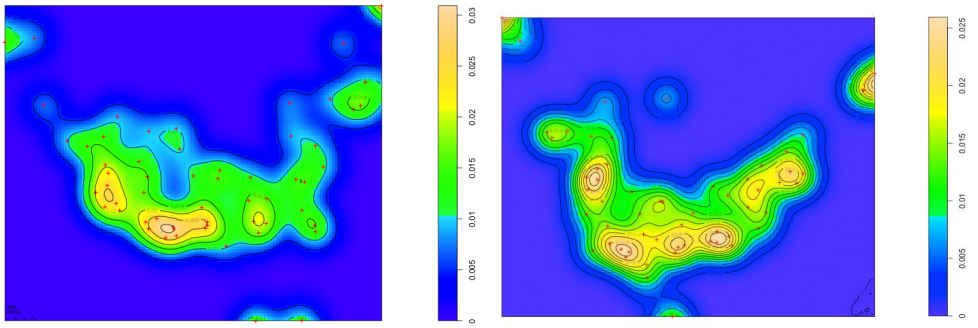


Fig. 2D. Kernel density clustering for the knowledge level on the technology to prevent transmission of HIV for (a) male and (b) female

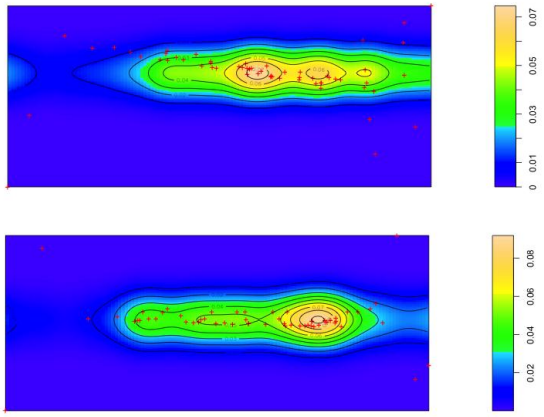


Fig. 2E. Kernel density clustering for HIV infection population for (a) male and (b) female

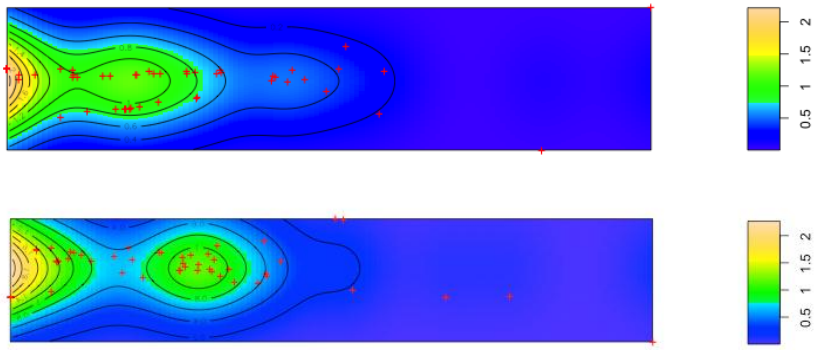


Fig. 2F. Kernel density clustering for connectivity

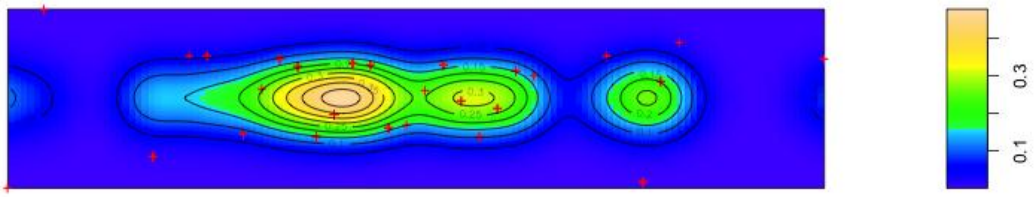


Table 1. Male results for the estimation of correlation in neighborhood structure between HIV infection conditions and the (A) Age group, (B) Education level, C) Income level, and (D) Knowledge level

	(A) HIV infection conditions			(B) HIV infection conditions			(C) HIV infection conditions			(D) HIV infection conditions		
	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
Age	0.129*	0.128*	0.108									
	(1.95)	(1.92)	(1.60)									
Age^2			0.00468									
			(1.37)									
Education				0.129*	0.129*	0.128*						
				(1.95)	(1.95)	(1.92)						
$Education^2$						0.00468						
						(1.37)						
Wealth							-0.0509*	-0.0520*	-0.0510*			
							(-1.96)	(-1.98)	(-1.98)			
$Wealth^2$									-0.00106			
									(-1.30)			
Knowledge										0.0167	0.0177	0.0129
										(0.56)	(0.59)	(0.40)
$Knowledge^2$												0.000272
												(0.32)
Euclidian distance		0.155			-0.155			-0.278			0.217	
		(0.29)			(-0.29)			(-0.52)			-0.39)	
Constant	3.38e-15	5.35e-11	0.670	-3.38e-15	-5.35e-11	-0.670	-3.86e-15	-9.59e-11	0.985	2.87e-11	1.05e-10	0.205
	(0.00)	(0.00)	(0.73)	(-0.00)	(-0.00)	(-0.73)	(-0.00)	(-0.00)	(0.90)	(0.00)	(0.00)	(0.20)
N	58	58	58	58	58	58	58	58	58	58	58	58
adj. R-sq	0.047	0.031	0.062	0.047	0.031	0.062	0.047	0.035	0.059	0.012	0.028	0.029

Note: t statistics in parentheses. * p<0.05, ** p<0.01, *** p<0.001. The large value of the 'Education' means highly educated population dominant in the *department*. The large value of the 'Wealth' means wealthier population dominant in the *department*. The large value of the 'Knowledge' indicates a large proportion of people in the *department* have correct knowledge on the technology to avoid HIV/AIDS transmission.

Table 2. Female results for the estimation of correlation in neighborhood structure between HIV infection conditions and the (A) Education level, (B) Income level, and (C) Knowledge level

	(A) HIV infection conditions			(B) HIV infection conditions			(C) HIV infection conditions			(D) HIV infection conditions		
	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
Age	0.0183	0.0189	0.0107									
	(0.20)	(0.21)	(0.12)									
Age^2			0.00868									
			(1.62)									
Education				0.0995***	0.0994***	0.0998***						
				(3.01)	(2.98)	(2.96)						
$Education^2$						0.0000634						
						(0.05)						
Wealth							-0.0711***	-0.0710***	-0.0675**			
							(-2.70)	(-2.67)	(-2.44)			
$Wealth^2$									0.000347			
									(0.47)			
Knowledge										0.0434	0.0433	0.0581
										(1.30)	(1.29)	(1.63)
$Knowledge^2$												0.00118
												(1.13)
Euclidian distance		0.0971			-0.0334			-0.00998			-0.0622	
		(0.17)			(-0.06)			(-0.02)			(-0.11)	
Constant	6.34e-16	3.35e-11	0.801	-7.96e-16	-1.15e-11	-0.0374	-7.48e-16	-3.44e-12	-0.331	-2.25e-09	-2.26e-09	-0.767
	(0.00)	(-0.00)	(-0.81)	(-0.00)	(-0.00)	(-0.03)	(-0.00)	(-0.00)	(-0.31)	(-0.00)	(-0.00)	(-0.70)
N	58	58	58	58	58	58	58	58	58	58	58	58
adj. R-sq	0.017	0.035	0.012	0.124	0.108	0.108	0.099	0.083	0.087	0.012	-0.006	0.017

Note: t statistics in parentheses. * p<0.05, ** p<0.01, *** p<0.001. The large value of the 'Education' means highly educated population dominant in the *department*. The large value of the 'Wealth' means wealthier population dominant in the *department*. The large value of the 'Knowledge' indicates a large proportion of people in the *department* have correct knowledge on the technology to avoid HIV/AIDS transmission.

Table 3. Relationships between structural similarity in connectivity and socio-demographic attributes

	Connectivity (Explanatory variables: Male's)				Connectivity (Explanatory variables: Female's)			
Age	0.00000259				0.00000336			
	(1.44)				(1.51)			
Education level		0.00000117				0.000000988		
		(1.23)				(1.11)		
Income level			0.000000324				0.000000212	
			(0.45)				(-0.30)	
Knowledge level				0.00000116				0.000000270
				(1.48)				(0.32)
Constant	6.16e-12	6.16e-12	6.16e-12	6.16e-12	6.16e-12	6.16e-12	6.16e-12	6.14e-12
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
N	58	58	58	58	58	58	58	58
adj. R-sq	0.019	0.019	0.014	0.021	0.022	0.004	0.016	0.016

Note: t statistics in parentheses. * p<0.05, ** p<0.01, *** p<0.001. The large value of the 'Connectivity' means the ratio of inter-*department* communication is large than that of intra-*department* communication.

Disease Outbreak Detection by Mobile Network Monitoring: a case study with the D4D datasets

Nicola Baldo, Pau Closas
Centre Tecnològic de Telecomunicacions de Catalunya (CTTC)
Av. Carl Friedrich Gauss 7, Castelldefels (Barcelona), Spain
{nbaldo, pclosas}@cttc.es

***Abstract*— We consider the possibility of realizing a cost-effective disease outbreak surveillance system based on mobile network monitoring. We perform a case study for Ivory Coast by investigating the correlation between the mobile network data made available for the D4D challenge and the virological data provided by FluNet. Though no definite conclusion is reached about the feasibility of such disease outbreak system, this study identifies some key requirements for the data that is needed to be able to progress in this particular line of research.**

I. Introduction

Seasonal infectious diseases like influenza are responsible of significant morbidity and mortality every year. The World Health Organization (WHO) estimates that annual influenza epidemics result in about 3 to 5 million cases of severe illness and about 250,000 to 500,000 casualties worldwide. Surveillance systems that timely detect disease outbreaks can play a key role in limiting the spread of the disease and its impact.

In developed countries, surveillance system like sentinel networks covering small portion of the population and, more recently, electronic medical records are successfully employed. However, these systems require very costly infrastructures. Developing countries such as Ivory Coast cannot deploy these systems because of their cost. On the other hand, if it were possible to build a disease outbreak surveillance system that uses existing non-medical monitoring infrastructures (such as a mobile network infrastructure), its cost would be lower and a developing country could afford it.

In this paper we discuss the design of a system for the automated detection of epidemics of infectious diseases based on the analysis of mobile network traffic and mobility patterns. The objective is to build an autonomous detection algorithm that is able to determine whether certain seasonal infectious diseases are active or not based on the measurement data which can be provided by a mobile network operator. We present a case study carried out with the mobile network datasets provided by Orange for the D4D Challenge [1], discussing how we processed the mobile network data, and what other data sources we considered for the identification of disease outbreaks. Finally, we describe the results that we obtained, drawing our considerations.

II. Case study based on the D4D and FluNet data for Ivory Coast

We now present a study that we performed for the realization of the disease outbreak detection system based on the of the mobile network datasets provided by Orange for Ivory Coast as part of the D4D Challenge [1]. The D4D datasets cover a 5 month period going from December 2011 to April 2012; a detailed description of these datasets can be found in [2].

A. Feature extraction

To design and train the disease outbreak detection system, we considered the virological data provided by FluNet [3] for Ivory Coast over the same period. FluNet gathers data by the National Influenza Centres (NICs) of the Global Influenza Surveillance and Response System (GISRS) as well as from other national influenza reference laboratories and from the WHO regional databases. In practice, the FluNet data for Ivory Coast consists of weekly counts of the number of specimens for which certain types of influenza viruses were detected. For our case study, we aimed at general influenza outbreak detection, hence we focused on the total number of influenza viruses detected; this data is summarized in Figure 1. No additional feature extraction was performed on this data, as we considered it to be already in a sufficiently concise and representative format.

As the first step in our case study, we processed the D4D data with the aim of extracting the features to be correlated with the FluNet data. To this aim, we focused on the mobile network data relative to those location within a certain range from known hospitals in Ivory Coast. The reason for this is that we expected an influenza outbreak to increase the presence of people in hospitals, which in turn we expected to increase the likelihood of both mobile phones being camped on a cell nearby an hospital and of a mobile call being performed to or from such cells. We considered the hospitals listed in Table II-A, which are all hospital among those we could find from an internet search for which we could find sufficiently detailed position information. Latitude and longitude are expressed in degrees.

Among the D4D datasets [2], we considered only the sets *Antenna-to-antenna* (SET1) and *Individual Trajectories: High Spatial Resolution Data* (SET2). We did not considered neither *Individual Trajectories: Long Term Data* (SET3), because

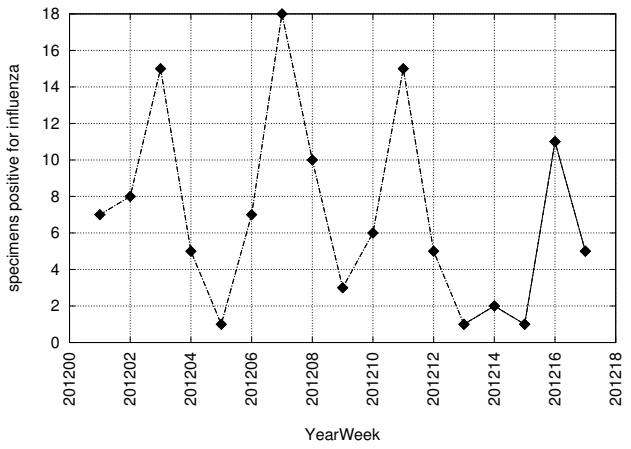


Fig. 1. FluNet data for Ivory Coast available for the time period covered by the D4D datasets.

TABLE I

LIST OF CONSIDERED HOSPITALS WITH THEIR POSITION

Hospital ID	Hospital name	Latitude	Longitude
1	Port Bouet	5.2637	-3.9621
2	L'Indenie	5.33849	-4,01931
3	Sainte Anne-Marie	5.33676	-4.01324

of its coarse spatial resolution that did not allow to single out users close to an hospital with sufficient precision, nor *Communication Subgraphs* (SET4), because we did not expect to see a strong correlation of this data with influenza outbreaks.

We performed feature extraction on the D4D datasets by the following operations. For SET1, we applied some preprocessing in order to aggregate the data by week, thus obtaining a record of the total number of calls and sum of their duration for every antenna pair in each week during the observation period. For SET2, we also applied some preprocessing, aggregating the data first by week and user identifier, and then aggregating the resulting data by week only; in this way we obtained a record of the number of users that were reported to be camped on each antenna on each week, together with the total number of distinct report events recorded for that antenna and week. Finally, for both SET1 and SET2, we selected the data corresponding to those antennas within a distance d from each of the hospitals of Table II-A. To this aim, we considered the antenna positions provided in the data set ANT_POS [2], and calculated their Euclidean distance to each of the hospitals after conversion from geographical coordinates to cartesian coordinates [4]. The resulting data for the particular case of $d = 1$ km is shown in Figures 2, 3 and 4. We note that D4D data relative to December 2011 was not considered due to the lack of FluNet data for Ivory Coast at that time.

B. Results

In order to evaluate the feasibility of the proposed system, we first evaluated the sample Pearson correlation r [5] between

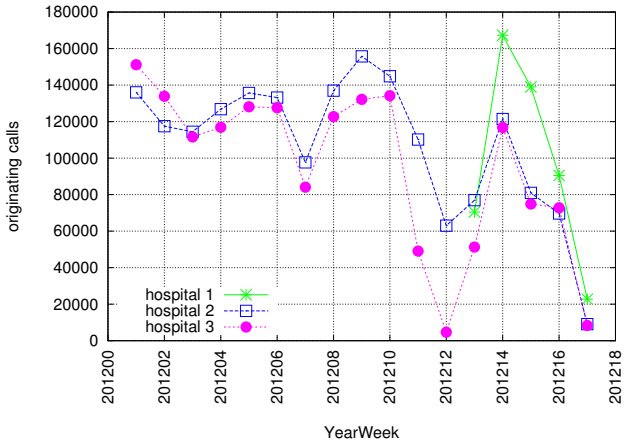


Fig. 2. Originated calls by week after feature extraction from SET1 for $d = 1$ km.

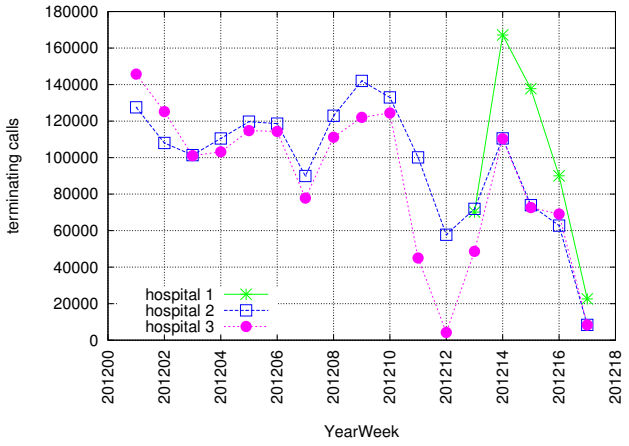


Fig. 3. Terminated calls by week after feature extraction from SET1 for $d = 1$ km.

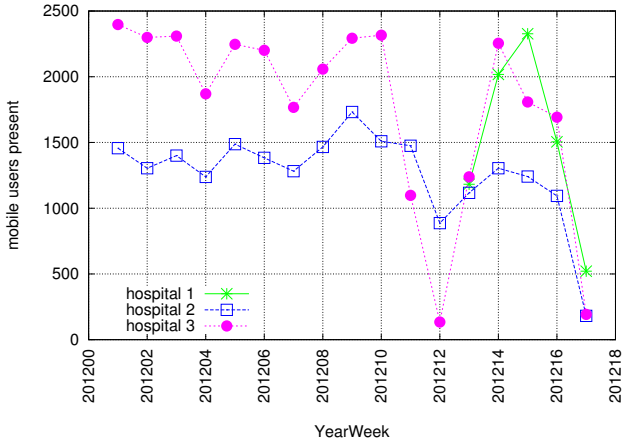


Fig. 4. Number of mobile users present after feature extraction from SET2 for $d = 1$ km.

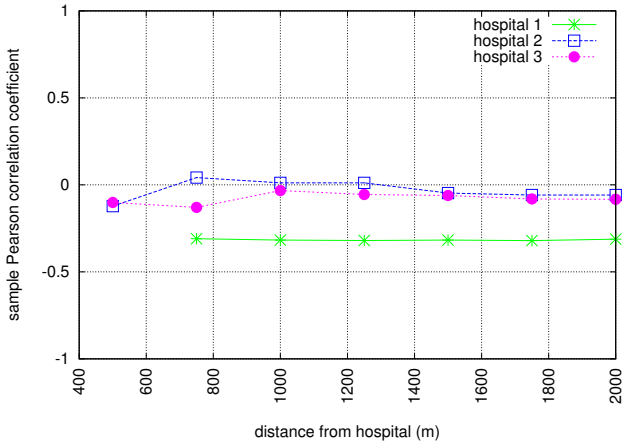


Fig. 5. Correlation between the number of mobile users present and the number of influenza cases detected as a function of the radius of the area around the hospital.

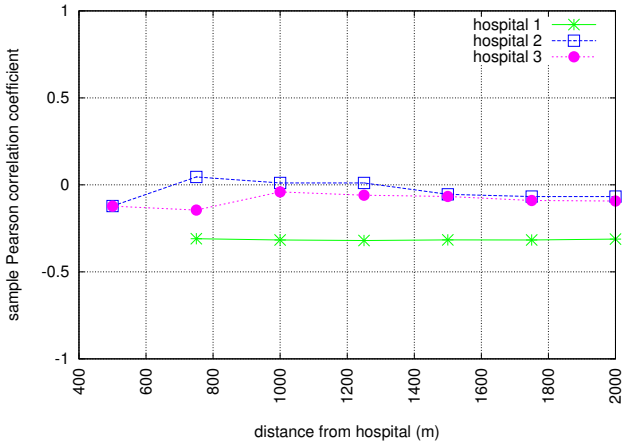


Fig. 6. Correlation between the number of terminated calls and the number of influenza cases detected as a function of the radius of the area around the hospital.

the FluNet data and the D4D data. Let $i = 1, \ldots, N$ be the index identifying the year and week in the data considered, let X_i denote the number of specimens resulted positive for influenza according to the FluNet data relative to week i, and let Y_i denote the value relative to week i of the specific variable obtained from the D4D dataset for which we are calculating the correlation. We first calculate the sample means $\overline{X}, \overline{Y}$ and sample standard deviations s_X, s_Y as

$$\begin{aligned} \overline{X} &= \frac{1}{N}\sum_{i=1}^{N} X_i \\ \overline{Y} &= \frac{1}{N}\sum_{i=1}^{N} Y_i \\ s_X &= \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \overline{X})^2} \\ s_Y &= \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \overline{Y})^2} \end{aligned} \tag{1}$$

(2)

The sample Pearson correlation is then calculated as

$$r = \frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{X_i - \overline{X}}{s_X}\right)\left(\frac{Y_i - \overline{Y}}{s_Y}\right) \tag{3}$$

We calculated r for different sets $\{Y_i\}$ corresponding to the number of originating calls, the number of terminated calls and the number of mobile users present relative to an area within a distance d from each hospital. The obtained values of the correlation r are reported in Figures 5, 6 and 7 as a function of the distance d. Unfortunately, as evident from the figures, the correlation resulted to be very small, thus suggesting that the data sets we were dealing with could actually be incorrelated.

This fact prompted us to try to perform some validation of the data to check whether it is meaningful for the study

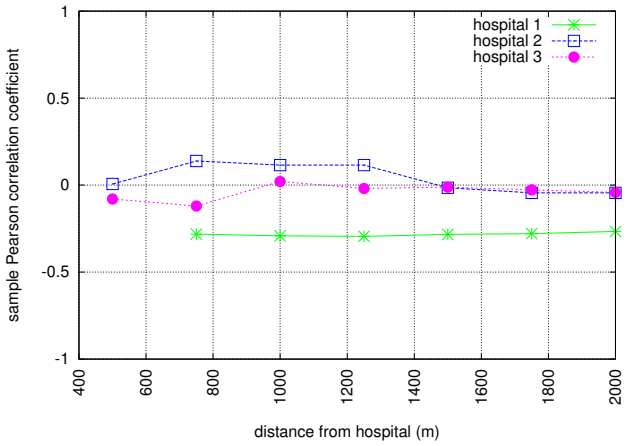


Fig. 7. Correlation between the number of originated calls and the number of influenza cases detected as a function of the radius of the area around the hospital.

that we are considering. To this aim, we evaluated the FluNet data available for other countries; as an example, in Figure 8 we show the data relatively to the years 2011 and 2012 for China, Italy, Spain, and the USA; the data of each country is normalized by the maximum value observed for that country, for an easier visualization. Looking at this data, some influenza outbreaks that last several weeks are evident for each country. Comparing this with Figure 1, it is evident that data that the FluNet data available for Ivory Coast for the period covered by the D4D datasets is not good: in fact, no influenza outbreak is evident in Figure 1, and the fluctuations in the number of positive specimens appears to be random. Additionally, the period of the available data (determined by the intersection between the D4D datasets and the FluNet datasets) is too short considering the typical duration of influenza outbreaks.

We note that there is good quality FluNet data available for many countries other than Ivory Coast; hence, if mobile network datasets were available for some of those countries, an

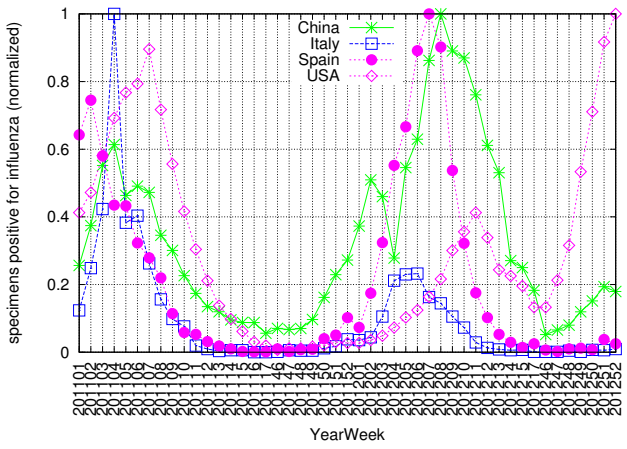


Fig. 8. FluNet data for some selected countries available for 2011 and 2012. The values reported have been normalized relatively to the maximum value seen for each country.

appropriate study on the correlation between mobile network data and seasonal disease outbreaks could be performed.

III. Conclusions

The case study presented in this paper does not allow to reach a definite conclusion on whether it is possible and convenient to realize a disease outbreak detection system based on mobile network monitoring. In particular, in spite of the good detail of the D4D datasets, the poor quality of the FluNet data for Ivory Coast prevented us to reach a meaningful conclusion. Therefore, the main conclusion that we draw from this study is that, in order to be able to evaluate appropriately the feasibility of disease outbreak detection systems, it is of primary importance to have good data available regarding both the mobile network and the disease outbreaks. In this respect, it would be very interesting to be able to access mobile network data for countries for which detailed seasonal disease data is available.

IV. Acknowledgements

N.B. was partially supported by the Spanish Ministry of Science and Innovation under grant number TEC2011-29700-C02-01 (project SYMBIOSIS), by the Catalan Regional Government under grant 2009SGR-940, by COST Action IC0902 “Cognitive Radio and Networking for Cooperative Coexistence of Heterogeneous Wireless Networks” and by the FP7-ICT Network of Excellence ACROPOLIS “Advanced coexistence technologies for radio optimization in licensed and unlicensed spectrum”.

P.C. was partially supported by the Spanish Ministry of Economy and Competitiveness project TEC2012-39143 (SOS-RAD), by the European Commission in the COST Action IC0803 (RFCSET), and the FP7 Network of Excellence in Wireless COMmunications NEWCOM# (contract n. 318306).

References

[1] “D4D Challenge.” [Online]. Available: http://www.d4d.orange.com/

[2] V. D. Blondel, M. Esch, C. Chan, F. Clérot, P. Deville, E. Huens, F. Morlot, Z. Smoreda, and C. Ziemlicki, “Data for Development: the D4D Challenge on Mobile Phone Data,” *CoRR*, vol. abs/1210.0137, 2012.

[3] “FluNet.” [Online]. Available: www.who.int/flunet

[4] G. Strang and K. Borre, *Linear Algebra, Geodesy, and GPS*. Wellesley–Cambridge Press, 1997.

[5] J. L. Rodgers and W. A. Nicewander, “Thirteen ways to look at the correlation coefficient,” *The American Statistician*, vol. 42, no. 1, pp. 59–66, 1988.

Applying Mobile Datasets in Computational Public Health Research

Jonathan P. Leidig*, Yuka Kutsumi*, Kurt A. O'Hearn*, Christine M. Sauer†, Jerry Scripps*, Greg Wolffe*
*School of Computing and Information Systems, Grand Valley State University
†Department of Economics, Grand Valley State University

***Abstract*—The developing world, particularly in Africa, has embraced mobile phone technology. The growth in the percentage of the population with mobile phones has been accompanied by research into mobile applications and other technical systems that have the potential to improve public health responses to disease outbreaks and vaccination campaigns. Computational models can be utilized by public health officials in identifying individuals for preventative measures. These models employ a two-tier approach: coarse-grained, countrywide probabilistic modeling that predicts contagion diffusion followed by fine-grained, community-based telecommunication data analysis for selecting individuals. Using this information, public health workers may identify targeted individuals for disease mitigation strategies such as vaccinating or sending via SMS real-time health updates on outbreaks of contagious diseases.**

I. Introduction

Cote d'Ivoire has a population of almost 22 million with over 17 million mobile phone users [12]. Given the high connectivity rate and the relatively young age of Ivoirians, using mobile phone data and creating computational models and algorithms to characterize certain individuals in communities has the potential to address development issues facing Cote d'Ivoire. In particular, the public health situation in Cote d'Ivoire is less than desirable. The physicians density is 0.14 doctors per 1,000 people, and the degree of risk for infectious diseases is classified as "very high" [12]. February 2012 saw a meningitis outbreak in the northern region of the country in which 40 people were infected and 11 people died [16]. In addition, there have been recent outbreaks of contagious and vector-borne disease such as cholera, measles, Typhoid fever, HIV/AIDS, schistosomiasis, and yellow fever along with many notable cases of malaria. The success of mobile health (mhealth) programs in African countries, such as Uganda and Tanzania [27], demonstrates that the intersection of communications technology and health can be exploited to improve the well-being of rural populations. As an example, residents in Kouto and Tengrela, where the 2012 meningitis outbreak affected the most people, could have been notified and tracked with their mobile phones in order to improve health outcomes.

In addition to examining the use of mobile technology to advance health initiatives, many studies have also focused on the positive effects of mobile phone usage on economic efficiency in developing countries. In neighboring Ghana, micro-scale enterprises (MSEs) have embraced the use of mobile telephony, reporting increased business efficiency and higher profit margins as a result of constant contact with business partners and customers and the reduction of transportation costs due to faster communication [10]. This is a significant development as MSEs comprise a substantial part of national economies. In some cases, the employment level in the MSE sector is twice as high as that in larger enterprises and the public sector. The economic well-being of the society directly impacts the availability and affordability of medical preventative, diagnostic, and treatment services [22]. Mobile phones can also drastically cut search costs and give business owners easier access to information, which may result in a disruption of monopolistic power and standardization of prices. Mobile technology can reduce risk by allowing individuals to respond quickly to economic shocks and epidemics [1].

Mobile phone utilization to improve an individual's or a family's socioeconomic standing has been observed in Cote d'Ivoire. Consistent with the social nature of Ivorian society, one phone is often shared among many people. Operators of mobile phone "booths", where phone cards may be sold, allow others to use their phones for a fee [17]. Individuals have also seen many of the benefits discussed. Mariama, a hairdresser in Abidjan, has successfully developed her business by relying on her mobile phone to contact, schedule, and organize appointments [17]. Mobile phone utilization in Ivorian culture allows for researchers, policymakers, and public health workers to analyze mobile datasets and design mobile applications to address pressing public health problems.

Countrywide mobile datasets, at the prefecture level, may be used to identify communities vulnerable to infectious disease outbreaks. Data for Development (D4D) content provided by Orange, a telecom operating in Cote d'Ivoire, assists in the identification of and search for important individuals to target for public health interventions.

The paper is organized as follows. Section 2 discusses the public health context of the project, Section 3 gives an overview of the use of modeling and simulation in the public health arena, and Section 4 discusses the techniques employed in identifying important individuals in the communication network. Section 5 discusses the findings and avenues for future research.

II. Public Health Context

Infectious diseases and their social and economic costs have a significant negative impact on the economy, health, and well being of developing countries. The cost of recovering from

and succumbing to infectious diseases often places the highest burden on poor and disadvantaged citizens. These individuals are unable to acquire sufficient preventative, diagnostic, and treatment services. The effects of an infection are generally more severe in high-risk groups, e.g., those immunocompromised, pregnant, HIV-positive, and elderly. Public health programs and interventions play a large role in preventing the emergence of an epidemic and allowing for future eradication of specific diseases.

The success of preventing and mitigating an epidemic is dependent on public health policies and personal recommendations made to the public. To identify and classify an emerging outbreak early enough to prevent an epidemic is a difficult task relying on appropriate surveillance and reporting policies. The local and national governments' response to an emerging outbreak are ideally guided by research on the predicted impact of a given set of policies and the associated economic costs. Previous scientific efforts have concentrated on modeling populations in developing countries [28], developing models of disease spread and bodily effects [13], [14], [21], and producing large-scale, computationally-intensive simulation applications [4], [7]. These applications are then used to run thousands of predictive simulations to determine the likely spread of a specific disease based on the population, governmental interventions, and actions of individual persons. With the results and analyses of these simulations, governmental agencies are then able to set public policies to best mitigate epidemics and optimally allocate limited resources.

The availability of mobile datasets allows for more accurate disease mitigation planning in developing countries. High-resolution datasets including individual social networks, numbers of communities a person contacts, and travel patterns have the ability to improve public and private actions. D4D datasets 3 and 4 lead to more accurate population models, population activities and movement, classification of individuals (e.g., student, long-traveller, elderly, city-dweller), healthcare provider and clinician capacities, distribution delays, and realistic potential responses. The datasets provide the ability to advance the current research by enabling individual-specific interventions. As an example, individuals that consistently travel between cities or high-risk areas might be encouraged by the government or an installed mobile application to receive a vaccination. During a wet-season, individuals often traveling through or living in specific malaria-intensive areas might be reminded through a public health application to use bed nets coated with insecticides. Another individual in a city who interacts with multiple large communities, e.g., large family, school system, workplace, and market, might be offered a free course of antivirals. The systematic identification and classification of an individual's role is only possible through high-resolution datasets regarding the travel and number of social contacts a person consistently maintains. This type of classification is not possible in developed countries due to a prioritization of privacy over disease avoidance, which has reduced the accuracy of existing population models in developed countries. With this information, public health officials, clinicians, and mobile applications can deliver targeted interventions, information dissemination, risk notification, and advice based on simulation-supported public policies.

III. Public Health Modeling and Simulation

Numerous studies have been conducted to investigate the spread of diseases. These studies have looked into the spread of contagious communicable diseases including smallpox, pertussis (whopping cough), H5N1 (swine flu), and H1N1 (avian flu) in addition to vector-borne diseases such as malaria [2], [3], [20], [28]. These studies have required the production of generic large-scale simulation applications that work with multiple diseases and population datasets scaling from small regions to large cities up to entire countries. Some of these epidemic prediction simulation software packages are available for public use including Flute [8], GSAM [24], and ABM++ [25]. Similar applications are well described in literature, see EpiSimdemics and EpiFast [4], [7].

Groups including public health officials, medical workers, and international health organizations are beginning to rely on computational approaches and experimentation to set policies. Through the utilization of the existing disease models and simulation applications, these organizations may study and analyze disease prediction, policy planning, and controversial response strategies *in silico* that would be unethical and infeasible to study in real-world populations. Simulation-based studies have proven valuable for groups including the World Health Organization, Bill and Melinda Gates Foundation, U.S. Centers for Disease Control, and U.S. National Institutes of Health.

A. Application to Cote d'Ivoire

While many of the existing studies have researched developed countries, the modeling and simulation approach has been previously applied in studies of developing countries within Africa and Asia. Experimental studies commonly attempt to determine optimal public and private strategies, e.g., interventions with insecticides and bed nets to reduce the spread of malaria in Tanzania [28]. To apply computational epidemiology approaches in Cote d'Ivoire, researchers require population datasets, models of population movements, models of local healthcare infrastructure, and the ability to design strategies appropriate for the resources within Cote d'Ivoire. D4D datasets 3 and 4 provide high-resolution data for producing activity models, population movement models, and ego classification as required by existing simulation applications. Data mining and analysis of D4D datasets 2 and 3 provide insights into the physical movements of individuals and may be used to produce models of population travel, movements, and shifts. Population modelers in developed countries attempt to produce similar population models based on small surveys of participants' reported locations and activities. The U.S. NIH MIDAS community and the Bill and Melinda Gates Foundation have produced extensive population datasets and models for the entirety of many developed countries (e.g., the

United States), large world-wide cities (e.g., Mumbai), and rural areas in developing countries (e.g., Ifakara, Tanzania) [23]. The D4D datasets provide real-world, verifiable, high-resolution data points for an extensive subset of the Ivorian population. In turn, more accurate, trustworthy, and verifiable population models may be produced for Cote d'Ivoire in comparison to other countries.

B. Early simulation results and visualization

Dataset 3 provides information on individual travel between areas of Cote d'Ivoire. This dataset was utilized to produce an activity model detailing how individuals travel between sub-prefectures for work and personal trips. Individuals were assigned a home sub-prefecture based on call patterns, and their travel patterns between sub-prefectures were tracked. Individual behavior was aggregated to the sub-prefecture level to produce general travel rates and patterns between sub-prefectures. This activity model was then paired with a population distribution dataset from a previous census [15] to approximate the home locations and movement of all Ivoirians. The simulation application Flute was then utilized to model, simulate, and predict the spread of an influenza outbreak starting with initial infections randomly distributed to ten separate individuals throughout the country. The application provides detailed results on the predicted spread of the disease in this scenario, broken down by the number of infected individuals by age.

One simulation predicted over 6 million people becoming infected from a generic influenza virus in the absence of any public interventions (e.g., vaccine distribution or school closures) or individual behavior changes (e.g., continuing to go to work or school if sick). Figure 1 visualizes the results from that simulation where the capital of each region, department, and sub-prefecture is marked with a red box indicating the number of infected individuals in that region. As an example, Abidjan is represented by a 10x10 set of pixels where each pixel is representative of 1,000 cases of infection, i.e., the solid red 10x10 box over Abidjan indicates that over 100,000 individuals became infected in the simulation. In other areas, the red box is proportionally shaded based on the number of cases in that region, e.g., a sparsely colored box for Kouto representing roughly 10,000 infections. Repeating this scenario thousands of times provides an expected worst-case scenario.

After the worst-case scenario has been determined, a range of potential interventions may be experimentally tested to determine the optimal combination of interventions that will mitigate or prevent the epidemic. The potential interventions possible to examine in Flute include vaccines distributed to different types of individuals (e.g., critical workers, pregnant women, families with infants, high-risk children, adults, and the elderly), vaccination boosts, antiviral treatments, antiviral prophylaxis, airport and long distance travel considerations, school closures, voluntary isolation, and quarantine. Another scenario was simulated to predict the effects of public health officials vaccinating 70% of the population after an epidemic was identified by surveillance. Figure 2 details the results of a simulation in this scenario. In comparing Figures 1 and 2, most areas have fewer red dots in each 10x10 square under the vaccination strategy, where each red dot that does not appear in Figure 2 indicates a set of 1000 individuals that did not become infected after the public health intervention. Table I compares the results of taking no governmental action (baseline scenario) versus vaccinating a portion of the population after a significant portion of the population has been identified as infected. The effect of the action was a complete prevention of infections for 12% of the population. See 'Appendix I: Simulation Results' for further details on the simulation results.

Complete studies require a full factorization of potential governmental strategies, execution of thousands of repetition of each scenario, and analyses on the optimal choice given a specific real-world surveillance scenario, i.e., the recognition of a given number of infections in a certain region. After setting public policies based on these simulations, local and national health officials may then utilize ego analysis to determine the identity of individuals to target and best method of disseminating information to those individuals.

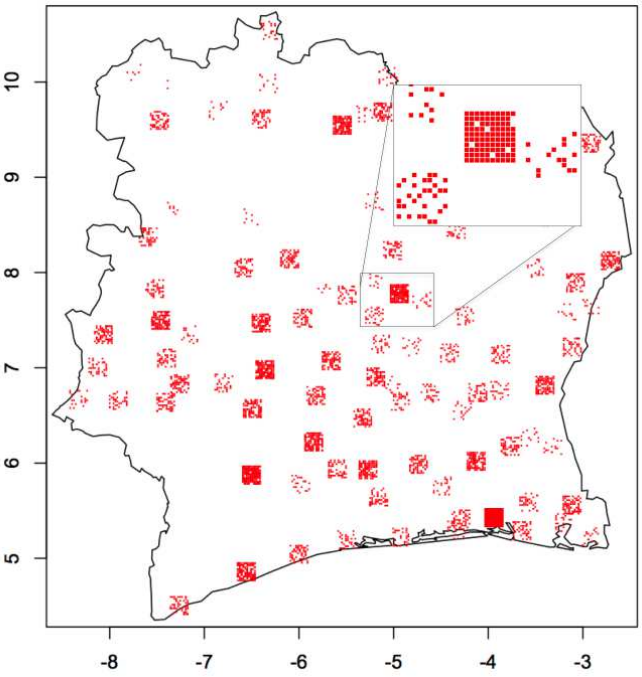


Fig. 1. Baseline predictive epidemic simulation of H5N1 with no governmental interventions or actions by individuals. Red dots indicate 1,000 cases of infections in nearby areas of major cities or regions.

IV. Ego Analysis

Governmental agencies use simulations of disease spread through networks to set public policies. In this section, D4D dataset 4 is studied as a viable network on which to base simulations and intervention analysis. In the first subsection, the high level view of the ego-centered graphs is presented

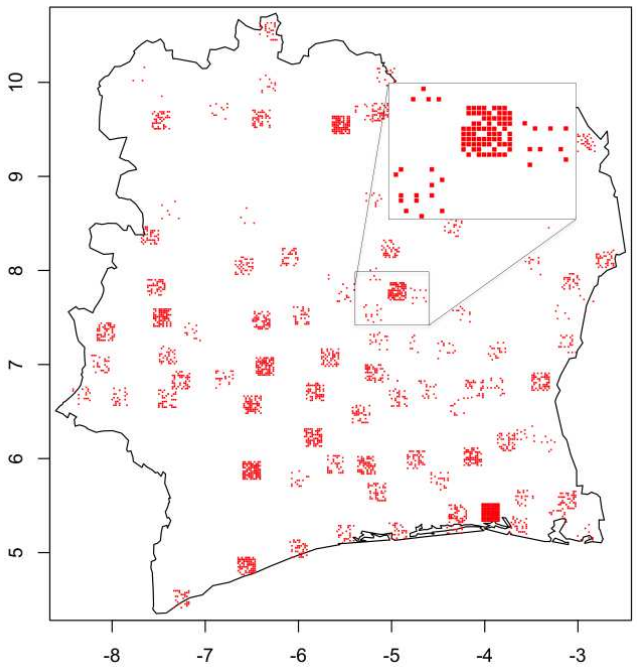


Fig. 2. Predictive epidemic simulation of H5N1 with a country-wide vaccination strategy that prevents 1.9 million infections in comparison to the base case. Red dots indicate 1,000 cases of infections in nearby areas of major cities or regions.

TABLE I
STATISTICS FOR A SINGLE RUN OF POTENTIAL SCENARIOS

Scenario	Attack Rate	Children Infected	Elderly Infected	Total Infected
Baseline	41.6%	412,679	539,173	6,414,609
Vaccination	29.5%	280,771	345,154	4,539,017

while considering ego graphs as part of a network. In that context, the characteristics of what (in our analysis) is called the individual ego graphs are analyzed in the second subsection.

In the following discussion, ego graphs of a node are repeatedly analyzed. The definition of ego graph for node A is the node A itself, all of A's neighbors (the nodes that are directly connected to A), all of the edges between A and its neighbors, and those edges between A's neighbors. Given a node with a degree of n (having n neighbors), the ego graph will contain a minimum of n and a maximum of $n(n+1)/2$ edges. An ego graph with only n edges is a star, see Figure 3 for an example of a star, and one with the maximum number of edges is a clique. The following definitions of density, extra edges, and extra density will be helpful in the discussion, considering a node A with degree n and a total of e edges in its ego graph:

$$density = \frac{e}{n(n+1)/2}$$

$$xtrEdges = e - n$$

$$xtrDens = \frac{xtrEdges}{n(n+1)/2}$$

A. Network Characteristics

One of the distinctive features of D4D dataset 4 is the star-like nature of the ego graphs. Many of the ego graphs are stars and those that are not often have very few links between the ego's neighbors. Many social networks have a high clustering coefficient value, meaning that nearly every node is part of a well-connected subgraph. A notable exception is the graph of a dating network [5]. D4D dataset 4 and dating networks do not contain edges representing the full set of individuals in the ego's social network, e.g., one does not always communicate with or date all of their family members and close friends. However, dating networks involve a completely different type of relationship than that between typical cell phone callers.

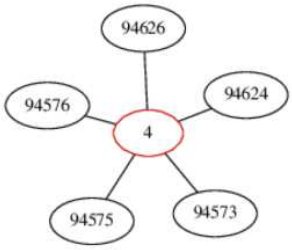


Fig. 3. Typical star ego graph consisting of unconnected one-hop neighbors.

TABLE II
STATISTICS FOR SET 4

time period	avg deg	extra edges	nbr of stars	xtr dens	max xtr dens
0	5.747	1.277	1859	0.043	0.6
1	4.629	1.079	2304	0.038	0.57
2	4.725	0.928	2209	0.036	0.60
3	5.029	0.940	2072	0.039	0.62
4	6.199	1.421	1578	0.046	0.53
5	5.147	0.952	2050	0.039	0.60
6	6.220	1.368	1651	0.044	0.50
7	4.542	0.758	2238	0.035	0.60
8	6.237	1.391	1609	0.046	0.52
9	4.512	0.722	2235	0.034	0.67
avg	5.299	1.084	1981	0.040	0.58

To support the claim of the star-like nature of the ego graphs, refer to Table II. For each time period, statistics were calculated regarding the following attributes (using the ego graphs of D4D dataset 4):

- avg deg of all egos
- extra edges: total edges minus degree
- nbr of stars: total number of stars of the 5000 egos
- xtr dens: average of the xtrDens for each ego graph
- max xtr dens: maximum extra density of the ego graphs

Note that the average degree ranges between 4.5 and 6.2. For an ego with a degree of 5, there are 6 total nodes in

the graph resulting in a maximum number of edges of 15. Excluding the 5 edges from the ego, the ego could have between 0 and 10 extra edges in comparison to a star graph. Assume that an ego with degree = 5 had one extra edge, then its extra density would be 0.067.

Approximately 40% of the egos are stars. With an average number of extra edges that hovers around 1, it appears that even the egos that do not have a star graph, have a very sparse, star-like, ego graph.

This does not appear to represent a typical social network. Many networks, particularly those representing friendship links between humans, have a rather dense clustering around each ego. The typical networks have the small world property that was first developed by Watts and Strogatz [29]. They identified two characteristics of small world networks. First, they exhibit a regular graph structure with clustering around each node. Second, they also have some random edges between disparate nodes.

The social network of Cote d'Ivoire appears to have, like most social networks, the small world properties. However it can be stratified into two parts; the dense, regular clusters and the longer distance, random edges. The face-to-face interactions of people within an ego's immediate neighborhood (family, village) have the tight clustering properties. The random edges are revealed by the cell phone connections.

To test for small world properties, a series of networks using the model for small world networks was generated. A network of 5000 nodes where each node was connected to its closest k neighbors was created. Then, with a probability of p, each link was re-wired from a neighbor to a random node in the network. According to Watts and Strogatz [29], with $p = 0$, the network is regular and as p gets larger the network becomes more random until at $p = 1$ it is a random graph. At values of $0.01 \leq p \leq 0.1$ the network has the characteristics of a small world network.

For each value of p and k we generated 100 networks. The extra density was calculated for each node and averaged across all nodes. We then averaged these across all 100 trials. Table III shows the results of our tests. The two right-most columns represent values for different size neighborhoods; the second column has the results for networks where nodes are connected to their 10 closest neighbors and the third is for networks with neighborhoods of size 20.

According to the results, a 5000 node small world network (where $p = .01$ or $p = .10$) has extra density values ranging from 0.58 to 0.68. The networks that have extra density values of 0.07 are generated from a value of $p = 0.90$. In comparing Tables II and III, the extra densities of 0.04 found in D4D dataset 4 are consistent with a p of 0.90 as determined by the experimets shown in Table III. As these values are close, the analysis lends support to the notion that the cell network data reveals the random layer of small world network of Cote d'Ivoire.

Because access to the dense clusters of networks was unavailable, the following simplifying assumption made: every node in the network has an identical face-to-face network.

TABLE III
EXTRA DENSITY FOR SYNTHETIC NETWORKS USING DIFFERENT VALUES OF p

	neighborhood size	
p	k=10	k=20
0.01	0.63	0.68
0.10	0.58	0.62
0.20	0.51	0.56
0.30	0.44	0.48
0.40	0.38	0.42
0.50	0.32	0.35
0.60	0.25	0.28
0.70	0.18	0.21
0.80	0.13	0.14
0.90	0.07	0.07

Even cellular providers would not have this information as it would require extensive, manual data collection and modeling of phone use in relation to face-to-face social networks. While naïve, this assumption allows for reasoning regarding the identification of influential nodes.

B. Nodes Characteristics

The well known problem of influence maximization [6], [9], [11], [18], [19] is based on finding influential nodes within a network assuming a particular diffusion model. These models are also used for the study of disease spread. It is appropriate to consider identifying the influential nodes in this network for targeted dissemination of health or social welfare information.

If the assumption is correct that the cell phone network does not in fact represent the entire social network of Cote d'Ivoire, but only the random portion of it, it would not make sense to try to find influential nodes using methods designed to work with the entire network.

One strategy for identifying influential nodes is to choose the high degree egos. While this would appear to maximize the influence throughout the network, another approach is available with a higher degree of precision. Some of the high degree egos are not stars but have extra edges in their ego graph. This indicates an overlap between the random layer of the network with the face-to-face layer. For example, if ego A has neighbors B and C, there is also a link between B and C. While not conclusive proof, it offers some evidence that B and C might be part of A's face-to-face network.

A more precise strategy selects nodes that have a combination of high degree and low density. [26] proposed a metric called *rawComm* that assigns a number to a node giving it a relative measure of the number of communities it belongs to. The two assumptions are that face-to-face graphs are considered identical and do not influence cellular call graphs and that extra edges reflect neighbors in the face-to-face graph. With these assumptions, selecting the egos that have the highest value of rawComm will be more likely to spread influence to more nodes than using just the highest degree egos.

1) Node Classification Metrics: The rawComm metric is defined as:

$$rawComm(u) = \sum_{v \in N(u)} \tau_u(v)$$

where $N(u)$ is the neighborhood of node u — that is all of the nodes that are directly linked to u — and $\tau_u(v)$ is given by

$$\tau_u(v_i) = \frac{1}{1 + \sum_{v_j \in N(u)} I(v_i, v_j)p + \bar{I}(v_i, v_j)(1-q)}$$

In the above formula for $\tau_u(v_i)$, $I(v_i, v_j)$ is defined as:

$$I(v_i, v_j) = \begin{cases} 1, & \text{if there is a link between nodes } v_i \text{ and } v_j \\ 0, & \text{otherwise} \end{cases}$$

Also, $\bar{I}(v_i, v_j)$ is defined as the compliment of $I(v_i, v_j)$. The quantity p denotes the probability that an edge exists between nodes v_i and v_j if the nodes are in the same community, while q represents the probability that an edge does not exist between nodes v_i and v_j if they are in different communities. For this work, the simplifying assumption was made that $p = q = 1$.

Given only the network structure, rawComm approximates the number of communities formed by the neighbors of a given node. The underlying premise here is that the communal information is hidden but that links provide evidence of this information. To uncover this association, rawComm identifies communities by comparing the neighborhoods of the neighbors of the node of interest to a clique — a maximal complete subgraph — and determining that neighbor's communal contribution. Thus, a group that forms a clique will be considered one community; another group that forms two non-overlapping cliques will be considered two communities.

In addition to utilizing rawComm to estimate the number of communities that a node is a part of, the density metric can be used in order to assess the confidence that the telecommunication network of an individual reflects their real-world network. In order to maximize this likelihood along with the assumption that the telecommunication network exhibits a small-world network, individuals whose telecommunication network has a high density are desirable and sought.

2) Identification of Influential Nodes: To increase the effectiveness of public health actions, the metrics defined in the previous section were used to find individuals who, when targeted, would lead to the greatest impact. The following process was developed for identification:

1) A distribution of rawComm-density pairs was constructed for all nodes for a time period and desirable values were selected.
2) The variability of average rawComm-density pair values was analyzed across the ten 2-week time periods in dataset 4 for these individuals, and the values with the lowest variability were selected in order to further maximize influence.
3) From these individuals, the country-wide Flute outbreak simulations may be used to restrict the set of individuals to particular geographic locations.

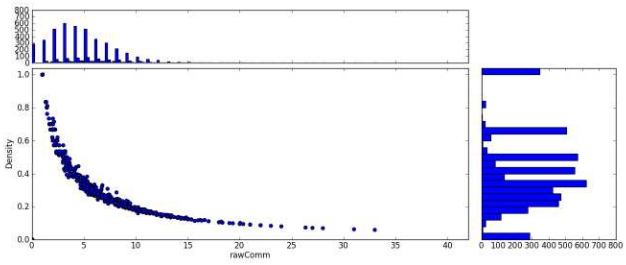


Fig. 4. Sample distribution plot for rawComm-density values for all egos for time period.

For the first step, distributions for rawComm-density pairings for all egos were constructed for a given time period. Figure 4 provides an example of one such time period. Given a distribution, individuals were then selected based on whether their rawComm and density values exceeded set thresholds. Sample values for these thresholds include rawComm values greater than 5 and density values greater than 0.2.

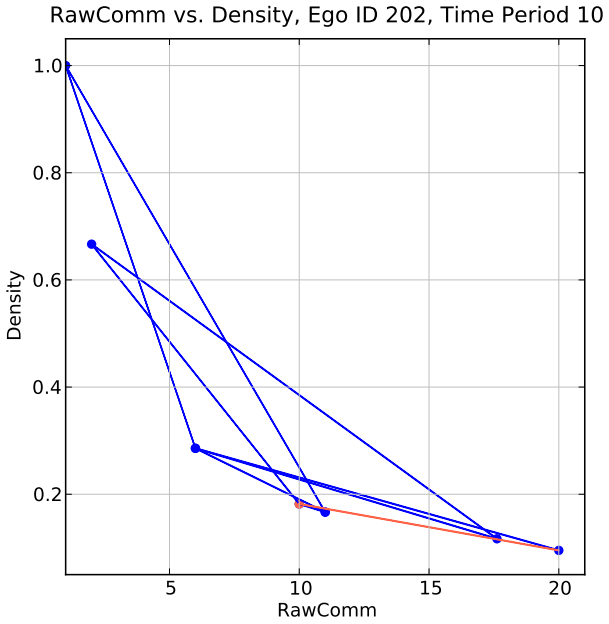


Fig. 5. Sample plot of the variability in rawComm-density values for ego ID 202. Each edge represents the change, in relation to the previous time period, of the ego's current number of active communities and density of the ego's network.

Following this, weighted averages of the mean rawComm-density values were computed and the minimal values were sought. Figure 5 provides a plot of the change in rawComm-density values over time.

The next step in an end-to-end simulation and ego analysis process would be to tie ego identification back into designing followup simulation studies that predict the effect of applying interventions to targeted individuals. Flute simulations may provide estimates for assessing the potential of disease spread

within regions of the country and the effectiveness of treatment based on targeted SMS and mobile communication. These individuals of interest could also be filtered based on their historical locale's simulation results.

V. Discussion and Future work

The datasets provided by D4D allowed for the generation of fine-grained population, activity, and social network models. The first two types of models were used in simulation-based experimentation to determine how contagions are spread throughout Cote d'Ivoire. Models were developed to represent the population of Cote d'Ivoire in a format that could be utilized by simulation applications. Additional study designs and complex scenarios may be designed further study the effects of public and private interventions on disease mitigation. The results of simulations, such as those detailed in this effort, provide scientific support for public health policy making. The social network analyses possible through D4D datasets were shown to have implications for the identification of egos with consistent, desirable characteristics. Individuals were deemed important based on the number of connections they had in the network as well as the number of communities to which they belonged. In order to identify the latter, novel community-finding algorithms were developed and employed. Approaches described in Section IV can identify notable egos with a large 'star' network that act as ambassadors to many unconnected social groups. These influential egos may be targeted for information dissemination and campaigns in order to make use of influence maximization and optimally apply limited public health resources. Other egos that travel with certain patterns to given regions of the country can also be identified with Section IV approaches and D4D datasets 2 and 3. These egos can be targeted with SMS or mobile application public health recommendations or offered pharmaceutical interventions and risk notifications. The approach demonstrated here may be utilized by national public health officials, local health providers and clinics, and international planning organizations in setting policies and identifying potential mobile applications for research and dissemination.

We envision a computational epidemiology tool that combines the content and visualizations of Figures 3, 4, and 5. This tool would allow health officials or mobile software to identify persons suitable for a given targeted intervention. Pairing this tool with the simulation environment would allow researchers to conduct studies with simulations in order to concentrate identification in appropriate regions of the country. This pairing would allow simulations to predict the effect of such targeted interventions on disease spread. 'Appendix II: Visualization' contains a visualization that combines the three previously described graphs based on density and rawComm metrics. A single ego (202) is analyzed and followed over the 10 time periods of D4D dataset 4. The graphs in this appendix visualize the evolution of the ego, demonstrating its changing behavior and roles over time. The visualization allows researchers to identify individual egos for further analysis and potential targeted intervention.

Future modeling and simulation efforts may pair the trajectories datasets and our derived population models with estimations of mosquito and disease prevalent areas to produce models of high-risk travellers. Further economic analysis of the cost of public health interventions in light of their predicted impact on disease mitigation is warranted. Comparisons of study results generated for Cote d'Ivoire and extensive existing studies for other well-studied areas of the world may lead to insights on the nature of contagion diffusion in this region.

Social network analyses consist of determining how to apply interpretations of individuals with consistent roles, social networks of specific structures, and the nature of unpredictable egos. Further analysis may determine if an ego with consistent behavior is actually in contact with the same people over time and if inconsistent behavior is impacted by egos that connect their friends into friends of friends relationships that bypass the ego.

Acknowledgment

We thank France Telecom-Orange and the Data 4 Development Challenge for providing access to mobile datasets regarding Cote d'Ivoire.

References

[1] J. Aker and I. Mbiti. Mobile Phones and Development in Africa. *Economic Perspectives*, 24(3):207–232, 2010.

[2] C. Barrett, K. Bisset, J. Leidig, A. Marathe, and M. Marathe. Economic and social impact of influenza mitigation strategies by demographic class. *Epidemics Journal*, 3:19–31, 2011.

[3] C. Barrett, S. Eubank, and J. Smith. If smallpox strikes Portland ... . *Scientific American*, 292, 2005.

[4] C. L. Barrett, K. R. Bisset, S. G. Eubank, X. Feng, and M. Marathe. EpiSimdemics: an efficient algorithm for simulating the spread of infectious disease over large realistic social networks. In *SC '08: Proceedings of the 2008 ACM/IEEE conference on Supercomputing*, pages 1–12, Piscataway, NJ, USA, 2008. IEEE Press.

[5] P. Bearman, J. Moody, and K. Stovel. Chains of affection: The structure of adolescent romantic and sexual networks. *American Journal of Sociology*, pages 44–91, 2004.

[6] S. Bharathi, D. Kempe, and M. Salek. Competitive influence maximization in social networks. In *Proceedings of WINE*, 2007.

[7] K. Bisset, J. Chen, X. Feng, A. Vullikanti, and M. Marathe. EpiFast: a fast algorithm for large scale realistic epidemic simulations on distributed memory systems. In *ICS '09: Proceedings of the 23rd international conference on Supercomputing*, pages 430–439, New York, NY, USA, 2009. ACM.

[8] D. L. Chao, M. E. Halloran, V. J. Obenchain, and I. M. Longini. FluTE, a Publicly Available Stochastic Influenza Epidemic Simulation Model. *PLoS Comput Biol*, 6(1), 2010.

[9] W. Chen, Y. Wang, and S. Yang. Efficient influence maximization in social networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2009.

[10] G. Essegbey and G. Frempong. Creating space for innovationThe case of mobile telephony in MSEs in Ghana. *Technovation*, 31(12):678–688, 2011.

[11] P. Estevez, P. Vera, and K. Saito. Selecting the most influential nodes in social networks. In *Neural Networks, 2007. IJCNN 2007. International Joint Conference on*, 2007.

[12] C. W. Factbook. Cote d'Ivoire, Jan. 2013.

[13] M. Halloran, N. Ferguson, and S. Eubank. Modeling targeted layered containment of an influenza pandemic in the United States. *Proceedings of the National Academy of Sciences*, 105(12):4639–4644, 2008.

[14] H. W. Hethcote. The Mathematics of Infectious Diseases. *SIAM Review*, 42(4):599–653, 2000.

[15] Institut National de la Statistique. Republic of Cote d'Ivoire, 1998 Census (Republished). http://www.geohive.com/cntry/coteivoire.aspx, 2012.

[16] IRIN-News2012Cote. COTE D'IVOIRE: Meningitis spreads as people scramble for vaccine, Feb. 2012.
[17] O. Kamga. Mobile Phone in Cote d'Ivoire: Uses and Self-Fulfillment. *Information and Communication Technologies and Development*, pages 184–192, 2006.
[18] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 137–146, 2003.
[19] M. Kimura and K. Saito. Approximate solutions for the influence maximization problem in a social network. In *Knowledge-Based Intelligent Information and Engineering Systems*, 2006.
[20] B. Lewis, R. Beckman, V. Kumar, J. Chen, P. Stretz, K. Bisset, H. Mortveit, K. Atkins, A. Marathe, M. Marathe, S. Eubank, and C. Barrett. Simulated Pandemic Influenza Outbreaks in Chicago. Technical Report 07-004, NIH DHHS Study Final report, 2007.
[21] I. M. Longini, A. Nizam, S. Xu, K. Ungchusak, W. Hanshaoworakul, D. A. Cummings, and M. E. Halloran. Containing pandemic influenza at the source. *Science*, 309(5737):1083–1087, 2005.
[22] D. Mead and C. Liedholm. The dynamics of micro and small enterprises in developing countries. *World Development*, 26(1):61–74, 1998.
[23] N.I.H. National Institutes of Health, MIDAS. http://www.nigms.nih.gov/Initiatives/MIDAS/, 2009.
[24] J. Parker and J. M. Epstein. A distributed platform for global-scale agent-based models of disease transmission. *ACM Trans. Model. Comput. Simul.*, 22(1):2:1–2:25, Dec. 2011.
[25] RTI International. ABM++ Distributed Computing Framework. http://parrot-farm.net/ABM++/, 2010.
[26] J. Scripps, P. N. Tan, and A.-H. Esfahanian. Node roles and community structure in networks. In *Proceedings of the Thirteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining Joint Workshop on Web Mining and Social Network Analysis*, 2007.
[27] M. Siedner, J. Haberer, M. Bwana, N. Ware, and D. Bangsberg. High acceptability for cell phone text messages to improve communication of laboratory results with HIV-infected patients in rural Uganda: a cross-sectional survey study. *BMC Medical Informatics and Decision Making*, 12(56), 2012.
[28] T. Smith, G. F. Killeen, N. Maire, A. Ross, L. Molineaux, F. Tediosi, G. Hutton, J. Utzinger, K. Dietz, and M. Tanner. Mathematical modeling of the impact of malaria vaccines on the clinical epidemiology and natural history of Plasmodium falciparum malaria: Overview. In *Am J Trop Med Hyg*, volume 75, pages 1–10, 2006.
[29] D. J. Watts and S. H. Strogatz. Collective dynamics of small-world networks. *Nature*, pages 440–442, Jun 1998.

APPENDIX I: SIMULATION RESULTS

Figures 6, 7, and 8 and Tables IV, V, and VI detail the simulation results for the study used in this report.

TABLE IV
TOTAL INFECTIONS AND ATTACK RATES FOR EACH SCENARIO BY AGE

	0-4	5-18	19-29	30-64	65+
Baseline Total infections	412679	1973384	731151	2758222	539173
Baseline Attack Rate	0.4089	0.5780	0.3734	0.3808	0.3024
Vaccination Total infections	280771	1588086	479675	1845331	345154
Vaccination Attack Rate	0.2782	0.4651	0.2450	0.2548	0.1936

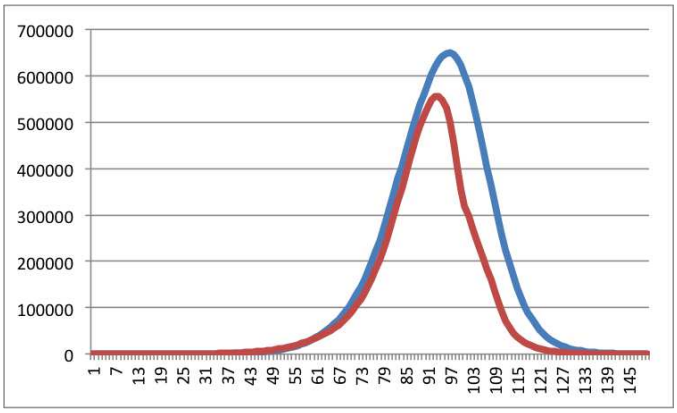


Fig. 6. Epicurves of the country-wide infection count for the baseline (blue) and vaccine strategy (red) showing the number of new incidents of infection by day.

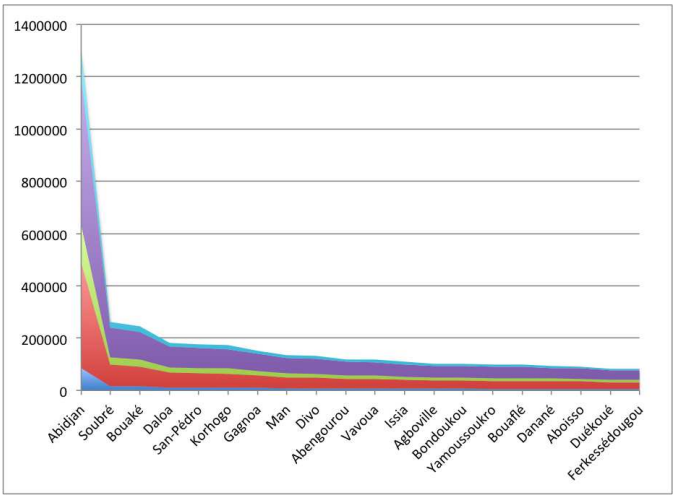


Fig. 7. Total number of infected individuals, by age group, for the 20 regions with the most cases of infection in the baseline simulation scenario (0-4 blue, 5-18 red, 19-29 green, 30-64 purple, and 65+ light blue).

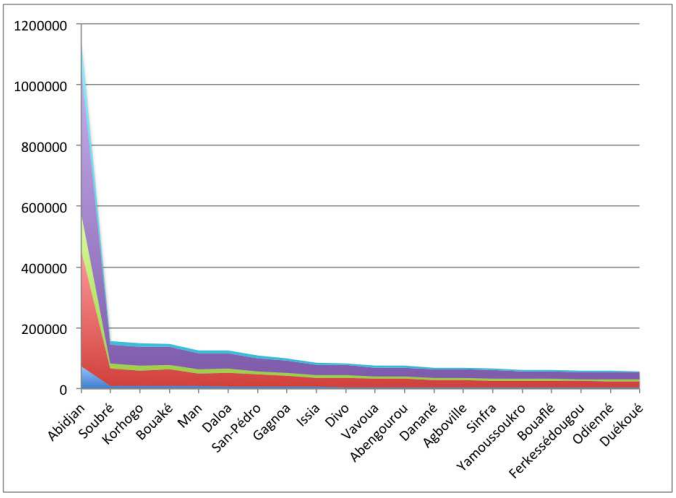


Fig. 8. Total number of infected individuals, by age group, for the 20 regions with the most cases of infection in the vaccination scenario (0-4 blue, 5-18 red, 19-29 green, 30-64 purple, and 65+ light blue).

TABLE V
BASELINE INFECTION COUNT BY AGE FOR 60 REGIONS WITH THE HIGHEST TOTAL INCIDENTS OF INFECTION

Region	0-4	5-18	19-29	30-64	65+
Abidjan	84321	400249	148631	558980	109560
Soubr	16885	80981	29520	112958	22196
Bouak	15699	75394	27815	105167	20634
Daloa	11719	56268	20840	78660	15435
San-Pdro	11358	54161	20047	75562	14729
Korhogo	10995	53249	19707	74137	14374
Gagnoa	9680	47036	17283	65547	12822
Man	8608	41534	15505	58471	11513
Divo	8478	40784	15022	57324	11208
Abengourou	7786	36597	13655	51477	9924
Vavoua	7634	36431	13562	50777	10003
Issia	6896	33853	12441	47090	9009
Agboville	6496	30948	11687	43671	8568
Bondoukou	6421	31021	11502	43684	8460
Yamoussoukro	6212	30269	11114	42453	8383
Bouafl	6297	29904	11186	42132	8198
Danan	6203	28685	10693	40057	7675
Aboisso	5800	28405	10409	39371	7775
Dukou	5462	25473	9436	35456	6775
Ferkessdougou	5267	25611	9458	35259	6934
Mankono	5175	24469	9196	33989	6551
Oum	4890	22380	8361	31628	6122
Tiassal	4670	22345	8395	31380	6353
Odienn	4634	22151	8349	31046	6058
Sinfra	4522	22110	8121	30558	5922
Tanda	4444	21384	7900	29969	5772
Guiglo	4521	21561	7852	29554	5929
Adzop	4414	21443	7940	29757	5859
Sassandra	4110	19563	7260	27880	5411
Sgula	4084	19758	7340	27393	5395
Bouna	4042	19217	7199	27327	5232
Zunoula	3977	19082	6892	26619	5163
Lakota	3896	18684	6997	26251	5228
Dabou	3714	17772	6631	24476	4611
Tabou	3647	17569	6502	24546	4750
Grand-Bassam	3700	17406	6495	24593	4686
Bongouanou	3622	17151	6456	24370	4872
Katiola	3440	17307	6325	23805	4668
Bangolo	3547	16991	6270	24004	4725
Boundiali	3421	17006	6283	23605	4717
Guitry	3489	16830	6252	23356	4542
Biankouma	3325	15752	5735	21868	4235
Boumi	3215	15145	5660	21217	4244
Touba	3196	14864	5654	21013	4151
Daoukro	3015	14447	5288	20159	3987
Agnibilkrou	2922	13651	5132	19095	3695
Toumodi	2860	13755	4915	19034	3706
Dabakala	2827	13390	5077	18791	3837
Alp	2464	12200	4613	17206	3305
Blolquin	2549	12001	4482	16926	3246
Zouan Hounien	2525	12074	4422	16825	3092
Grand-Lahou	2216	10970	3975	15375	3033
Zoukougbeu	2240	10829	3910	15148	3015
Bocanda	2162	10687	3986	15202	2917
Fresco	2134	10788	3876	14863	2864
Dimbokro	2182	10322	3898	14416	2808
Sakassou	2175	9982	3854	14177	2913
Agboville	1965	9597	3507	13258	2528
Tibissou	2016	9196	3354	12795	2505
M'Bahiakro	1716	8527	3088	11910	2381
Arrah	1679	8452	3016	11511	2236
Tengrla	1673	8252	3046	11421	2298
Adiak	1537	7595	2744	10595	2112
Sikensi	1521	7468	2675	10500	2010
Guyo	1478	7046	2648	9708	1904

TABLE VI
VACCINE STRATEGY INFECTION COUNT BY AGE FOR 60 REGIONS WITH THE HIGHEST TOTAL INCIDENTS OF INFECTION

Region	0-4	5-18	19-29	30-64	65+
Abidjan	72501	378060	124265	473452	90932
Soubr	9322	57602	15600	61777	11238
Korhogo	9249	49377	16101	61610	11837
Bouak	8898	54118	15120	58438	10747
Man	8132	40797	13884	53049	10224
Daloa	7544	44889	13220	50701	9467
San-Pdro	6604	39786	10858	42606	7813
Gagnoa	5973	35636	10167	39923	7193
Issia	5412	29367	8933	34346	6601
Divo	5088	30314	8624	33589	6082
Vavoua	4516	27461	7524	29823	5278
Abengourou	4542	27076	7707	29607	5487
Danan	4256	24030	7323	27784	5063
Agboville	4097	24498	7212	27135	5055
Sinfra	4257	21465	7154	27356	5239
Yamoussoukro	3572	22188	6243	24334	4489
Bouafl	3733	21894	6288	23957	4302
Ferkessdougou	3589	20886	6094	23817	4530
Odienn	3647	19791	6253	24281	4600
Dukou	3440	20458	5992	22994	4191
Dabou	3679	17610	6503	24116	4597
Aboisso	3063	19475	5321	20563	3770
Boundiali	3339	16572	5756	22073	4175
Tiassal	3125	17997	5504	21011	3908
Biankouma	3294	15700	5690	21237	4218
Adzop	2961	17213	5065	19704	3659
Guitry	2902	15324	5048	18924	3680
Oum	2786	16593	4686	18263	3222
Katiola	2811	15499	4770	18310	3556
Touba	2833	14426	4880	18429	3569
Sassandra	2657	15299	4571	17858	3303
Bondoukou	2430	16659	3993	15633	2858
Guiglo	2574	15281	4294	16210	2954
Sgula	2298	14454	3997	15234	2843
Lakota	2245	14064	3999	15565	2907
Mankono	2162	13964	3650	14106	2485
Tanda	2204	13783	3681	14082	2501
Bangolo	2159	13050	3707	14456	2680
Grand-Bassam	2172	12493	3527	14049	2497
Zunoula	1930	12864	3498	13306	2409
Grand-Lahou	2080	10940	3691	14294	2802
Toumodi	2062	11868	3498	13683	2505
Bongouanou	2060	12370	3363	13121	2279
Bouna	2009	12173	3301	12388	2205
Tabou	1822	11774	3064	12250	2101
Zouan Hounien	1877	10067	2988	11669	2159
Fresco	1728	9674	2907	11381	2121
Tengrla	1657	8153	3064	11147	2275
Blolquin	1570	9160	2569	10068	1832
Zoukougbeu	1527	8883	2547	10166	1932
Dabakala	1469	9380	2548	9662	1767
Alp	1323	8593	2390	8984	1673
Sikensi	1335	7184	2345	9187	1706
Jacqueville	1316	6659	2343	9112	1788
Agnibilkrou	1250	8139	1988	8060	1464
Dimbokro	1345	7554	2150	8053	1453
Daoukro	1167	7879	1984	7777	1328
Tibissou	1136	6789	1858	7336	1405
Akoupe	1048	6856	1908	7289	1268
Boumi	1061	7018	1813	6735	1197
Ouangolodougou	1123	5957	1984	7340	1362
Bocanda	920	6390	1594	6545	1130
Toulpleu	976	5692	1778	6697	1198
Sakassou	964	6010	1703	6290	1180
Arrah	962	6081	1663	6191	1116

APPENDIX II: VISUALIZATION

A visualization is required to sort through millions of egos for the identification of individual egos for further analysis and potential targeted intervention.

The main panel of the top graph is a plot of density versus rawComm, see the discussion of Figure 4. Above it is a histogram of rawComm values, to the right is a histogram of density values. The lower left graph is a trace of the changing rawComm-density pair values. At time period 1, there is a single red dot indicating the values for ego 202. At time period 2, the new rawComm-density pair value is plotted as a red dot; the time 1 value is blue. At time 3, the new value is plotted as a red dot; old values are blue. Following the red line and the red dot traces the changing values for that ego over time. The lower right graph is the corresponding one-neighborhood subgraph for ego 202 for the indicated time period.

The visualization allows for the characterization of the ego node. For example, at time 1, ego 202 is a star node. The lower left graph illustrates this by plotting a high rawComm value (# of communities it belongs to), and a very low density (few of its neighbors are linked to each other). This implies a relatively "important" ego state, as it connects with many other egos who do not connect with each other. On the other hand, at time 2 ego 202 is a loner. It belongs to only 1 community, hence it has a very low rawComm value; and it has a density of 1.0, as it only has one neighbor. This represents a less important ego, in terms of the potential amplification effect of targeting this ego.

Components are in place to transform these visualization into an interactive, web-based tool. Beginning with the aggregate plot (the scattergram), it could be possible to select individual nodes, to view their specific rawComm-density pair values, to observe those values changing over time, and to view the corresponding ego graphs.

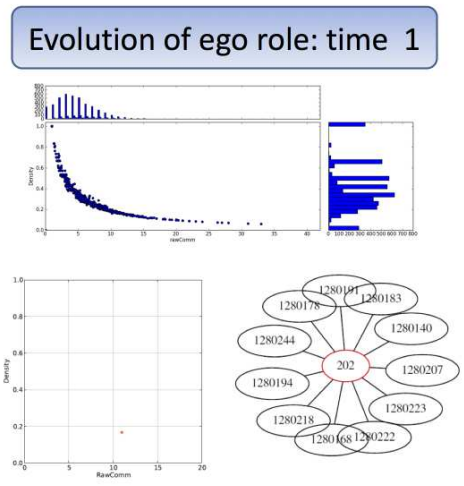


Fig. 9. Visualization of ego analyses by time period.

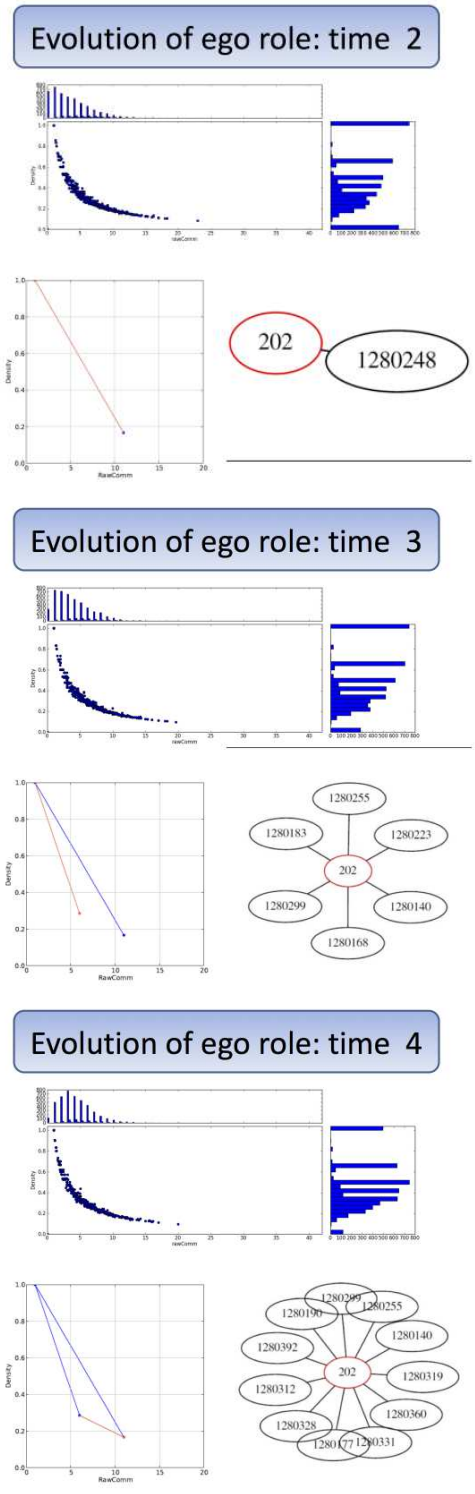


Fig. 10. Visualization of ego analyses by time period.

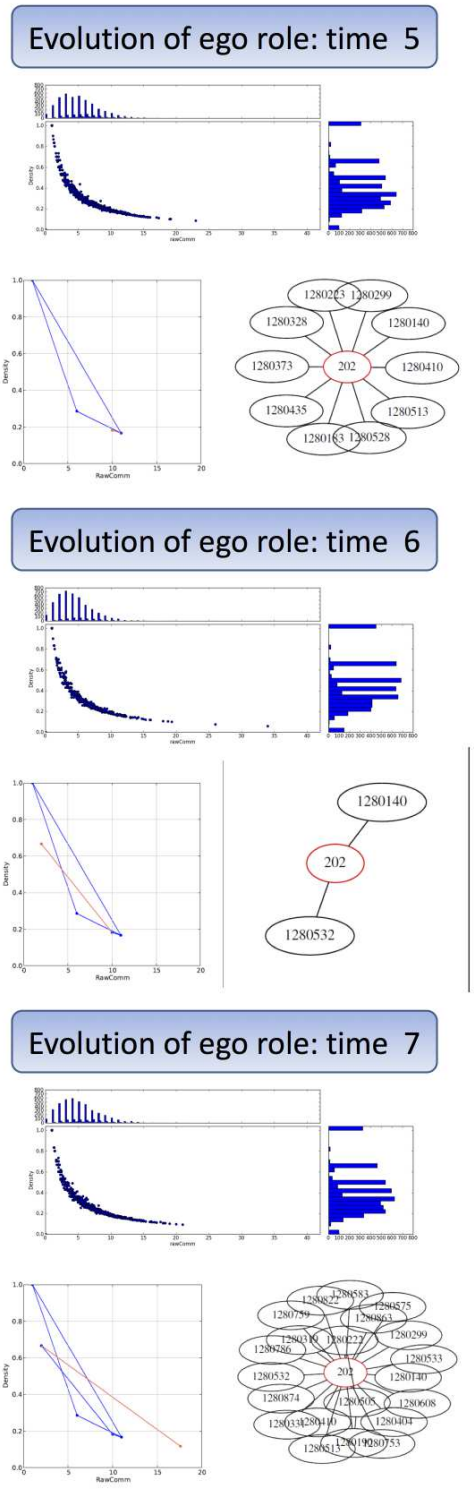


Fig. 11. Visualization of ego analyses by time period.

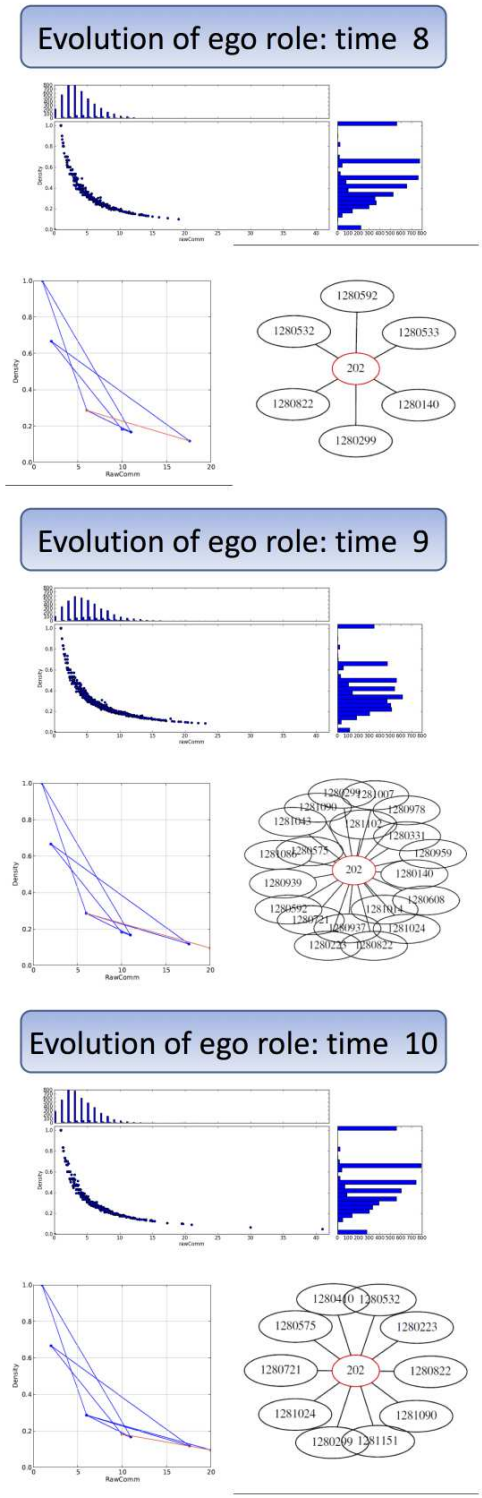


Fig. 12. Visualization of ego analyses by time period.