NetMob 2011

Workshop on the Analysis of Mobile Phone Networks

MIT (Cambridge, USA)
October 10-11, 2011

Book of abstracts

Editors: Vincent Blondel
Gautier Krings

Scientific Committee:

Co-chair: Vincent Blondel	UCLouvain (Belgium) and MIT
Co-chair: Alex (Sandy) Pentland	Media Lab, MIT
Rein Ahas	University of Tartu (Estonia)
Samuel Arbesman	Harvard University
Laszlo Barabasi	Northeastern University
Nicholas Christakis	Harvard University
Rob Claxton	British Telecom (UK)
Massimo Colonna	Telecom Italia (Italy)
Nicolas de Cordes	Orange Group Strategy (France)
Nathan Eagle	txteagle
Kenth Engø-Monsen	Telenor (Norway)
Alexandre Gerber	AT&T Research
Marta Gonzales	MIT
Cesar Hidalgo	Media Lab, MIT
Kimmo Kaski	Aalto University (Finland)
János Kertész	Budapest University of Technology (Hungary)
Renaud Lambiotte	University of Namur (Belgium)
Juha Laurila	Nokia Research (Switzerland)
David Lazer	Northeastern University
Franck Legendre	ETH Zurich (Switzerland)
Esteban Moro Egido	Universidad Carlos III de Madrid (Spain)
Nuria Oliver	Telefonica Research (Spain)
Jukka-Pekka Onnela	Harvard University
Dino Pedreschi	Università di Pisa (Italy)
Daniele Quercia	University of Cambridge (UK)
Carlo Ratti	Senseable City Lab, MIT
Jari Saramäki	Aalto University (Finland)
Zbigniew Smoreda	Orange Labs (France)
John Tsitsiklis	MIT
Paul Van Dooren	UCLouvain (Belgium)
Alexander Varshavsky	AT&T Labs

Organizing Committee:

Chair: Vincent Blondel	UCLouvain (Belgium) and MIT
Francesco Calabrese	IBM Research (Ireland) and MIT
Yves-Alexandre de Montjoye	Media Lab, MIT
Gautier Krings	UCLouvain (Belgium)
Dashun Wang	Northeastern University

Table of contents

Program overview

Monday October 10, 2011

8:30-9:00 *Registration*

9:00-9:15 *Conference overview: V. Blondel*

9:15-10:30 SESSION A: Individual mobility

10:30-11:00 *Coffee break*

11:00-12:15 SESSION B: Individual mobility

12:15-13:30 *Lunch break*

13:30-14:00 *Keynote talk: A.-L. Barabási*

14:00-15:45 SESSION C: Social networks

15:45-16:15 *Coffee break*

16:15-17:30 SESSION D: Population mobility

18:00 - 19:30 *Dinner*

starting at 19:30 Progressive Rock event

Tuesday October 11, 2011

9:30-10:30 SESSION E: Social structure

10:30-11:00 *Coffee break*

11:00-12:15 SESSION F: Dataset mining

12:15-13:30 *Lunch break*

13:30-14:00 *Keynote talk: A. Pentland*

14:00-15:30 SESSION G: Churn and value

NetMob 2011

Second conference on the

Analysis of Mobile Phone Datasets and Networks

Program

Monday October 10, 2011

8:30-9:00 *Registration*

9:00-9:15 *Conference overview: trends in mobile phone network analysis*

V. Blondel, UCLouvain

SESSION A (9:15-10:30). Individual mobility

9:15-9:30 *Identifying Important Places in People's Lives from Cellular Network Data*

S. Isaacman (1), R. Becker (2), R. Cáceres (2), S. Kobourov (3), M. Martonosi (1), J. Rowland (2), A. Varshavsky (2)

1. Princeton University
2. AT & T Labs - Research
3. Dept. of Computer Science, University of Arizona

9:30-9:45 *Measuring repeated visitations with mobile positioning data. Applications for marketing.*

R. Ahas (1), M. Tiru (2), A. Kuusik (3)

1. Department of Geography, University of Tartu
2. Positium LBS
3. Faculty of Economics, University of Tartu

9:45-10:00 *Classifying Routes Using Cellular Handoff Patterns*

R. Becker, R. Cáceres, K., J. M. Loh, S. Urbanek, A. Varshavsky, C. Volinsky

AT&T Labs - Research

10:00-10:15 *Characterizing Urban Road Usage Patterns With a New Metric*

P. Wang (1,2), T. Hunter (3), A. Bayen (3), M. González (1)

1. Department of Civil and Environmental Engineering, MIT

3. Department of Civil and Environmental Engineering, University of California Berkeley

10:15-10:30 *Individual Mobility Networks from Clusters of Human Activity*

S. Jiang (1), J. Ferreira (1), M. González (2)

1. Department of Urban Studies and Planning, MIT
2. Department of Civil and Environmental Engineering, MIT

10:30-11:00 *Coffee break*

SESSION B (11:00-12:15). Individual mobility

11:00-11:15 *Structure and seasonality in human motion*

J. Bagrow (1,2), Y.-R. Lin (3,4)

1. Center for Complex Network Research, Northeastern University
2. Center for Cancer Systems Biology, Dana-Farber Cancer Institute
3. College of Computer and Information Science, Northeastern University
4. Institute for Quantitative Social Science, Harvard University

11:15-11:30 *Exploring mobility of mobile users*

B. Csáji (2,3), A. Browet (1), V. Traag (1), J.-C. Delvenne (1), E. Huens (1), P. Van Dooren (1), Z. Smoreda (4), V.D. Blondel (1,5)

1. Department of Mathematical Engineering, Université catholique de Louvain
2. Department of Electrical and Electronic Engineering, University of Melbourne
3. Computer and Automation Research Institute (SZTAKI), Hungarian Academy of Sciences
4. Sociology and Economics of Networks and Services Department, Orange Labs
5. Laboratory for Information and Decision Systems, MIT

11:30-11:45 *Chatty Mobiles: Individual mobility and communication patterns*

T. Couronné, Z. Smoreda, A.-M. Olteanu

Sociology and Economics of Networks and Services department, Orange Labs R&D

11:45-12:00 *Distance Matters: Geo-social Metrics for Mobile Location-based Social Networks*

S. Scellato, C. Mascolo

Computer Laboratory, University of Cambridge

12:00-12:15 *A tale of many cities: universal patterns in human urban mobility*

A. Noulas (1), S. Scellato (1), R. Lambiotte (2), M. Pontil (3), C. Mascolo (1)

1. Computer Laboratory, University of Cambridge
2. Department of Mathematics, University of Namur
3. Computer Science Department, University College London

12:15-13:30 *Lunch break*

13:30-14:00 *Keynote talk: Human Dynamics and Cell Phones: From mobility to predictability*

A.-L. Barabási, Northeastern University

SESSION C (14:00-15:45). Social networks

14:00-14:15 *Finding Meaningful Usage Clusters From Anonymized Mobile Call Detail Records*

R. Becker (1), R. Cáceres (1), C. Han (2), K. Hanson (1), J. M. Loh (1), S. Urbanek (1), A. Varshavsky (1), C. Volinsky (1)

1. AT&T Labs - Research
2. Virginia Tech University

14:15-14:30 *Using Cellular Network Data for Urban Planning*

R. Becker, R. Cáceres, K. Hanson, J. M. Loh, S. Urbanek, A. Varshavsky, C. Volinsky

AT&T Labs - Research

14:30-14:45 *Segmentation of towns using call detail records*

R. Guigourès, M. Boullé

Orange Labs

14:45-15:00 *Understanding the evolution of spatial structure in urban communities*

F. Walsh, A. Pozdnoukhov

National Centre for Geocomputation, National University of Ireland Maynooth

15:00-15:15 *Trace-driven analysis of data forwarding in opportunistic networks*

M. Karaliopoulos, P. Pantazopoulos, E. Jaho, I. Stavrakakis

Department of Informatics and Telecommunications, National & Kapodistrian University of Athens

15:15-15:30 *Statistical detection of information flow traces in a large phone call dataset*

L. Tabourier (1), F. Peruaniy (2)

1. LIP6, Université Pierre et Marie Curie
2. Max Planck Institute for the Physics of Complex Systems

15:30-15:45 *Socially-mediated diffusion via cell phones: an analysis of ringback diffusion in a cell phone network*

I. Kiani (1), D. Ruths (2), T. Shultz (3)

1. John Molson School of Business, Concordia University
2. School of Computer Science, McGill University
3. Department of Psychology, McGill University

15:45-16:15 *Coffee break*

SESSION D (16:15-17:30). Population mobility

16:15-16:30 *Estimating Dynamic Urban Population Through Assimilated Mobile Phone Data*

T. Horanont (1), R. Shibasaki (2)

1. Institute of Industrial Science, University of Tokyo
2. Center for Spatial Information Science, University of Tokyo

16:30-16:45 *Modelling City Population Dynamics from Cell Phone Usage Data Streams*

C. Kaiser, A. Pozdnoukhov

National Centre for Geocomputation, National University of Ireland Maynooth

16:45-17:00 *Exploiting cellular network MSC counters for the analysis of temporary populations*

P. Tagliolato, F. Manfredini, C. Di Rosa

Dipartimento di Architettura e Pianificazione, Politecnico di Milano

17:00-17:15 *Mobility and Predictability of Population Movements after the Haiti 2010 Earthquake*

X. Lu (1,2), L. Bengtsson (1), P. Holme (3)

1. Department of Public Health Sciences, Karolinska Institutet
2. Department of Sociology, Stockholm University
3. IceLab, Department of Physics, Umea University

17:15-17:30 *Anonymizing Location Data Does Not Work*

H. Zang (1), J. Bolot (2)

1. Sprint
2. Technicolor

18:00 - 19:30 *Dinner*

starting at 19:30 Progressive Rock event at the Medialab in the 3rd Atrium. Joe Paradisio will showcase amazing tech, including a 2 story video wall.

Tuesday October 11, 2011

SESSION E (9:30-10:30). Social structure

9:30-9:45 *Using Randomization Methods to Identify Social Influence in Mobile Networks*

R. Belo, P. Ferreira

Carnegie Mellon University

9:45-10:00 *Quantifying and Modelling Good Communicators in Dynamic Phone Networks*

A. Mantzaris, D. Higham

Department of Mathematics and Statistics, University of Strathclyde

10:00-10:15 *Examining the Social Decomposition of Mobile Call Graphs*

D. Doran (1), V. Mendiratta (2), C. Phadke (2), H. Uzunalioglu (2)

1. Dept. of Computer Science & Engineering, University of Connecticut
2. Bell Laboratories, Alcatel-Lucent

10:15-10:30 *Time allocation in social networks: correlation between social structure and human dynamics*

G. Miritello (1,2), R. Lara (2), E. Moro (1,3,4)

1. Departamento de Matemáticas & GISC, Universidad Carlos III de Madrid
2. Telefónica Research
3. Instituto de Ciencias Matemáticas CSIC-UAM-UCM-UC3M
4. Instituto de Ingeniería del Conocimiento, Universidad Autónoma de Madrid

10:30-11:00 *Coffee break*

SESSION F (11:00-12:15). Dataset mining

11:00-11:15 *Correlated bursty behaviour in human communication*

M. Karsai (1), K. Kaski (1), A.-L. Barabási (2,3), J. Kertész (3,1)

1. BECS, School of Science, Aalto University
2. Center for Complex Networks Research, Northeastern University
3. Institute of Physics, Budapest University of Technology and Economics

11:15-11:30 *Ethnic Segregation in the Area of Residence, Work, and Free-time Evidence from Mobile Communication*

O. Toomet (1,2), S. Silm (2), E. Saluveer (2), R. Ahas (2)

1. Department of Economics, Tartu University
2. Department of Human Geography, Tartu University

11:30-11:45 *Risk and Reciprocity Over the Mobile Phone Network: Evidence from Rwanda*

J. Blumenstock (1), Nathan Eagle (2), M. Fafchamps (3)

1. University of California, Berkeley
2. Santa Fe Institute
3. Oxford University

11:45-12:00 *Composite Social Network for Predicting Mobile Apps Installation*

W. Pan, N. Aharony, A. Pentland

MIT Media Laboratory

12:00-12:15 *Mobile phone and motorway traffic: a panel data perspective*

E. Tranos, J. Steenbruggen, P. Nijkamp, H. Scholten

Dept. of Spatial Economics, VU University

12:15-13:30 *Lunch break*

13:30-14:00 *Keynote talk: Influence Networks*

A. Pentland, MIT

SESSION G (14:00-15:30). Churn and value

14:00-14:15 *Churn Analysis in Mobile Telecom Data using Hybrid Paradigms*

V. Yeshwanth, M. Saravanan

Ericsson R & D

14:15-14:30 *Relational Learning for Customer Churn Prediction: The Complementarity of Networked and Non-Networked Classifiers*

W. Verbeke (1), T. Verbraken (1), B. Baesens (1,2,3)

1. Department of Decision Sciences and Information Management, Katholieke Universiteit Leuven
2. School of Management, University of Southampton
3. Vlerick, Leuven-Gent Management School

14:30-14:45 *Network neighbor effects on customer churn in cell phone networks*

P. Krivitsky, P. Ferreira, R. Telang

iLab, H. John Heinz III College, Carnegie Mellon University

14:45-15:00 *Subscriber Behaviour in a Cellular Network implementing Dynamic Pricing for Voice Calls*

H. Wang, L. Kilmartin

Electrical & Electronic Engineering, National University of Ireland

15:00-15:15 *What is the Economic Value of Cell Phone Location Data?*

F. Baccelli (1), J. Bolot (2)

1. ENS/INRIA
2. Technicolor

15:15-15:30 *Collective response of human populations to large-scale emergencies (invited talk)*

J. Bagrow

Center for Complex Networks Research, Northeastern University

Workshop on the Analysis of Mobile Phone Networks

MIT (Cambridge, USA)
October 10-11, 2011

Abstracts

Identifying Important Places in People's Lives from Cellular Network Data

Sibren Isaacman[◇], Richard Becker[†], Ramón Cáceres[†],
Stephen Kobourov[⋆], Margaret Martonosi[◇], James Rowland[†], Alexander Varshavsky[†]

[◇] Princeton University, Princeton, NJ, USA
[†] AT&T Labs – Research, Florham Park, NJ, USA
[⋆] Dept. of Computer Science, University of Arizona, Tucson, AZ, USA

[◇] {isaacman,mrm}@princeton.edu
[†] {rab,ramon,jrr,varshavsky}@research.att.com
[⋆] kobourov@cs.arizona.edu

While people travel further and faster than ever before, it is still the case that they spend much of their time at a few important places. Identifying these key locations is thus central to understanding human mobility and social patterns. Such understanding can, in turn, inform solutions to large-scale societal problems in fields as varied as telecommunications, ecology, epidemiology, and urban planning. As an example, knowing how large populations of people move about would help determine their carbon footprint and in turn help guide policies intended to reduce that footprint.

Wireless cellular networks hold great potential for providing the necessary information to identify important places in people's lives. The growing ubiquity of cellular phones means that a large percentage of people keep a phone with them most of the time. In addition, the networks need to know roughly where each phone is in order to provide the phones with voice and data services.

In this work, we explore the use of anonymized Call Detail Records (CDRs) from a cellular network to estimate the locations of important places in the lives of large populations of people. CDRs are routinely collected by cellular network providers to help operate their networks, for example to identify congested cells in need of additional bandwidth. Each CDR contains information such as the time a voice call was placed or a text message was received, as well as the identity of the cell tower with which the phone was associated at that time. This information can serve as sporadic samples of the approximate locations of the phone's owner.

CDRs are an attractive source of location information for two main reasons. One, they are collected for all active cellular phones, which number in the hundreds of millions in the US and in the billions worldwide. Two, they are already being collected to help operate the networks, so that additional uses of CDR data incur little marginal cost. Contrast this low cost, for example, with the expense of carrying out surveys to ask people where they spend their time. This high expense generally limits other data collection methods to orders of magnitude fewer participants.

However, CDRs have two significant limitations as a source of location information. One, they are sparse in time because they are generated only when a phone engages in a voice call or text message exchange. Two, they are coarse in space because they record location only at the granularity of a cell tower. It is an interesting research question whether CDRs can be used to identify important places in people's lives.

We show that applying clustering and regression techniques to CDR data can indeed identify important places in people's lives. First, we present an algorithm for identifying important places. Then, we describe two other algorithms for selecting home and work locations from among those important places. We validate all three algorithms by comparing their results to ground truth provided by a group of volunteers. We then apply these algorithms to much larger anonymous populations in the Los Angeles (LA) and New York City (NY) areas. Our LA and NY dataset spans two months of activity for hundreds of thousands of phones, yielding hundreds of millions of location samples.

Finally, we present two example applications of these techniques. We start by using the home and work locations identified by our algorithms to calculate the distribution of commute distances per postal code in our Los Angeles and New York dataset. We then estimate the carbon footprints of those commutes, also aggregated by postal code.

Privacy Measures.

Given the sensitivity of the data, we took several steps to ensure the privacy of individuals in our CDR dataset. First, only anonymous records were used in this study. In particular, personally identifying characteristics were removed from our CDRs. CDRs for the same phone are linked using an anonymous unique identifier, rather than a telephone number. Second, all our results are presented as aggregates. No individual anonymous identifier was singled out for the study. Finally, each CDR only included location information for the cellular towers with which a phone was associated at the beginning and end of a voice call or at the time

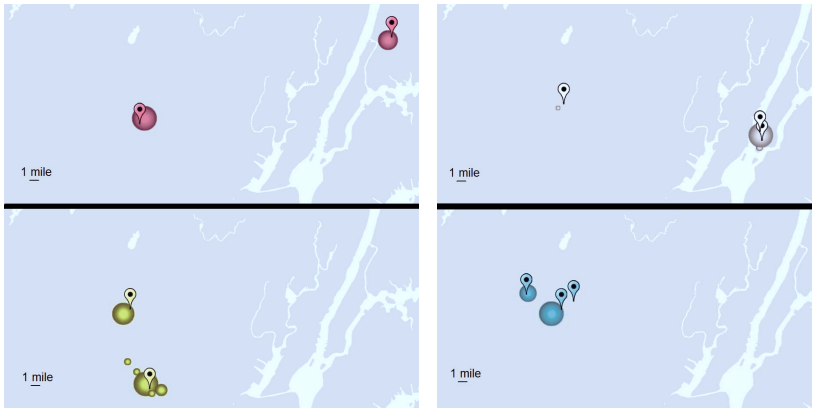


Figure 1: **True important locations vs. discovered important clusters for volunteers. Paddles are the important locations provided by the volunteers. Circles are the important clusters discovered by our algorithm, with their radii signifying days of use. The four examples are drawn to the same scale.**

of a text message. The phones were effectively invisible to us aside from these events. Furthermore, we could estimate the phone locations only to the granularity of the cell tower coverage radius, which is often larger than 1 mile [5].

In addition to obtaining anonymized CDRs for hundreds of thousands of phones in LA and NY, we recruited several dozen volunteers who gave us the true locations of important places in their lives, as well as permission to inspect their CDRs. An important place was defined as a location where a person spends a lot of time or visits frequently.

Finding Important Places.

We demonstrate an accurate and efficient algorithm for identifying *Important Places* from CDRs.. Our algorithm is the first to operate on the majority of cellular phone users, rather than relying either on more continuous and fine-grained tracking (e.g. GPS) or focusing on high-call-rate users whose mobility is easier to analyze.

Our algorithm proceeds in two stages. In the first stage, we spatially cluster the cell towers that appear in a user's trace. We pre-sort the towers in descending order according to the number of days a tower was contacted ("tower-days"), then we cluster them according to Hartigan's leader algorithm [3]. In the second stage, we identify which of the clusters are important using a model derived from a logistic regression of volunteers' CDRs.

To determine if a cluster is important, we use logistic regression. We considered five observable factors and ten derived variables. Of the 15 factors, three were statistically significant. First, the percentage of total "tower-days" this cluster represents. Second, the number of days between the first and last contact with any cell tower in the cluster. Third, the number of days on which any cell tower in the cluster was contacted compared to other clusters from this user.

$$Prob(x_1, ..., x_n) = \frac{1}{(1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n})} \quad (1)$$

Equation 1 shows the general form of the logistic regression formula that we use to estimate the likelihood of the importance of a cluster. In this formula, $Prob(x_1, ..., x_n)$ is the probability that a cluster with factors x_i is the closest cluster to an important location and the β_js are coefficients that are discovered during the regression. Once we determine the likelihood that a particular cluster is "important," we discard clusters with a low probability of being important.

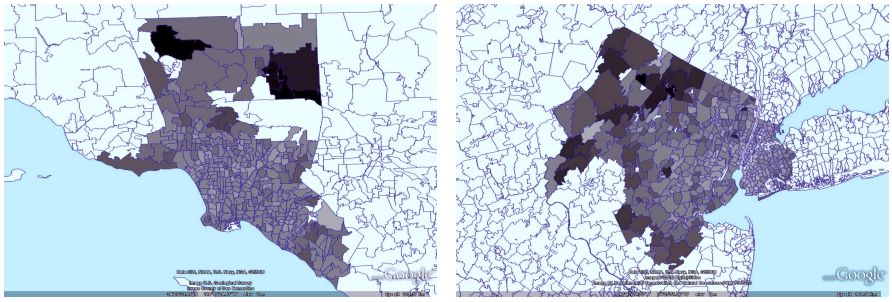
The effects of applying the *Important Places* algorithm can be seen in Figure 1 along with true important locations provided by volunteers. We come within 3 miles of the truth for 88% of the important locations of our volunteers.

Finding Home and Work.

We propose and evaluate two algorithms for applying semantic meaning to important locations, namely *Home* and *Work*, using models also derived via logistic regression. From the important clusters identified by the Important Places algorithm described above, our *Home* and *Work* algorithms select the clusters that correspond to where a person lives and works, respectively. The algorithms are independent and Home and Work may end up in the same cluster.

To select Home or Work, the relevant algorithm (i.e., either the Home or Work algorithm) calculates a score for each important cluster using coefficients obtained from a logistic regression. The algorithm then assigns the cluster with the highest score to be Home or Work. To calculate a score for a cluster, we use the logistic regression formula shown in Equation 1. In this case, $Prob(x_1, ..., x_n)$ is the score calculated for each cluster, x_i is the value of the ith factor and β_js are regression coefficients fitted during training. Home and Work, then, are chosen as the clusters with the highest probability as computed by the coefficients given by the logistic regression.

For the Home algorithm, the dominant factor is the cluster with the largest number of events during the "home" hours, which is selected as Home. For the Work algorithm, there are two dominating factors. The first is the ranking of the cluster based on the number of events occurring during "work" hours. The second factor is the percentage of the events occurring during "home" hours in the cluster. A cluster is assigned a higher score by the Work algorithm if the percentage of the home hour events in the cluster is low. Once again, we apply our algorithms to a test set of volunteers. We estimate home and work locations with median



(a) Los Angeles (b) New York

Figure 2: Heat maps of median carbon emitted per person for each direction of a commute in the ZIP codes in our study. Darker ZIP codes denote larger carbon footprint. Note that all NY and LA ZIP codes are colored according to the same scale.

errors of less than 1 mile.

Calculating Commutes.

We test our approaches on a dataset that is more universal than prior work in several ways. First, it is simply larger than prior work in terms of CDRs and number of users. Second, it covers multiple distinct geographic areas. Third, it considers users with a wide variety of call/text rates, from as low as a few calls/texts per week up to dozens of calls/texts per day.

We apply the Home and Work algorithms to two large metropolitan areas in the United States (US): Los Angeles (LA) and New York (NY). By defining commute distance as the distance as the crow-flies between home and work, we compare commute distances as calculated by our Home and Work algorithms to those derived from US Census statistics. In particular, we estimate the average commute distance for residents of the 891 ZIP codes in our CDR dataset to be 21 and 20 miles for the Los Angeles and New York areas, respectively. Using tables of where people live and work published by the US Bureau of Transportation Statistics [2], we calculate the average commute for residents of the same ZIP codes to be 21 and 19 miles for the Los Angeles and New York areas, respectively. This very close match between our algorithms and the census results further validates our approach. It is also important to note that the low cost of our approach makes it practical to regenerate current statistics much more frequently than with a census, for example every few months instead of every ten years.

Carbon Footprints.

Finally, we show of how technology providers and policy makers might use our algorithms in their work. In particular, we combine the commute numbers we have calculated with publicly available data to estimate the carbon footprints of those commutes in two major metropolitan areas.

We determine the mode of transportation of commuters at the ZIP code level using US census data. Specifically, we used Table P30 from the 2000 US census (Summary File 3): "Means of Transportation to Work for Workers 16+ Years." [6] to calculate the percentage of commuters that uses a particular mode of transportation per ZIP code. Further, combining this information with the amount of carbon dioxide emitted per person by each mode of transportation [1] allows us to compute the rough amount of carbon dioxide emitted per commuter. Aggregating commuters at the ZIP code level allows us to generate a distribution of carbon dioxide emissions per commuter in each ZIP code.

Figure 2 shows heat maps of LA and NY, where shading corresponds to the median carbon emission per person in each ZIP code in each direction of a commute. In the New York area, increasing distance from Manhattan correlates with increasing carbon footprint. In contrast, Los Angeles is fairly uniform throughout, with the exception of certain parts of Antelope Valley (in the northeast part of the map), which are separated from downtown LA by a mountain range that must be driven around. These patterns match well with what would be expected from both cities.

Generating carbon footprint estimates is a good example of how our technique for computing commuting distances can be combined with already available data to produce new and previously difficult to obtain information.

Conclusion.

Overall, our clustering and location algorithms form a foundation for a range of accurate, low-overhead analyses of human movement and social patterns. Our work demonstrates that call and text records from cellular networks represent an unobtrusive and accurate way to gather large-scale mobility data. Furthermore, the large degree of aggregation and anonymization allows us to usefully employ this data without unduly impinging on the privacy of any individual. For more details on our data set, our analysis methodology, and our results, we refer the reader to our full paper [4].

1. References

[1] M. J. Bradley and Associates. Comparison of energy use & CO_2 emissions from different transportation modes. Report to American Bus Association, 2007.

[2] US Bureau of Transportation Statistics. Downloaded from http://www.transtats.bts.gov.

[3] J. A. Hartigan. *Clustering Algorithms*. John Wiley & Sons, New York, 1975.

[4] S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, M. Martonosi, J. Rowland, and A. Vasharsvky. Identifying important places in peoples lives from cellular network data. In *Intl. Conference on Pervasive Computing*, 2010.

[5] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási. Limits of predictability in human mobility. *Science*, 327, 2010.

[6] US Census Data. Downloaded from http://www.census.gov.

Measuring repeated visitations with mobile positioning data. Applications for marketing.

Rein Ahas[1], Margus Tiru[2], Andres Kuusik[3]

[1] Department of Geography, University of Tartu , rein.ahas@ut.ee
[2] Positium LBS, 9 Õpetaja St., Tartu , margus.tiru@ut.ee
[3] Faculty of Economics, University of Tartu, andres.kuusik@ut.ee

Abstract
The aim of this paper is to elaborate a mobile positioning based methodology to measure repeated visitations and place loyalty. Repeated visitations show behavioural loyalty of people, ties to certain places, attractions or events. Loyal visitors are highly appreciated in marketing, since they are of greater benefit than random customers; they need different marketing strategies than those required to attract new visitors.
We developed a model to determine repeat visits of tourists and domestic people to a destination using passive mobile positioning data. Model uses the number, duration, frequency and geography of visits in order to identify repeat visitors in a particular area. We tested the algorithm using a database of foreign and local visitors during past 5 years. The algorithm was applied in destination marketing solutions for Estonian tourism authorities. Further, it has potential for various LBS can solutions.

Classifying Routes Using Cellular Handoff Patterns

Richard A. Becker, Ramon Caceres, Karrie Hanson, Ji Meng Loh,
Simon Urbanek, Alexander Varshavsky and Chris Volinsky
AT&T Labs - Research
{rab, ramon, karrie, loh, urbanek, varshavsky, volinsky} @research.att.com

Urban planners are interested in understanding the mobility patterns of the people who live in and use their cities. This understanding facilitates effective solutions to problems with traffic congestion, parking, vehicular and pedestrian safety, and other aspects of urban living. To gain some knowledge of mobility patterns, planners currently use a combination of census data and vehicle counting. However, the expense of these methods typically results in infrequent data collection and/or small population samples.

Cellular telephone networks have the potential to provide near real-time information about human mobility on a large scale and at a low cost. These networks must know the approximate locations of all affiliated cell phones in order to provide the phones with voice and data services. Since people usually carry their cell phones with them, the location of a phone is a good proxy for the location of its owner.

This abstract explores the use of cellular handoff patterns to identify which routes people take into and out of a city. A handoff pattern is the sequence of cellular antennas that a moving phone uses while engaged in a voice call. The main challenge in translating these patterns into useful route information is location inaccuracy due to the large geographic areas covered by individual antennas: often more than one square mile. Therefore, knowing which antenna a phone is connected to does not immediately reveal what route the phone's owner is traveling on. However, it is possible that knowing sequences of antennas can yield enough information to reveal these routes.

In this abstract, we investigate the use of handoff patterns extracted from anonymized Call Detail Records (CDRs). CDRs are collected when a phone is involved in a call, and may contain the full sequence of antennas used by the phone during that call. CDRs are routinely collected by network operators for all active cellular phones, which number in the hundreds of millions in the US and billions worldwide. Furthermore, CDRs are already used for network operation and planning, so additional uses incur little marginal cost. Another advantage of CDRs is that they are generated inside the network and thus do not place any extra burden on the limited resources of mobile phones. We note, however, that our route identification techniques are independent of how cellular handoff patterns are recorded, whether by the network or by the phones.

Our work explores the following research questions: (a) Are handoff patterns stable across a wide enough range of conditions to be used for identifying routes? (b) Can we devise algorithms that reliably match handoff patterns to routes? (c) Can we derive reliable route utilization statistics from cellular network data?

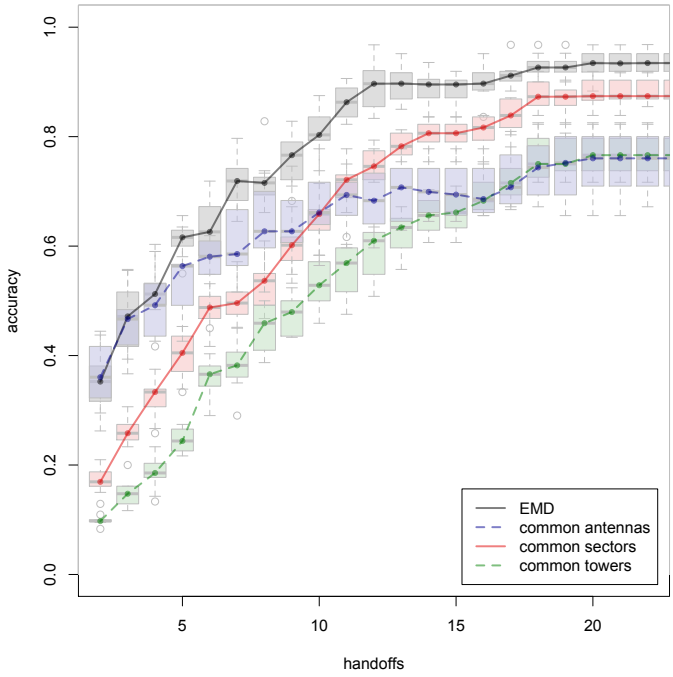


Fig. 1. A comparison of nearest-neighbor classification using four distance metrics, with four handoff patterns as training and four as testing per route. Accuracy is shown as a function of the number of handoffs per drive. The boxplots show the range of accuracy over 10 randomly chosen training sets and the colored lines connect the medians of those sets. The EMD metric out-performs the others.

To answer these questions, we conducted a study of 15 driving and train routes leading into Anytown, a suburban city with roughly 20,000 residents. We used cellular phones to maintain active voice calls while we drove a car and rode the train on these routes. Then, we used the corresponding handoff patterns (obtained from CDRs) to evaluate different route classification techniques. Finally, we applied our best performing techniques to 60 days of anonymized CDRs for all calls handled by one cellular carrier in the Anytown area. Due to the lack of space, we only summarize our key contributions here. Please see our full paper [1] for details.

A. *Stability of Handoff Patterns*

We found empirically that cellular handoff patterns on a given route are stable across phone models, weather conditions, traffic conditions and time. This stability allows captur-

ing the "typical" handoff pattern for a given route at one time and use it for route classification at another time.

B. Route Classification Algorithms

We developed two algorithms for matching handoff patterns to routes and showed that they have comparable accuracy. The first uses nearest neighbor classification based on Earth Mover's Distance [3]. The second uses signal strength data to compute the likelihood that a given handoff pattern occurs on a particular route. Due to lack of space, we discuss only the first algorithm in this abstract.

We collected data on 13 commuter routes and 2 train routes leading into Anytown. These routes represent all major ways to get in and out of the city. The route lengths vary from 3 to 6 miles. We travelled each route four times, two in each direction, primarily in the Fall of 2010, with a few fill-in drives and train rides in March of 2011. During each drive, there were two phones of different models in the car, one calling the other. In total, we collected $4 \times 2 \times 15 = 120$ handoff patterns.

Our classification is done via a nearest-neighbor algorithm. For each route, we split the 8 test drives randomly into equal sized training and test sets. For each instance in the test set, we assign the route label of the nearest instance from the training set. The choice of distance metric is crucial for determining the nearest neighbor. In the following three subsections, we describe the four distance metrics we evaluated and discuss their relative performance.

1) Common Subset Distances: Distances between two handoff patterns can be defined by measuring how much the two patterns have in common. These distances are based on attributes of antennas in the handoff patterns. The larger the intersection between these sets of attributes, the more similar the handoff patterns. We refer to these distances as *common-subset* distances. We defined three common-subset distances that compare these attributes at different levels of granularity: cell towers, sectors, and antennas. A cell tower is a physical structure that holds radio antennas. A sector corresponds to a direction from a given cell tower. Each sector is covered by one or more antennas, the physical devices that communicate with mobile phones and service a particular cellular technology (e.g., UMTS) and frequency (e.g., 2.1 GHz). The *Common Antennas* distance between two handoff patterns is the number of antennas that appear in both handoff patterns. Similar definitions apply to *Common Sectors* and *Common Towers* distances.

2) Earth Mover's Distance: Although common subset distances are good for baselines, they do not account for three important characteristics of the handoff pattern. First, the sequential nature of the handoff pattern is lost, basically reduced to an unordered set of entities. Second, temporal information on how long the call spends on each tower is not used by these algorithms. Finally, the cell tower locations are not accounted for. Two patterns that differ only by towers that are close to each other should be considered close patterns.

We propose a variant of *Earth Mover's Distance* (EMD) as a distance metric that accounts for all of these characteristics. EMD is traditionally used to measure the differences between images. Conceptually, imagine a pair of two dimensional images, where each pixel's brightness value is represented by a pile of dirt on that pixel. Pixels with similar brightness have similar amounts of dirt. Now, consider the energy needed to transform one image into the other by moving the piles of dirt. EMD is defined as the minimal energy needed to move the mass of dirt of one image into the locations that result in the target image. Similarly, EMD is defined for arbitrary probability distributions as the mass of probability that needs to be moved to turn one distribution into another. Applying EMD in a nearest-neighbor classification task is a straightforward calculation of EMD between the test instances and those in the training set, then selecting the route label of the nearest one.

3) Performance of the Classification Algorithms: We are interested in measuring how much information is needed on a single drive to classify it correctly. Since all of our routes emerge from the center of Anytown, we assessed the accuracy of our algorithms for a call starting in the center of Anytown as a function of the number of handoffs the algorithm is allowed to observe. To generate results, we split our data into training and test sets, with four randomly selected drives for each route in each set. We fit each test instance to its nearest neighbor in the training set using our four distance metrics, EMD and the three common subset distances, repeating this procedure for 10 different random selections of the training set.

Figure 1 shows our results, with boxplots representing the accuracy across all replications for each distance metric and number of handoffs. Colored lines connect medians of the boxplots. The figure shows that, in general, the prediction accuracy increases as the number of handoffs increases up to a saturation point for all metrics. That is, the farther away the phone moves from the Anytown center, the easier it is to differentiate between routes. The EMD metric performs the best, achieving a median classification accuracy of 90% after 12 handoffs (corresponding roughly to 2 miles). The Common Towers metric performs the worst because it cannot differentiate between handoffs occurring between antennas on the same cell tower. Interestingly, the Common Antennas metric outperforms the Common Sectors metric for up to 10 handoffs, but then performs worse because sometimes phone calls on the same route are handled by different antennas pointing in the same direction.

C. Estimating Relative Traffic Volumes

We showed how CDRs, in combination with our algorithms, can be used to determine the relative traffic volumes on roads and validate these results against statistics published by a state transportation authority.

1) CDR Data Collection: We collected anonymized CDRs from calls carried by the 35 cell towers (about 300 antennas) located within 5 miles of the center of Anytown. Our goal was to capture cellular traffic in and around the city and choosing the 5-mile radius allowed us to cover both Anytown proper and its neighboring areas. We collected voice traffic for 60 days between November 29, 2009, and January 27, 2010, resulting in 15 million voice CDRs for 475,000 unique phones.

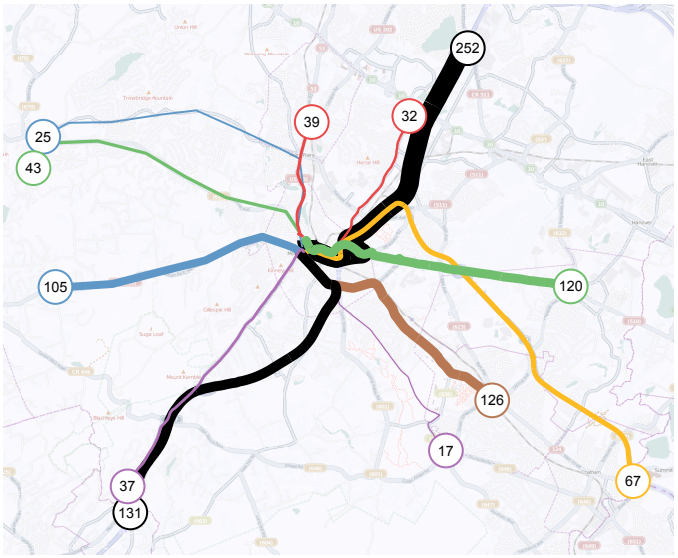


Fig. 2. Predicted traffic distribution of 1000 calls. The EMD algorithm was used to estimate traffic volumes on the 13 commuter routes into Anytown-which are shown as flows into Anytown, with the line width proportional to the estimated volume; the Signal Strength algorithm results are similar.

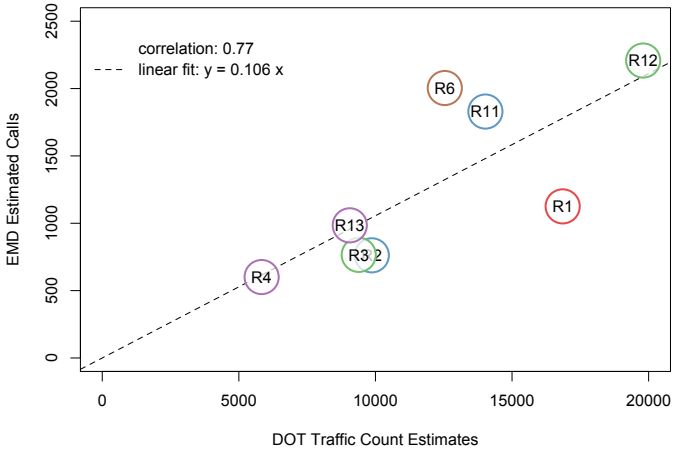


Fig. 3. Comparison of traffic volumes estimated by our EMD algorithm against corresponding values obtained from the Anystate Department of Transportation.

2) Privacy: Given the sensitivity of CDR data, we took several steps to ensure the privacy of individuals. First, only anonymous records were used in this study. The data was collected and anonymized by a party not involved in the data analysis. Personally identifying characteristics were removed from our CDRs. CDRs for the same phone are linked using an anonymous unique identifier, rather than a telephone number. No demographic data is linked to any cell phone user or CDR. Second, all our results are presented as aggregates. That is, no individual anonymous identifier was singled out for the study. By observing and reporting only on the aggregates, we protect the privacy of individuals. Finally, each CDR only included location information for the cellular antennas associated with a phone during a voice call. The phones were effectively invisible to us outside those times, and we only knew those phone locations at the granularity of an antenna's coverage area, often greater than one square mile.

3) Analysis: Figure 2 plots relative traffic volumes as estimated by the EMD algorithm, overlaid on our map of routes into and out of Anytown. The counts are normalized to a count per 1000 calls. The thickness of the line represents the volume estimated on that route. The plot allows comparing the relative number of people who access the town from north and south on the interstate (the black lines) vs. the relative number of people who enter and leave Anytown on secondary state or county roads.

4) Validation: We present our results as relative volumes instead of absolute volumes since there are many factors that play into whether we will see a particular traveling vehicle: the phone must be active, the user must be a customer of the cellular provider that supplies the data, the call must use five unique towers (to ensure that it is moving, not stationary), and so on. Because of this, validation of our numbers against readily available government traffic count data is challenging. An additional challenge is that the government traffic data is typically collected using in-street traffic meters or human car counters. Both methods give a count at a static point, while our methods are estimating traffic along a particular route made up of many sequential points.

Nonetheless, we attempted to validate our data with available traffic count information from the Anystate Department of Transportation (DOT). The DOT has a multitude of data available at strategic places around town from traffic meters over the years 2004-2010 [2]. Most of our routes had multiple traffic measurements available at different locations along the route.

Because Anytown is a local hub, we made the simplifying assumption that all of the traffic on the secondary roads heading into town either originated or terminated at the town center. Since this assumption might not hold for the main interstate running through town, we did not include those routes in this analysis. Also, we did not have any traffic counts for any part of one of our routes, so we excluded that route as well. Additionally, if a route had multiple measurements of daily traffic at different locations, we selected the minimum of these counts. In most cases, this minimum count was at the furthest point from Anytown, and hence was the best estimate of the number of cars that had travelled the entire length of the route. Finally, we removed a single data point from the DOT data that was highly suspicious as an outlier; it was incongruous with nearby data points in a way that made it appear to be erroneous. Figure 3 shows a scatterplot of EMD-estimated traffic counts vs. the DOT-supplied traffic counts. The figure shows that the two estimates are closely related, with a correlation coefficient of 0.77. We believe that this result validates our methodology.

REFERENCES

[1] R. Becker, R. Cáceres, K. Hanson, J. M. Loh, S. Urbanek, A. Vasharsvky, and C. Volinsky. Route classification using cellular handoff patterns. In *Proc. of the 13th ACM International Conference on Ubiquitous Computing*, 2011.

[2] NJ Department of Transportation. Roadway information and traffic counts. www.state.nj.us/transportation/refdata/roadway/traffic.shtm.

[3] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *Int. J. Comput. Vision*, 40:99–121, November 2000.

Characterizing Urban Road Usage Patterns with a New Metric

Pu Wang[1,2], Timothy Hunter[3], Alex Bayen[3] and Marta C. González[1]

[1]*Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, 02139, USA*

[2]*School of Traffic and Transportation Engineering, Central South University, Changsha, Hunan, 410000, P.R. China*

[3]*Department of Civil and Environmental Engineering, University of California, Berkeley, Berkeley, CA, 94720, USA*

Email: Pu Wang*: puwang@mit.edu,* Timothy Hunter*: tjhunter@eecs.berkeley.edu,* Alex Bayen*: bayen@berkeley.edu,* Marta C. González*: martag@mit.edu*

Global communication through mobile phones is a massive phenomenon in urban centres from the entire planet. This generates petabytes of information that contains fingerprints of individual human activity from remote locations. This sea of information has revolutionized the opportunities for sensing and modelling human travel, offering us a real-time large scale 'survey' incessantly recording people's everyday communication (*1*), community (*2*) and mobility information (*3, 4*). Comparing to the survey and traffic count data, mobile phone data is less expensive and widely available in most regions. These advantages have attracted many attentions in using mobile phone data to study urban transportation, yet most of the works only estimated relatively small OD matrices (*5, 6*) on some small or conceptual road networks (*7, 8*). To date it is still missing the link to quantify the interactions between the individual mobility extracted from mobile phone data and the real and detailed road infrastructures at a city or even larger scale.

We use mobility data from half million anonymized mobile phone users to study the road usage patterns in San Francisco Bay Area. Through our modelling framework each trip's route is predicted and recorded, surprisingly, we find that averagely 60% of vehicles passing through a road segment come from 1% of Bay Area drivers' home locations, suggesting a high predictability of vehicle sources. For each road segment we apply Gini coefficient to quantify the heterogeneous traffic contribution by drivers living in different neighbourhoods. We find that Gini coefficient ranges from 0.77 to 1, hinting

that a road segment's traffic flow is highly unequally produced by drivers living in different neighbourhoods. Counter intuitively, a road segment's Gini coefficient is lowly correlated with its basic properties such as traffic volume and volume over capacity, suggesting that Gini coefficient is a new metric on top of the traditional measures, quantifying road usage patterns in the perspective of drivers' demographical distribution.

REFERENCES:

1. J.-P. Onnela et al., *Proc. Natl. Acad. Sci. U.S.A.* **104**, 7332 (2007).

2. G. Palla, A.-L. Barabási, T. Vicsek, *Nature* **446**, 664 (2007).

3. M. C. González, C. A. Hidalgo & A.-L. Barabási, Understanding Individual Human Mobility Patterns. *Nature* **435**, 779-782 (2008).

4. C. Song, Z. Qu, N. Blumm & A.-L. Barabási, Limits of Predictability in Human Mobility. *Science* **327**, 1018-1021 (2010).

5. N. Caceres, J.P. Wideberg & F.G. Benitez, Deriving origin–destination data from a mobile phone network, *IET Intell. Transp. Syst.*, **1**, (1), pp. 15–26 (2007).

6. J. Schlaich, T. Otterstätter & M. Friedrich, Generating Trajectories from Mobile Phone Data, *Transportation Research Part B* Annual Meeting CD-ROM (2010).

7. C. Ratti, D. Frenchman, R. M. Pulselli & S. Williams, Mobile Landscapes: using location data from cell phones for urban analysis, *Environment and Planning B: Planning and Design* **33**, 727 -748 (2006).

8. Y. Zhang, X. Qin, S. Dong & B. Ran, Daily O-D Matrix Estimation using Cellular Probe Data, *Transportation Research Part B* Annual Meeting CD-ROM (2010).

Individual Mobility Networks from Clusters of Human Activity

Shan JIANG [1], Joseph FERREIRA [1], Marta GONZÁLEZ [2]

[1] Department of Urban Studies and Planning, MIT, Cambridge, MA

[2] Department of Civil and Environmental Engineering, Cambridge, MA

Email: shanjang@mit.edu, jf@mit.edu, martag@mit.edu

Introduction

Recent innovation in both data sources and analytical approaches have inspired new studies about the dynamics of human activities (Gonzalez et al. 2008; Song et al. 2010). While deploying these new sources of data (such as mobile phone data), researchers have been facing significant challenges—due to privacy and legal constraints, no or very limited information on socioeconomics and demographics of the human subjects being studied and/or types of activities being conducted can be accessed in these circumstances. Despite the fact that these new datasets may allow researchers to study social relationship and networks (Eagle et al. 2009), they still have limited capacity in revealing underlying reasons driving human behavior (Nature Editorial 2008).

Compared with the aforementioned urban sensing data, survey data is disadvantaged by high cost, low frequency, and small sample size. However, in terms of socioeconomic and demographic information, survey data provides much richer and detailed information. We exploit the richness of survey data using data mining techniques, which have not been applied in this context before. Since the survey collected over the metropolitan area is conducted by the metropolitan planning organization (MPO) for the regional transportation planning purposes, it is a representative sample of the total population of the region (Chicago Metropolitan Agency for Planning 2008). In this study, we integrate and compare mobile phone activity with travel survey information, both collected in the Chicago Metropolitan Area.

Methodology

We first perform a time series analysis of each of the 30,000 surveyed individuals, dividing the entire day into 288 five-minute intervals; for each time interval we know the type of activity an individual is performing. We divide the self-reported activities into the following nine categories: (1) Home, (2) Work, (3) School, (4) Transportation Transitions, (5) Shopping/ Errands, (6) Personal Business, (7) Recreation/Entertainment/Friends, (8) Civil/ Religious, and (9) Other. To detect characteristic groups, we compute the principal components of the temporal activity matrix (30,000 x 2,592), in which each row represents an individual, and the columns are the 288 five-minute intervals of a day combined with the 9 types of activities.

The differences between our study and the traditional time-use studies on human activities lie in our methods. We do not superimpose any predefined social classification on the observations; we

group the results with K-means clustering. We let the inherent activity structure inform us of the patterns and clusters of individual temporal activities in the metropolitan area (Jiang et al. 2011). Figure 1 exhibits the K-means clustering results (with K=8) of individuals' temporal activity patterns on an average weekday along with the social demographic characteristics of those individuals grouped into each cluster based on their activity signatures.

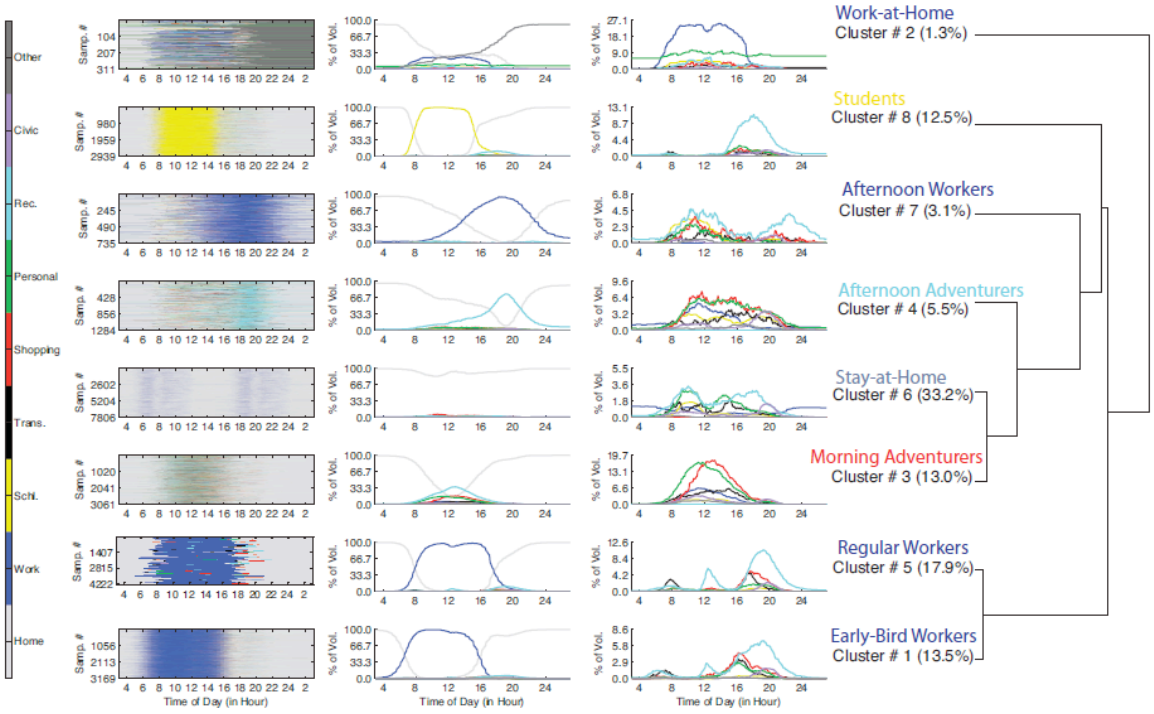


Figure 1: Clusters of individuals' weekday temporal activity patterns.

[Note: The color bar indicates the nine reported activity types (see previous text). Left panel shows the activity matrix of each cluster, each row is an individual one-day itinerary, the columns representing time of day and colors indicating the activity. Central panels show the percentage of individuals performing a particular activity vs. time of day, and left side plots is a zoom from the one on the right— the colors of each curve represent the activity types. Right panel represents the name assigned to the cluster given its particular behavior and the percentage of total population belongs to it.]

From each cluster we construct individual mobility networks (see examples in Figure 2). These networks define dynamic networks in which nodes represent the location of a particular activity and the links the trips between them. We extract the signatures of eight network groups that can be used to model urban trips. The first challenge is to see if we may find characteristic networks within each cluster that can be expressed with simple dynamic rules. Next, we explore individual mobility networks as extracted by the phone users and use the findings of the survey to try to complete some aspects of the daily activity of the mobile phone users. By placing those networks in real space, one can build the aggregated mobility network or OD matrix of the entire city.

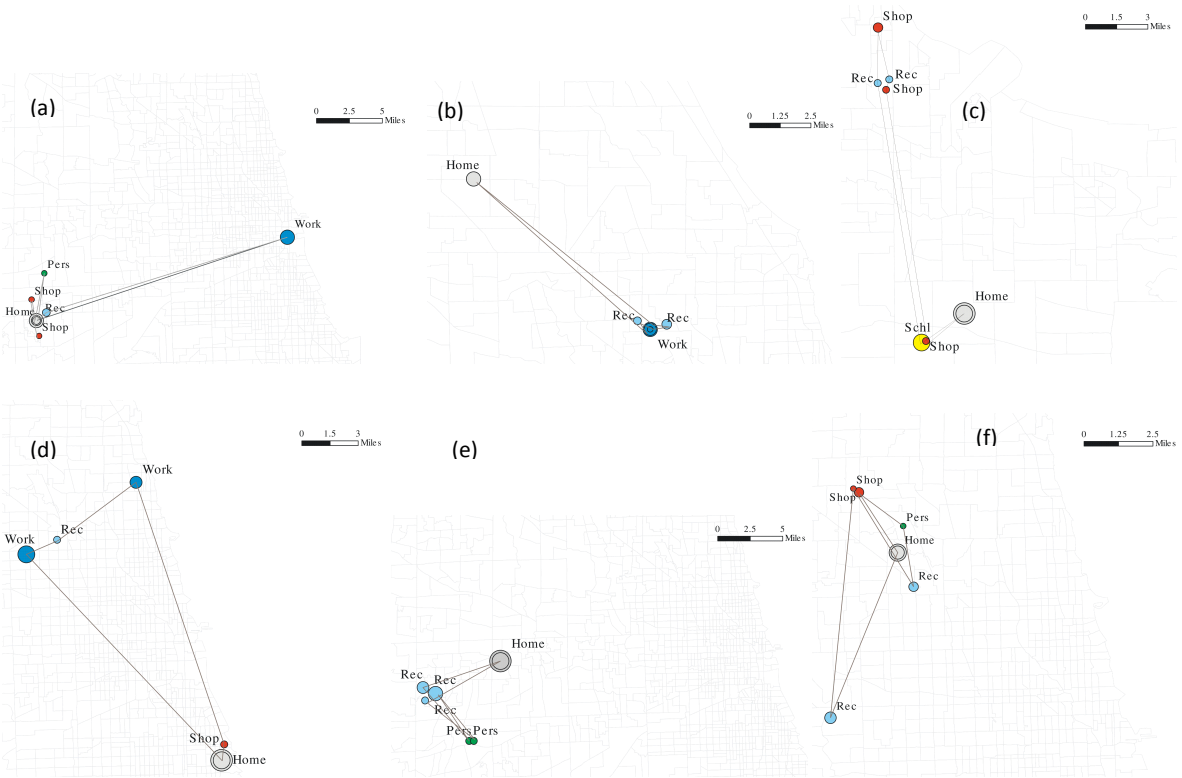


Figure 2 Examples of individual mobility networks of (a) an early-bird worker, (b) a regular worker, (c) a student, (d) an afternoon worker; (e) a morning adventurer, and (f) an afternoon adventurer. [Note: The examples of individual mobility network are extracted from samples of the 8 classified clusters defined in Figure 1 from the survey data. Nodes in each graph represent activity destinations, colors represent the activity types, and sizes of the nodes are proportionate to activity duration at a particular destination.]

Research Contributions

Similar data mining techniques, to the ones here proposed have successfully being applied to characterize human behavior, as for example, the extraction of patterns of temporal variations in on-line media content(Yang and Leskovec 2011) and activities within a college campus (Eagle and Pentland 2009). Daily activity data of groups of individuals in a city must have underlying structures that can be extracted using similar techniques. K-means and eigen decompositions combined with network analysis are particularly useful because they provide a low dimensional characterization of complex phenomena. The obtained classification related to the spatial information of the activities will provide clear information on where, when and how the individuals interact with places in the city, providing a clear framework for urban and transportation planning.

References

Chicago Metropolitan Agency for Planning. 2008. Chicago travel tracker household travel inventory (2007).

Eagle, N., and Pentland, A., 2009. Eigenbehaviors: Identifying structure in routine. *Behavioral Ecology and Sociobiology,* 63 (7), 1057-1066.

Eagle, N., Pentland, A., and Lazer, D., 2009. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*.

Gonzalez, M. C., Hidalgo, C. A., and Barabasi, A.-L., 2008. Understanding individual human mobility patterns. *Nature,* 453 (7196), 779-782.

Jiang, S., Ferreira, J., and Gonzalez, M., 2011. *Understanding the link between urban activity destinations and human travel patterns* Paper presented at the Computers in Urban Planning and Urban Management 2011 Conference, Lake Louise, Canada.

Nature Editorial. 2008. A flood of hard data. *Nature,* 453 (7196), 698-698.

Song, C., Qu, Z., Blumm, N., and Barabási, A.-L., 2010. Limits of predictability in human mobility. *Science,* 327 (5968), 1018-1021.

Yang, J., and Leskovec, J., 2011. Patterns of temporal variation in online media. *In: Proceedings of the fourth ACM international conference on Web search and data mining,* Hong Kong, China: ACM.

Structure and seasonality in human motion

James P. Bagrow[1,2,*] and Yu-Ru Lin[3,4,†]

[1]Center for Complex Network Research, Northeastern University, Boston, MA 02115.
[2]Center for Cancer Systems Biology, Dana-Farber Cancer Institute, Boston, MA 02115, USA.
[3]College of Computer and Information Science, Northeastern University, Boston, MA 02115, USA
[4]Institute for Quantitative Social Science, Harvard University, Cambridge, MA 02138, USA
[*]bagrowjp@gmail.com [†]yuruliny@gmail.com

July 15, 2011

Recently a number of intriguing results have emerged regarding the long-term properties of human motion, such as its high predictability and ultra-slow diffusive growth. Using a country-wide mobile phone dataset, we wish to explore the underlying structures of human motion. We follow a population of approximately 80,000 users over a 30-week period, using their individual trajectories to extract a mobility network between cellular towers. These networks show highly non-random structure, from heavily clustered central locations to long chains and loops. The evolution of the distribution of user relocations over long times allows us to explore seasonal effects.

Interest in human mobility has exploded due to the recent availability of new cutting edge datasets, such as dollar bill-tracking websites [1] and cell phone call records [2]. It has been shown that people can move over very broad spatial ranges. It is also known that people are very recurrent in their motion, showing strong daily periodicities; users are highly predictable [3]. Models have emerged to explain the ultraslow exploratory growth of human mobility. Interestingly, it has also been shown that a simple Markovian random walk can well approximate how people move between knowns locations [4]. (This paradoxical result is resolved when one realizes that this random walk requires a data-dependent transition matrix, and is not capable of generating it on its own. For that one needs a proper generative model, such as [5].)

We wish to explore further the underlying causes of these phenomena. Are their long-term seasonalities in human motion? What are the structures of such trajectories? For example, if users cluster around specific locations such as home and work, how often do they move between these clusters? [6] Do there exist extended periods of motion away from these clusters, where users anomalously prefer less visited—and thus less predictable—locations? Do users move more or less frequently during winter compared with summer? What demographic effects are present, such as due to age or gender?

We begin with a mobile phone dataset, from which we sample approximately 80k users, following the criteria of [3]. We follow each user's trajectory over a period of 30 weeks, considerably longer than previous studies. Each user i provides a trajectory τ_i based on the time-ordered sequence of cellular towers they visit, $\tau_i = [L(t_1), L(t_2), \ldots]$, where $L(t_i)$ represents the location that user i visited at time t_i. We take time at an hourly resolution. In Fig. 1(a) we draw an example mobility network, where each node represents a cellular tower and links represent relocations from one tower to another. Highly non-random structures such as several cores and a number of long-range, low-degree loops are present. In Fig. 1(b) we study the degree distributions of these Mob-Nets. They are well approximated by a power-law with an exponential cutoff. In Fig. 1(c) we show the temporal evolution, as a function of time of year and hour of week, of the call patterns

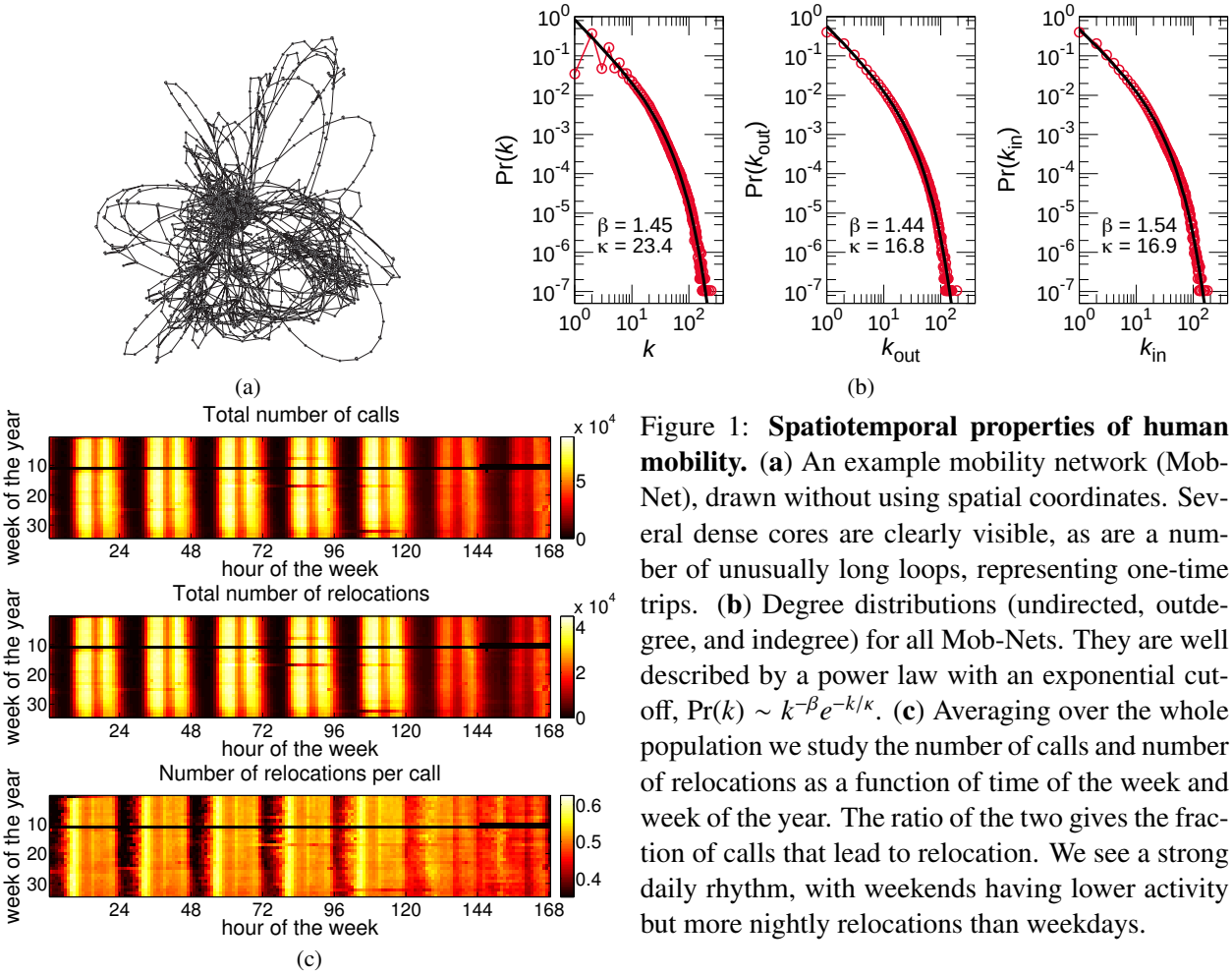


Figure 1: **Spatiotemporal properties of human mobility.** (**a**) An example mobility network (Mob-Net), drawn without using spatial coordinates. Several dense cores are clearly visible, as are a number of unusually long loops, representing one-time trips. (**b**) Degree distributions (undirected, outdegree, and indegree) for all Mob-Nets. They are well described by a power law with an exponential cutoff, $\Pr(k) \sim k^{-\beta} e^{-k/\kappa}$. (**c**) Averaging over the whole population we study the number of calls and number of relocations as a function of time of the week and week of the year. The ratio of the two gives the fraction of calls that lead to relocation. We see a strong daily rhythm, with weekends having lower activity but more nightly relocations than weekdays.

and relocations for the entire sample. We see a striking weekly pattern, whereas there is little dependence on time of year, indicating that human mobility is relatively independent of weather or temperature effects, at least for this population. We also observe that users move more frequently during weekend nights than weekday nights.

The wealth of new data recently available opens avenues for a number of studies of human mobility, as well as the interplay between mobility, demographics, and social network effects. Further work includes quantifying the effects of age and gender on human mobility, and understanding how mobility correlates with social network effects such as the number of friends.

[1] D. Brockmann, L. Hufnagel, and T. Geisel. The scaling laws of human travel. *Nature*, 439(7075):462–465, 2006.

[2] M.C. González, C.A. Hidalgo, and A.L. Barabási. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[3] C. Song, Z. Qu, N. Blumm, and A.L. Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018, 2010.

[4] J. Park, D.S. Lee, and M.C. González. The eigenmode analysis of human motion. *Journal of Statistical Mechanics: Theory and Experiment*, 2010:P11021, 2010.

[5] C. Song, T. Koren, P. Wang, and A.L. Barabási. Modelling the scaling properties of human mobility. *Nature Physics*, 2010.

[6] J.P. Bagrow and T. Koren. Investigating bimodal clustering in human mobility. In *International Conference on Computational Science and Engineering, 2009. CSE'09.*, volume 4, pages 944–947. IEEE, 2009.

EXPLORING MOBILITY OF MOBILE USERS

B.Cs. Csáji[†] [‡], A. Browet[*], V.A. Traag[*], J.-C. Delvenne[*], E. Huens[*],
P. Van Dooren[*], Z. Smoreda[§], V.D. Blondel[*¶]
[*]Department of Mathematical Engineering, Université catholique de Louvain, Belgium
[†]Department of Electrical and Electronic Engineering, University of Melbourne, Australia
[‡]Computer and Automation Research Institute (SZTAKI), Hungarian Academy of Sciences
[§]Sociology and Economics of Networks and Services Department, Orange Labs, France
[¶]Laboratory for Information and Decision Systems, MIT, Cambridge (MA), USA

I. INTRODUCTION

Mobile phone technology has been an important source of inspiration to sociology, especially these last years. Firstly, because it influences the behavior of people, which is a subject of study in itself (e.g. [1] [5]). Secondly, because they can be used to investigate human behavior. Finally, they are useful for the providers of these technologies, most notably for marketing purposes. One example of this is the detection of 'social leaders' [2].

More recently, the mobile phone data available to researchers have been enriched with new information: the position of the antenna used by the caller/receiver. This allows us to know the approximate position of people when they give/receive a call. For instance, it has been exploited to check assumptions on the statistical laws governing mobility of people in their everyday life [3]. In that research the mobile phone is only a detector of position, and no use is made of the social network of the user.

Here we present work based on mobile phone data enriched with geographical information. We first perform a statistical analysis of the data in order to provide a better understanding of it. We observe that most people's mobility is concentrated on a small set of geographical locations. Then we introduce a probabilistic model aimed at the analysis of these frequent locations and the understanding of their temporal dynamics.

II. STATISTICAL ANALYSIS

The anonymous data, obtained from a mobile phone company, include all the communications between the mobile phones of this company in Portugal, over a period of 15 months. For every communication, we know the time of initiating and ending it; which users were involved; what antenna picked it up from the sender; and what antenna relayed it to the receiver. In addition, we also know the exact coordinates (longitude, latitude) of all antennas.

The antenna relaying the communication is not necessarily the nearest one. Therefore, the data contain some noise. For the statistical analysis, we have filtered out this noise by applying a linear weighted moving average. We then define fifty features based on different aspects of users' behavior, such as: (1) time patterns of the calls (average duration, etc.); (2) location patterns of the user (most frequently used antenna, etc.); (3) location patterns of the correspondent (average distance of the correspondent, etc.); (4) user statistics (number of calls, number of callers, etc.); and (5) mobility statistics (gyration, diameter of convex hull, etc.). These mobility features are displayed in Fig. 1, next to a map showing the density of users across the country.

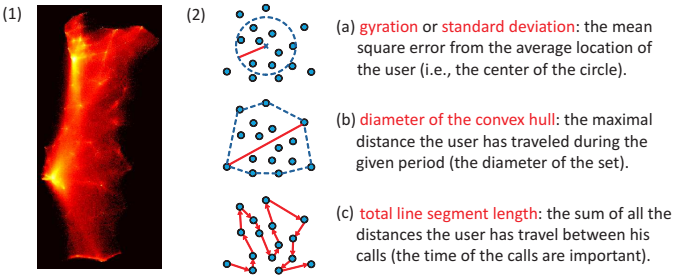


Figure 1. (1) The average locations of the users. Brighter colors indicate that a higher number of users have their average locations in that area. (2) Three different ways to measure the size of the movements of a customer.

We first perform a basic correlation analysis. It appears that the data are highly redundant and can be presented much more efficiently. In particular, by selecting only five new features, combinations of the original ones by using principle component analysis (PCA) or clustering analysis, a negligible amount of information is lost. Those five new features are mostly determined by location and mobility patterns, the positions of the two most frequently used antennas being especially important. This points to the fact that users spend most of their time in only a few locations.

III. FREQUENT LOCATION ANALYSIS

In order to investigate these frequent locations, we performed a more detailed analysis. Often, various antennas could be used for calls made in a single location, due to noise on the signal, such as reflections, buildings, etc. Therefore, we

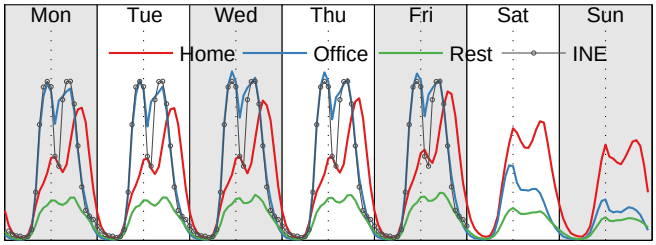


Figure 2. Weekly dynamics of the three detected clusters using k-means clustering: home, office, and the remainder, compared to the independent statistics of the *Instituto Nacional de Estatistica* (INE). Using more clusters yields relatively similar results. The dotted line indicates noon of every day.

first group antennas around the most frequent antennas, based on Delaunay neighborship. We will only consider groups that represent more than 5% of the calls. We refer to these groups of antennas as 'frequent locations'.

For each frequent location we can identify the weekly calling pattern. We use this pattern to determine the type of location we are dealing with. In particular we perform k-means clustering, where each cluster represents a type of frequent location. We find that most of the frequent locations can be represented by only a few clusters, mainly associated to home and office, as displayed in Fig. 2. We find excellent agreement with independent statistics from the *Instituto Nacional de Estatistica*[1] (INE) in terms of time spent at work, as displayed in fig.2. This implies that only the office and the home location have a clearly identifiable calling pattern, while other frequent locations cannot be so easily classified.

Since the frequent locations are represented by a multitude of antennas, we propose a model to estimate a more precise position (i.e. to determine more precisely the home and office). We consider a simplified version of the model proposed in [6], also used in [7]. The idea is that users connect to antennas that have the highest signal strength. This signal strength $S_i(x)$ of antenna i at position x is stochastic, and is given by

$$S_i(x) = p_i L_i(x) R_i, \tag{1}$$

where p_i is the power of the antenna, $L_i(x)$ is the loss of the signal over distance (which decays as a power law with exponent β) and finally R_i represents the stochastic inference present in the environment (i.e. Rayleigh fading). In particular, the random variable R_i is an exponentially distributed variable with mean 1, and independent for all i.

Given a certain location x, we assume the user connects to the antenna that has the highest total signal strength of all antennas in the neighborhood. We denote by $\mathcal{X}$ the set of antennas and, for a user to connect to an antenna i, require that $S_i(x) \geq S_j(x)$ for all $j \in \mathcal{X}$. Since R_i is a random variable, the signal strength of antenna i is larger than all other antennas $j \in \mathcal{X}$ with some probability. Then the probability that the connected antenna a is antenna i, given position x, can be written as $\Pr(a = i|x) = \Pr(S_i(x) \geq S_j(x)\ \forall j \in \mathcal{X})$. Given a specific realization of R_i we can then write

$$\Pr(a = i|x, R_i = r) = \prod_{j \neq i} \Pr\left(R_j \frac{L_j(x)}{L_i(x)} \leq r\right). \tag{2}$$

Using the law of total probability, we can write

$$\Pr(a = i|x) = \int_0^\infty e^{-r} \Pr(a = i|x, R_i = r) dr. \tag{3}$$

In fact, this can be seen as a smooth approximation of the Voronoi tessellation, in which a user will always connect to the closest antenna.

For each frequent location, we know the number of calls k_i made using antenna i. The probability there were k_i calls using antenna i given position x is then $\Pr(a = i|x)^{k_i}$. Hence, the log likelihood of observing the call frequencies for a certain location x can then be given as

$$\log \mathcal{L}(x|k) = \sum_{i \in \mathcal{X}} k_i \log \Pr(a = i|x). \tag{4}$$

The Maximum Likelihood Estimate (MLE) $\hat{x}$ of the position for a frequent location is then given by

$$\hat{x} = \arg\max_x \log \mathcal{L}(x|k). \tag{5}$$

For finding the MLE, we employ a derivative-free optimization scheme, since the gradient of the likelihood function is costly to evaluate. In particular, we used the Nelder-Mead algorithm.

Acknowledgments

This paper presents research results supported by the Concerted Research Action (ARC) "Large Graphs and Networks" of the French Community of Belgium and the Belgian Network DYSCO (Dynamical Systems, Control, and Optimization), funded by the Interuniversity Attraction Poles Programme, initiated by the Belgian State, Science Policy Office. This work has also been supported by the Orange Labs R&D Research Grant 46143202. The scientific responsibility rests with its authors.

References

[1] T. de Bailliencourt, T. Beauvisage, F. Granjon and Z. Smoreda, Extended Sociability and Relational Capital Management: Interweaving ICTs and social relations, in *Mobile Communication: Bringing Us Together or Tearing Us Apart?*, R. Ling, S. Campbell, eds. Transaction Publishers, 2011.

[2] C. de Kerchove, E. Huens, P. Van Dooren and V. Blondel, Social Leaders in Graphs, in Lecture Notes in Control and Information Sciences, *Positive Systems*, Vol.341, pp.231–237, 2006.

[3] Marta C. González, César A. Hidalgo and Albert-László Barabási, Understanding individual human mobility patterns, *Nature*, Vol.453, pp.779–782, June 2008.

[4] R. Lambiotte, V.D. Blondel, C. de Kerchove, E. Huens, C. Prieur, Z. Smoreda, P. Van Dooren, Geographical dispersal of mobile communication networks, *Physica A*, Vol.387, pp.5317–5325, 2008.

[1] http://www.ine.pt

[5] C. Licoppe, Z. Smoreda, Are social networks technologically embedded? *Social Networks*, Vol.27(4), pp.317–335, Oct. 2005.
[6] H. Zang, F. Baccelli, and J. Bolot, Bayesian inference for localization in cellular networks, *Proc 29th Conf Info Comm*, pp. 1963–1971, IEEE Press, 2010.
[7] V.A. Traag, A. Browet, F. Calabrese and F. Morlot Social Event Detection in Massive Mobile Phone Data Using Probabilistic Location Inference, submitted to *IEEE SocialCom'2011*.

Chatty Mobiles: Individual mobility and communication patterns

Thomas Couronné, Zbigniew Smoreda, Ana-Maria Olteanu
Sociology and Economics of Networks and Services department
Orange Labs R&D, Paris

Introduction

Human mobility analysis is an important issue in social sciences, and mobility data are among the most sought-after sources of information in urban studies, geography, transportation and territory management. In network sciences mobility studies have become popular in the past few years, especially using mobile phone location data [1,2,3,4,5]. For preserving the customer privacy, datasets furnished by telecom operators are anonymized. At the same time, the large size of datasets often makes the task of calculating all observed trajectories very difficult and time-consuming. One solution could be to sample users. However, the fact of not having information about the mobile user makes the sampling delicate. Some researchers [1] select randomly a sample of users from their dataset. Others try to optimize this method, for example, taking into account only users with a certain number or frequency of locations recorded [2,3]. At the first glance, the second choice seems to be more efficient: having more individual traces makes the analysis more precise. However, the most frequently used CDR data (Call Detail Records) have location generated only at the moment of communication (call, SMS, data connection). Due to this fact, users' mobility patterns cannot be precisely built upon their communication patterns. Hence, these data have evident shortcomings both in terms of spatial and temporal scale.

In this paper we propose to estimate the correlation between the user's communication and mobility in order to better assess the bias of frequency based sampling. Using technical GSM network data (including communication but also independent mobility records - as in [6]), we will analyze the relationship between communication and mobility patterns.

Data

One weekday GSM data of the Paris Region territory (12,012 km² - 4,638 sq mi) were used. The dataset covers 4 million of French mobile phone users and more than 94 million records. Data are issued from Orange GSM network probes, they are anonymous (a secure, random network attributed temporary identity) and contain both cell localized communication events (calls and SMS) and mobility events (handover (HO) and location area update (LAU)).

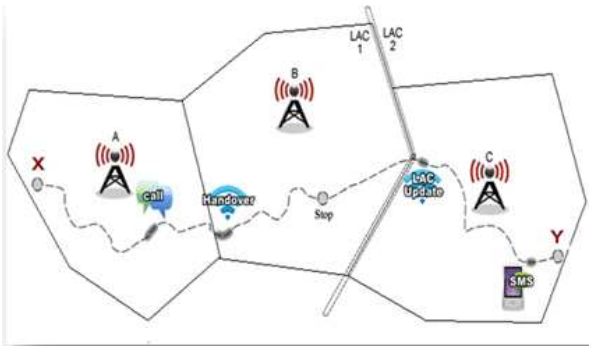


Figure 1. An example of mobile phone network localization data types for one user travelling from X to Y

Concerning mobility records, there are two types of data: HO data are generated during a communication, when a mobile phone changes position and is transferred to a new antenna; LAU records are generated when a device changes location area (in Paris Region, a location area groups on average 150 cells). The LAU is generated also when a mobile moves from one location area to the next while not on a call. It is exactly the information we need for our analysis

as it is independent of the user communication behavior.

Results

Our analysis was conducted using two separate record types: communication data and communication-independent itinerancy data. Communication frequency was plotted against the median number of mobility records (LAU) for each user (see: figure 2).

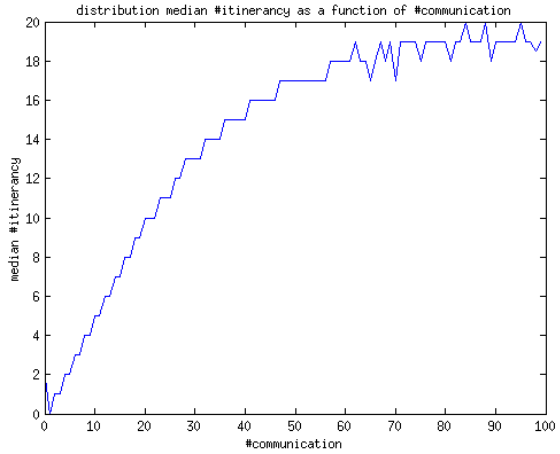


Figure 2. Median number of local area change as a function of communication frequency

Ninety percent of users have less than 30 communication events (calls or SMS) during the observed day. For this group, we notice a clear, almost linear correlation between the frequency of communications and the median number of location area changes (daily mobility indicator).
The curve reaches a plateau at about 50 communications per day and then the communication - mobility link disappears. People who communicate extremely frequently can no longer be distinguished by their median mobility.
The relationship between the number of itinerancy events (LAU) and the median communication frequency has also been studied. Again this correlation is nearly a perfect one: the more mobility records are, the more frequent mobile communications are.

To analyze conjointly user's mobility and communication, we constructed 8 equal frequency ordered classes for each event type (where class 1 is the lowest mobility/communication, and the class 8 is the most intensive itinerancy/communication). Then we populated an 8x8 matrix with users having all specific combinations of itinerancy and communication events. The result of this operation is showed in figure 3.

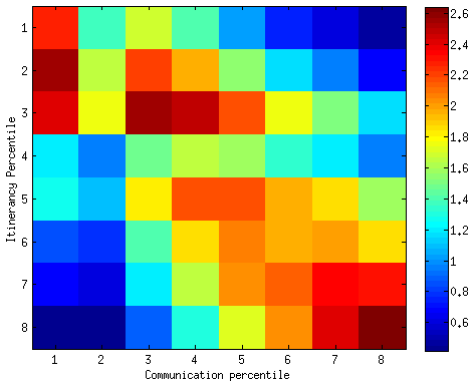


Figure 3. Conjoint distribution of users into communication and itinerancy intensity classes (1-low, 8-high)

The red color signifies the most probable combination and the deep blue color designates the less probable one. As we can observe, red squares are distributed on the top-right (both infrequent communication and mobility), on the bottom-left (both intense communication and mobility) as well as in the center of the matrix (average values). All other combinations, and in particular high-low combinations, are less frequent. The matrix indicates that in our data the relationship between user mobility and user communication frequency is really strong.

To complete this approach, the daily displacement distance per user was calculated using all localized records (calls, SMS, HO, and LAU) and compared to communication events distribution (see: figure 4).

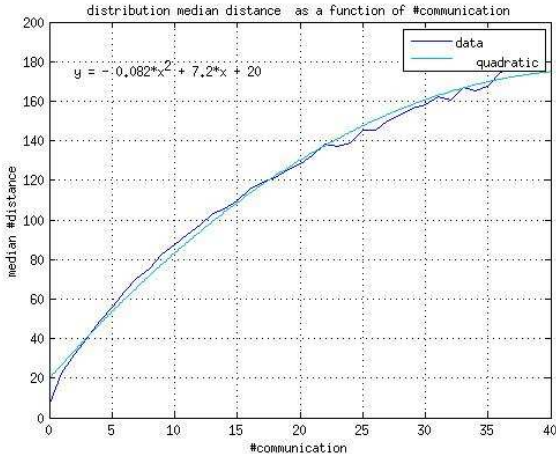


Figure 4. Median daily distance traveled (km) as a function of communication events frequency

We looked for a best regression model to fit our data, where y is the median daily distance in km and x is the number of communication events (call, SMS). It appears that the best model is a quadratic function:

$$y = -0.82x^2 + 7.2x + 20$$

This analysis confirms our observation showed in figure 2: the higher the number of communication events, the less strong the increase of the mobility distance.

Conclusion

A significant correlation between user mobility events and communication frequencies confirms our intuition that in mobile phone usages both phenomena are interrelated. A highly mobile person has in fact greater probability to use a mobile phone than someone who only commutes between a few places where s/he can also communicate *via* a fixed telephone. As his/her correspondents learn with time which is the most adapted communication channel to reach this person, they will also contribute to reinforce the observed correlation.

From the point of view of human mobility analysis using data from mobile phone, such as CDRs, our results ask for a very careful examination of sampling methods which are used. Selecting users with frequent communication traces, i.e., with many cell localizations, seems to introduce a clear bias because people having more mobile communications are also in a more mobile class of the general population.
Definitely, to better calibrate mobile phone data analysis in this domain a cross-analysis of mobile phone and survey-like data is needed.

References

[1] González MC, Hidalgo CA, Barabási AL, 2008, Understanding individual human mobility patterns, Nature 453: 779–782.

[2] Song C., Qu Z, Blumm N, Barabási AL, 2010, Limits of Predictability in Human Mobility. Science 327, 1018 (2010); DOI: 10.1126/science.1177170.

[3] Onnela JP, Arbesman S, González MC, Barabási AL, Christakis NA, 2011 Geographic constraints on social network groups, PLoS ONE 6(4): e16939.

[4] Calabrese F, Di Lorenzo G, Liu L, Ratti C, 2011, Estimating Origin-Destination flows using opportunistically collected mobile phone location data from one million users in Boston Metropolitan Area, IEEE Pervasive Computing 99, DOI: 10.1109.

[5] Calabrese F, Smoreda Z, Blondel V, Ratti C, 2011, Interplay between telecommunications and face-to-face interactions - a study using mobile phone data, PLoS ONE 6(7): e208814.

[6] Caceres N, Wideberg J, Benitez F, 2007, Deriving origin-destination data from a mobile phone network, Intelligent Transport Systems, IET 1(1): 15 – 26.

Distance Matters: Geo-social Metrics for Mobile Location-based Social Networks

Salvatore Scellato
Computer Laboratory
University of Cambridge
ss824@cam.ac.uk

Cecilia Mascolo
Computer Laboratory
University of Cambridge
cm542@cam.ac.uk

ABSTRACT

The widespread adoption of mobile smartphones has led to a significantly large portion of users continuously accessing online social services in their daily lives. Furthermore, these devices offer geolocation capabilities: the ability to share your location, to generate location-tagged information and to search for it adds a crucial dimension to these social services. These new features make possible to investigate how space and geographic distance affect the social connections created by mobile users on these services.

Here we study the socio-spatial properties of 3 large-scale online location-based social networks. We observe strong heterogeneity across users, with different characteristic spatial lengths of interaction across both their social ties and social triads. Finally, we describe new metrics able to characterize how geographic distance affects social structure and to capture socio-spatial heterogeneity across mobile users.

1. INTRODUCTION

Online Location-based Social Networks (LBSNs) have recently attracted millions of users, experiencing a huge popularity increase over a short period of time. Thanks to the widespread adoption of location-sensing mobile devices, users can share information about their location with their friends. As a consequence, these services offer a groundbreaking opportunity to understand and exploit the spatial properties of the social networks arising among online users, but also a potential window on real human socio-spatial behavior.

Spatial networks have been extensively studied, particularly when dealing with transportation networks, power grids, urban road networks and other systems where nodes are embedded in a metric space [2]. In general, metric distance directly influences the network structure of such systems by imposing higher costs on the connections between distant entities. Social networks, instead, have been largely studied from a purely topological perspective, focusing on the structural position of their nodes and on structural mechanisms that describe their evolution. Indeed, the connection cost that heavily affects other types of spatial networks may not be as important in social systems, particularly when focusing on online interactions. However, some initial results on the spatial properties of social networks suggest that distance still matters. In fact, there is universal agreement that the probability of having a social connection between two individuals decreases with distance [4, 3, 1].

Dataset	N	$\langle k \rangle$	$\langle C \rangle$	$\langle D \rangle$	$\langle l \rangle$
Brightkite	54,190	7.88	0.181	5,651	2,041
Foursquare	258,706	22.07	0.191	8,494	1,442
Gowalla	122,414	9.48	0.254	5,663	1,792

Table 1: Properties of the mobile datasets: number of nodes N in the social network, average node degree $\langle k \rangle$, average clustering coefficient $\langle C \rangle$, average geographic distance between nodes $\langle D \rangle$ [km], average social link length $\langle l \rangle$ [km].

In this work we study the socio-spatial properties of three different popular mobile LBSNs: Brightkite, Foursquare and Gowalla. We have collected data about all of them and extracted the social networks among their users. We are able to assign a "home location" to each user, in order to embed the nodes in a 2-dimensional metric space. We analyze their global properties, observing robust universal features across them. Furthermore, we discuss how mobile users are affected in a heterogeneous way by geographic distance, with some individuals exhibiting mainly short-range rather than long-distance social ties and clusters. Finally, we describe two geo-social metrics which are able to capture this heterogeneity across social links and social triangles.

2. MOBILE DATASETS

We study three spatial social networks acquired from different popular mobile online LBSNs. For each service we extract the social network arising among users and a single *geographic home location* for each user, defined as the place where he/she has more check-ins overall.

Brightkite was founded in 2007 as a social networking website which allows users to share their location with their friends: it is available worldwide and it is based on the idea of making check-ins at places, where users can see who is nearby and who has been there before. Brightkite users can establish mutual friendship links and they can push their check-ins to their Twitter and Facebook accounts. We study a dataset collected in September 2009 which includes the whole Brightkite user base at that time, with information about 54,190 users [5].

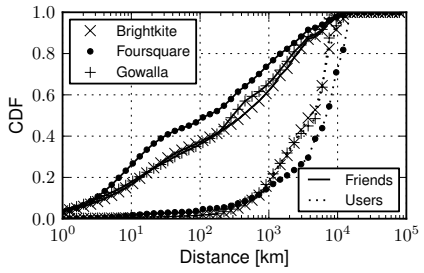


Figure 1: Empirical Cumulative Distribution (CDF) of the geographic distance between all users (dotted line) and between connected friends (solid line) for the three datasets.

Foursquare was created in 2009 and it has quickly risen as the most popular location-based service. Users utilize the Foursquare application on their mobile devices, which allows them to check-in and share with their friends the place where they are. Many Foursquare users choose to automatically push their check-in messages to Twitter, which provides a public API to search and download these messages. Thus, we have recorded both check-in messages and friendship ties among Foursquare users on Twitter, where they are publicly available. Our dataset includes about 250,000 users: we estimate that our sample contains approximately 20% to 25% of the entire Foursquare user base at collection time.

Gowalla is a location-based social network created in 2009: its users check-in at places through their mobile devices. Check-ins are shared with friends: as a consequence, friends can check where a user is or has been; conversely, it is possible to see all the users that have recently been in a given place. We have collected a complete snapshot of Gowalla data in August 2010, including check-ins and social connections, totaling about 120,000 active users.

3. GLOBAL PROPERTIES

We first address the spatial properties of the mobile social networks under analysis, focusing on the main topological and geographic measures, reported also in Table 1. The average degree is lower in Brightkite and Gowalla, respectively 7.88 and 9.48, than in Foursquare, where users have on average 22.07 friends. These networks also exhibit high values of average clustering coefficient, between 0.18 and 0.26.

The average geographic distance between users $\langle D \rangle$ is consistently larger than the average distance between friends $\langle l \rangle$ across all the datasets: while the first value ranges between 5,600 and 8,500 km, the latter has much shorter values, between 1,400 and 2,000 km. This already provides evidence that the probability of having a social link between two users decreases with distance: we will further investigate this relationship later. The distribution of social link length is comparable across the three datasets, as shown in Figure 1: about 40%-50% of all couples of friends are within 100 km, with more than 3% of all links being shorter than 1 km. Instead, the distribution of distances among users, also shown in Figure 1, has a different behavior: about 50% of users are at distances larger than 4,000 km across all the networks.

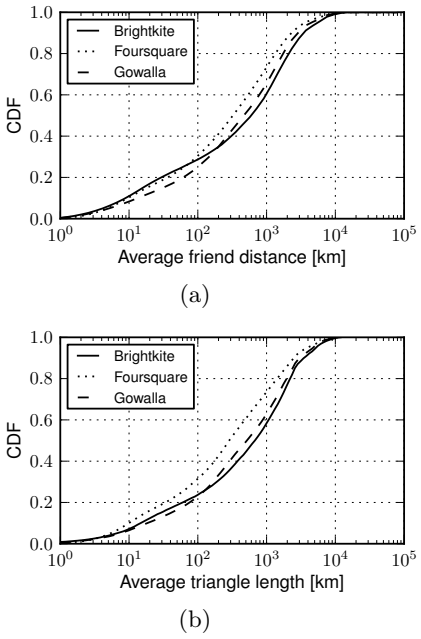
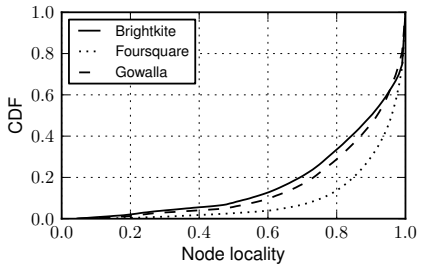


Figure 2: Empirical Cumulative Distribution (CDF) of the average friend distance (a) and average triangle length (b) for each user in the social networks.

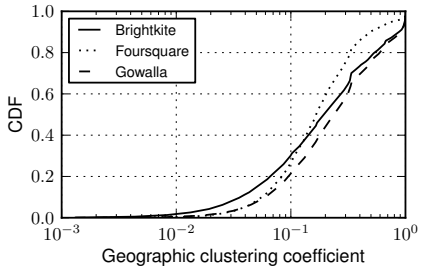
4. USER PROPERTIES

We now focus on individual users, studying how their social ties stretch across space. We plot the distribution of the average friend distance of each user in Figure 2(a). The existence of values over all geographic scales is due to the existence of users with different characteristic lengths of interaction. For instance, about 10% of users have connections with an average length of just 10 km, whereas about 20% of users have average friend distances above 2,000 km. In particular, since this distribution closely matches the aggregated link length distribution in Figure 1, links with different geographic lengths do not appear homogeneously across all users. Instead, there is heterogeneity between users, with some of them with only short-range connections and others with long-distance ties.

We now study the geographic properties of social triangles. In fact, users tend to belong to several triads, resulting in high values of clustering coefficient: our networks exhibit clustering values between 0.18 and 0.26. To assess whether geographic heterogeneity arises also for social triangles, we compute the geographic average triangle length of the three links of each triangle and then we compute the average triangle geographic length for each user by considering all the triangles he/she belongs to. Our aim is to assess the geographic span of a user's social triangles, whatever their number might be. In Figure 2(b) we show the distribution of the average triangle length for each user: triangles with different geographic span are not equally arising among all users, but instead there are users with smaller triads and users with wider ones. For example, there are at least 20% of users with an average triangle length less than 100 km, while the top 20% have values above 2,000 km.



(a)



(b)

Figure 3: Empirical Cumulative Distribution (CDF) of node locality (a) and geographic clustering coefficient (b) for each user in the social networks.

5. GEO-SOCIAL MEASURES

Finally, we define two new geo-social network measures which are able to capture, respectively, the geographic heterogeneity we have observed in social ties and social triangles. We represent a spatial social network as a graph G with N nodes and K links: nodes represent users and a link among two nodes exists if there is a social tie between them (e.g., a person lists another user as one of his/her friends). Nodes are embedded in a 2-dimensional metric space where the distance between two nodes i and j is given by the geographic distance D_{ij} between their locations on the planet. This distance is used as the length of the link l_{ij} between nodes i and j.

Then, we define a metric to capture the geographic closeness of the neighbors of a certain node. Let us consider an undirected geographic social network, a node i with a particular geographic position and the set Γ_i of its neighbors. The node degree k_i is the number of these neighbors, that is $k_i = |\Gamma_i|$. Then, the *node locality* of i can be defined as a measure of how much geographically close its neighbors are and it is computed as follows:

$$NL_i = \frac{1}{k_i} \sum_{j \in \Gamma_i} e^{-l_{ij}/\beta} \quad (1)$$

where β is a scaling factor which avoids extremely small values of node locality when links have large lengths. By definition, NL_i is always normalized between 0 and 1. The value of β can be chosen so that networks with different geographic size can still be compared with each other.

Users exhibit an overall high average node locality: Brightkite has an average value of 0.82 and Gowalla of 0.85, while in FourSquare this value goes up to 0.90. The distributions of node locality for the three networks are shown in Figure 3(a): node locality is able to capture how different users have heterogeneous spatial properties, with values spanning the entire range. For example, in Brightkite and Gowalla about 40% of users have a node locality higher than 0.90, whereas in the FourSquare dataset this phenomenon is more evident, with 70% of users above 0.90.

Similarly, we define the *geographic clustering coefficient* as an extension of the clustering coefficient used for complex networks. The clustering coefficient measures the proportion of triangles among the neighbors of a given node: the geographic clustering coefficient of node i is thus defined in the same way as the clustering coefficient, but each existing triangle between nodes i, j and k is assigned a weight w_{ijk} defined as:

$$w_{ijk} = e^{-\frac{\Delta_{ijk}}{\beta}} \quad (2)$$

where Δ_{ijk} is the maximum length among the three links, that is $\Delta_{ijk} = max(l_{ij}, l_{ik}, l_{jk})$. We define $w_{ijk} = 0$ if there is no link between j and k. Since this measure uses the maximum weight among all the links of a triangle, it focuses on nodes which are all close to each other: when just one of the three nodes is not close to the other two, the weight will immediately decrease. This emphasizes social triangles where users are extremely close to each other. Again, the parameter β is used to scale the values of the measure.

The three networks exhibit different values of geographic clustering coefficient: while Brightkite has an average value of 0.165 and Gowalla of 0.171, FourSquare scores a much higher 0.209. The geographic clustering coefficient is close to the standard clustering coefficient, signaling how triangles tend to form at shorter distances. The probability distributions of the geographic clustering coefficient are shown in Figure 3(b): again, the three networks exhibit a non-negligible portion of users with a coefficient close to 1.

6. CONCLUSION

In this work we have studied the socio-spatial properties of users of mobile location-based services. We observe and discuss heterogeneity in user socio-spatial behavior: users exhibit friendship connections across a wide range of geographic distance, showing similar variability in the social triads they belong to. Finally, we have defined and discussed two novel geo-social measures that capture such user heterogeneity.

7. REFERENCES

[1] L. Backstrom, E. Sun, and C. Marlow. Find me if you can: improving geographical prediction with social and spatial proximity. In *Proceedings of WWW '10*, pages 61–70, 2010.

[2] M. Barthélemy. Spatial Networks. *Physics Reports*, 499:1–101, 2011.

[3] R. Lambiotte, V. Blondel, C. Dekerchove, E. Huens, C. Prieur, Z. Smoreda, and P. Vandooren. Geographical dispersal of mobile communication networks. *Physica A*, 387(21):5317–5325, September 2008.

[4] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins. Geographic routing in social networks. *PNAS*, 102(33):11623–11628, August 2005.

[5] S. Scellato, C. Mascolo, M. Musolesi, and V. Latora. Distance Matters: Geo-social Metrics for Online Social Networks. In *Proceedings of WOSN'10*, June 2010.

A tale of many cities: universal patterns in human urban mobility

Anastasios Noulas
Computer Laboratory
University of Cambridge
an346@cl.cam.ac.uk

Salvatore Scellato
Computer Laboratory
University of Cambridge
ss824@cl.cam.ac.uk

Renaud Lambiotte
Department of Mathematics
University of Namur
renaud.lambiotte@fundp.ac.be

Massimiliano Pontil
Computer Science Department
University College London
m.pontil@cs.ucl.ac.uk

Cecilia Mascolo
Computer Laboratory
University of Cambridge
cm542@cl.cam.ac.uk

July 23, 2011

Since the seminal works of Ravenstein [1], the movement of people in space has been an active subject of research in the social and geographical sciences. It has been shown in almost every quantitative study and described in a broad range of models that a close relationship exists between mobility and distance. People do not move randomly in space, as we know from our daily lives. Human movements exhibit instead high levels of regularity and tend to be hindered by geographical distance. The origin of this dependence of mobility on distance, and the formulation of quantitative laws explaining human mobility remains, however, an open question, whose answer would lead to many applications [2], e.g. improve engineered systems such as cloud computing and location-based recommendations [3, 4], and yield insight into a variety of important societal issues, such as urban planning and epidemiology [5, 6, 7].

In classical studies, two related but diverging viewpoints have emerged. The first camp argues that mobility is directly deterred by the costs (in time and energy) associated to physical distance. Inspired by Newton's law of gravity, the flow of individuals is predicted to decrease with the physical distance between two locations, typically as a power-law of distance [8, 9, 10]. These so-called "gravity-models" have a long tradition in quantitative geography and urban planning and have been used to model a wide variety of social systems, e.g. human migration [11] and traffic flows [12]. The second camp argues instead that there is no direct relation between mobility and distance, and that distance is a surrogate for the effect of *intervening opportunities* [13]. The migration from origin to destination is assumed to depend on the number of opportunities closer than this destination. A person thus tends to search for destinations where to satisfy the needs giving rise to its journey, and the absolute value of their distance is irrelevant. Only their ranking matters. Displacements are thus driven by the spatial distribution of places of interest, and thus by the response to opportunities rather than by transport impedance as in gravity models. The first camp appears to have been favoured by practitioners on the grounds of computational ease [14], despite the fact that several statistical studies have shown that the concept of intervening opportunities is better at explaining a broad range of mobility data [15, 16, 17, 18, 19].

This long-standing debate is of particular interest in view of the recent revival of empirical research on human mobility. Contrary to traditional works, where researchers have relied on surveys,

small-scale observations or aggregate data, recent research has taken advantage of the advent of pervasive technologies in order to uncover trajectories of millions of individuals with unprecedented resolution and to search for universal mobility patterns, such to feed quantitative modelling. Interestingly, those works have all focused on the probabilistic nature of movements in terms of physical distance. As for gravity models, this viewpoint finds its roots in Physics, in the theory of anomalous diffusion. It tends to concentrate on the distributions of displacements as a function of geographic distance. Recent studies suggest the existence of a universal power-law distribution $P(\Delta r) \sim \Delta r^{-\beta}$, observed for instance in cell tower data of humans carrying mobile phones $\beta = 1.75$ [21] or in the movements of "Where is George" dollar bills $\beta = 1.59$ [22]. This universality is, however, in contradiction with observations that displacements strongly depend on where they take place. For instance, a study of hundreds of thousands of cell phones in Los Angeles and New York demonstrate different characteristic trip lengths in the two cities [20]. This observation suggest either the absence of universal patterns in human mobility or the fact that physical distance is not a proper variable to express it.

In this work, we address this problem by focusing on human mobility patterns in a large number of cities across the world. More precisely, we aim at answering the following question: "Do people move in a substantial different way in different cities or, rather, movements exhibit universal traits across disparate urban centers?". To do so, we take advantage of the advent of mobile location-based social services accessed via GPS-enabled smartphones, for which fine granularity data about human movements is becoming available. Moreover, the worldwide adoption of these tools implies that the scale of the datasets is planetary. Exploiting data collected from public *check-ins* made by users of the most popular location-based social network, Foursquare, we study the movements of about 800,000 users around the globe over a period of about six months, and study the movements across more than 5 million places in metropolitan cities that span five continents.

After discussing how at larger distances we are able to reproduce previous results of [21] and [22], we also offer new insights on some of the important questions about human urban mobility across a variety of cities. We first confirm that mobility, when measured as a function of distance, does not exhibit universal patterns. The striking element of our analysis is that we observe a universal behavior in all cities when measured with the right variable. We discover that the probability of transiting from one place to another is inversely proportional to a power of their *rank*, that is, the number of intervening opportunities between them. This universality is remarkable as it is observed despite cultural, organizational and national differences. This finding comes into agreement with the social networking parallel which suggests that the probability of a friendship between two individuals is inversely proportional to the number of friends between them [23], and only indirectly on physical distance. More importantly, our analysis is in favour of the concept of intervening opportunities rather than gravity models, thus suggesting that trip making is not explicitly dependent on physical distance but on the accessibility of objectives satisfying the objective of the trip. Individuals thus differ from random walkers randomly exploring physical space because of the motives driving their mobility.

Our results are confirmed with a series of simulations verifying the hypothesis that the density is the driving force of urban movement. By using only information about the distribution of places of a city as input and by coupling our model with a rank-based mobility preference we are able to reproduce the actual distribution of movements observed in real data.

References

[1] E.G. Ravenstein (1885) The Laws of Migration, *Journal of the Royal Statistical Society* **48**,167227.

[2] P. Hui and J. Crowcroft (2008) Human Mobility Models and Opportunistic Communication System

Design, *Royal Society Philosophical Transactions B* **366**, 2005-2016

[3] D. Quercia, N. Lathia, F. Calabrese, G. Di Lorenzo and J. Crowcroft (2010) Recommending Social Events from Mobile Phone Location Data, *Proceedings of IEEE ICDM 2010*

[4] V. W. Zheng, Y. Zheng, X. Xie, and Q. Yang (2010) Collaborative location and activity recommendations with GPS history data, *WWW 10*, 1029-1038.

[5] K. Nicholson and R.G. Webster, Textbook of Influenza (Blackwell, Malden, Massachusetts, 1998)

[6] L. Hufnagel, D. Brockmann and T. Geisel (2004) Forecast and control of epidemics in a globalized world, *Proc Natl Acad Sci USA* **101**, 15124.

[7] V. Colizza, A. Barrat, M. Barthlemy and A. Vespignani (2007) Predictability and epidemic pathways in global outbreaks of infectious diseases: The SARS case study, *BMC Med.* **5**, 34.

[8] V. Carrothers (1956) A Historical Review of the Gravity and Potential Concepts of Human Interaction, *J. of the Am. Inst. of Planners* **22**, pp. 94-102.

[9] A.G. Wilson (1967) A statistical theory of spatial distribution models, *Transportation Research* **1**, pp. 253-269. **54**, pp. 68-78

[10] S. Erlander and N. F. Stewart (1990) The Gravity Model in Transportation Analysis: Theory and Extensions, Brill Academic Publishers, Utrecht.

[11] M. Levy (2010) Scale-free human migration and the geography of social networks, *Physica A* **389**, 4913-4917.

[12] W.-S. Jung, F. Wang and H.E. Stanley (2008) Gravity model in the Korean highway, *Europhys. Lett.* **81**, 48005.

[13] S. Stouffer (1940) Intervening opportunities: A theory relating mobility and distance, *American Sociological Review* **5**, 845-867

[14] S.M. Easa (1993) Urban Trip Distribution in Practice. I: Conventional Analysis, *Journal of Transportation Engineering* **119**, 793-815

[15] E. Miller (1972) A note on the role of distance in migration: costs of mobility versus intervening opportunities, *J. Reg. Sci.* **12**, 475478.

[16] K.E. Haynes, D. Poston and P. Sehnirring (1973) Inter-metropolitan migration in high and low opportunity areas: indirect tests of the distance and intervening opportunities hypotheses, *Econ. Geogr.* **49**, 68-73.

[17] W.J. Wadycki (1975) Stouffer's Model of Migration: A Comparison of Interstate and Metropolitan Flows, em Demography **12**, 121-128.

[18] R.H. Freymeyer and P.N. Ritchey (1985) Spatial Distribution of Opportunities and Magnitude of Migration: An Investigation of Stouffer's Theory, *Sociological Perspectives* **28**, 419-440.

[19] C. Cheung and J. Black (2005) Residential location-specific travel preferences in an intervening opportunities model: transport assessment for urban release areas, *Journal of the Eastern Asia Society for Transportation Studies* **6**, 3773-3788.

[20] S. Isaacman, R. Becker, R. Cceres, S. Kobourov, J. Rowland, A. Varshavsky (2010) A Tale of Two Cities, *11th Workshop on Mobile Computing Systems and Applications*

[21] M. C. Gonzlez, C. A. Hidalgo and A.-L. Barabsi (2008) Understanding individual human mobility patterns, *Nature* 453, 779-782. D. Brockmann and F. Theis (2008) Money Circulation, Trackable Items, and the Emergence of Universal Human Mobility Patterns, *IEEE Pervasive Computing* **7**, 28 - 35

[22] D. Brockmann, L. Hufnagel and T. Geisel (2006) The scaling laws of human travel, *Nature* **439**, 462-465.

[23] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan and A. Tomkins (2005) Geographic routing in social networks, *PNAS* **102**,11623-11628.

Finding Meaningful Usage Clusters From Anonymized Mobile Call Detail Records

Richard A. Becker, Ramon Caceres, Chao Han*, Karrie Hanson, Ji Meng Loh,
Simon Urbanek, Alexander Varshavsky and Chris Volinsky
AT&T Labs - Research (* Virginia Tech University)

A. Introduction

Given the ubiquity of cell phones and their frequent use, cell phone calling patterns can be used to study the location and movements of large numbers of people for cellular network operation, urban planning and event management. In this abstract, we investigate ways to cluster mobile phone usage signatures derived from Call Detail Records (CDRs) in order to identify common usage patterns. Specifically, we apply unsupervised clustering algorithms to summaries derived from anonymized CDRs to identify groups of users who share the same patterns of calling and texting intensity. Typical behavior of a cluster is represented by the mean behavior in that group. The clustering result can also be combined with cell antenna location to better understand the mobility patterns for urban planning. Due to the lack of space, we only summarize our key contributions here. Please see our full paper [2] for details.

B. Data and Methods

We analyze anonymized CDRs collected over a 60 day period (November 29, 2009 to January 27, 2010) from 35 cell towers located within 5 miles of the center of a suburban city in the Northeast United States with approximately 20,000 residents. The CDRs include 475,000 unique handsets that generated around 15 million voice calls and 26 million text messages. We studied these cell phone records to group city dwellers into categories useful to urban planners. Each CDR records the starting time and duration of a voice call or Short Messaging Service (SMS) transaction, and the locations and azimuths of the cell tower antennas associated with the event.

Given the sensitivity of CDR data, we took several steps to ensure the privacy of individuals. First, only anonymous records were used in this study. The data was collected and anonymized by a party not involved in the data analysis. Personally identifying characteristics were removed from our CDRs. CDRs for the same phone are linked using an anonymous unique identifier, rather than a telephone number. No demographic data is linked to any cellphone user or CDR. Second, all our results are presented as aggregates. That is, no individual anonymous identifier was singled out for the study. Finally, each CDR only included location information for the cellular towers with which a phone was associated during a call or text message. The phones were effectively invisible to us aside from these events.

Our interest is in cell phone usage by time of day and day of week, so we preprocess the CDRs for each user by aggregating the calls and text messages into separate bins. Each bin represents a particular hour of the day and day of the week, resulting in $24 \times 7 = 168$ bins. In order to make voice minutes and SMS counts comparable, we normalize SMS usage so that both groups have the same global mean, making one SMS message correspond to 1.5 minutes of voice calling. The final step of the preprocessing concatenates the voice and SMS usage bins for each user to form a vector of 336 elements, which is normalized to sum to 1. Thus the data for each individual is the relative fraction of all calling made by voice or SMS in each hour of each day of the week.

We are now ready to cluster, but there are many clustering algorithms we might use, each with parameters to set or estimate. For example, in K-means clustering, we need to supply a value for k, the desired number of clusters. In addition, many users in our dataset have very little activity; for example, the median number of SMS events per user per day is 2. We anticipate that it may be difficult to cluster them, so it will be important to have techniques that allow us to evaluate the performance of the algorithms.

In order to assess whether we can cluster low usage users well, we rank users by their usage volume and then split the users into 10 slices, with around 15,000 users in each slice. We perform clustering algorithms on the slices one by one and observe how clustering results change. The clustering performance is evaluated by two fold cross-validation which can also help investigate the prediction accuracy for new data. Particularly, within each slice, we randomly sample 50% of the data to be used as training set for clustering, and use the rest to test the result. Since we do not have any prior knowledge about the user classification, there is no group label available as in usual cross-validation approach. Thus for each user in the test set, instead of aggregating the usage vector over the entire two months, we aggregate the activities over December and January separately such that every test set user has two usage vectors. Each vector is assigned to a cluster according to its proximity to the cluster centers derived from clustering the training set. If we assume that the usage pattern for each user is consistent over time, the cluster assignment for both vectors should match. The vector agreement is measured by Cohen's Kappa index [3], which evaluates the concordance using relative observed agreement adjusted by the hypothetical probability of agreement by chance. We perform the random sampling and the cross-validation three times for all slices and get similar agreement index values across the three runs, which indicates that the consistent usage assumption is valid.

We have tried several widely used clustering algorithms which include: 1. K-means which is an iterative procedure that refines initial estimates of the k cluster centroids into final centroids in an attempt to minimize the sums of squares of distances from each point to the nearest cluster centroid. 2. Hierarchical clustering (with 'average', 'complete' and 'ward' agglomeration methods) which starts with every observation data in its own cluster and then repeatedly merges the closest pair of clusters until all of the data are in one cluster. 3. Partition Around Medoids (PAM) [5] which is similar to K-means but takes medoid as the representative point such that the algorithm is robust to outliers. 4. Fuzzy C-Means [4] which allows each data point to belong to more than one cluster with different membership degrees (between 0 and 1), where lower membership values may indicate data noisiness. However, Fuzzy clustering usually takes much longer time than hard clustering methods such as K-means. For the clustering methods which are sensitive to initialization (e.g., K-means, PAM, Fuzzy), we run the algorithms with different starting points and pick the solution that provides the best Calinski index [7]. The Calinski index measures the dispersion of the data points within a cluster and between the clusters by taking the ratio of the between cluster sum of squares and within cluster sum of squares (adjusted by degrees of freedom). A higher Calinski index indicates better cluster separation.

We refer to the Gap Statistic [9] and the Silhouette Statistic [5] for determining the number of clusters k, because the clustering literature shows they have good performance and because they are applicable to any clustering technique and distance metric. The Gap Statistic compares the change in W_k (the pooled within-cluster dispersion) as k increases for the original data with that expected for data generated from a reference null (uniform) distribution. The optimal number of clusters is estimated as the value of k for which $\log(W_k)$ falls the farthest below its expected curve. The Silhouette width for each point compares the average distance of a point to all other points in nearest cluster besides its own cluster with the average distance of the point to all other points in the same cluster. Observations with a large Silhouette width are well clustered. We can make the judgment about k by checking the Silhouette plot visually which shows the within-cluster compactness and between-cluster separation.

If we take Euclidean distance as the distance metric, we find that K-means has a better performance than the other algorithms in the top 7 slices (top 70% quantiles ordered by usage). For the low volume users in the bottom 3 slices (less than 30 minutes of calls/SMS over the 60 day period), the clustering approaches all perform poorly and generate similar Kappa values around 0. Since low volume users generate very few calls or text messages, their usage vectors are quite sparse, typically consisting of isolated spikes. Regardless of the number of clusters k, the conventional clustering scheme based on Euclidean distance always results in one large cluster and many small clusters consisting entirely of individual bins from the distribution. The location of the bins is random, based solely on the initial cluster seeds. The reason for such odd clusters is that the Euclidean metric does not take into account any additional topology of the bins; also, in a high-dimensional space, all sparse vectors are roughly equidistant from each other.

To find meaningful clustering structure for the low volume users, we selected a different distance metric: the Earth Mover's Distance (EMD) [8], which can incorporate the notion of nearness between bins naturally. EMD is an intuitive metric between distributions if we think of them as piles of sand sitting on the ground (underlying domain). Each pile of sand is an observed sample. To quantify the difference between two distributions, we measure how much sand must be moved to turn one distribution into the other. EMD is the minimal total ground distance travelled weighted by the amount of sand moved. A major advantage of using EMD is that it allows us to define a topology on the bins. Since the boundaries of bins by hour are in a sense arbitrary, we would consider neighboring bins in time as 'close'. Also human activities typically follow a daily cycle, so that there is a similarity between adjacent days of the week at the same time. Finally, we have two types of activities, voice and SMS, which are distinct, so we model them as being at a constant distance from each other. EMD allows us to specify this complex relationship between the bins. The biggest drawback of using EMD for clustering is its high computational complexity (runs in $O(n^3 \log(n))$ time). When dealing with large datasets as in our case, the exact computation of EMD becomes impractical. To alleviate this problem, we use a modified version [1] which reduces the cost to linear time. We applied hierarchical clustering with EMD to the bottom 30% of the usage distribution. Even with very sparse vectors, EMD is capable of separating users into coherent groups.

C. Results

We illustrate how to cluster data from CDRs to arrive at meaningful groups by applying K-means to a subset of 30,000 high-volume users (at least 7.5 hours of combined activity over the 60 day period). Plots of the Gap and Silhouette statistics confirmed that the optimal number of clusters was in the range of 5 to 7. The Kappa value generated by K-means with 7 clusters. achieves 0.6 (which indicates substantial agreement [6]) for the top 20% of the data. The resulting clusters (Figure 1) are distinct and informative. For example, cluster 1 consists of voice call activity just before and after work hours; cluster 2 consists of heavy voice users during weekday business hours; cluster 3 is primarily weekday usage of both voice and SMS; cluster 4 is voice calling, primarily in the evening; cluster 6 is after-work voice calling; cluster 7 is a business-hour texting.

Let's explore cluster 5, the largest cluster, representing 22% of the 30K clustered users, or about 6600 users. We hypothesize that cluster 5 consists of students because of the heavy SMS use on weekdays (3pm-10pm) and the early beginning and late ending on weekends (10am-1am). We then attempt to see if locations of use, represented approximately by antennas, can help confirm the hypothesis. Figure 2 shows the activities in each of the clusters on the antenna that points in the direction of the city high school. The figure shows a matrix of 'lip plots', where the two rows represent weekends and weekdays, and the columns represent the different clusters.

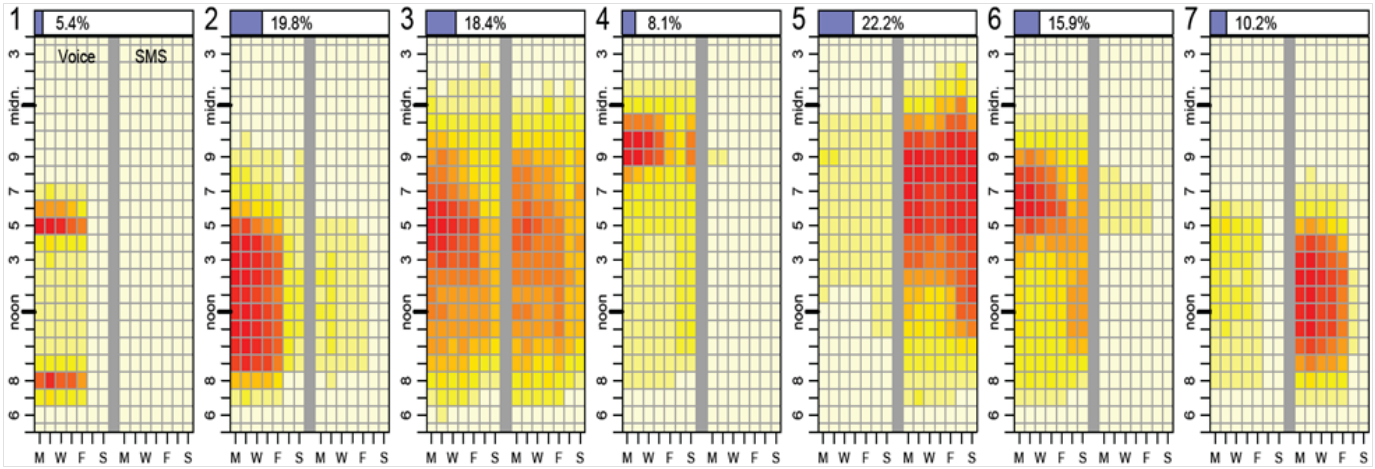


Fig. 1. Heat maps of the seven cluster means, each representing a specific usage profile. Each profile consists of the voice usage matrix (left of the gray strip) and the SMS usage matrix (right of the grey strip) across hour of the day (24 rows) and day of the week (7 columns). Darker colors indicate higher usage. The blue bar at the top denotes the relative size of the cluster.

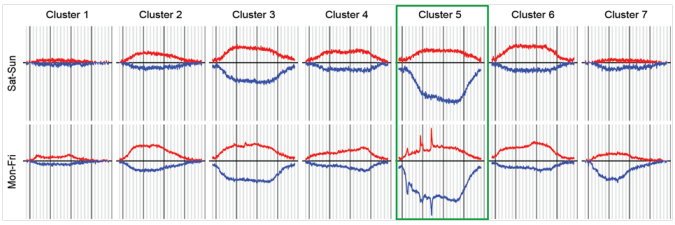


Fig. 2. Lip plot of activities in each of the clusters on the high school direction antenna. Lines show the average volume of voice (red) and SMS (blue). The plot for cluster 5 on weekdays stands out with spikes at 7:30am, 11:30am and 2:30pm, which match the times of school start, lunch, and dismissal.

Each plot shows the average volume of voice (red) and SMS (blue, plotted below the y-axis) from 6AM to 6PM during our study period. Cluster 5 stands out because of activity spikes that appear linked to the school day; the spikes occur weekdays at 7:30AM when school begins, at 11:30AM during an open lunch period, and at 2:30PM dismissal. Usage for all other clusters show smooth changes over time and no similar spikes, indicating that students are not in the other clusters. In a similar manner, we investigated cluster 1, which we tentatively identified as commuters driving to and from work; it turned out to have a much greater proportion of calls involving a moving cell phone (connecting to more than 5 different cell towers).

D. Conclusions

Understanding cell phone usage patterns is an important step towards creating applications and services useful for urban communities. Our study involved clustering of usage patterns found in call detail records gathered in a small city. We analyzed anonymized versions of these records, ran a series of clustering algorithms to find phones that exhibited similar usage patterns, and relied on cross-validation to check the clustering performances. We found that for relatively high-volume users, K-means with Euclidean distance provides an informative grouping. However, K-means failed in clustering the sparsely populated usage vectors of low volume users. For these, we chose hierarchical clustering with EMD to take into account the underlying topology of the distributions.

REFERENCES

[1] D. Applegate, T. Dasu, S. Krishnan, and S. Urbanek. Unsupervised clustering of multidimensional distributions using earth mover distance. In *KDD*, San Diego, CA, 2011.

[2] R. Becker, R. Cáceres, K. Hanson, J. M. Loh, S. Urbanek, A. Varshavsky, and C. Volinsky. Clustering anonymized mobile call detail records to find usage groups. In *1st Workshop on Pervasive Urban Applications (PURBA)*, 2011.

[3] J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:38–46, 1999.

[4] J. C. Dunn. *A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters*, 1973.

[5] L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data*. New York, 1990.

[6] J. Landis and G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33:159—-174, 1977.

[7] Milligan and Cooper. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50:159—-179, 1985.

[8] Y. Rubner, C. Tomasi, and L. J. Guibas. A metric for distributions with applications to image databases. In *Proceeding of International Conference On Computer Vision*, pages 59—66, 1998.

[9] R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of data clusters via the gap statistic. *Journal of the Royal Statistical Society B*, 63:411–423, 2001.

Using Cellular Network Data for Urban Planning

Richard A. Becker, Ramon Caceres, Karrie Hanson, Ji Meng Loh,
Simon Urbanek, Alexander Varshavsky and Chris Volinsky
AT&T Labs - Research
{rab, ramon, karrie, loh, urbanek, varshavsky, volinsky} @research.att.com

With the continuing urbanization of the world's population and the rapid growth of cities, urban planners are faced with many challenges, including heavily congested roads, overzealous development, and increasing pollution. To efficiently address these problems, urban planners need to develop a better understanding of the dynamics of modern cities. In this paper, we explore the use of anonymized Call Detail Records (CDRs) to capture city dynamics.

To conduct our study, we collected anonymized CDRs from the cellular network of a large US communications service provider. We captured transactions carried by the 35 cell towers (roughly 300 antennas) located within 5 miles of the center of Morristown, NJ, a suburban city in the greater New York City metropolitan area. In place of the phone number of the person involved in a transaction, each CDR was assigned an anonymous identifier consisting of the 5-digit billing zip code and a unique integer. Each CDR also contains the starting time of the voice or SMS event, the duration of the event, and the locations and azimuths of the cell tower antennas associated with the event. We collected voice and SMS traffic for 60 days between November 29, 2009 and January 27, 2010. In total, we collected 15 million voice CDRs and 26 million SMS CDRs for 475,000 unique phones.

Given the sensitivity of the data, we took several steps to ensure the privacy of individuals. First, only anonymous records were used in this study. In particular, personally identifying characteristics were removed from our CDRs. CDRs for the same phone are linked using an anonymous unique identifier, rather than a telephone number. No demographic data is linked to any cellphone user or CDR. Second, all our results are presented as aggregates. That is, no individual anonymous identifier was singled out for the study. By observing and reporting only on the aggregates, we protect the privacy of individuals. Finally, each CDR only included location information for the cellular towers with which a phone was associated during a voice call or at the time of a text message. The phones were effectively invisible to us aside from these events, and we only knew those phone locations at the granularity of an antennas's coverage area, often greater than one square mile.

In this paper, we demonstrate that cellular network data can be used to determine the geographical distributions of home locations of workers (*laborshed*) and partiers (*partyshed*) in a city and validate our methodology by comparing our results to the 2000 Census. We also present a novel visualization technique, called *lip plots*, and show how it can be used to capture the 'life beat' of a city. Due to the lack of space, we only summarize our key contributions below. Please see our full paper [1] for details.

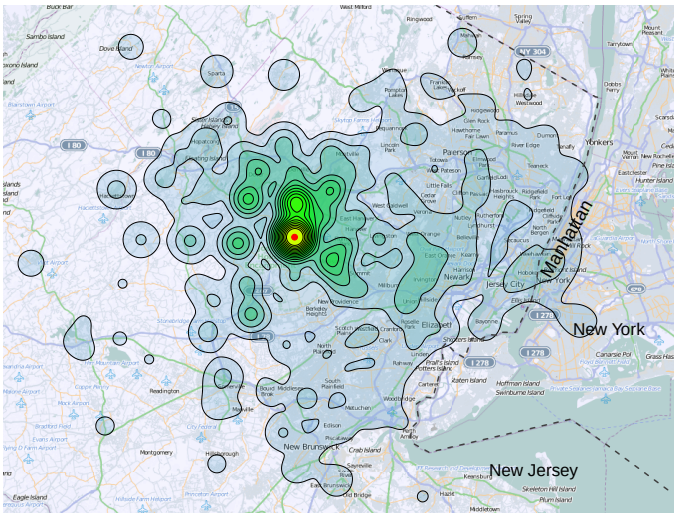


Fig. 1. Morristown laborshed maps calculated from CDR data. The red dot represent the center of Morristown.

Calculating and Validating the Laborshed

Morristown is a regional center of commerce and shopping, with a developed downtown area, many office complexes, a large hospital, and the county courthouse. It draws a large worker base from the nearby suburbs and even some from the much larger New York City. The geographical area representing where a city's workers live is known as its *laborshed*.

We used our CDR data to calculate the laborshed for Morristown. We classified a cellphone user as a worker if she is frequently observed in Morristown during business hours (9am to 5pm, Monday to Friday). More specifically, a worker needs to satisfy two conditions. First, she needs to engage in an average of at least 4 calls/messages per week during business hours, involving one of the Morristown cell towers. Second, she needs to make those calls/messages on an average of at least 2 unique weekdays per week. We derived these thresholds experimentally and observed that moderate changes to these values did not affect our results. We used account billing ZIP codes to identify place of residence.

Figure 1 shows a contour map of the Morristown laborshed as calculated from the CDR data. Studying the map reveals a high-density region centered directly over Morristown, indicating a large concentration of people who both live and work in

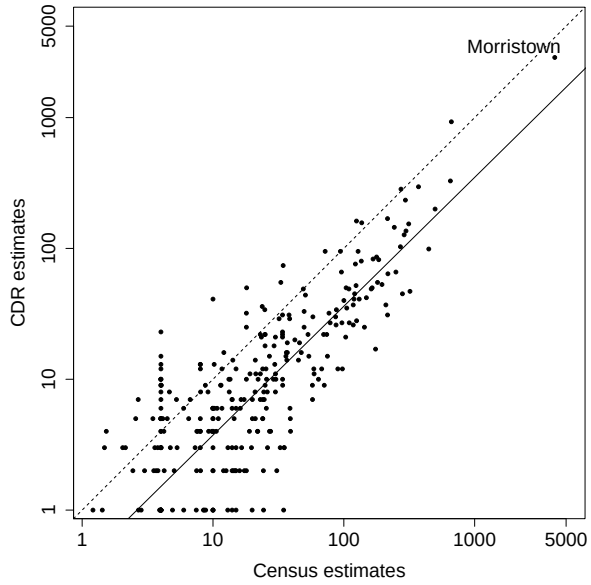


Fig. 2. Scatterplot showing agreement between Morristown laborshed numbers from CDR data and US Census data. Each point represents one ZIP code. The solid line shows the best linear fit, where the CDR count equals 0.387 of the Census count. If the CDR estimates exactly matched the Census numbers, the points would fall on the dotted line.

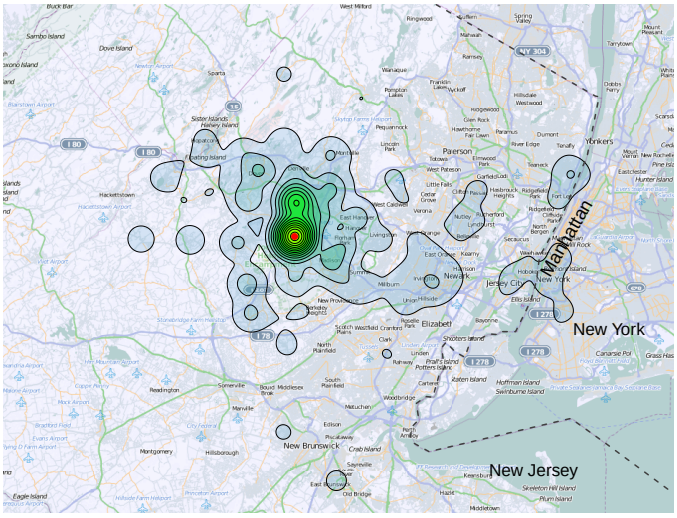


Fig. 3. Morristown partyshed map showing the home locations of people who used their cellphones during weekend late nights in downtown Morristown. Comparing to the earlier laborshed maps, partiers' homes are concentrated in areas closer to Morristown than workers' homes.

Morristown. We also see that many more workers come from areas north of Morristown than south, and that there seems to be a cluster of workers to the east, in the more heavily populated areas close to New York City. Additionally, there are some pockets of workers who come from towns west of Morristown.

We validate our CDR-based laborshed results by comparing them to publicly available US Census data. We used the 2000 Census Transportation Planning Package (CTPP), specifically the "Journey to Work" package which includes detailed information on commuting patterns including counts of commuters to and from specific census tracts [2]. We mapped census tracts to ZIP codes by calculating the proportion of a census tract that fell within each ZIP code of interest.

Figure 2 compares our laborshed estimates against the corresponding Census CTPP data. We draw each ZIP code as a point reflecting the relationship between our estimates with the Census numbers. We do not expect our estimates to have perfect agreement with the Census numbers because our CDR records only show those who are actively generating calls or SMS records, and reflect only the activity of the part of the population on our company's network. Additionally, the census data we had access to is a decade old, and commuting patterns might have changed significantly in that time. Still we do expect our points to align with theirs up to a multiplicative factor and hence should fall close to a straight line in this plot. For comparison, we show the $y = x$ equality line as a dotted line, and the best linear fit, $y = .387x$ as a solid line. The correlation coefficient is 0.81. There does seem to be a clear correspondence between the CDR- and census-based numbers, and if we want to roughly estimate numbers of people, we would multiply the CDR numbers by $1/.387$.

To summarize, we argued that understanding where workers live is important for urban planning and showed that the laborshed generated from CDRs matches closely that obtained from census data. These results give confidence in the validity of our approach. Our approach has significant advantages over the census because its low cost makes it practical to generate laborshed results much more frequently, for example every few months instead of every ten years.

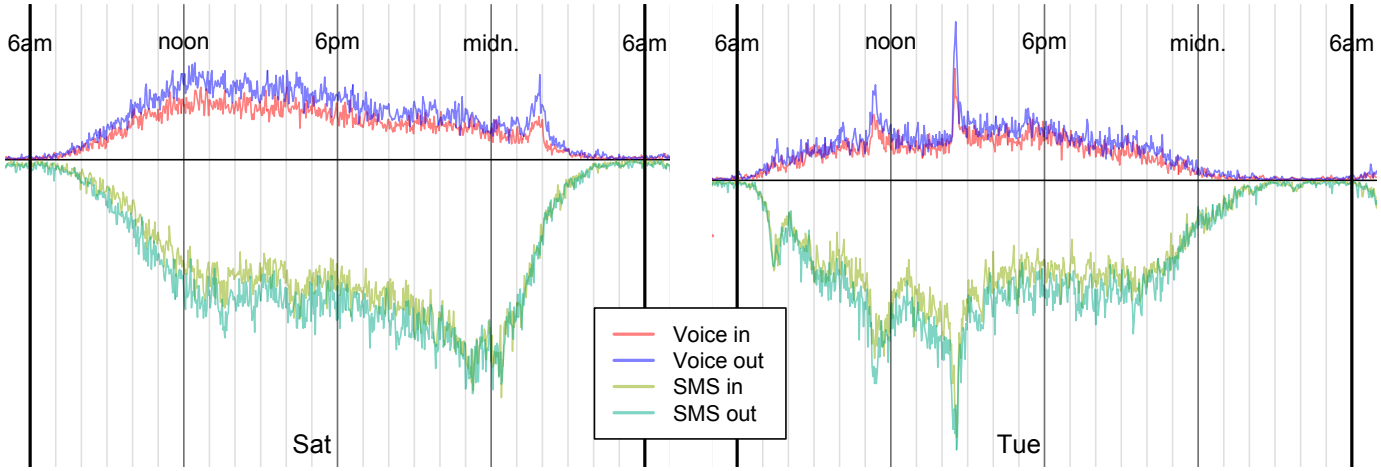
Calculating the Partyshed

Similarly, we can apply these techniques to other groups of people besides daytime workers. Like many cities, Morristown has a lively bar and restaurant scene that attracts people from other communities as well as locals. We refer to the geographical distribution of where this group lives as the *partyshed*. We look for cellphone users who have voice call or text messaging activity on late weekend nights (10pm to 3am, Fridays to Sundays). The resulting partyshed is shown in Figure 3. It appears that the distribution of partiers is considerably more concentrated in and near Morristown than the distribution of workers. Nonetheless, there is still some representation of people who live far away, even as far as New York City. Knowing where groups of revelers come from and where they return at the end of the night could allow towns to tailor services such as late-night shuttle buses intended to keep inebriated drivers off the road.

Capturing the Lifebeat of a City

In this section, we demonstrate that it is possible to identify patterns of human activity in different parts of a city by observing cell phone usage in different cell tower antenna coverage regions. We refer to these patterns as the *lifebeat* of a city.

To analyze data from multiple cell tower antennas simultaneously, we developed a novel visual display capable of representing the multivariate nature of the data. We first



(a) Downtown Antenna

(b) High School Antenna

Fig. 4. *Lip plots* of voice call and SMS volumes show unusual spikes highlighting local patterns or events in Morristown. Call volume (plotted upwards; inbound: red; outbound: blue) and SMS volume (plotted downwards; inbound: light green; outbound: dark green) on two antennas are shown. The antenna in (a) points towards the commercial and restaurant district and the antenna in (b) points towards the high school. A voice peak occurs Saturday at 2AM when the bars close. Both voice and SMS peaks occur Tuesday when the school lets out.

aggregated the underlying data in two-minute intervals and then aggregated by the day of the week, which allows the study of day-specific patterns. In total, we ended up with 720 two-minute bins for each day of the week for both voice calls and SMS messages. We then used the principle of small multiples to display the data for all combinations of the partitioned variables. (a high-resolution, zoomable version of this complete display is available at `http://bit.ly/BigLipPlot`).

Here we present plots for two specific antennas in Morristown in detail. Figure 4 shows usage plots captured on two different days of the week for two antennas located on the same cell tower but pointing in different directions. The x-axis represents time, starting and ending at 6am. The height of the plot shows the amount of traffic: height above the axis represents voice call volume, while height below the axis represents SMS volume. By using these opposite directions of the axes we avoid overplotting and retain the ability to recognize shapes at a glance. In addition, our visual cognitive system is good at evaluating symmetry quickly, thus we can quickly assess whether voice usage strongly deviates from SMS usage. For both types of traffic, color is used to distinguish inbound vs. outbound traffic. The resulting shape of the plot resembles lips and hence we call this type of visualization *lip plots*.

The patterns of these two plots are strikingly different. Figure 4(a) shows data from a Saturday in one part of town. The SMS traffic dominates the voice traffic, and the volumes keep rising throughout the day with maximal usage between 11 PM and 1 AM. The voice traffic, despite being dominated by the SMS traffic, has a noticable spike at 2AM. This is the cell tower antenna that points to the downtown area including the majority of the restaurants and bars in town. The spikes might represent late night revelers, and in particular the 2AM voice spike might reflect the fact that the bars close at 2AM, and patrons are looking for a ride home.

The plot in Figure 4(b) has quite a different pattern. This plot is from a weekday and also shows a majority of data from SMS traffic as opposed to voice. But here the majority of traffic is in the morning and in particular, we see spikes in SMS usage at 7AM, 11AM and 2PM. Only at 2PM do we see a similar spike in the voice traffic. This cell antenna points towards the town's high school, and could reflect the communication patterns of the students there, texting before and after school and during lunch. The larger 2PM spike at the end of the school day might reflect calls between students and parents, where voice channels would be more likely.

In general (see `http://bit.ly/BigLipPlot`), we found that there is a large heterogeneity in the patterns of antennas pointing to different directions, reflecting differing usage patterns in different parts of the city. We also found a wide variance in the volume covered by the different directions. Certain directions simply have more traffic than others. In addition, small volume antennas have high variance, and often result in giving the lips a 'fuzzier' look. Finally, the relationship between SMS and voice changes by direction, especially on the weekends.

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. *IEEE Pervasive Computing, special issue on Large-Scale Opportunistic Sensing*, 10(4), October-December 2011.

[2] Census Transportation Planning Package (CTPP) 2000: Part 3. Downloaded from http://www.transtats.bts.gov.

Segmentation of towns using call detail records

Romain Guigourès* and Marc Boullé*
*Orange Labs, 2 av. Pierre Marzin, 22300 Lannion, France
romain.guigoures@orange-ftgroup.com, marc.boulle@orange-ftgroup.com

***Abstract*—In this paper, we deal with the segmentation of towns using call detail records. The data can be viewed as a directed bipartite graph wherein the source nodes correspond to the towns, origins of the calls, and the target nodes that are the towns, destinations of the calls. A nonparametric method based on a Bayesian Approach is proposed to determine the finest segmentation of these two sets. Instead of directly clustering the nodes, we propose here to make a coclustering on the edges defined as bidimensionnal items described by two features : the source and target nodes. Once the finest clustering is obtained, the clusters are successively merged on the two sets until only one cluster remains, in such a way that the loss of information is minimal. The initial segmentation is optimally coarsened in order to enable a hierarchical exploratory analysis of the data. Thus, it is possible to get insights either nationalwide or locally. A study of the telephone areas of Belgium by exploring the coclustering structure at different grain levels demonstrates the interest of the method.**

***Keywords*-Community detection; Clustering; Bayesianism; Model Selection; Density estimation**

I. Introduction

Graph partitioning has long been studied in the operational research field. One of the oldest approaches is the minimum-cut method, where the graph is divided into a predetermined number of disjoint subsets, usually of approximately the same size, chosen such that the number of edges between the clusters of nodes is minimized.

With the recent availability of many network data, such as world wide web, social networks, phone call networks, science collaboration graphs [1], there is a renewed interest for the graph partitioning problem, especially for the automatic discovery of community structures in large networks. Many approaches have been studied for the problem of graph clustering, including hierarchical clustering, divisive clustering, spectral methods, random walk [2]. To evaluate the quality of a clustering regardless of the cluster number, the modularity criterion proposed in [3] is now widely accepted in the literature, and has even been treated as an objective function in clustering algorithms [4]. The modularity is a measure ranging from -1 to 1, being all the more high that the clusters have more internal edges than the expected edges number if the connections were made randomly, with the same nodes degrees.

In this paper, we present a way of analyzing and summarizing the structure of large graphs, based on piecewise constant edge density estimation. The approach extends the stochastic block modeling approach [5] in that the modeling method is fully non-parametric with the number of clusters as a free parameter, and exploits a statistical model selection technique and scalable optimization algorithms. Data grid models [6] are applied to graph data, where each edge is considered as a statistical unit with two variables, the source and target nodes. The objective is to find a correlation model between the two variables, owing to a data grid model, which in this case turns to be a coclustering of both the source and target nodes of the graph. The cells resulting from the cross-product of the two clusterings summarize the edge density in the graph. The best correlation model is selected using the MODL (Minimum Optimized Description Length) approach [6], and optimized by the means of combinatorial heuristics with super-linear time complexity.

Then, a post processing technic is introduced consisting in merging successively the clusters in the least costly way, from the finest clustering model down to one single cluster containing all the towns. It appears that the cost of the merge of two clusters is a weighted sum of Kullback-Leibler divergences from the merged clusters to the created cluster which can be interpreted as a dissimilarity measure between the two clusters that have been merged. Thus, the post-processing technique can be considered as an agglomerative hierarchical clustering [7].

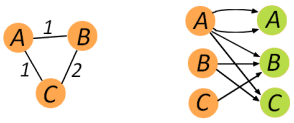
The rest of the paper is organized as follows. In Section 2, we present the MODL approach for data grid models applied to the edge density estimation in graphs and the postprocessing enabling the exploratory analysis. Then in Section 3, experiments on Belgian call detail records illustrate the property of the method. Finally, Section 4 concludes the paper and introduces the future works.

II. The segmentation

A. Graph clustering using MODL

Unlike making clustering on a simple graph like the modularity-based method do, the graph we deal with is directed, bipartite and with multiple edges. The source nodes are the towns, origins of the call, the target nodes the towns, destinations of the calls and the edges the calls. Figure 1 illustrates the different data representations.

The towns are grouped if the distributions of the calls are similar. This means that instead of making groups of towns that frequently call each other, the method brings together the towns that call the same towns and in the same



(a) (b)

	A	B	C
A	2	1	1
B	0	1	1
C	0	1	0

(c)

Figure 1: Representation of the tabular data displayed in Fig.(c) as a simple weighted graph (Fig.(a)) and as a directed bipartite multigraph (Fig.(b))

ratio. The objective of the method is to estimate the density of the edges owing to a coclustering of the sources and target nodes. Figure 2 illustrates such a coclustering with two source clusters and two target clusters. In this example, the probability of edges from Brussels or Liège to Brussels or Namur is 50%.

	Brussels	Namur	Liège	Charleroi
Brussels	1	1	0	0
Liège	1	1	0	0
Charleroi	0	0	1	1
Namur	0	0	1	1

	{Brussels,Namur}	{Liège,Charleroi}
{Brussels,Liège}	4	0
{Charleroi,Namur}	0	4

Figure 2: Example of coclustering

Formally, a model M of edge density estimation is defined by :

- the number of source and target clusters
- the partition of source (resp. target) nodes into source (resp. target) clusters.
- the edges distribution on the coclusters defined as the cross-products of the source and target clusters.
- for each source (resp. target) cluster, the edges distribution on the node of the cluster.

The coarsest model is based on one single cluster of towns, whereas the finest one exploits one cluster per town. Coarse grained models tend to be reliable, whereas fine grained are more informative. The issue is to find a trade-off between the informativeness of the edge density estimation and its reliability, on the basis of the granularity of the coclustering. Applying a Bayesian model selection approach, the best model M^* is defined as being the most probable given the data D, obtained by maximizing a criterion built on prior terms $P(M)$ which give priority to simple models, i.e. with a low number of clusters, and on the likelihood which favours the informative models, i.e. fine grained models :

$$M^* = \underset{M}{\operatorname{argmax}}\, P(M|D) = \underset{M}{\operatorname{argmax}} \left(\frac{P(M)P(D|M)}{P(D)} \right)$$

$$\Rightarrow M^* = \underset{M}{\operatorname{argmin}}(-\log P(M) - \log P(D|M))$$

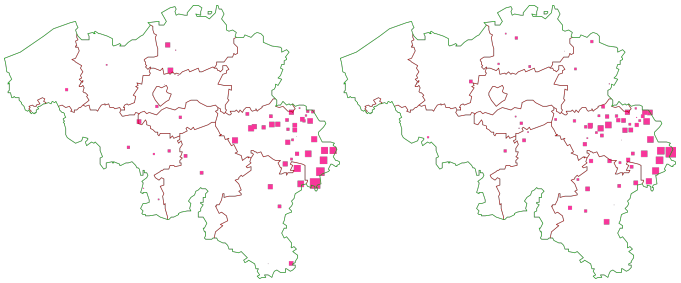
The detailed criterion is left out for brevity. The full criterion is described in [8]. As for the optimization algorithm, we have used the optimization heuristics detailed in [9], which have practical scaling properties, with $\mathrm{O}(m)$ space complexity and $\mathrm{O}(m\sqrt{m}\log m)$ time complexity, with m the number of edges. The main heuristic is a greedy bottom-up heuristic, which starts with a fine grained model, considers all the merges between clusters and performs the best merge if the criterion decreases after the merge. The process is reiterated until no further merge decreases the criterion. This heuristic is enhanced with post-optimization steps (moves of towns across clusters), and embedded into the variable neighborhood search (VNS) meta-heuristic [10], which mainly benefits from multiple runs of the algorithm with different random initial solutions. The optimization algorithms summarized above have been extensively evaluated in [9], using a large variety of artificial datasets, where the true data distribution is known.

B. *Merging the clusters*

In case of large datasets, i.e. with a huge number of edges, the edge density converges to the true edge distribution. This means that, for each town, the distribution of the calls is fine enough to be differentiated. Thus the method yields one town per cluster, that is too fine for an easy interpretation. To overcome this issue, a post-processing technique is proposed. It consists in merging successivly the clusters so as to worsen the least the criterion. By studying in detail the variation of the criterion due to the merge, it appears that the merge of two clusters is all the more likely that the distributions of their in/outcoming calls are similar. Figure 3 illustrates two towns very likely to be merged because of their similar distributions. Technically, this variation is a sum of Kullback-Leibler divergences from the merged clusters to the resulting one, weighted by the size of each of them. In brief, this process is equivalent to making a hierarchical agglomerative classification, whose dissimilarity measure is based on probability divergences.

III. Experiments

Experiments have been conducted on call detail records of the Belgian telecommunications company Mobistar aggregated on 6 months. There are 217 millions calls between 589 towns. Another approach has been applied on the same dataset in [11], which results in 17 clusters. Like in this study



(a) (b)

Figure 3: Similar distributions of the calls of two german-speaking Belgian towns.

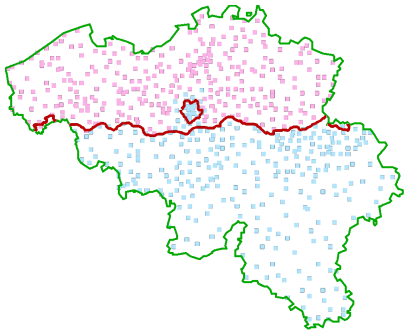


Figure 5: Segmentation of Belgium into two clusters

our clusters are geographically connected. However our clustering results enable a multiscale exploratory analysis of the Belgian telephone areas.

A. The Finest Clustering

The finest clustering highlights 588 groups over Belgium. Hence, each cluster is made up of one town, except one cluster that groups two towns together. Given the huge number of calls (217 millions for 589 cities), the finest clustering on this dataset has reliably approximated the true distribution. This is shown on Figure 4, the clustering is all the more fine that there are edges.

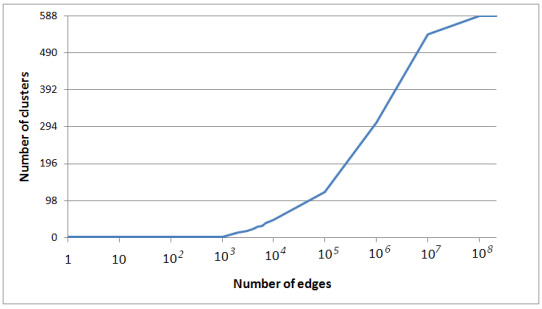


Figure 4: Number of clusters retrieved by the method for a given subset of randomly selected edges

B. Two linguistic communities

In this experiment, the clusters have been merged successivly until obtaining two clusters. This segmentation in two groups highlights the two linguistic communities of Belgium: Flanders and Wallonia. This reveals that the distribution of the calls of a town is denser in the areas with the same linguistic characteristics. The case of Brussels is particular, because the majority of the inhabitants of the city are french-speakers despite it is included into the Flemish territories. That is why the region of Brussels has been clustered into Wallonia.

C. Eleven clusters that do not match with the provinces boundaries

There are eleven provinces in Belgium, five in Flanders and five in Wallonia, the eleventh being the province of Brussels-Capital. In order to compare the delimitation of the telephone areas and the boundaries of the provinces, we have studied the clustering with eleven clusters. The clusters are displayed on a map in Figure 6.

For Antwerp, East and West Flanders, the clusters fit well the provinces territories.

The provinces of Hainaut and Liège are splitted into three clusters. For the first one, it can be explained by the presence of some major cities in the same province (Mons, La Louvière and Charleroi). For the second one, we can notice the sphere of influence of Liège while the area east of the city corresponds to the arrondissement of Verviers where more than 25% of the inhabitants are German-speakers.

There are also clusters that straddles some provinces, like the cluster grouping the arrondissement of Leuven and the province of Limburg or the one grouping the province of Luxembourg and a part of the provinces of Namur and Liège. These telephone areas are consequently vast, that is why a finer and local study would yield enough clusters to make a relevant exploratory analysis.

The case of Brussels highlights the correlation between the calls and the sphere of influence of the city including a little part of the Flemish Brabant and almost all the Walloon Brabant. This can be explained by the current trend of expansion of the suburbs to a southern direction, towards Walloon-Brabant, the inhabitants of Brussels being attracted by more peaceful areas with the same linguistic characteristics [12].

D. A local study of Brussels Region

Brussels is a particular city. The capital of Belgium is very cosmopolitan. In spite of the predominance of the french-speaking community, it is included into the Flemish territories. A study of the calls from the towns of the province of Brussels-Capital to all the Belgian towns allows the segmentation of Brussels and its suburb according to the

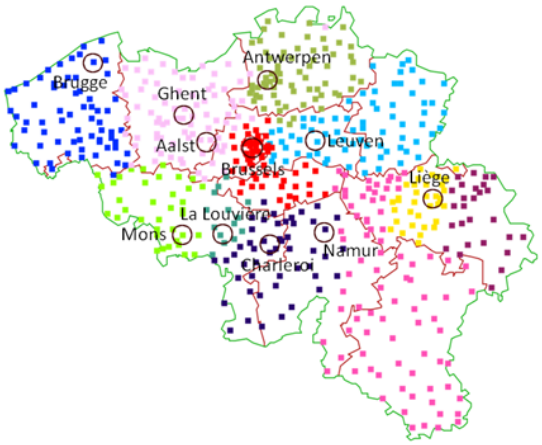


Figure 6: Segmentation of Belgium into eleven clusters

calls the users made all over Belgium. Merging until three clusters (over 19 towns) highlights interesting groups. The first group that is colored in pink in Figure 7 is located on the West side of the downtown and globally corresponds to the disadvantaged neighborhoods of Brussels while the green group highlights the privileged south-east quadrant of Brussels [12]. As for the two towns colored in Orange, Uccle and Ixelles, the higher education institutions are relativly concentrated there.

Figure 7: Segmentation of Brussels-Capital into three clusters

IV. Conclusion

In this paper, we have focused on graph clustering applied to a telephone dataset. The method allows the discovery of structures in graphs. By clustering the source and target nodes while selecting the best model according to a Bayesian approach, the method behaves as a nonparametric estimator of the edge density. In case of large graphs, the best model tends to be too fine grained for an easy interpretation. To overcome this issue, a post-processing technique is proposed. This technique aims at merging successivly the clusters until obtaining a simplified clustering while worsening the least the model.

Experimentations on a Belgian dataset show the variety of the possible analysis. The finest study yields almost one town per cluster. In order to make a global analysis, the model is coarsened by merging the clusters. With two clusters, the linguistic communities are well segmented while the province boundaries do not match with the clusters delimitations when the clusters are merged until eleven groups. Local groupings based on the call made all over Belgium are also possible and illustrated by the example of Brussels and its suburbs.

Because the method is based on a density estimation, the future works will be extended to the dynamic graphs by adding a third temporal variable. This would enable a study of the temporal evolution of social networks and yield the optimal discretization into time slots.

References

[1] R. Albert and A.-L. Barabasi, "Statistical mechanics of complex networks," *Reviews of Modern Physics*, vol. 74, p. 47, 2002.

[2] S. E. Schaeffer, "Graph clustering," *Computer Science Review*, vol. 1, no. 1, pp. 27 – 64, 2007.

[3] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences*, vol. 99, no. 12, pp. 7821–7826, 2002.

[4] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," 2008.

[5] S. Wasserman, G. Robins, and D. Steinley, *Statistical Network Analysis: Models, Issues, and New Directions*, 2007.

[6] M. Boullé, "A bayes optimal approach for partitioning the values of categorical attributes." *Journal of Machine Learning Research*, vol. 6, pp. 1431–1452, 2005.

[7] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer, New York, 2002.

[8] M. Boullé, "Nonparametric edge density estimation in large graphs," Orange Labs, Tech. Rep., 2011.

[9] ——, *Data grid models for preparation and modeling in supervised learning*. Microtome, 2010.

[10] P. Hansen and N. Mladenovic, "Variable neighborhood search: principles and applications," *European Journal of Operational Research*, vol. 130, pp. 449–467, 2001.

[11] V. D. Blondel, G. Krings, and I. Thomas, "Regions and borders of mobile telephony in belgium and in the brussels metropolitan zone," *the e-journal for academic research on Brussels*, 2010.

[12] C. Kesteloot, C. Vandermotten, and B. Ippersiel, "Dynamic analysis of troubled neighbourhoods in the belgian urban regions," 2007.

Understanding the evolution of spatial structure in urban communities

Fergal Walsh and Alexei Pozdnoukhov

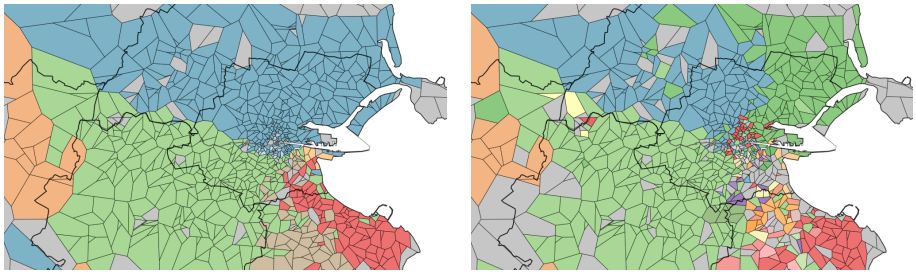
National Centre for Geocomputation
National University of Ireland Maynooth
{fergal.walsh,alexei.pozdnoukhov}@nuim.ie
http://ncg.nuim.ie

This work explores the evolution and spatial organization of urban communities in a daily cycle of a city using scalable social network analysis methods. Using mobile phone records we investigate calling and movement patterns of about 500.000 citizens of Dublin in the course of a typical weekday. The analysis reveals spatial patterns in community structures at fine spatial scales which are most likely related to the activities typical to the different functional regions of a city. To our knowledge this is the most detailed analysis, in terms of both spatial and temporal resolution, carried out to date on this type of data.

We combine two methods adopted from the study of complex networks to examine the time varying spatial structure of communities in an urban environment [1], [2]. The spatial structure of these communities is observed from the analysis of the short-term aggregates of traffic on a mobile telephone network. The analysis is based on an origin destination matrix of calls and text messages between customers.

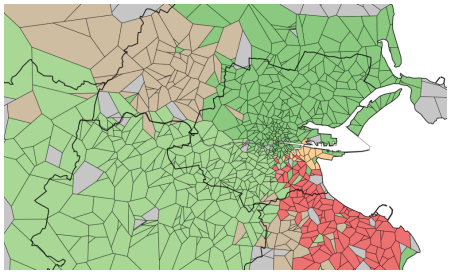
In our case study we use the cells of the network as the spatial areas rather than administrative boundaries. Cells in urban areas are typically very small due to high population densities and so offer high spatial resolution. Rather than using the customer's home address, each end of the call is geo-located to the cell used at the time of the call. The mobile phone network is by definition dynamic, with phone users constantly moving. We use time-of-call locations in an attempt to capture these dynamics. Finally, and most importantly, we perform community detection at multiple snapshots in time, and then analyse the change in community composition [2], [3] . We use a combination of geographical, temporal and graphical layouts linked together in a single view to visualise and interpret these changes (Figure 1).

We observe that different communities are found at different times of day, especially in urban areas where the population and communication habits vary with the time of day. A distinct pattern of transitions and transformations of the communities at the temporal scales of several hours and sub-kilometer distances is observed [4].

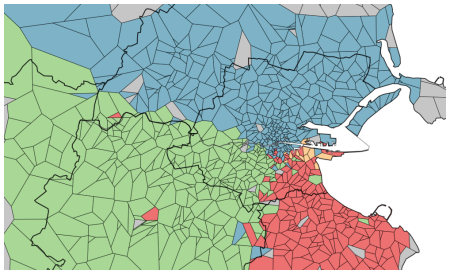


(a) 00:00 - 01:59

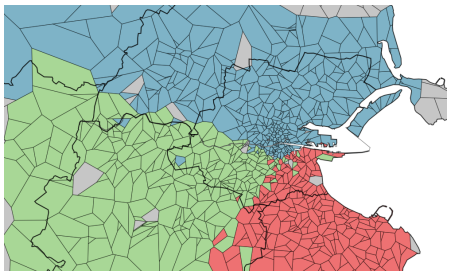
(b) 02:00 - 07:59



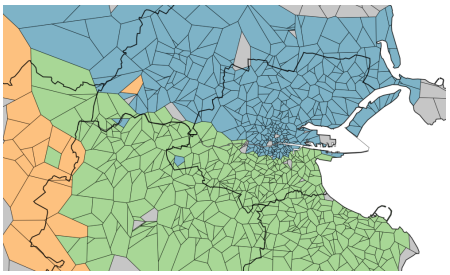
(c) 08:00 - 09:59



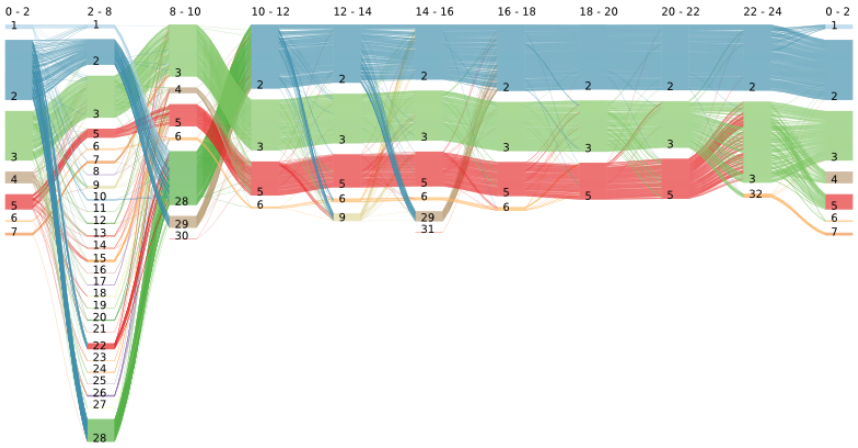
(d) 10:00 - 11:59



(e) 20:00 - 21:59



(f) 22:00 - 23:59



(g) Alluvial timeline diagram of community membership. Each line represents a single cell. Each column represents a time period, as labeled.

Fig. 1: The daily cycle of community structure in the Dublin region.

It is important to understand the forces behind these changes. Are we seeing these changes because people are moving in space or is it because people communicate with different people and different places over the course of the day? Of course the answer is a bit of both, but we attempt to quantify this by doing a combined analysis of communications and physical movements. We show exactly how much physical movement there is between each time step. We also quantify the influence of time of day on the calling patterns of each user. From these two measures we can try to understand the causes of change in spatial structure of communities.

Our combined fine resolution movement-communication analysis has several implications for understanding urban structure. It is able to capture the fundamental properties of human behavior in terms of communication and transportation habits, as well as opening ways to aid decision making in introducing social policies, optimise resource allocation and refine planning of the city infrastructures. Particularly, we observe a distinct North-South divide that splits the city into two parts, separated both in terms of communication and physical movement. We observe the existence of small distinctive communities showing poor communication with the outside world that might indicate ongoing alienation and segregation processes. Generally, these findings also show a further need to extend the typology of community evolution events used in state-of-the-art network evolution models to implicitly incorporate spatial relations at a variety of temporal scales.

Acknowledgments. Research presented in this paper was funded in part through Stokes Lectureship programme and Strategic Research Cluster grant (07/SRC/I1168) by Science Foundation Ireland, and faculty research awards from Google and IBM. The authors gratefully acknowledge this support. We would also like to gratefully acknowledge Dr. R. Farrell and the support of Meteor for providing the data used in this paper, in particular Helene Graham, John Bathe and Adrian Whitwham.

References

1. Blondel, V., Guillaume, J., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment 2008, P10008 (2008)
2. Greene, D., Doyle, D., Cunningham, P.: Tracking the evolution of communities in dynamic social networks. In: Proc. International Conference on Advances in Social Networks Analysis and Mining (ASONAM'10) (2010)
3. Rosvall, M., Bergstrom, C.: Mapping change in large networks. PLoS One 5(1), e8694 (2010)
4. Walsh, F., Pozdnoukhov, A.: Spatial structure and dynamics of urban communities (June 2011), the First Workshop on Pervasive Urban Applications (PURBA)

Trace-driven analysis of data forwarding in opportunistic networks

Merkourios Karaliopoulos Panagiotis Pantazopoulos Eva Jaho Ioannis Stavrakakis
Department of Informatics and Telecommunications
National & Kapodistrian University of Athens
Ilissia, 157 84 Athens, Greece
Email: {mkaralio, ppantaz, ejaho, ioannis}@di.uoa.gr

***Abstract*—We summarize undertaken and ongoing work on the direct and exclusive use of mobile phone traces for assessing the performance of different opportunistic forwarding schemes. Our methods draw on graph-expansion techniques and circumvent the need for more custom simulation software packages. They address a wide range of opportunistic dissemination schemes including controlled flooding and socioaware protocol variants. We outline the general approach and exemplify it with an assessment of centrality metrics as drivers of data dissemination decisions. Finally, we report results on the benefits of identifying community structure out of the similarity of interests across the opportunistic network and discuss their implications for trace-based evaluation.**

I. Introduction

Both the motivation and concerns for the use of real data traces in evaluating protocols and algorithms relate exactly to the words 'real' and 'trace'. On the one hand, they promise realistic performance evaluation and credible results when compared to synthetic input data. On the other hand, they always raise concerns about their representativeness and the generality of the evaluation results. Nevertheless, the use of such traces has become the de facto approach to the evaluation of data dissemination in user-oriented network paradigms such as the opportunistic networking.

We report herein work we have been carrying out on trace-based performance evaluation of different opportunistic forwarding schemes. Our methods draw on graph-expansion techniques and circumvent the need for more custom simulation software packages. They address a wide range of opportunistic dissemination schemes coming under the controlled flooding family of protocols and can be extended to socioaware protocol variants. We focus, in particular, on two main directions that socioaware protocol design in opportunistic networks has taken: the introduction of social metrics into the individual node-oriented relaying utility functions (*e.g.,* [1]), and the explicit a priori assumption that such networks avail community structure that can be detected and exploited in disseminating data (*e.g.,* [2]).

II. Computation of shortest paths over traces

The computation of shortest path(s) for a given message proceeds in three sequential processing steps that differentiate depending on the forwarding scheme and whether shortest paths correspond to minimum-delay or minimum-hopcount space-time paths. In every case, input to this process are time-ordered traces of node contacts, *i.e.,* sequences of contact records with the general format shown in Fig. 1(a). Each contact record $c = (n_1, n_2, t_s, t_e)$ includes four fields: the two nodes that meet, the time their encounter starts, t_s, and the time the encounter ends, t_e; the difference of the last two fields corresponds to the contact duration. The fourth field becomes redundant under the infinite link capacity assumption, *i.e.,* messages need minimal (zero) time to traverse a link between two nodes once this becomes available to them.

1) From the original contact trace to the forwarding contacts: In this step, the original full contact trace is filtered with criteria that account for the different opportunistic schemes so that only those contacts that can result in forwarding of data, hereafter called *forwarding contacts*, are retained. If t_s is the time a message becomes available at source node s for destination node d, then the filtering step first excludes all contact records up to the first one $c_0 = (s, n, t_s)$ involving node s after time t_s. It then initializes an ordered list, hereafter called forwarding list, with the nodes s and n. The forwarding list stores at each time *potential* forwarders of the message, nodes that have acquired the message and may, depending on who they encounter, forward it further.

Contact records after c_0 are scanned sequentially. These contacts may belong to one of three typologies, depending on the encountered nodes: (a) neither node lies in the forwarding list; (b) both nodes are already listed in the forwarding list; and (c) one of the two nodes is in the forwarding list (1-entry contacts). Contacts of the first type do not contribute to the forwarding process and are ignored. On the other hand, contacts of the second type do not represent real additional forwarding opportunities since the assumption in all schemes is that nodes with a message copy will forward it to a node that does not have it and is eligible to acquire it upon the first encounter with it. The most interesting type of contacts is the third one, whose manipulation directly depends on the opportunistic forwarding scheme.

For example, let us consider the manipulation of 1-entry contacts for the two-hop scheme, where nodes other than the message source availing a message copy cannot forward it but only to the destination node. Therefore, two types of 1-entry contacts are retained: those $(s, *, t), t \geq t_s$ involving the source node s as the single node already logged down in the forwarding list, as well as the first 1-entry contact involving the destination node. All other 1-entry contacts are filtered out of the trace (contacts C_1-C_6 in Fig. 1(b)). In fact, when we

The work referred to in this abstract has been supported in part by the European Commission IST-FET project RECOGNITION (FP7-IST- 257756) and the Marie Curie grant RETUNE (FP7-PEOPLE- 2009-IEF-255409).

are after minimum-delay space-time paths, the filtering step terminates upon the first appearance of the destination node d in an 1-entry contact.

2) Building the forwarding contact graph: Outcome of the first processing step is the reduced set of contact records corresponding to forwarding contacts. The next step is to derive the graph construct $G_c = (V_c, E_c)$ that can capture these contacts and their timing relationship. The construct draws on the "temporal" graph representation in [3].

For forwarding schemes under the controlled flooding category, the graph construct is built out of the first contact c_0 and 1-entry forwarding contacts occurring thereafter. Each one of them adds to the graph: a) a pair of vertices, one for each node involved in the contact; b) one *space-spanning* directed edge connecting the two encountered nodes; and c) one *time-spanning* directed edge towards the node that is already included in the forwarding list, originating from the vertex that represents its most recent forwarding contact. Hence, every time a node $v \in V$ appears in an 1-entry forwarding contact, it generates a new vertex $v_c \in V_c$ for construct G_c. As a result, each network node $v \in V$ is eventually identified with a single global index in $[1, |V|]$ for the node set V and the forwarding list, and multiple non-successive indices for the construct V_c.

The graph edge set E_c is weighted. When we are interested in minimum-delay space-time paths, time-spanning edge weights equal the time differences between successive occurrences of the node in forwarding contacts and express the time over which a message may be stored and carried by a given node. Space-spanning edges express the time it takes to forward a message upon a contact and, under the infinite link capacity assumption, are assigned zero weights. On the contrary, when we are interested in minimum-hopcount space-time paths, time-spanning edges are assigned zero weights and space-spanning ones unit weights. The G_c constructs resulting from the contact trace 1(a) for the two-hop forwarding scheme is given in Fig. 1(c).

3) Computing shortest paths: The last processing step consists in the computation of shortest space-time paths over the expanded graph G_c. The size of the graph depends on whether we want to compute minimum-delay or minimum-hopcount space-time paths. When we are after the minimum-delay path(s), the parsing of contacts ends upon the first appearance of the message destination node d in an 1-entry contact. This may be anywhere from the first till the $(|V|-1)^{th}$ contact that is retained in the set of forwarding contacts. On the contrary, to include all possible minimum-hopcount paths, the parsing should continue until either all network nodes enter the forwarding list or the source nodes contact the destination node directly, whatever happens first. It can be shown that the graph constructs G_c are directed acyclic graphs (DAGs) and running Dijkstra will yield the s-d minimum-hopcount space-time path in $O\big((|V_c|+|E_c|)log_2|V_c|\big) = O(|V|^2 log_2|V|)$ time [4].

III. Centrality-based data dissemination

Our trace-based performance evaluation approach can be applied to socioaware opportunistic forwarding schemes. SimBetTS [1] and BubbleRap [2] protocols compute metrics borrowed from Social Network Analysis (SNA) over contact graphs, which effectively aggregate the sequence of node encounters over certain time windows T. Both protocols have identified betweenness centrality (BC) as the dominant user-centric metric, even when it is combined with more metric when making forwarding decisions.

Centrality computation caveats: There are three main concerns regarding the realization of a user-centric approach relying solely on BC. First of all, node centrality values are destination-agnostic; namely, the node relaying utilities are averages computed across all node pairs in the network. Secondly, SNA metrics are computed over graphs. The derivation of these graphs out of the sequence (history) of contacts has been shown to be highly sensitive to the time window T during which all past contacts are aggregated into a contact graph [5]. Thirdly, the original centrality metrics need to be approximated by egocentrically computed centrality variants [6], which, in principle, offer only limited views of the node's utility in the network. We have employed real human mobility traces of pairwise node encounters to experimentally study these three factors and their impact on the BC-based data dissemination [7].

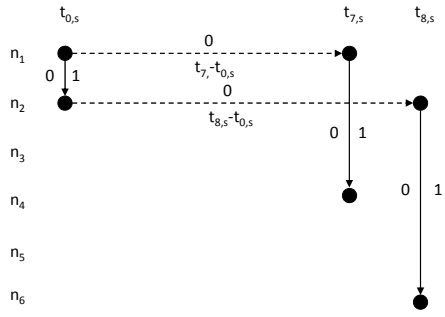
Mobile traces: We have used five well-known experimental traces, part of the iMote-based traceset available in [8]. The traces cover a rich diversity of environments with an experimental period from few days to almost one month. All traces, gathered over the last five years, include Bluetooth sightings of users carrying iMotes. Each Bluetooth sighting is assumed

contact id	involved nodes		contact start time	contact end time	additional fields
...	...	...	...	...	...
C_0	n1	n2	$t_{0,s}$	$t_{0,e}$	...
C_1	n3	n4	$t_{1,s}$	$t_{1,e}$	...
C_2	n4	n5	$t_{2,s}$	$t_{2,e}$	...
C_3	n2	n5	$t_{3,s}$	$t_{3,e}$	...
C_4	n5	n3	$t_{4,s}$	$t_{4,e}$	...
C_5	n3	n2	$t_{5,s}$	$t_{5,e}$	...
C_6	n3	n6	$t_{6,s}$	$t_{6,e}$	...
C_7	n1	n4	$t_{7,s}$	$t_{7,e}$	...
C_8	n2	n6	$t_{8,s}$	$t_{8,e}$	...
...	...	...	...	...	...

(a) Original trace.

contact id	involved nodes		contact start time	contact end time	additional fields
...	...	...	...	...	...
C_0	n1	n2	$t_{0,s}$	$t_{0,e}$	...
C_1	n3	n4	$t_{1,s}$	$t_{1,e}$	...
C_2	n4	n5	$t_{2,s}$	$t_{2,e}$	...
C_3	n2	n5	$t_{3,s}$	$t_{3,e}$	...
C_4	n5	n3	$t_{4,s}$	$t_{4,e}$	...
C_5	n3	n2	$t_{5,s}$	$t_{5,e}$	...
C_6	n3	n6	$t_{6,s}$	$t_{6,e}$	...
C_7	n1	n4	$t_{7,s}$	$t_{7,e}$	...
C_8	n2	n6	$t_{8,s}$	$t_{8,e}$	...
...	...	...	...	...	...

(b) Forwarding contacts: two-hop forwarding



(c) Two-hop forwarding graph.

Fig. 1. Original contact trace, forwarding contacts (entries in bold), and resulting graph construct for message $m = (1, 6, t_{0,s})$: two-hop forwarding scheme.

TABLE I
CHARACTERISTICS OF EMPLOYED DATASETS

Configuration	Intel	Cambridge	Infocom05	Content	Infocom06
Device type	iMote	iMote	iMote	iMote	iMote
Network type	B/T	B/T	B/T	B/T	B/T
Duration (days)	6	6	4	24	4
Scan time (sec)	5-10	5-10	5-10	5-10	5-10
Granularity (sec)	120	120	120	120-600	120
Mobile Devices	8	12	41	36	78
Stationary Dev.	1	0	0	18	20
External Dev.	119	211	233	11368	4421
Average internal contacts/pair/day	9.09	12.09	8.60	0.66	9.03
# of Contacts	2766	6732	28216	41330	227657

to be a contact where nodes can exchange information. In Table I scan time is the time needed by iMotes to perform a complete scan for Bluetooth devices and takes approximately 5 to 10 seconds; time granularity represents the idle time between two consecutive scans and affects significantly the measurement accuracy. We analyze only the contacts between iMotes (*i.e.*, internal contacts), which represent the data transfer opportunities among participants.

Emulation of optimal routing over the traces: The theoretically optimal paths (of minimum delay and hopcount) to the destination, have been computed directly out of the dataset sequence of encounters according to the expanded graph technique, introduced in Section II. The outcome values are naturally considered as performance benchmarks.

Emulation of BC-based routing over the traces: To account for the relative social standing of each node we need to aggregate the encounters' history to an unweighted or a more "informed" weighted static graph with link weights equal to the frequency of contacts, over which the centrality values are computed. The trace is again replayed (sequentially read) but now we compute five different centrality-variants for each contact record; a message is forwarded provided that the encountered node exhibits higher value of the corresponding variant than the one of the current holder. These values may be either the sociocentric (including the destination-aware BC variant called Conditional Betweenness Centrality (CBC) [9]) or their egocentric counterparts [6] computed over the full set of contacts within the past T time window.

Summary of results: We have generated messages with randomly chosen source and destination and emulated their paths over the traces. The message delivery delay and number of forwarding hops have been computed and compared with those of the optimal (opt) scheme described above. Our findings are summarized in the following discussion. The centrality-based approaches perform considerably worse than the optimal method both in terms of message delay and hops to the destination. Forwaring performance across the different traces depends on the way nodes' mobility patterns mix with each other. Less intuitively, replacing BC with its destination-aware counterpart (CBC) does not give benefits in terms of message delay but results in significant energy savings by reducing the message hops. When computing BC over weighted graphs our study reports that the performance does not consistently improve for all traces. Nevertheless, the routing protocol is more resilient to variations of the time window used for contact aggregation. Finally, we have found that using the egocentric BC variant penalizes the forwarding performance only marginally, when computed over unweighted and even less when computed over weighted graphs. This is further supported by the observed strong positive correlation between socio- and egocentric BC in almost all traces. A more detailed discussion as well as explanatory plots appear in [7].

IV. ENRICHING TRACES WITH USERS' INTERESTS

In [10], we have enhanced primitive push mechanisms for data dissemination in opportunistic settings with social information concerning the preferences and interests of network nodes. It is shown that *interest-based forwarding* can improve considerably the information dissemination process. Inspired by this result, in [11] we have proposed a framework called *ISCoDe*, which identifies communities of nodes with similar interests. We have applied *ISCoDe* to the Delicious (www.delicious.com) platform, showing how end-user interests can be inferred out of a real online social networking (OSN) application.

Adding further realism to this thread would call for data traces that, besides encounters, provide information about the preferences of users. Enriching mobile phone datasets with such information is one option that could be considered in future datasets to become available. This information encoded in the form of user preference distributions (profiles) over a set of certain thematic areas (such as music, sports, art), can be inferred out of tags annotating data that users save in their mobile phones. A more demanding alternative would be to indirectly infer such information from online social networks, trying to correlate traces of encounters with OSN user profiles (*e.g.*, [12]).

REFERENCES

[1] E. M. Daly and M. Haahr, "Social network analysis for information flow in disconnected delay-tolerant manets," *IEEE Trans. Mob. Comput.*, vol. 8, no. 5, pp. 606–621, 2009.

[2] P. Hui *et al.*, "Bubble rap: Social-based forwarding in delay tolerant networks," *IEEE Trans. Mob. Comput.*, (To Appear) 2011.

[3] V. Kostakos, "Temporal graphs," *Physica A-statistical Mechanics and Its Applications*, vol. 388, pp. 1007–1023, 2009.

[4] M. Karaliopoulos, "Trace-based performance analysis of opportunistic forwarding under imperfect cooperation conditions," University of Athens, Tech. Rep., July 2011.

[5] T. Hossmann, T. Spyropoulos, and F. Legendre, "Know thy neighbor: Towards optimal mapping of contacts to social graphs for dtn routing," in *Proc. IEEE Infocom 2010*, San Diego, CA, USA, Mar 2010.

[6] P. Marsden, "Egocentric and sociocentric measures of network centrality," *Social Networks*, vol. 24, no. 4, pp. 407–422, October 2002.

[7] P. Nikolopoulos, T. Papadimitriou, P. Pantazopoulos, M. Karaliopoulos, and I. Stavrakakis, "How much off-center are centrality metrics for routing in opportunistic networks," in *ACM MobiCom 2011 CHANTS Workshop)*, Las Vegas, NV, USA, Sep 2011.

[8] J. Scott, R. Gass, J. Crowcroft, P. Hui, C. Diot, and A. Chaintreau, "CRAWDAD data set cambridge/haggle (v. 2009-05-29)," Downloaded from http://crawdad.cs.dartmouth.edu/cambridge/haggle, May 2009.

[9] P. Pantazopoulos, I. Stavrakakis, A. Passarella, and M. Conti, "Efficient social-aware content placement for opportunistic networks," in *IFIP/IEEE WONS*, Kranjska Gora, Slovenia, February, 3-5 2010.

[10] S. Allen, M. Chorley, G. Colombo, E. Jaho, M. Karaliopoulos, I. Stavrakakis, and R. Whitaker, "Exploiting user interest similarity and social links for microblog forwarding in mobile opportunistic networks," in *Elsevier Pervasive and Mobile Comuting (submitted)*, 2011.

[11] E. Jaho, M. Karaliopoulos, and I. Stavrakakis, "ISCoDe: A framework for interest similarity-based community detection in social networks," in *Proc. 3rd Int'l Workshop on Network Science for Communication Networks*, 2011.

[12] G. Bigwood and T. Henderson, "Bootstrapping opportunistic networks using social roles," in *Proc. 5th IEEE WoWMoM Workshop on Autonomic and Opportunistic Communications (AOC)*, 2011.

Statistical detection of information flow traces in a large phone call dataset

Lionel Tabourier*

LIP6, Université Pierre et Marie Curie, Paris.

Fernando Peruani†

Max Planck Institute for the Physics of Complex Systems, Dresden.

Abstract

The content of phone calls is in general out of reach for confidentiality reasons. In addition, the massive amount of information exchange in large mobile phone networks makes in practice impossible to analyze every conversation in order to identify how information flows. That is why it would be of great importance to design *'content-free'* methods to identify where and when information propagation takes place.

Here we propose a method to address this issue that combines dynamical measurement tools with appropriate comparison models. Our method relies on the assumption that intentional information propagation implies strong causality effects. In our analysis, phone calls are intended to transmit information and involve a causality directionality: from the caller to the callee. The transmission of information in the opposite direction, while clearly possible, does not imply an intentionality and consequently causality. This simple assumption leads us to take into account the directed character of the mobile phone data. By implementing the same measures on a real dataset and null models, we are able to identify when and where causality effects impact the information spreading statistics.

We used for this study a large cellphone record providing who calls whom and when. It consists of calls among 1 million subscribers of the same mobile phone company over a period of a month. As we are interested in the broadcasting of information, we focus on actual calls where the caller indeed reach the callee, restricting the dataset to around 14 million events.

We run on the database a set of statistical tools and collect statistics on several dynamical patterns. A central measurement of this work is the amount of cascade-like patterns: if a node is 'activated' during a period τ after being called and may activate other nodes by calling them, we describe the obtained

*lionel.tabourier@lip6.fr

†peruani@pks.mpg.de

patterns as cascades — or causality trees — which may be seen as proxies of information propagation processes.

We first lay stress on the importance of taking into account the directedness of such data: by combining statistics on causality trees with theoretical calculations, we prove that the spreading processes are highly dependent on the in-out degree correlations exhibited by the users. We also learn that a given information, e.g. a rumor, would require users to retransmit it for more than 30 hours in order to cover a macroscopic fraction of the system.

Then, we build models that are obtained by randomizing a part of the original dataset so that we keep the ingredients which are essential for the understanding of the communication behavior over the whole network. Two comparison models are proposed: the first one consists in randomizing the timestamps of events over the whole dataset. It is a well-known fact that diffusion phenomena widely depends on temporal patterns such as day-night periodicity or bursting activity of users. However, this trivial model yields results on statistics which roughly reproduces the behavior in the real dataset, so that it can be used as a baseline to compare measurements on the real network and on a more precise model.

The second comparison model aims specifically at detecting traces of diffusion phenomena. In both its purposes and features, it has to our knowledge no equivalent in the literature: it is supposed to keep all the characteristics of the original dataset except for the causality link that may exist between a phone call received and a phone call given. The calling activity of each individual remains indeed unchanged as well as the destinations of the calls but the correlation possibly existing between the phone calls given by two different users are broken.

By focusing on subtle features of the dynamical data, this model reveals that causality effects are only visible as local phenomena and during short time-scales — in this context no more than a few hours — making information flow traces in such dataset hard to detect. However, we show that node-node correlations of the underlying social network, while allowing the existence of information loops, promote information spreading at short range and during short time-scales. It is indeed possible to discover very specific motifs, such as star-like and cycle-like communication patterns, which are underestimated by the model, and we can assess the probability that such patterns are related to a diffusion process.

The method that we propose is thus a step forward in locating where and when information flow happens in a phone call dataset without resorting to the content of the calls. As it only relies on having a large sequence of timestamped events, we suggest that such analysis could be used in other contexts of information spreading over large communication datasets, such as emails exchanges or instant messaging networks.

Socially-mediated diffusion via cell phones: an analysis of ringback diffusion in a cell phone network

Isar Kiani[1], Derek Ruths[2], Thomas Shultz[3]

[1] John Molson School of Business, Concordia University
[2] School of Computer Science, McGill University
[3] Department of Psychology, McGill University

The ringback is a recent innovation introduced by some cell phone service providers which permits a cell phone owner to select the ringtone another individual hears when calling that person's cell number. Like many features introduced in the mobile service market, the ringback provides a way for customers to personalize the cell phone experience that they and members of their social circles have.

In some implementations of such a system, the primary way to acquire a ringback is to copy it from another user who already has it. This social rule for ringback diffusion combined with the singular utility of ringbacks as an acoustic social gesture from the callee to the caller makes the spread of ringbacks through a cell phone calling network a useful phenomenon for understanding the larger issue of how social content propagates through mobile networks.

In this project, we used a data set from a leading telecommunication company in the Middle East to characterize the diffusion profiles of 3,434 ringbacks. The dataset was collected over a 2.5 month period, from August 19 to October 31, 2010. It consists of 944,271 ringback adoption events among 1,261,044 distinct users. The service provider specifies a classification of each ringback into one of four categories: pop, film and TV, sports, and religion.

We find that, despite ringback content and introduction times being different, most ringback diffusion curves can be categorized into a small number of well-supported curves. We use Kullback-Leibler divergence to obtain pairwise diffusion curve distances and, after clustering, explore several ringtone features that might explain diffusion patterns.

In addition, we also consider the individual ringback diffusion networks. Here, too, we can cluster ringbacks according to those whose networks share similar topological features including degree distribution, motif densities, and diameter. We find that ringback network structure also can be classified into a relatively small number of classes.

Our preliminary evidence suggests the presence of diffusion mechanisms that are well-conserved across subsets of ringbacks and that attributes of individual ringbacks influence the mechanisms that are responsible for their eventual diffusion through the cell phone network.

Estimating Dynamic Urban Population Through Assimilated Mobile Phone Data

Teerayut Horanont
Institute of Industrial Science, University of Tokyo
E-mail: teerayut@iis.u-tokyo.ac.jp

Ryosuke Shibasaki
Center for Spatial Information Science, University of Tokyo
E-mail: shiba@csis.u-tokyo.ac.jp

***Abstract*—Up to this day, the understanding of real-time population distribution remains highly questionable and problematic. Therefore, this research provides a novel method in order to capture of large-scale quantitative data related to human mobility. We developed techniques using mobile phone Call Detail Records (CDRs) and trip survey data to provide an estimated population at high temporal resolution. The trip surveys are served as unbiased movement samples while mobile phones CDRs indicate the current state of the observations. The assimilation method gives a prediction of present population by incorporating the actual appearance of people over time from the concurrent generated mobile phone activities. We demonstrated and tested this methodology on the simulation-based system and the results provide timely and more spatially precise population estimation in grid square basis. (Abstract)**

Keywords- UbiGIS; Data assimilation; Population dynamics; Urban mobility; Mobile simulation

I. INTRODUCTION

The concept of the daytime population refers to the number of people who are present in an area during the day or during normal business hours. This is in contrast to the census data that represented resident or nighttime population. In 1989, Stanley K. Smith [9] proposed a methodology to estimate temporary residents. The limitation of this study is laid on the lack of data sources that provide complete, consistent coverage of temporary movement and migration. Instead, estimates have to be cobbled together from a variety of administrative records, business statistics, and sample surveys. There have been other considerable studies done to understand human mobility and improve the spatial resolution of static population counts through the use of census data [1][3]. They introduced a project call "LandScan", a population distribution model created by Oak Ridge National Laboratory, that seeks to overcome the limitations of aggregated and static population counts by estimating high spatial resolution population distribution for both day and night. The project suggested a multi-dimensional dasymetric modeling approach, which has allowed the creation of a high-resolution population distribution data over the area. However knowing the average day time population is insufficient in many particular cases since spatial distribution of people changes over time dependent on the activities and events that occurred at various times of the day.

II. MOBILE PHONES AS NEW URBAN SENSORS

In recent years, the large deployment of mobile and wireless technologies has provided new means to understand the dynamics of a city. The footprint of mobile phone use is overwhelmingly, and has a side incredible benefit. It is therefore not surprising that considerable research has gone in this new and interesting directions, in particular, developing techniques to bring out and explore large scale human mobility [2][4][5]. The previous research [7] has show that how erlang data, which is a measurement unit of a telecommunication network, can provide an important new way of looking at the city as a holistic, dynamic system. This approach can complement traditional data collection techniques, which are often lack of present information and outdated. Other studies have suggested that erlang measured at particular cells seem to be a decent indicator of actual presence of people [6][8]. However, without additional information, erlang data itself could not yield the existing of people who are not using the mobile phone. A recent study [10] Song used cell phone billing data of 50,000 people in a European country to show that people's travel patterns are extremely predictable. He found that most people travel very little on a daily basis, for instance, 5 to 10 kilometers or so. There were a few individuals, who on a daily basis travel hundreds of kilometers. This finding suggests that, for the vast majority of the people, there is an average of 93 percent predictability across the user base. It is also means that by analyzing the amount of mobile phone footprints, we could possible predict the existence of people and their commuting path in a daily basis. In contrast with other work in utilizing mobile phone data for human mobility research, this work is focusing on measuring spatial redistribution of population over the metropolitan area on a 24-hour cycle.

REFERENCES

[1] Bhaduri, B., E. Bright, P. Coleman, and J. Dobson. (2002) LandScan: Locating People is What Matters, Geoinformatics Vol. 5, No. 2, 34-37.

[2] Candia, J., Gonzalez, M.C., Wang, P., Schoenharl, T., Madey, G., Barabasi. (2008) A.: Uncovering individual and collective human dynamics from mobile phone records. Journal of Physics A: Mathematical and Theoretical Vol. 41(22), 1–16

[3] Dobson, J. E., E. A. Bright, P. R. Coleman, R.C. Durfee, AND B. A. Worley (2000) LandScan: A global population database for estimating populations at risk. Photogrammetric Engineering & Remote Sensing Vol. 66(7), 849-857

[4] Eagle, N., Pentland, A., Lazer, D. (2009) Inferring social network structure using mobile phone data. Proceedings of the National Academy of Sciences (PNAS) Vol. 106(36), 15274-15278.

[5] Gonzalez, M.C., Hidalgo, C.A., Barabasi, A.L. (2008) Understanding individual human mobility patterns. Nature Vol. 453(7196), 779–782

[6] Horanont, T., Shibasaki, R. (2008) An Implementation of Mobile Sensing For Large-Scale Urban Monitoring, UrbanSense08, 4 November 2008.

[7] Ratti, C. (2007). Go with the flow. The Economist Technology Quarterly, March, 12-13

[8] Reades, J., Calabrese, F., Sevtsuk, A., and Ratti, C. (2007) Cellular Census: Explorations in Urban Data Collection, IEEE Pervasive Computing, Vol. 6(3), 30-38

[9] Smith, S.K. (1989) Toward a Methodology for Estimating Temporary Residents, Journal of the American Statistical Association, Vol. 84, 430-436.

[10] Song, C., Qu, Z., Blumm, N., Barabsi, A.L. (2010) Limits of predictability in human mobility. Science, Vol. 327(5968), 1018–1021.

Modelling City Population Dynamics from Cell Phone Usage Data Streams

Christian Kaiser and Alexei Pozdnoukhov

National Centre for Geocomputation
National University of Ireland Maynooth
{christian.kaiser,alexei.pozdnoukhov}@nuim.ie
http://ncg.nuim.ie/i2maps/

Telecommunication systems with their high penetration into modern society provide huge volumes of streaming data on human activities. Data streams are often aggregated and referenced to an areal spatial unit such as a polygon rather than to a precise point in space. This is the case when geo-referencing is done by user IP addresses or from a mobile phone cell ID covering some geographical area, or data are initially aggregated within an area and attributed to an extended region due to privacy issues. In these cases it is important to adapt the processing methods to correctly account for an extended spatial support. A typical example is a spatial interpolation problem when a continuous surface (Figure 1, right) has to be produced from areal data (Figure 1, left) for downscaling, data homogenization and interoperability in further processing, or simply for visualization. Many recent applications have demonstrated interpolated human activity heatmaps from areal data such as mobile phone cells [4], however, delt with such data in an ad-hoc way with no regard to its areal support at interpolation step. It limits the usefulness of such maps for rigorous high-fidelity spatial analysis and decision making as intra-cell areal aggregates and inter-cell flows are not preserved.

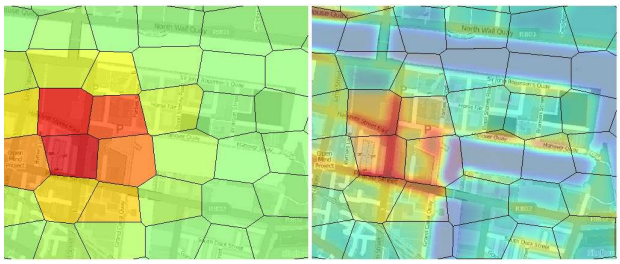


Fig. 1: An example of the area-to-point regression: people density interpolation from data aggregated over cell polygons.

Spatial statistics offers various approaches to overcome this problem in part, including the so-called pycnophylactic interpolation [5] and area-to-point kriging [3]. Surprisingly, there is a theoretical framework that can provide similar functionality for a wider class of methods in machine learning, and kernel methods in particular. It is known as the Vicinal Risk Minimisation principle [6]. We develop this framework for spatial data, derive analytical expressions for

Gaussian-like vicinities and introduce an efficient numerical integration scheme that can be used for performing various learning tasks with arbitrary kernels and polygonal units. The framework is generally applicable to various content-rich data types such as user-generated geo-referenced texts or images.

Particularly, we implement these ideas as an incremental learning algorithm based on kernel reqursive least squares (KRLS) [1] and apply the derived methods on streaming data. We further enhance the scalability of the method with a MapReduce-based distributed implementation [2] available as an open source project.

Population density sensing. Population dynamics data are available in various forms of spatial aggregates streaming from fixed hardware installations. In case the polygonal spatial areas within which streaming data is collected are known, one can define density $p(\mathbf{x}|\mathbf{x}_i, r_i)$ within these polygons depending on the type of a measurement system at hand: (1) $p(\mathbf{x}|\mathbf{x}_i, r_i) \sim const$ corresponding to simple aggregation over the polygon area, or: (2) $p(\mathbf{x}|\mathbf{x}_i, r_i) \sim 1/S_i$ corresponding to intensity-type measurements, and modulate it according to the spatial alighnment of physical infrastructures within a cell. We used the OpenStreetMap geometries to compute prior occupancy probabilities and enhance spatial fidelity of population estimates. An animation illustrating the temporal dynamics of the estimates from streaming data is available online[1].

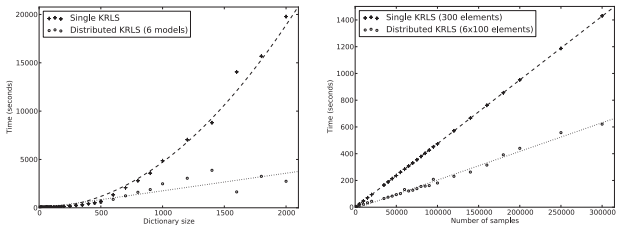


Fig. 2: Processing time with respect to the number of entries in a dictionary ℓ and a number of samples processed in a stream.

Scaling properties and wide-area deployment. We used an implementation introduced in [2] to investigate the properties of the method when applied to high volume stream processing. It operates with a distributed ensemble of kernel predictors each trained incrementally and stored at local nodes. The implementation is built using the MapReduce framework. Two baseline scaling properties we would like to demonstrate (see Figure 2) is a constant processing time per sample in a stream and a quadratic growth of time with respect to the number of kernels used in a distributed kernel dictionary.

We also report on experience gained in our current work of implementing this approach in a real-life setting using a dataset of CDR mobile phone records

[1] `http://www.youtube.com/watch?v=NJB5Cv_WfMM`

for wide area real-time population estimates in Greater Dublin region in Ireland. We are performing spatial data pre-procesing for parallel kernel computations over units defined by cell tower geometries using the computational facilities of the Irish Centre for High-End Computing. The implementation of a streaming component on mobile phone data records runs locally on a conventional mid-range 2.44GHz 8-core server.

Conclusions. Geospatial data analysis often requires dealing with data available at different supports. While spatial statistics provide suitable tools to approach this problem, state-of-the-art applications nevertheless often deal with such data in an ad-hoc way. For example, this concerns non-intrusive population density sensing by leveraging data streams from existing telecommunication infrastructures. The knowledge of physical environment and various constraints available from high-resolution spatial databases is often either neglected or incorporated at the post-processing stage via simple thresholding.

We introduced a mathematically rigorous and consistent framework to incorporate this knowledge into spatial modelling. This framework is applicable for a broad class of machine learning algorithms. We derived a particular area-to-point regression method and adapted an incremental training algorithm to apply it for streaming data. The validity of the method for real-time sensing of city population density from a typical data stream available from telecommunication infrastructures was demonstrated. We investigated the scaling properties of the algorithm and found it to be promising approach to take in our current work which extends it into a larger-scale system useful to uncover high-fidelity patterns of city dynamics.

Acknowledgments. Research presented in this paper was funded in part through Stokes Lectureship programme and Strategic Research Cluster grant (07/SRC/I1168) by Science Foundation Ireland, and faculty research awards from Google and IBM. The authors gratefully acknowledge this support.

References

1. Engel, Y., Mannor, S., Meir, R.: The kernel recursive least-squares algorithm. IEEE Transactions on Signal Processing 52(8), 2275–2285 (2004)
2. Kaiser, C., Pozdnoukhov, A.: Enabling real-time city sensing with kernel stream oracles and mapreduce (June 2011), the First Workshop on Pervasive Urban Applications (PURBA)
3. Kyriakidis, P.C.: A geostatistical framework for Area-to-Point spatial interpolation. Geographical Analysis 36(3), 259–289 (2004)
4. Ratti, C., Pulselli, R., Williams, S., Frenchman, D.: Mobile landscapes: Using location data from cell phones for urban analysis. Environment and Planning B 5(33), 727–748 (2006)
5. Tobler, W.: Smooth pychnophylactic interpolation for geographical regions. Journal of the American Statistical Association 367(74), 519–530 (1979)
6. Vapnik, V.N.: Statistical Learning Theory. Wiley-Interscience (September 1998)

Exploiting cellular network MSC counters for the analysis of temporary populations

Paolo Tagliolato, Fabio Manfredini, Carmelo Di Rosa
Dipartimento di Architettura e Pianificazione, Politecnico di Milano
Via Bonardi, 9 - 20133 Milano, Italy
{paolo.tagliolato, fabio.manfredini, carmelo.dirosa}@polimi.it

Introduction

Visitors and tourists are among the more important urban populations for their impact on local economy, on global fluxes dimensions and on urban vitality and attractiveness.

Despite the fact that this assumption is widely shared within the urban studies community, there is a lack of available and real-time data on such significant dimensions of contemporary cities. Tourism statistics contain data about the activity of registered accommodation establishments, hotels, motels, hostels, youth hostels, holiday villages and camping sites. Data are collected monthly, electronically or with questionnaires or as computer printouts from the accommodation establishments within the scope of the statistics.

Urban attractiveness is moreover characterized by other forms of tourism and visitors, which are not adequately taken into account by traditional surveys and statistics, e.g. the huge amount of people attracted by cities for every reason and moving beyond the boundary of official channels (black market accommodation, hospitality from friends, couch-surfing, etc.).

In recent years we assisted at an extraordinary increase in mobile communications, a phenomenon referred to as the next social revolution [7]. In this scenario, mobile phone is the widest adopted technology and, according to several studies, it is changing the way people organizes its daily life [3][5]. New ICT offers users new time-spatial coordination tools and it influences human activity-travel behavior, tending to increase people's spatial mobility [8] [4], and enabling people to carry on several activities while on the move [6]. Mobile phone traffic data are promising sources for large scale surveys due to the high pervasiveness of cellular phones in contemporary societies [1]: users provide information about the (use of the) territories by simply using a technology that is more and more ubiquitous, or, in the words of Weiser [9], a technology that disappeared "woven into the fabric of everyday life". Moreover, these data are automatically generated from the telecommunication networks, so that there is no need for huge investment such as those necessary for the acquisition of traditional data banks.

In this paper we will introduce and exploit a particular kind of cellular network data, namely HLR MSC counters, novel in literature at our knowledge. What could be the contribute of MSC data to extend conventional sources was the core topic of our research.

NetMob 2011 MIT (Media Lab), Cambridge, MA, October 10-11, 2011

Mobile switching center data

We had the opportunity to use data about switched-on mobile phone rather than on mobile phone traffic (calls received or made). The data refer to the mobile switching center (MSC), which is the primary service delivery node for GSM, responsible for routing voice calls and SMS. It also records information on the mobility of subscribers by updating the position of mobile devices in the HLR (Home Location Register), which contains all subscriber information.

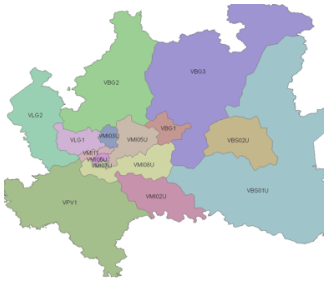


Figure 1: Map of the Lombardia region's MSC Service Area and patterns of registered users

The service area of each MSC has a variable size depending on the number of GSM tower cells served and on the intensity of mobile phone traffic generated inside it. MSC service areas are small in dense urban areas and they can be very large in suburban areas (even thousands of square kilometers). Despite their different sizes, the spatial distribution of the MSC serving areas appears to be very interesting in terms of urban dynamics and land use.[1]

The data at our disposal were, for each MSC service area of the Lombardia region , the number of registered users (i.e. having the phone turned on and logged on TIM's network) distinguished by the nationality of the SIM at a hourly base. The interest in this type of information, despite of the low spatial resolution, is due to the possibility to have access to information directly related to the number of active phones, which may correspond approximately to hourly people presence in the different MSC service area.

The further possibility to access information on the na-

[1]For example: the central area of Milan is divided into two MSC service areas; other important cities like Bergamo and Brescia are defined by a single MSC service area; the sparse and mountainous northern side of the region is also divided in two big MSC service areas; the dynamic northern side of Milan is shared by 3 MSC service areas.

tionality of the SIM of active customers has also suggested to make an assessment on the potential of this data for monitoring foreign presences in different MSC.

Data analysis

We carried out some analysis on the variability of the number of active users in each MSC for two periods of time (September 7 20, 2009; April 1 30, 2010). We tried to determine whether and how these data might provide new and useful understanding of urban temporary population dynamics at different temporal scales.We considered the density of telephone contacts (active customers per square km), in order to take into account the substantial differences in size between the MSC, and performed several analyses on its variation over time in the MSC service areas of the Lombardia region .

Monitoring urban populations

The variability of active clients during working days highlights, for the urban dense Milan MSC service areas, the attractive role of the city for jobs and services. In fact, during working days, the city of Milan attracts a large amount of people from a vast territory that goes beyond the municipal boundaries, namely Milan Urban region. In other terms, this phenomenon can be referred to the daily mobility and its dimension. We show (figure 2) this result by comparing, at an hourly basis, the amount of active clients during a Wednesday and during a Saturday in two areas. We observe that the ratio Wednesday / Saturday sharply increases in Milan during the first hours of the day until 8 AM (marked in black). Afterwards, the amount of contacts is steady. After 4 PM the number of active clients decreases very slowly, until it reaches the same value of the morning.

The number of contacts during Wednesday is more than a half than that of Saturday. Milan municipality's official data inform that each working day almost 600.000 people enter in Milan, a value that is compatible with what emerges from MSC service area statistics.

The dotted line refers to a mountainous MSC service area in the Northern Lombardy region, This area is characterized by a dependent profile in terms of people attractiveness during the working days, compatible with the activities present in the area.

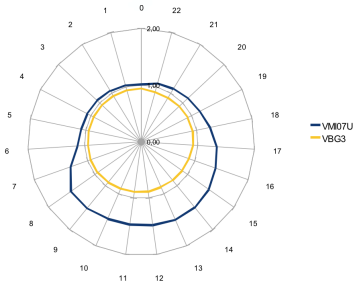


Figure 2: Hourly Active clients: working day / Saturday. Service area central Milan (7-20/9/2009)

The monthly Milan city rhythm is shown in figure 3, which represents the hourly variation of active clients in one MSC service area for the period April, 1st 2010 April, 30th 2010.

The graph highlights different temporal patterns of daily population:

- huge decrease of active clients during the weekend due to reduction of job and study travels
- characteristic inflection in the Eastern period. Weekend April 4 5 more pronounced than that of other April weekends (inhabitants of Milan leave the city)

The possibility to obtain an indirect measure of the temporal variation of presences in the city is a great value of this data, relevant for policy and decision makers.

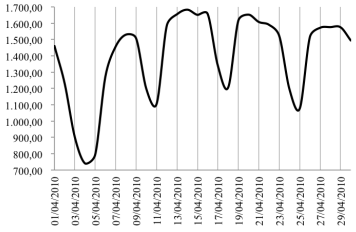


Figure 3: Active clients density in the MSC service area of Milan (1-30/4/2010)

Foreigners trends and evidences

MSC counters register the nationality of the SIM connected to the system. We tried to determine the potential of this information for monitoring tourists and visitors, studying their variability in Lombardia Region MSC service areas.

Nowadays, tourism statistics are related to capacity and occupancy in collective tourist accommodation, and they are collected via surveys filled in by accommodation establishments. Statistics on the occupancy of collective tourist accommodation refer to the number of arrivals (at accommodation establishments) and the number of nights spent by residents and non-residents, separated into establishment type or region; annual and monthly statistical series are available.

This information is available at the provincial spatial scale and underestimates the dimension of phenomena, because it considers only tourists or visitors who go to official structures. A large proportion of temporary visitors are therefore excluded. It is estimated that in Milan there are about 500,000 visitors per year who are not counted by statistics because they spend nights by friends or by informal touristic structures (Observatory of Tourism - Milan).

On the other hand, a strong need to know the dimension of tourism is expressed by several urban stakeholders, such as municipalities, public agencies, industry and trade organizations: they are interested to increase the attractiveness of urban regions, providing new services and activities targeted for this type of temporary population (tourists, business people, etc.).

We focused on the most numerous nationalities registered in MSC in September '09. The first, German (23%), was mostly concentrated in the big MSC service area covering the western side of the Lombardia plain and the Garda lake (fig. 4), one of the main touristic attractor in inland Northern Italy, in particular for German population. The trend during September 2009 is clearly due to the presence of German tourists in the Garda Lake area for the summer.

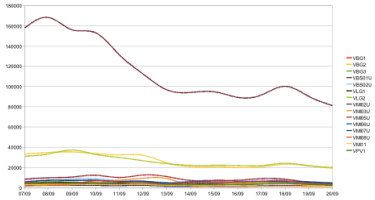


Figure 4: German active clients in MSC service area of Lombardia region - Sept.2009

The National Statistic survey on Tourism indicates, in September 2008, in the province of Brescia, the presence of about 915,000 tourists, most of which (82%) were concentrated in the lake and the mountain areas, entirely included in one MSC service area. In the same year, Germans scored more than 2,000,000 admissions, representing 41% of total foreign tourists and 25% of the total. Although the two data sources differ in level of detail, period, nature of the data, it seems that the data on the variability of active clients MSC can effectively describe the dynamics of tourist arrivals.

Swiss visitors (15%) have a rather different trend. Their distribution is indeed more concentrated in the MSC area near the border between Switzerland and Italy, in the northern side of the region and it is characterized by a positive peak on Saturdays and Sundays.

Americans are mainly concentrated in the center of Milan and show a growing trend on working days, from Monday onwards, and a subsequent decrease in the weekend, a phenomenon that seems to be compatible with business travel rather than tourism travel.

Monitoring visitors during a big event

We performed specific investigations regarding the period of the *Milan International Design Week* (April 14 19, 2010), a leading event that takes place in the Rho-Pero Exhibition District and in dozens of places inside the city where art galleries and museums, shops, industrial areas and showrooms host events and exhibitions, parties and special initiatives (named "Fuorisalone").

For a better comprehension of foreign dynamics, Italians were excluded from the analysis. The foreign active clients present a trend that is very related with the International Design Week development. In fact, it is immediate to observe that since the day before the beginning of the Event (April 13), the number of foreigners increases sharply until it achieves its peak on Saturday April, 17. Afterwards the curve decreases slowly and reaches a more typical course.

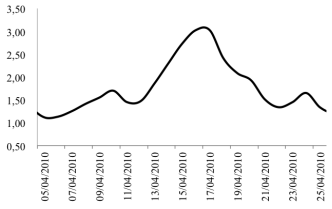


Figure 5: Foreign active clients density in the MSC service area of Milan - The International Design Week (April 14 to 19, 2010).

Among the more than one hundred registered nationalities, during the week of the event, the most represented ones were France, Switzerland, United Kingdom, Spain, German, Russian Federation and USA, accounting for more than the 50% of the total foreigners.

The spikes of attendances are during Thursday and Friday for all the nationalities except for the Swiss, which had its peak on Sunday.

Conclusions

Despite of the dimension of the MSC service area, the data were able to catch the foreign trends as a phenomenon visible at a urban scale. For major events, active clients can provide an important tool for monitoring visitors and tourists, for evaluating the attractiveness of a big event, for evaluating its economic impact and for providing new touristic and business services directed to specific nationalities. On these topics there is a strong need of updated knowledge on urban attractiveness, a major component of global cities competition which is hardly intercepted by standard surveys. Innovative methods are therefore required for measuring and monitoring tourists and visitors at different urban scales. MSC active clients data seems to be promising, providing informations that currently institutional sources do not offer.

Acknowledgments

The authors would like to acknowledge Piero Lovisolo, Dario Parata and Massimo Colonna, Tilab - Telecom Italia for their collaboration during the research project. This work was supported by Telecom Italia.

References

[1] N. Caceres, J. Wideberg, and F. Benitez. Deriving origin destination data from a mobile phone network. *Intelligent Transport Systems, IET*, 1(1):15–26, March 2007.

[2] M. Goodchild. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69:211–221, 2007. 10.1007/s10708-007-9111-y.

[3] J. Katz. *Connections: Social and cultural studies of the telephone in American life*. Transaction Pub, 2003.

[4] M. Kwan. Mobile communications, social networks, and urban travel: Hypertext as a new metaphor for conceptualizing spatial interaction*. *The Professional Geographer*, 59(4):434–446, 2007.

[5] R. Ling. *The mobile connection: The cell phone's impact on society*. Morgan Kaufmann Pub, 2004.

[6] G. Lyons and J. Urry. Travel time use in the information age. *Transportation Research Part A: Policy and Practice*, 39(2-3):257–276, 2005.

[7] H. Rheingold. *Smart mobs: the next social revolution*. Basic Books, Cambridge, MA, 2002.

[8] A. M. Townsend. Life in the Real-Time City: Mobile Telephones and Urban Metabolism. *Journal of Urban Technology*, 7(2):85–104, 2000.

[9] M. Weiser. The computer for the 21st century. *SIGMOBILE Mob. Comput. Commun. Rev.*, 3:3–11, July 1999.

Mobility and Predictability of Population Movements after the Haiti 2010 Earthquake

Xin Lu[1 2*], Linus Bengtsson[1 †], Petter Holme[3 2 ‡]

[1] Department of Public Health Sciences, Karolinska Institutet, 17177 Stockholm, Sweden
[2] Department of Sociology, Stockholm University, 10961 Stockholm, Sweden
[3] IceLab, Department of Physics, Umeå University, 90187 Umeå, Sweden*†‡

Emails: *lu.xin@sociology.su.se; †linus.bengtsson@ki.se; ‡petter.holme@physics.umu.se.

Abstract:

In 2010, 385 natural disasters caused up to 300,000 deaths worldwide, affected over 217 million and resulted in US$ 120 billions worth of damage. Large population movements are a common consequence of large-scale disasters and itself a cause of secondary problems. To understand and predict the movement of people in disasters is one of the keys to effective humanitarian relief as well as long-term societal recovery.

We analyzed the movements of 2.9 million people during the period 42 days before to 342 days after the tragic Haiti earthquake of January 12, 2010. While the earthquake caused 25% of the mobile phone users in the earthquake-affected capital of Port-au-Prince to leave the capital 19 days after the earthquake the predictability of movements over the longer term remained high and the entropy of movements returned to the pre-quake patterns already three months after the earthquake. On the other hand, the time to return to Port-au-Prince after having fled the disaster has a heavy-tailed distribution, illustrating how heterogeneously affected individuals are. The destinations of the large number of phones that left the capital within the first 19 days of the earthquake were highly correlated with their pre-earthquake movements. Specifically, 52% of the displaced Port-au-Princians also left the capital during the preceding Christmas holiday and out of these 85% returned to the same province as during Christmas.

Studying entropy-based predictability measures, we find that the travel pattern of the affected, are highly predictable. This means that historical trajectories can be used to predict the movement in a disaster, to aid decision makers. These analyses also show that, as a fast, convenient method, the tracking of mobile phones can be used as an emergency response system to provide useful information during nature disasters.

Anonymizing Location Data Does Not Work

Hui Zang
Sprint
Burlingame, CA 94010, USA
hui.zang@sprint.com

Jean Bolot*
Technicolor
Palo Alto, CA 94301, USA
jean.bolot@technicolor.com

1. INTRODUCTION

We examine a very large-scale data set of more than 30 billion call records made by 25 million cell phone users across all 50 states of the US and attempt to determine to what extent anonymized location data can reveal private user information. Our approach is to infer, from the call records, the "top N" locations for each user and correlate this information with publicly-available side information such as census data. For example, the measured "top 2" locations likely correspond to home and work locations, the "top 3" to home, work, and shopping/school/commute path locations. We consider the cases where those "top N" locations are measured with different levels of granularity, ranging from a cell sector to whole cell, zip code, city, county and state. We then compute the anonymity set, namely the number of users uniquely identified by a given set of "top N" locations at different granularity levels.

We find that the "top 1" location does not typically yield small anonymity sets. However, the top 2 and top 3 locations do, certainly at the sector or cell-level granularity. We consider a variety of factors that might impact the size of the anonymity set, for example the distance between the "top N" locations or the geographic environment (rural vs urban). We also examine to what extent specific side information, in particular the size of the user's social network, decrease the anonymity set and therefore increase risks to privacy.

Our study shows that sharing anonymized location data will likely lead to privacy risks and that, at a minimum, the data needs to be coarse in either the time domain (meaning the data is collected over short periods of time, in which case inferring the top N locations reliably is difficult) or the space domain (meaning the data granularity is strictly higher than the cell level). In both cases, the utility of the anonymized location data will be decreased, potentially by a significant amount.

A more complete report of this study is published as [1].

2. LOCATION PRIVACY

Call Data Records (CDRs) contain information of every call carried by the cellular network, including time, location, and identities of both parties involved in the call. There is an increasing need or desire to share CDRs, or location traces in general, to third parties and the standard approach is to anonymize those traces, which raises important issues of privacy which can be summed up in the question: **Is it safe to share these anonymized CDRs or location traces?**

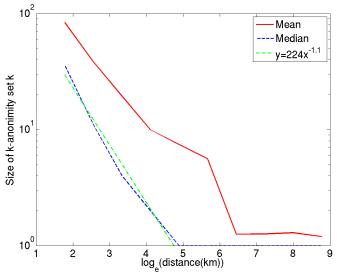


Figure 1: Median and mean size of anonymity set vs. distance between top locations

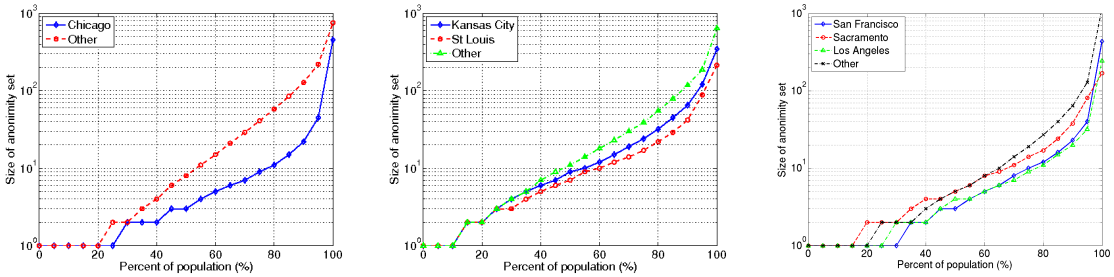
A privacy breach occurs when users are re-identified from anonymized data. Earlier work has shown that a fraction of the US working population can be uniquely identified by their home and work locations, even when those locations are not known accurately. Given that the top locations visited by a mobile user often correspond to the home and work locations, the risk in releasing locations traces of mobile phone users appears very high.

We use a well-known metric, k-anonymity to quantifies the degree of privacy of an anonymized data set. With k-anonymity, each individual will be in distinguishable from at least $k-1$ others, i.e., is "hiding in crowd of k". When applied to location traces, k-anonymity means that the mobility behavior of a user is similar to that of at least $k-1$ other users. Equivalently, k is the size of what we refer to as the anonymity set.

We study a data set of 30 billion CDRs from a nationwide cellular service provider in the United States which contains location information of about 25 million mobile phone users over a three-month period. We examine important factors that affects the privacy of users when their anonymized "top N" locations are shared, in particular:

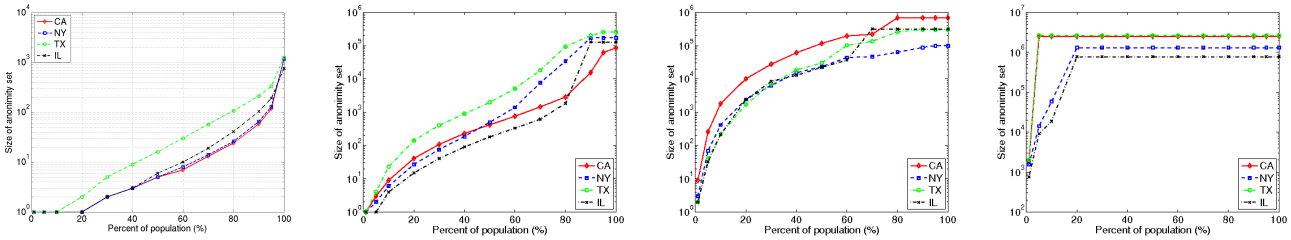
- The value of N: Tables 1 through 3 summarize the 1^{st}

*Part of this work was done while the author was working at Sprint.



(a) Illinois (b) Missouri (c) California

Figure 2: k-anonymity for people whose top location is in the specific states, in the specific cities and the rest of the state.



(a) Cell (b) City (c) County (d) State

Figure 3: Size anonymity set on different granularity levels for people whose top location is in the specific state: California (CA), New York (NY), Texas (TX), and Illinois (IL).

percentile, the 5^{th} percentile, the 10^{th} percentile and the median of the k-anonymity values at six granularity levels: sector, cell, zip-code, city, county, and state, for $N = 1, 2, 3$, respectively. Entries in the tables with a "1" indicate the fraction of users that are uniquely identified. We can see that the larger N is, the smaller the anonymity sets are. In other words, one is more identifiable when more top locations are known.

- The location granularity: As shown by Tables 1 through 3, the number of individuals that can be re-identified decreases as the top N locations are known with increasingly coarse granularity, i.e. vary from the sector to the state level. For example, according to Table 2, at the "sector" and "cell" levels, between 10% and 50% of the users are uniquely identifiable; at the "zip code" level, 5% and at the "city" level, 1%. In other words, if we know the top two locations for each user at the granularity of a cell or of a city, and given that the total population under study is 20 million, we can uniquely identify 2,000,000 users (cell level) and 200,000 users (city level), respectively.
- The distance between locations: As shown in Fig. 1, the longer distance between the top two locations, the smaller the anonymity set. The figure also shows that the size of the anonymity set decreases approximately inversely to the distance between the two locations.
- The geographical regions: We find that different geographical areas have different levels of privacy risks, and at a different granularity level this risk may be higher or lower than in other areas. We investigate users whose top location is in selected states or in/out-of specific cities. The results are shown in Figs. 2 and 3. In Figure 2(a) for example, users in Chicago have much smaller anonymity sets than users in the rest of Illinois. This observation suggests a difference in k-anonymity linked to urban/rural lifestyles: in urban environments, we expect users' lifestyles to be more diversified and hence the anonymity sets to be smaller. Fig. 2(b) shows that users in St. Louis are quite different from the those in the other areas of Missorri, whereas users in Kansas City fall somewhere in the middle. In California (Fig. 2(c)), the anonymity sets of users in San Francisco and Los Angeles are similar. The rest of California is less at re-identification risk than these two cities. Figure 3 shows the distribution of k-anonymity for users in the four states at the cell, city, county, and state levels, respectively. One state may have higher anonymity (larger k) at one level, but lower anonymity (smaller k) at another level. For example, users in California have by far the highest anonymity at the county level (Fig. 3(c)), but have the lowest anonymity at the cell level (Fig. 3(a)). Users

Table 1: k-anonymity with top 1 location

Location granularity	Size of anonymity set (k)			
	1^{st} %ile	5^{th} %ile	10^{th} %ile	Median
Sector	28	71	111	372
Cell	92	220	331	967
Zip code	184	557	909	3125
City	162	487	874	7638
County	802	2972	6272	55649
State	60139	1.5e+05	2.6e+05	7.2e+05

Table 2: k-anonymity with top 2 locations

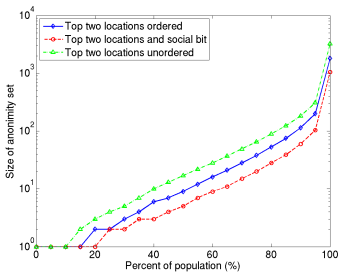
Location granularity	Size of anonymity set (k)			
	1^{st} %ile	5^{th} %ile	10^{th} %ile	Median
Sector	1	1	1	2
Cell	1	1	1	9
Zip code	1	1	2	75
City	1	2	6	437
County	2	23	143	15628
State	530	6912	51291	6.8e+05

Table 3: k-anonymity with top 3 locations

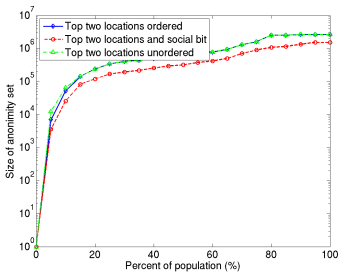
Location granularity	Size of anonymity set (k)			
	1^{st} %tile	5^{th} %tile	10^{th} %tile	Median
Sector	1	1	1	1
Cell	1	1	1	1
Zip code	1	1	1	2
City	1	1	1	24
County	1	2	7	3407
State	40	1074	5671	4.6e+05

in Texas have higher anonymity than the other states at the cell level, but lower anonymity than the others at the county level.

- Additional knowledge about social behavior: We consider the case when more information is released about users than just location information, in particular information about the size of their social network (for example measured by the number of unique individuals they call during a month) and speficially whether the network is large or small. We use 20 as the threshold size between large and small (the reason for 20 is explained in the reference below) and we introduce an extra bit of information to the top N locations indicating the size of an individual's social network relative to this threshold (i.e. whether it is above or below 20). We observe that with this extra bit of information, the size of anonymity sets drops about 50% at every granularity level, which indicates that social behavior is usually orthogonal to mobility behavior and that additional knowledge about the users' social patterns is helpful to re-identify those users. We demonstrate this impact in the case of top 2 locations, at cell and state granularity, in Fig. 4 with the red curves and blue curves denoting the size of anonymity sets with and without this extra bit, respectively. Please refer to [1] for impact at other location granularity and for meaning of the green curves.



(a) Cell



(b) State

Figure 4: k-anonymity with additional social information and when quasi-identifier is an unordered pair of locations.

3. CONCLUSION

In this paper we conducted a large scale study on the risk of re-identification attacks with published location data obtained through call records. Our study shows that publishing or sharing anonymized location data will likely lead to privacy risks and that, at a minimum, the data needs to be coarse in either the time domain (meaning the data is collected over short periods of time of the order of a day, in which case inferring the top N locations reliably is difficult) or the space domain (meaning the data granularity is strictly higher than the cell level). In both cases, the utility of the anonymized location data will likely be decreased by a significant amount.

4. REFERENCES

[1] H. Zang and J. Bolot, "Anonymization of location data does not work: A large-scale measurement study", *Proc. ACM Mobicom 2011*, Las Vegas, NV, Sept. 2011 (to appear).

Using Randomization Methods to Identify Social Influence in Mobile Networks

Rodrigo Belo*, Pedro Ferreira†

Carnegie Mellon University

July 15, 2011

1 Introduction

Identification of social influence in observational data is not trivial; endogeneity issues such as homophily, correlated unobservables, or simultaneity, that frequently pose challenges to the researcher interested in estimating the magnitude of a causal relation in the diffusion process (Shalizi and Thomas, 2010). Commonly used identification strategies include the definition of structural models (e.g., Ma et al., 2009), the use of instrumental variables (e.g., Tucker, 2008), propensity score matching (e.g., Aral et al., 2009), and more recently, randomization (e.g., Anagnostopoulos et al., 2008; La Fond and Neville, 2010).

Randomization techniques consist of generating pseudo-samples based on the original sample by selectively permuting the values of some variables among observations (Noreen, 1989), allowing for the estimation of empirical distributions of a parameter of interest under the null hypothesis (e.g., Anagnostopoulos et al., 2008; La Fond and Neville, 2010). We apply these methods to assess the magnitude of peer influence in the adoption of products and services in a mobile network setting.

We use a comprehensive panel of data from a large European mobile network provider. The data are comprised of detailed information about all the subscribers, including call and SMS detail records, pricing plans, products and promotion adoptions, and handset information over an 11-month period. We estimate the influence effect in a carefully selected set of products from a total of 1200 supplementary services offered by the network provider.

We provide preliminary evidence for the existence of positive and negative social influence in the adoption process, depending on the products analyzed, and we aim to explore how social influence is affected by product characteristics.

2 Randomization methods to identify peer influence in social networks

Randomization tests are a technique that have been used for non-parametric hypothesis testing based on permutations of values among observations (Noreen, 1989). The key idea is that under the null hypothesis these permutations correspond only to random disturbances in the data and should not alter the statistics we're interested in estimating. The test is conducted as follows. The original data is altered several times by permuting some attributes among individuals, each permutation originating a pseudo-sample. A test score is calculated for each pseudo-sample, and from these test scores an empirical distribution is estimated. The test score of the original data is compared to the estimated distribution, and its significance is assessed. Two recent studies outline randomization strategies to identify peer influence effects when information about the timing of actions is available.

Anagnostopoulos et al. (2008) propose a test to identify influence as a source of correlation in a social network, the shuffle test. This tests consists of shuffling the adoption date among the edges that eventually

*Rodrigo Belo, CMU, rbelo@cmu.edu.
†Pedro Ferreira, CMU, pedrof@cmu.edu.

will become adopters. The test is based on the idea that under the *no-influence* null hypothesis, adoption dates should be *independent*, and therefore the obtained correlation coefficient should be the same whether or not adoption dates have been shuffled. In the case the original correlation coefficient is *different* from the shuffled one, the null hypothesis can be rejected and we can conclude we are in the presence of influence.

La Fond and Neville (2010) describe a general randomization method to identify homophily and influence in a two-period setting. They define a correlation measure that increases with the the number of connected edges with similar attribute values and with the number of unconnected edges with different attribute values. They calculate empirical distributions for this auto-correlation statistic under the null hypothesis that there is no influence (or homophily). Finally, they compare the statistic obtained from the actual change with the empirical distribution and determine whether the null hypothesis can be rejected or not.

Notably, their randomization strategy is not suitable in the case of adoption of a single product in which the only relevant attribute change is binary. The shuffle test proposed by Anagnostopoulos et al. (2008) is suitable for the adoption case, but requires a longitudinal view of the data, i.e., when performing the analysis the researcher needs to know exactly which individuals end up adopting the product, so that the shuffling can be performed among all eventual adopters.

In sum, the application of randomization methods to identify influence in social networks is still in an emergent stage and needs further research. We set out to explore these methods and develop a deeper understanding on their potential and limitations.

3 Data

We have secured access to a 11-month anonymized panel of data comprising of detailed information about all subscribers in a large mobile European network provider. The data includes detailed information on demographics (age, gender and zip code), call and SMS data records, pricing plans, product and promotion adoptions, as well as on subscribers' handsets.

The data are comprised of detailed information about every call and SMS originated and received by roughly 4 million subscribers during the period of analysis. These details include, among others, origin, destination, start time and duration of a call. On an average day subscribers generate about 4 million calls and exchange 40 million SMSs. Additionally, the data contain information about subscribers' pricing plans and supplementary services. At a given moment in time each subscriber is associated with one pricing plan, and possibly several supplementary services. Supplementary services are *à la carte* add-on services that subscribers can acquire, and that can be virtually anything the network operator decides, such as a pack of 1000 SMSs at a discounted rate, free calls on the weekends for a given period of time, or simply voice-mail activation. The operator has more than 1,200 distinct services, and there are a total of 10 million supplementary services subscriptions during the period of analysis, which corresponds to an average of 2.5 services per subscriber.

We estimate influence effects on a carefully selected set of services. Follows a brief description of three of them that have been analyzed so far: (1) a one time promotion in which subscribers get their phone account credited by the amount corresponding to the minutes they have spent during the month of December. To benefit from this promotion subscribers must pay a fixed entry fee; (2) an ongoing promotion in which subscribers pay a fixed fee to be able to perform video calls at discounted rates; (3) a one time promotion in which subscribers can call free at night during the month of July. To benefit from this promotion subscribers must pay a fixed entry fee.

We summarize and extract information about the network on a monthly basis. We calculate monthly summaries for each subscriber, including number of call neighbors, SMS neighbors, total number of calls made and received, and total conversation time. Also, in a first stage, we limit our analysis to a sample of 10,000 randomly selected subscribers and their direct neighbors.

4 Methods

We follow Anagnostopoulos et al. (2008) and define our problem as follows. Consider relational data represented as an undirected time-changing graph, $G = (V, E)$, where V is a set of nodes and E a set of undirected edges, e_{ij}, each edge connecting two nodes. Each node $v \in V$ has a time-changing attribute $\mathbf{W}_t^v$ indicating whether the node has adopted the product at time t or before.

Additionally, assume that at each time period, t, the probability of a node to adopt a given product follows a distribution that depends on the number of connected nodes v' that have already adopted the product, $a_{it} = \sum_{v_j:e_{ij}\in E} \mathbf{W}_t^{v_j}$, and on her own characteristics, $\mathbf{X}_i$: $p(a_{it}, \mathbf{X}_i)$.

Due to computational restrictions, and because our interest lies mainly in the first-order marginal effects of having one more adjacent adopter node, we estimate a linear probability model (LPM):

$$p(a_{it}, \mathbf{X}_i) = \alpha a_{it} + \mathbf{X_i}\beta + \varepsilon_{it}$$

where α corresponds to our parameter of interest and β is a parameter vector.

As mentioned above, α might be capturing not only the influence effect but also other sources of correlation, such as homophily and unobserved confounding variables. Thus, we estimate an empirical distribution for α under the null hypothesis of no influence, and then compare the parameter obtained from the original data with the empirical distribution.

The no influence hypothesis can be stated as follows:

H_0 : The probability of node i adopting a given product at time t is not determined by the number of adjacent nodes that have already adopted, a_{it}.

We generate randomized versions of the data as close as possible to the original, only without assuming that the time of adoption is important for the diffusion process: we shuffle the time of adoption among eventual adopters as suggested by Anagnostopoulos et al. (2008). This transformation has the advantage of preserving the network-level statistics, and thus avoiding potential problems of changing too much the original data. To calculate the empirical distribution we run the LPM model and obtain an estimate of α for each randomized version of the data. We can reject the null if α in the original data does not fall in the 95% confidence interval of the empirical distribution. Moreover, we identify social influence as the difference between the average value of the empirical distribution and the coefficient obtained with the original data.

5 Results, Contribution and Challenges

We identify social influence on three products with distinct characteristics. Table 1 shows the total number of adopters for each of these products (column (a.)), the coefficient obtained from the original data (column (b.)), the average coefficient obtained from randomization and respective standard deviation (column (c.)), and our estimate of social influence (column (d.)). For product (1) the average coefficient obtained from the randomization is .0054. Given that the coefficient obtained using the original data is .0064, outside the 95% confidence interval of the empirical distribution (see Figure 1), we conclude that social influence plays a role in the diffusion of this product. Its total effect corresponds to 5% of the total observed adoption.[1] We find no evidence of social influence in product (2), and for product (3) we find a negative effect. This result might be due to a substitution effect: there is no incentive for an adopter's contacts to subscribe the product since they will benefit from talking with their interlocutor for free anyways, as long as the adopter is the one who initiates the calls.

The different characteristics of the products analyzed might provide some insight on why we see differences in magnitude and sign of social influence on mobile network products. We need to further explore this topic by including more products in our analysis and by formulating hypotheses about specific product characteristics that affect social influence.

We characterize peer influence in the context of mobile network services, assess its impact in the diffusion process and evaluate its importance for marketing-related decisions in this context. We provide preliminary evidence for the existence of different social influence effects given different product characteristics. We also shed some light on the advantages and limitations of randomization methods in the identification of influence and, by applying these methods in a different context, we try to contribute to their consolidation as a valuable alternative to other identification methods.

[1] Given that in our sample each person is exposed to .09 adopters per day on average, the total influence effect corresponds to 28 extra adopters (from a total of 534 adopters) over the 32 days adoption occurs, which corresponds to 5%.

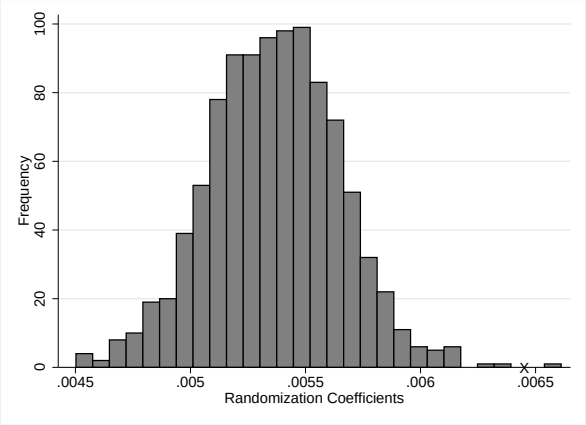


Figure 1: Product (1) coefficients over 1000 adoption date shuffles. The '×' mark represents the coefficient obtained from the original data.

Product	(a.) Adopters	(b.) Original Data Coeff. (α) $\times 10^3$	(c.) Randomization Avg. Coeff. $\times 10^3$ (s.d.)	(d.) Influence Est. $\times 10^3$
(1)	534	6.4	5.4 (.29)	1.0***
(2)	298	.11	.10 (.006)	.01
(3)	110	.30	.60 (.075)	-.3***

*** p<0.01, ** p<0.05, * p<0.1

Table 1: Influence estimates for selected products.

References

A. Anagnostopoulos, R. Kumar, and M. Mahdian. Influence and correlation in social networks. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 7–15, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-193-4. doi: http://doi.acm.org/10.1145/1401890.1401897.

S. Aral, L. Muchnik, and A. Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, 106(51):21544, 2009.

T. La Fond and J. Neville. Randomization tests for distinguishing social influence and homophily effects. In *Proceedings of the 19th international conference on World wide web*, pages 601–610. ACM, 2010.

L. Ma, R. Krishnan, and A. Montgomery. Homophily or Influence? An Empirical Analysis of Purchase within a Social Network, 2009.

E. Noreen. Computer Intensive Methods for Testing Hypothesis- An Introduction. *JOHN WILEY & SONS*, (229), 1989.

C. Shalizi and A. Thomas. Homophily and contagion are generically confounded in observational social network studies. *Arxiv preprint arXiv:1004.4704*, 2010.

C. Tucker. Identifying formal and informal influence in technology adoption with network externalities. *Management Science*, 54(12):2024, 2008.

Quantifying and Modelling Good Communicators in Dynamic Phone Networks

Alexander Mantzaris
alexander.mantzaris@strath.ac.uk

Desmond J. Higham
d.j.higham@strath.ac.uk *

Abstract

Dynamic networks are those with a connectivity structure that changes over time. In this setting, the ability of a node to transmit or receive information cannot be accurately measured, in general, by considering only simple snapshots or time-averages. In this presentation, in the context of cell phone communication, we describe some recent ideas in (a) measuring node centrality and (b) defining explanatory mathematical models, for dynamic networks. Our focus is on *dynamic communicators*: individuals who, either through status, intelligence or foresight, are able to communicate extremely effectively, in a manner that could not be discovered from static centrality measures.

1 Dynamic Centrality

Modelling and analysis of static networks has a long and distinguished history, and there are many concepts and tools available [6]. However, information about communication networks generally evolves dynamically, and although a static analysis can proceed by aggregating over time, we argue here that value can be added by respecting time's arrow. We will show that a fully dynamic viewpoint can uncover useful information that is lost with static summaries.

Edges between nodes (participants in the network) can be created by making phone calls, sending emails, or from any type of exchange of information that has finite duration. Consider a fixed set of N nodes in a network producing data of the form $\{A^{[k]} \in \mathbb{R}^{N\times N}\}$ for $k = 0, 1, 2, \ldots, M$, representing the adjacency matrix at each time t_k. Hence, $\left(A^{[k]}\right)_{ij} = 1$ if there is an edge from node i to node j at time t_k and $\left(A^{[k]}\right)_{ij} = 0$ otherwise. The time points $t_0 < t_1 < \cdots < t_M$ are ordered but not necessarily equally spaced.

To develop a dynamic measure for the centrality of the nodes, the downstream contribution (knock-on effect) of an edge can be calculated via the concept of *dynamic walks*, where edges are traversed in a manner that respects the time ordering. In [3] it was shown how to compute a matrix $\mathcal{Q} \in \mathbb{R}^{N\times N}$ for which $(\mathcal{Q})_{ij}$ is a weighted count of the number of dynamic walks of length w from node i to node j. Here, length is measured as the number of edges crossed. To model the loss of influence as walk length in-

*Department of Mathematics and Statistics, University of Strathclyde, Glasgow, UK. This work was supported by the Engineering and Physical Sciences Research Council and the Research Councils UK Digital Economy Programme, through the project MOLTEN: Mathematics Of Large Technological Evolving Networks.

creases, walks are weighted by a factor α^w where $0 < \alpha < 1$ is a fixed parameter. One intuition for down-weighting the contribution of longer walks is that information may become less relevant or reliable with length. Alternatively, α may be interpreted as the independent probability that a message successfully traverses an edge. The matrix $\mathcal{Q}$ may be written in the form

$$\mathcal{Q} = \left(I - \alpha A^{[0]}\right)^{-1} \dots \left(I - \alpha A^{[M]}\right)^{-1}.$$

In practice, if our aim is to rank nodes, it is equivalent and computationally safer to iterate with the normalized form

$$\hat{\mathcal{Q}}^{[k]} = \frac{\hat{\mathcal{Q}}^{[k-1]}\left(I - \alpha A^{[k]}\right)^{-1}}{\left|\left|\hat{\mathcal{Q}}^{[k-1]}\left(I - \alpha A^{[k]}\right)^{-1}\right|\right|}.$$

Overall, $(\mathcal{Q})_{ij}$ summarizes how well node i can communicate with node j using dynamic walks. The corresponding row and column sums

$$C_n^{\text{broadcast}} := \sum_{k=1}^{N} \mathcal{Q}_{nk} \quad \text{and} \quad C_n^{\text{receive}} := \sum_{k=1}^{N} \mathcal{Q}_{kn}$$

introduced in [3] are therefore centrality measures that quantify how effectively node n can *broadcast* and *receive* dynamic messages across the network. These measures generalize the classic Katz centralities [4, 6] in the sense that they become equivalent for a single time point.

We give here some new computations on a voice call dataset from MIT [1], comparing for each node the total number of edges (total degree) against the broadcast centrality measure. The data was sumarized into one-day windows, so $(A^{[k]})_{ij} = 1$ means that person i talked to person j at least once on day t_k. We collected data over a period of 365 days and the results for two 30 consecutive day periods are presented. Circled in red in Figure 1 are nodes which rank in the top 3 in terms of their broadcast centrality but do not rank within even the top 10 nodes in terms of their degree. Nodes of this type are named *dynamic communicators* in [5], since a static measure (in this case, aggregate degree) fails to allocate high centrality to them despite their ability to communicate across the network.

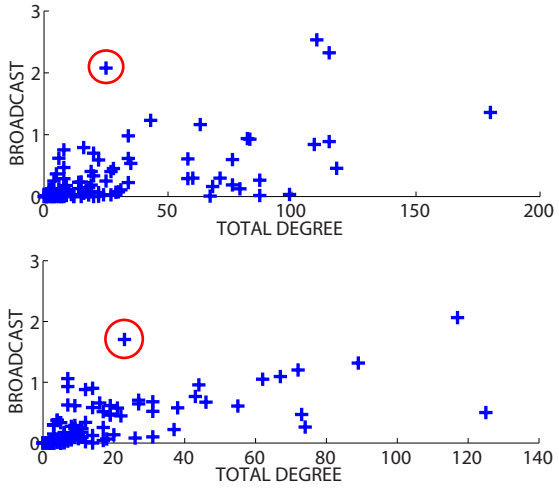


Figure 1: Total degree (horizontal axis) versus broadcast centrality (vertical axis) for voice call data of participants within non-overlapping 30 day periods. A *dynamic communicator* is circled in red.

2 A Model for Dynamic Communicators

To explain a possible mechanism for the emergence of dynamic communicators, a model has recently been developed to simulate such behavior [5]. This model uses the general discrete-time

discrete-space Markov chain framework of [2]. Here, given the current state of the network, a particular model specifies the probability of each edge existing at the next time point. This gives a natural and flexible way to incorporate whatever rules are felt to be appropriate.

To generate dynamic communicators, we will quantify a hierarchy amongst the nodes. Hierarchy, either explicitly imposed or implicitly self-organized, exists, for example, in businesses, educational establishments, military units, social groups and criminal networks. Our new model ([5]) represents the hierarchy of communicators by allocating a fixed level of importance for each node. The model is based on the idea that nodes near the top of the hierarchy have a greater level of 'influence' or 'prescience' in the sense that those who receive their messages tend to pass that message on. A key ingredient of the model is that nodes respond to messages they receive by sending out messages to other nodes. The level of response is based partly upon the ratio of the importance of the nodes who sent the messages and the largest importance value in the network. In this way, we have the possibility of producing 'low-bandwidth' nodes that do not appear to be active when viewed over single snapshots or across an aggregate summary, but have relatively high dynamic broadcast centrality because of their importance in the hierarchy. Such nodes represent the dynamic communicators that we illustrated above on the MIT data.

The new computation in Figure 2 shows the results of an average of 40 independent simulations with the model for a network of 40 nodes sending messages for 365 time points, using $\alpha = 0.3$. The node with the largest importance in the dynamic network is circled in red. This red circled node is a dynamic communicator—it has taken part in a very modest number of interactions (as measured by its total degree) and yet has the greatest ability to communicate through the network (as measured by broadcast centrality).

In summary, we have illustrated a computational tool for discovering good communicators in dynamic network data, and outlined an intuitively reasonable model that reproduces the 'dynamic communicator' effect seen in mobile communication data.

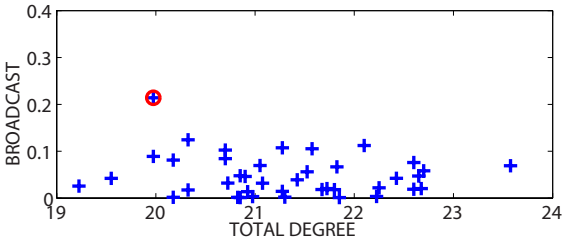


Figure 2: Total degree (horizontal axis) versus broadcast centrality (vertical axis) for the average of 40 independent simulations of 40 nodes over 365 time points.

References

[1] Nathan Eagle, Alex S. Pentland, and David Lazer. Inferring friendship network structure by using mobile phone data. *Proc. Nat. Acad. Sci.*, 106(36):15274–15278, September 2009.

[2] Peter Grindrod and Desmond J. Higham. Evolving graphs: Dynamical models, inverse problems and propagation. *Proc. Roy. Soc. A*, 466:753–770, 2010.

[3] Peter Grindrod, Desmond J. Higham, Mark C. Parsons, and Ernesto Estrada. Communicability across evolving networks. *Physical Review E*, 83:046120, 2011.

[4] L. Katz. A new index derived from sociometric data analysis. *Psychometrika*, 18:39–43, 1953.

[5] A. Mantzaris and D. J. Higham. A model for dynamic communicators. *submitted*, 2011.

[6] M. E. J. Newman. *Networks an Introduction*. Oxford Univerity Press, Oxford, 2010.

Examining the Social Decomposition of Mobile Call Graphs

Derek Doran*
Dept. of Computer Science & Engineering
University of Connecticut
Storrs, CT, 06269
Email: derek.doran@engr.uconn.edu

Veena Mendiratta, Chitra Phadke, and Huseyin Uzunalioglu
Bell Laboratories
Alcatel-Lucent
Murray Hill, NJ, 07974
Email: {veena.mendiratta, chitra.phadke, huseyin.uzunalioglu}@alcatel-lucent.com

***Abstract*—In this paper, we examine the social decomposition of a mobile call graph using a new approach to measure the social tie strength between users. This approach improves upon the existing measures by considering any number of observations about the calls made and the local social structure. The results suggest that the dyadic hypothesis is valid in the context of mobile call graphs, and that strong ties play a critical role in the structure of the graph.**

I. Introduction

Social Network Analysis (SNA) is a powerful approach used to better understand the behaviors and relationships of users. SNA is traditionally applied in the context of online social networks (OSNs) such as Facebook, Flickr, and Twitter, where users can directly establish ties, share information, and join groups to connect to users with similar interests. In these networks, SNA operates over attributes that directly imply a social connection between users. For example, the fact that two users are friends on an OSN, that they belong to the same groups, or that they share information with each other can each be used individually to infer that a social tie exists.

Such OSNs contain *causal* information, that is, data attributes which imply the existence of a social tie. There exist other social networks, however, where only the *effects* of a social tie are observable. Each effect, taken alone, does not directly suggest tie strength. Mobile call graphs are an example of such a social network. In a mobile call graph, the effects of a strong social tie may include a large number of calls placed, a long time spent talking, and many calls during weekend and evening hours. By themselves, however, none of the attributes directly imply the tie strength. For example, a user may call a bank to check balances and pay bills more times than they call a friend, even though friendship is a stronger social tie.

Mobile call graphs represent the way in which a large number of users communicate with each other, and these patterns of communication are related to the social ties between people. Thus, studies that apply SNA to mobile call graphs are rising in popularity [1], [2], [3], [4], [5]. Such studies, however, only pick a single feature about the calls made between two users to define a social relationship. As a result, the conclusions drawn by these studies are based only on a single *effect* of a possible social relationship that exists. In order to make observations about a call graph that more faithfully considers the social relationship between users, an improved measure of tie strength is needed.

In this paper, we propose an algorithm to quantify social tie strength through the synthesis of many calling attributes whose values are the effects of a social tie. We demonstrate that the resulting tie strengths support the dyadic hypothesis, which states that tie strengths depend primarily on the nature of the relationship between the two users and is independent on local social network structure [6]. Local social structure is shown to still play a role, however, when the amount of interaction between two users is small. We then apply the algorithm to a call graph provided by a major mobile service provider to study the relationship between tie strength and call graph structure. We decompose the mobile call graph by eliminating different levels of the weakest or strongest social ties. Through the new definition of social tie strength, we make observations that further supports the dyadic hypothesis and emphasizes the importance of strong social ties. Specifically, we find that:

- Users connected by strong social ties have a positively correlated out-to-in degree distribution, while the out-to-in degree distribution for nodes with weak ties are uncorrelated.
- There exist critical tie strength values where most of the connected components of the graph disintegrate.
- The massive connected component, which over 84% of all vertices are a member of, quickly fragments when strong social ties are eliminated.
- The structure of the massive connected component is highly robust to the failure of edges with weak social ties.

The layout of this paper is as follows. In Section II, we introduce our algorithm to derive social tie strength. Section III applies the algorithm to a mobile call graph and examines its social decomposition. Section IV concludes the paper and outlines our future work.

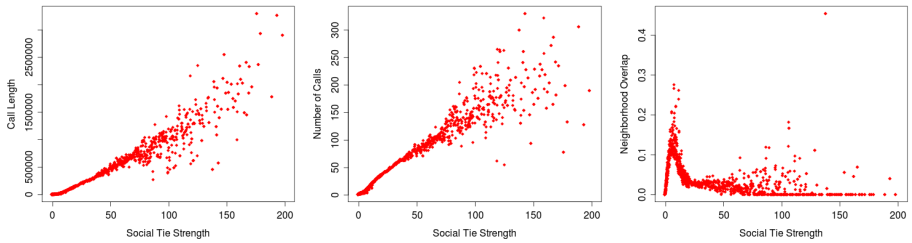
II. Tie Strength Algorithm

We define a mobile call graph as a simple directed graph $G = (V, E)$ where the set of vertices V represent mobile phone users, and an edge $e = (a, b) \in E$ iff $a, b \in V$ and a placed a call to b. G represents the $|E| = m$ calls that are placed between the $|V| = n$ users. The objective is to define a weighting function $\mathcal{S} : E \to \mathbb{R}$ that maps every directed edge to a value quantifying the strength of the social tie between the users incident on the edge.

In our proposed algorithm, we represent each edge as a vector of k attributes and compose them into the $m \times k$ matrix $\mathbf{E}$, where a row of $\mathbf{E}$ corresponds to an edge of G and $[\mathbf{E}]_{ij}$ is the value of attribute j for edge i. In order to map the row vectors of $\mathbf{E}$ to a value representing social strength, we first apply an approach, inspired by principle component analysis (PCA) [7], that projects the data onto a subspace which better represents the variation existing within the data. This projection is motivated by recognizing that the true social tie strength affects the value of the attributes that exhibit the greatest variation the most. The projection uses an orthogonal basis set of vectors that point in the directions where the variation in the data is the largest (referred to as principle components (PC)). This set is given by the eigenvectors of the covariance matrix Σ of $\mathbf{E}$ [7].

The sum of the eigenvalues of Σ is equal to the total variance within the data, which is the same as the dimensionality of the data if it has zero mean and unit variance. This means that the eigenvalues of Σ relate the amount of variation that is explained by each dimension of the projected data to the variation along the dimensions of the

*Derek Doran performed this work as a summer intern at Bell Labs.



(a) Call length (b) Number of calls (c) Neighborhood overlap

Fig. 1: Correlation between social tie strength and selected attributes

original data. We thus take each component of the projected data, multiply it by the corresponding eigenvalue, and then sum these weighted components to get a value for social tie strength. This gives the dimensions where the data exhibits very large variation additional influence in the tie strength value. The entire algorithm is summarized as follows.

1) Set $[\mathbf{E}]_{ij} = [\mathbf{E}]_{ij} - \frac{1}{m}\sum_{i=1}^{m}[\mathbf{E}]_{ij}$ for all j.
2) Set $[\mathbf{E}]_{ij} = [\mathbf{E}]_{ij}/\sigma_j$, with σ_j^2 as the variance of attribute j.
3) Find the covariance matrix Σ of $\mathbf{E}$.
4) Find Λ, a $k \times 1$ column vector where Λ_i is the i^{th} largest eigenvalue of Σ.
5) Find $\mathbf{U}$, a $k \times k$ matrix whose i^{th} column is the right eigenvector corresponding to Λ_i.
6) The social tie strength for edge e_i is given by i^{th} component of the vector $\mathcal{S} = \mathbf{E}\mathbf{U}\Lambda$.

III. Social Decomposition

In this section, we apply the social tie strength algorithm to a large mobile call graph and analyze the way its structure decomposes as edges with weak social ties are removed. The mobile call graph contains over 3M vertices and over 4.5M edges. The graph was extracted from a dataset of approximately 10M calls.

A. Social tie strength

To apply the social tie strength algorithm we chose three attributes, namely, the total length of all calls between two users, the total number of calls placed between two users, and the *neighborhood overlap* of two users (defined as the Jaccard similarity between the neighbor sets of each user). Figure 1 shows the relationship between the derived social tie strength value and the attributes used. Figure 1a and 1b demonstrate that the social tie strength value increases with the number of calls placed and the total amount of time users spend talking to each other, which are attributes that reflect the direct relationship between the users. We also see that the neighborhood overlap plays a role for the large number of edges with a small social tie strength. Since these edges place very few calls and have a small total call length, their social tie strengths are differentiated based on their neighborhood overlap. For edges that have a large social tie strength, the neighborhood overlap losses it's influence. This is because the neighborhood overlap contains a smaller amount of variation compared to the call length and total number of calls attributes.

The positive correlation between tie strength and the attributes reflecting the direct relationship between two users support applying the dyadic hypothesis to the social ties of users in a mobile call graph. Intuitively, the dyadic hypothesis should apply to call graphs because the act of calling someone by phone is a deliberate, directed action that is generally not influenced by the structure of the social network between the caller and callee. It is only when the calling interactions between two users are weak that the local social network structure is given a significant amount of influence over the tie strength value.

Next, we analyze the structural characteristics of the mobile call graph as the network is decomposed by removing edges with weak social ties.

B. Degree correlation

The in- and out-degree of a node in a mobile call graph reflects the number of unique people that call a user and the number of unique users that a person calls. It has been observed that the degree distribution of many types of social networks, including mobile call graphs, follow a power-law distribution [1], [8], [9], [10]. This indicates that while the mean degree of a node may be small, the distribution across the entire network exhibits a very large variance.

Equally important for telecom operators is understanding the relationship between the in-degree and out-degree of a node in a mobile call graph. For example, it may the case that nodes with a large in- but small out-degree represent call centers that receive a large number of calls but do not place any. Nodes with large out- but small in-degrees may represent users that treat use their mobile phone number as a business line, placing many phone calls that they do not expect to get call backs from.

In Figure 2, we plot the mean out- or in-degree for all nodes with a given in- or out-degree as we remove edges from the call graph, keeping only the top percent of edges that have the highest social tie strength. For example, the upper-left most plot considers the call graph with just top 95% of the edges with highest social tie strength, while the bottom-right plot includes only the top 5% of edges. In every plot, we find that the nodes maintain a positive in-to-out degree correlation. Thus, despite the strength of their social relationships, users that receive calls from many people also tend to make calls to many people. When examining the out-to-in degree correlation, however, we see that when the weak social ties are included there is no relationship but as the weak social ties are removed the correlation becomes positive. This suggests that when a user places a call along a strong social tie it is likely to reciprocate a call back, but calls made over weak social ties will not encourage the caller to be contacted by other users. This matches our intuition about the influence of the social ties between people. If two users have a strong social tie, a call made in one direction will likely lead to a call in the other direction in the future.

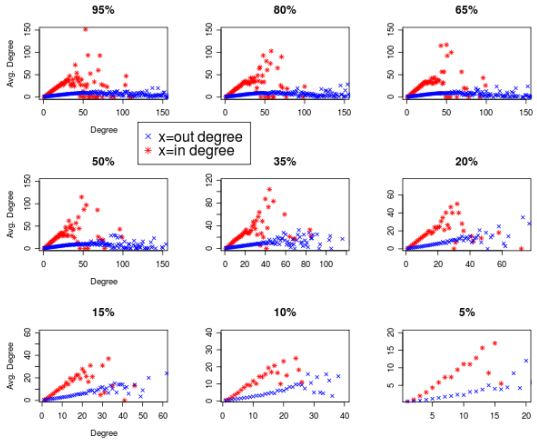
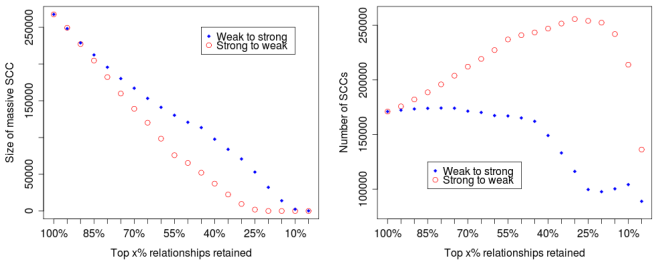


Fig. 2: In- and out-degree distributions as weak ties are removed



(a) Massive SCC size (b) Number of SCCs in graph

Fig. 3: Size and fragmentation of the massive connected component

C. Connected components

Next, we examine the relationship between strong social ties and the connected components of a call graph. The distribution of the sizes of connected components are indicative of the reachability between users, and is important when analyzing the diffusion of information across a mobile call graph. It has been demonstrated that mobile call graphs contain a single massive connected component [1], [2], which means that information introduced into the network has the potential to reach most users. Whether or not strong or weak social ties are critical to the structure of this massive connected component is an important consideration when studying the spread of information across a mobile call graph.

Figure 3 demonstrates the way in which the massive connected component breaks down as we gradually eliminate the weakest and strongest social ties from the graph. In Figure 3a, we observe that the number of nodes in the massive connected component decrease faster when the edges are removed in the order of strongest to weakest tie strength. In Figure 3b we observe that the total number of connected components in the entire network rise quickly when the strongest ties are removed first. This sudden rise in the number of connected components, combined with the observed decrease in the size of the massive connected component as the strongest ties are removed first, suggest that the loss of the strongest social ties breaks apart the massive component into multiple, smaller connected components.

When the weakest ties are removed first, the number of remaining connected components exhibits a contrasting pattern. The number of connected components remain constant, and it is not until the 45% strongest ties remain that the number decreases. It is only after the strongest 25% of ties are left that the number of connected components begin to increase. The knee in the trend at 25% likely represents the point where the structure of the massive connected component finally shatters into multiple, smaller components. The decrease in the number of connected components starting at 45%, then, represents a critical tie strength where the other, non-massive components begin to break down.

These observations suggest that the connected components of a mobile call graph are structurally robust against the loss of edges with weak social ties. Furthermore, we find that the edges most critical to the structure of the massive connected components are those with strong social ties. This suggests that weak social ties are, generally, *not* the critical connections that maintain the reachability of most nodes in a mobile call graph.

IV. Conclusions and Future Work

This paper presented a preliminary study on the social decomposition of a mobile call graph. The study uses a novel approach to quantify the social tie strength between two users that offers higher fidelity by considering any number of observations. The social ties suggest that the dyadic hypothesis is valid for mobile call graphs. By examining the decomposition of the call graph, we find that strong social ties are critical to maintain the structure of the network. Future work seeks to examine additional structural properties, including shortest paths, clustering coefficients, and neighborhood distributions. Additional attributes will also be incorporated for deriving the strength of social ties.

References

[1] K. Dasgupta, R. Singh, B. Viswanathan, D. Chakraborty, S. Mukherjea, and A. Nanavati, "Social Ties and their Relevance to Churn in Mobile Telecom Networks," in *Proc. of 11th ACM Intl. Conference on Extending Database Technology*, 2008.

[2] 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.

[3] Y. Richter, E. Yom-Tov, and N. Slonim, "Predicting customer churn in mobile networks through analysis of social groups," in *Proc. of SIAM Intl. Conference on Data Mining*, 2010.

[4] M. Seshardi, S. Machiraju, A. Sridharan, J. Bolot, C. Faloutsos, and J. Leskovec, "Mobile Call Graphs: Beyond Power-Law and Lognormal Distributions," in *Proc. of 14th ACM Conference on Knowledge Discovery & Data Mining*, 2008.

[5] A. Nanavati, S. Gurumurthy, G. Das, D. Chakraborty, K. Dasgupta, S. Mukherjea, and A. Joshi, "On the Structural Properties of Massive Telecom Call Graphs: Findings and Implications," in *Proc. of 15th ACM Conference on Information and Knowledge Management*, 2006.

[6] M. Granovetter, "The Strength of Weak Ties," *American Journal of Sociology*, vol. 78, pp. 1360–1380, 1973.

[7] J. E. Jackson, *A User's Guide to Principal Components*. John Wiley & Sons, 2004.

[8] J. Kunegis, A. Lommatzsch, and C. Bauckhage, "The Slashdot Zoo: Mining a Social Network with Negative Edges," in *Proc. of 18th ACM Conference on the World Wide Web*, 2009, pp. 741–750.

[9] A. Mislove, H. S. Koppula, K. Gummadi, P. Druschel, and B. Bhattacharjee, "Growth of the Flickr Social Network," in *Proc. of ACM SIGCOMM Workshop on Online Social Networks*, 2008.

[10] A. Mislove, M. Marcon, K. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and Analysis of Online Social Networks," in *Proc. of the ACM Internet Measurement Conference*, 2007.

Time allocation in social networks: correlation between social structure and human dynamics

Giovanna Miritello[1,2], Rubén Lara[2], and Esteban Moro[1,3,4]

[1]Departamento de Matemáticas & GISC, Universidad Carlos III de Madrid, 28911 Leganés, Spain

[2]Telefónica Research, Madrid, Spain

[3]Instituto de Ciencias Matemáticas CSIC-UAM-UCM-UC3M

[4]Instituto de Ingeniería del Conocimiento, Universidad Autónoma de Madrid, 28049 Madrid, Spain

July 15, 2011

Abstract

Recent research has shown the deep impact of the dynamics of human interactions on the spreading of information, opinion formation, etc. [1, 2, 3, 4, 5]. In general, the bursty nature of human interactions lowers the interaction between people to the extent that both the speed or reach of information diffusion is diminished [2, 3]. Using a large database of 20 million users of mobile phone calls we show evidence that there is a large correlation between this effect and the social topological structure around a given interaction. In particular we show that social relations of hubs in a network are relatively weaker from the dynamical point than those that are poorer connected. This means that, dynamically, hubs have a relatively lower importance on information transmission than poorer connected people. A detailed analysis shows that this happens because hubs tend to allocate time in an efficient way so that they manage a (bounded) number of social interactions within a given time interval. We propose a model of time allocation to explain the observed phenomena and discuss the importance of our results in general problems of information diffusion, coordination, opinion formation etc. in social networks.

References

[1] A. Vazquez *et al.* Impact of non-Poissonian activity patterns on spreading processes. Phys. Rev. Lett. (2007) vol. 98 pp. vol. 98 (15) pp. 158702

[2] G. Miritello, E. Moro, and R. Lara, Dynamical strength of social ties in information spreading. Phys. Rev. E (2011) vol. 83 (4) pp. 045102

[3] M. Karsai *et al.*, Small but slow world: How network topology and burstiness slow down spreading. Phys. Rev. E (2011) vol. 83 (2) pp. 025102

[4] A. Gautreau, A. Barrat, and M. Barthélemy, Microdynamics in stationary complex networks. P. Natl Acad Sci Usa (2009) vol. 106 (22) pp. 8847-52

[5] C. Cattuto *et al.* Dynamics of person-to-person interactions from distributed RFID sensor networks. PloS one (2010) vol. 5 (7) pp. e11596

Correlated bursty behaviour in human communication

Márton Karsai,[1] Kimmo Kaski,[1] A.-L. Barabási,[2,3] and János Kertész[3,1]

[1] *BECS, School of Science, Aalto University, Helsinki, Finland*

[2] *Center for Complex Networks Research, Northeastern University, Boston, MA 02115*

[3] *Institute of Physics, Budapest University of Technology and Economics, Budapest, Hungary*

Temporal inhomogeneities occur frequently in human dynamics as a result of decision making by individuals and of various kinds of correlations due to the social environment. These systems are characterized by intermittent switching between periods of low activity and those of high activity bursts as observed, e.g., in digital records of human communication through different channels (see Fig.1) where competing tasks with different priorities have been proposed as partial explanation of the origin of burstiness and of the scale-invariance of the inter-event time distribution [1, 2]. It has also been claimed that the broad inter-event time distribution can be explained by aggregated random cascading behaviour of individuals where homogeneous Poisson cascades evolve on a short time scale with a time dependent rate induced by a long-term non-homogeneous random process following circadian periodicities [3]. However, the circadian and weekly fluctuations explain only partially the heterogeneous dynamics of human communication since by removing them the broad distribution of inter-event times remains unchanged with a similar shape as the original event sequence [4].

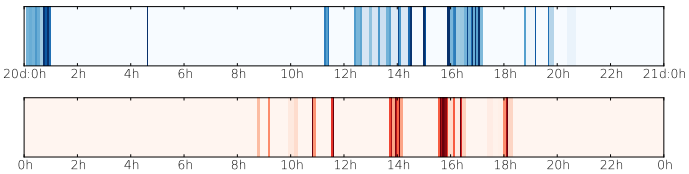


FIG. 1: Mobile-call activity of single individuals with color-coded inter-event times.

For systems with discrete events dynamics it is usual to characterize the observed temporal inhomogeneities by the inter-event time distributions, $P(t_{ie})$, where $t_{ie} = t_{i+1} - t_i$ denotes the time between consecutive events. A broad $P(t_{ie})$ reflects large variability in inter-event times. Yet $P(t_{ie})$ alone tells nothing about the presence of correlations, which is usually characterized by the autocorrelation function, $A(\tau)$, or equivalently by the power spectrum density. However, for temporally heterogeneous signals of independent events with power-law fat-tails in $P(t_{ie})$ the Hurst exponent and the autocorrelation function can indicate false positive correlations. To understand

the mechanisms behind these phenomena, it is important to know whether there are true correlations in these systems. Therefore, for the systems showing fat-tailed inter-event time distributions, there is a need to develop new measures that are sensitive to correlations but insensitive to fat tails.

A sequence of events separated by inter-event times smaller than Δt form a bursty period. Let $P(E)$ be the distribution of the number of bursty events E in such a bursty cascade. For uncorrelated events we have $P(E) \sim e^{-E/E^*}$ irrespective of the inter-event time distribution.

Our Analysis of mobile phone data shows that besides the fat tailed inter-event time distributions we also have a power-law distribution in $P(E) \sim E^{-\beta}$ indicating intrinsic correlations in human communication activity (see Fig.2).

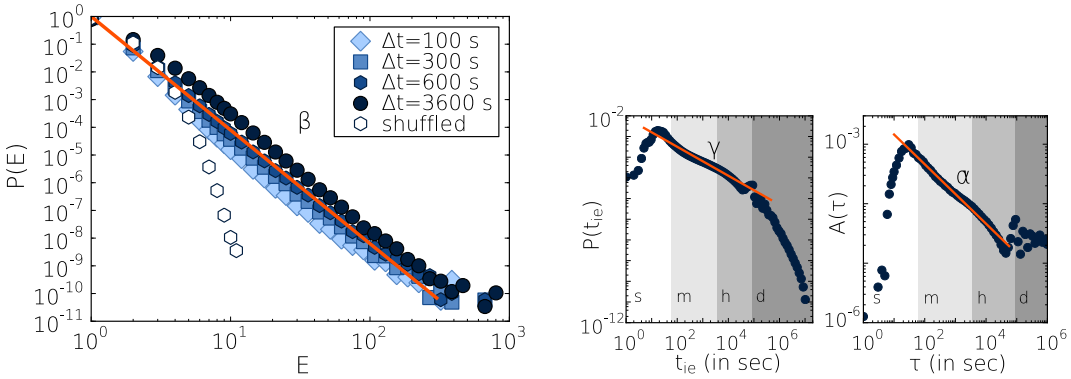


FIG. 2: Average $P(E)$, $P(t_{ie})$ and $A(\tau)$ characteristic functions calculated form the mobile-call data.

These correlations can be attributed to memory effects. We define a memory function as the probability of having the $n+1$-st event in a bursty period, provided it has already contained n events:

$$p(n) = \frac{\sum_{n=E+1}^{\infty} P(E)}{\sum_{n=E}^{\infty} P(E)} \qquad \text{which scales as:} \qquad 1 - p(n) \sim n^{-(\beta-1)} \tag{1}$$

This relationship has also been verified using mobile phone data.

[1] Barabási A.-L. The origin of bursts and heavy tails in human dynamics. *Nature* **435** 207-211 (2005).

[2] Vázquez, A. *et al.* Modeling bursts and heavy tails in human dynamics *Phys. Rev. E* **73**, 036127 (2006).

[3] Malmgren, R.D. *et al.* A Poissonian explanation for heavy tails in e-mail communication. *Proc. Natl. Acad. Sci.* **105**, 18153-18158 (2008).

[4] Jo, H-H., Karsai, M., Kertész, J, Kaski, K. Circadian pattern and burstiness in human communication activity. *arXiv*:1101.0377 (2011).

Ethnic Segregation in the Area of Residence, Work, and Free-time
Evidence from Mobile Communication

Ott Toomet*†‡ Siiri Silm‡ Erki Saluveer‡ Rein Ahas‡

July 15, 2011

Abstract

This paper analyzes spatial ethnic segregation using cellphone data. The data allows us to differentiate between place of residence, work, and freetime; and between ethnic majority and minority groups. We focus on individual pairwise meeting potential (copresence) between ethnic majority and minority groups in an European city (Tallinn, Estonia). We show that the potential number of meetings with the other population group at workplace is only weakly related to the ethnic composition of the neighborhood of residence, and that in freetime is almost completely unrelated. These results suggest that physical separation of ethnic groups in segregated neighborhoods is of a less concern than suggested by place of residence data only.

*Corresponding author. e-mail: otoomet@gmail.com
[2]Department of Economics, Tartu University, Narva 4, Tartu 51009, Estonia
[3]Department of Human Geography, Tartu University

1 Introduction

Everyday observations suggest that immigrants are fairly separated from the native population. A number of studies indicate that this is (partially) causing their inferior labor market and educational outcomes (Clark and Drinkwater, 2002; Card and Rothstein, 2007). Mainly due to data availability, the bulk of the literature focuses on the place of residence and related segregation. More recently, availability of matched employer-employee data has allowed to perform similar analyzes on workplaces (see, for instance Åslund and Skans, 2005; Hellerstein and Neumark, 2008).

However, although the existence of substantial segregation both at home and work is well established, the importance of these facts is unclear. Anecdotal evidence suggest that our most important contacts (besides of the close family) are neither neighbors nor colleagues. In that case, the above mentioned segregation might be of a less concern.

Unfortunately, there is very little evidence about segregation in dimensions other than place of residence and work. In particular, there is very little data available about the social life in free-time. The existing studies, based on friendship ties of school-age children suggest a substantial segregation as well, where members of many minority groups tend to socialize with friends of similar type (Currarini, Jackson, and Pin, 2009; Martinovic, van Tubergen, and Maas, 2009; Currarini, Jackson, and Pin, 2010). Survey-based evidence also indicates that leisure time contacts with the majority population does not depend on the neighborhood ethnic composition (Martinovic, van Tubergen, and Maas, 2009).

The current paper complements this literature by analyzing the segregation based on cellphone usage data. We observe the location (sender) and time of every call in a cellular network in a bilingual city (Tallinn, Estonia) of about 500 000 inhabitants. In addition, we also observe the preferred language (majority or minority) of all the callers. This allows us to analyze segregation (based on proximity in space and time) in the area of residence, work, and outside of these two regions. The previous research indicates that time-spacial copresence is a good indicator for social ties (Crandall, Backstrom, Cosley, Suri, Huttenlocher, and Kleinberg, 2010) in certain circumstances.

Our analysis indicate that the workplace segregation is only moderately associated with the ethnic composition of place of residence, and segregation in freetime activities is virtually unrelated to it. The results are robust with respect to choice of spatial units. Hence the results suggest that despite of the substantial residential segregation, both minority and majority group members have the potential to meet each other at work or in free time.

The paper continues as follows: the next sections discusses the concept of copresence and homophily, and explains how these are computed. Section 3 is devoted to the data description, Section 4 presents the results, the following section provides several robustness results, Section 6 discusses the main findings, and the last section concludes.

References

Card, D., and J. Rothstein (2007): "Racial segregation and the black-white test score gap," *Journal of Public Economics*, 91(11-12), 2158 – 2184.

Clark, K., and S. Drinkwater (2002): "Enclaves, neighbourhood effects and employment outcomes: Ethnic minorities in England and Wales," *Journal of Population Economics*, 15, 5–29, 10.1007/PL00003839.

Crandall, D. J., L. Backstrom, D. Cosley, S. Suri, D. Huttenlocher, and J. Kleinberg (2010): "Inferring social ties from geographic coincidences," *Proceedings of the National Academy of Sciences of the US*, 107(52), 22436–22441.

Currarini, S., M. O. Jackson, and P. Pin (2009): "An Economic Model of Friendship: Homophily, Minorities, and Segregation," *Econometrica*, 77(4), 1003–1045.

——— (2010): "Identifying the roles of race-based choice and chance in high school friendship network formation," *Proceedings of the National Academy of Sciences of the US*, forthcoming.

Hellerstein, J. K., and D. Neumark (2008): "Workplace Segregation in the United States: Race, Ethnicity, and Skill," *Review of Economics and Statistics*, 90(3), 459–477.

Martinovic, B., F. van Tubergen, and I. Maas (2009): "Dynamics of Interethnic Contact: A Panel Study of Immigrants in the Netherlands," *European Sociological Review*, 25(3), 303–318.

Åslund, O., and O. N. Skans (2005): "Will I see you at work? Ethnic workplace segregation in Sweden 1985-2002," Working Paper 2005:24, IFAU, P.O. Box 513, 751 20 Uppsala, Sweden.

Risk and Reciprocity Over the Mobile Phone Network: Evidence from Rwanda

Joshua Blumenstock
University of California, Berkeley

Nathan Eagle
Santa Fe Institute

Marcel Fafchamps
Oxford University

DRAFT VERSION

June 2011*

Abstract:

A large literature describes how local risk sharing networks help individuals deal with idiosyncratic economic shocks. The majority of the empirical evidence focuses on the benefits of risk sharing within small, local communities. When entire communities face the same shock, and when the transaction costs of transfers are high, these risk sharing networks are likely to be less effective. In this paper, we document how a new technology – mobile phones – reduces transaction costs and enables Rwandans to share risk quickly over long distances. Examining usage logs obtained from the monopoly telecommunications provider, and exploiting an earthquake that devastated the Lake Kivu region of Rwanda, we show that a large and anomalous volume of an early form of "mobile money" was sent to people affected by the earthquake in the days immediately following the quake. Accounting for increases in adoption over time, we estimate that between $22,000 and $30,000 would be transferred in response to a current-day earthquake. Though the effect is large, the benefits are heterogeneous: after controlling for a variety of factors, it is the wealthy individuals, and those individuals with contacts living in a disperse geographical region, who are most likely to receive a transfer after the earthquake. We further show that the patterns of transfers sent in response to the earthquake are most consistent with a model of risk sharing, rather than alternative models of remittances or pure altruism.

Keywords: Risk Sharing; Mobile Phones; Information and communications technologies; Development; Earthquakes; Rwanda; Africa.

*email: jblumenstock@berkeley.edu. Address: 102 South Hall, Berkeley, CA 94720, USA. Telephone: 1 (510) 642-1113. The authors are grateful for thoughtful comments from Alain de Janvry, Frederico Finan, Ethan Ligon, Jeremy Magruder, Edward Miguel, and to seminar participants at the Berkeley Development Lunch for helpful comments. We gratefully acknowledge financial support from the International Growth Center, the National Science Foundation, and the Institute for Money, Technology and Financial Inclusion. All errors are our own.

1 Introduction

In developing countries, where social safety nets are rare and formal markets for credit and insurance are thin, people frequently rely on friends and family for support in times of trouble. A large literature describes how such risk sharing networks can allow individuals to effectively smooth consumption over time and in the face of short-term economic shocks (Fafchamps & Gubert 2007, Townsend 1994, Udry 1994, Deaton 1992). However, these networks work best at insuring against idiosyncratic shocks that are uncorrelated across members of the same network. When a large covariate shock affects an entire community, local risk sharing networks are less effective.[1] While individuals could in principle receive support from friends and family living outside of the affected region, sending money over distance is costly and individuals are quite sensitive to the cost of remitting (Yang 2008). More critically, affordable mechanisms for transferring money over long distances often do not exist. In much of East Africa, for instance, formal money transfer systems such as Western Union are only available in major urban areas, and informal methods (such as sending money with a public bus driver) are slow, intermittent, and expensive. Thus, the empirical evidence indicates that in-kind and monetary transfers typically occur between friends and family within small, local communities.[2]

In an increasing number of developing countries, the mobile phone network has begun to provide a new mechanism for interpersonal transfers which could potentially remove the geographic constraint from risk sharing relationships. "Branchless banking" systems, with over 80 deployments worldwide, allow individuals to transfer "mobile money" from one phone to another at a fraction of the cost of existing alternatives (McKay & Pickens 2010). Typically, a mobile subscriber types in the phone number of the recipient and the amount to be transferred, and the balance is deducted from the sender's account and added to the recipient's. The transaction takes a few seconds to complete, and costs at least 50 percent less than what it would cost to send money through traditional channels (Ivatury & Mas 2008). Beyond the convenience and reduction in transaction costs, mobile banking systems are noteworthy for their increasing ubiquity. For instance, a recent study in Kenya found that although only 23 percent of adults owned a bank account, over 50 percent of adults were registered users the mobile banking system (FSD Kenya 2009).[3] Worldwide, it is estimated that by 2012 there will be 1.7 billion people with a mobile phone but no bank account (CGAP and GSMA 2009).

[1] See, for instance, evidence on limited giving in response to famines in India (Sen 1983, Dreze & Sen 1991).

[2] Udry (1994), for instance, observes that 75 percent of surveyed Nigerian households made informal loans, but that almost all loans occurred within a village. Fafchamps & Gubert (2007) similarly observe that geographic proximity is a major determinant of sharing patterns: when two households live near each other, it is more likely that the one will help the other. Kurosaki & Fafchamps (2002) and de Weerdt & Fafchamps (2010) obtain similar findings for Pakistan and Tanzania, respectively.

[3] Over $200 million dollars is transferred over the Kenayan mobile phone network *each day*. Pulver (2009) estimates that 47% of the Kenyan population uses mobile phones as the primary method of sending money. Similarly, in surveys conducted by the first author in Rwanda in July 2009, we found that 97.3% of Rwandan phone users had heard of the Rwandan mobile transfer service, and that nearly 80% had used it within the last year.

In this paper, we explore one mechanism by which such mobile money systems may have a meaningful economic impact on the lives of their users. We measure the extent to which individuals transfers funds over the network in order to help friends and family cope with severe economic shocks. Specifically, we test whether a large earthquake in Rwanda caused people in unaffected parts of the country to transfer a rudimentary form of "mobile money" to people living close to the earthquake's epicenter.[4] Observing the entire universe of mobile-based transfers occurring before and after the earthquake, we find that the earthquake caused a large and significant influx of transfers to the people close to the epicenter. The effect is highly significant on the day of the earthquake and the following day, but not on a number of "placebo" days. We find similar, albeit muted, effects following a number of floods. Our results are robust to different estimation strategies. Though the total volume of money sent following the earthquake was small in absolute terms – primarily because the banking service was launched shortly before the earthquake occurred – simple calculations indicate that if a similar earthquake were to occur today, the current value of mobile money sent would be roughly USD$22,000 to $30,000. This is particularly striking given the fact that, at the time of the earthquake, the liquidity of airtime transfers was rather limited. As the capabilities of such mobile banking systems expand and phone-based transactions become the norm, we would expect the volume (and utility) of such transfers to increase.

From a public policy perspective, we are also interested in identifying which types of people are most likely to receive a mobile-based transfers, both in general and in response to exogenous shocks such as the earthquake. To measure this heterogeneity, we exploit a rich source of data obtained from Rwanda's dominant mobile phone operator. Using the transaction records of *all mobile phone activity* occurring in Rwanda from 2005 until 2009, we measure the mobile usage patterns and a number of properties of the social networks of each of Rwanda's 1.5 million mobile subscribers. Combining this data with survey data collected by the Rwandan government and phone surveys we conducted, we are further able to compute a noisy measure of each individual's economic status. We find that wealthier phone users are significantly more likely to receive a transfer after the earthquake. And while individuals with a large number of contacts are more likely to receive transfers on normal days, they are not significantly more likely to receive a transfer in the day of the earthquake. As discussed later in the paper, this finding implies that the distributional effects of phones – and mobile banking in particular – may be regressive.

Finally, we analyze patterns of giving after the earthquake to test between two stylized models of *why*

[4]During the period we analyze, mobile subscribers were able to transfer balance from one phone to another using a service called "Me2u." This airtime could be used to make calls, could be resent to other subscribers, or could be sold informally for a small commission. However, there were no formal outlets at which the airtime could be converted to cash, and at the time of the earthquake it could not be used to purchase goods. In February 2010, the telecommunications operator launched a fully-fledged Mobile Money service, similar to the M-PESA system in Kenya, which allows subscribers to convert airtime to cash, and which will soon allow for over-the-counter purchases with airtime, as well as interest-bearing airtime accounts.

people are transferring funds over the mobile network. First, the transfers could be motivated primarily by charitable feelings, by which we refer to a variety of motives for assisting others that do not rely on quid-pro-quo.[5] Included in this type of model are remittance-like relationships where one person consistently depends on the other for support. If this general altrusim/remittance behavior were dominant, there are several implications: we would expect to see transfers increasing in the wealth of the sender relative to that of the recipient, and increasing with the distance between sender and recipient (because the sender is less likely to be affected by the same shock). We would also expect to observe transfers flow from urban areas to rural areas, and to occur more often in pairs of people where one person consistently sends and the other receives. On the other hand, if transfers are largely motivated by expectations of reciprocity, as emphasized in the risk sharing literature, the empirical predictions are quite different. In such arrangements, we would expect transfers to decrease with the distance between sender and recipient, as the increase in distance may both impinge a person's ability to assess another's true need (Vreyer et al. 2010), and may increase the difficulty of monitoring and enforcing informal, reciprocal commitments (Ligon et al. 2002, Ligon 1998). We would further expect to find evidence that earthquake-related transfers occur more often in reciprocal relationships where, for instance, at some point prior to the earthquake funds had been sent in the opposite direction (i.e. from receiver to sender). By examining individual-level heterogeneity, and by comparing edges (dyads) of the directed call graph, we test these predictions on the data from Rwanda. We find that while the data are broadly consistent with the latter model of risk sharing, there is little evidence that people give for purely charitable motives in response to the earthquake.

The remainder of the paper is organized as follows. In Section 2 we outline our estimation strategy to measure the effects of the earthquake and potential heterogeneity. In Section **??**, to help motivate the empirical analysis, we develop a simple model of phone-based interpersonal transfers that distinguishes between "charitable" transfers and transfers with a quid-pro-quo expectation. The data are described in Section 3, together with a basic background of mobile phone services in Rwanda. The empirical results are discussed in Section 4, and a series of robustness checks are presented in Section 5. Section 6 concludes.

[5]These include pure altruism (Becker 1976), inequality aversion (Fehr & Schmidt 1999), warm glow (Andreoni 1990), and subjective reputational rewards (Bénabou & Tirole 2006).

Composite Social Network for Predicting Mobile Apps Installation

Wei Pan and **Nadav Aharony** and **Alex (Sandy) Pentland**
MIT Media Laboratory
20 Ames Street
Cambridge, Massachusetts 02139
{panwei,nadav,pentland}@media.mit.edu

Introduction

We are interested in studying the network-based prediction for mobile applications (referred as "apps") installation, as the mobile application business is growing rapidly (Ellison 2010). This work also benefits the general theme of social network-based personalization and recommendation. However it was very difficult to adopt existing tools from large-scale social network research to model and predict the installation of certain mobile apps for each user due to the following facts:

1. The underlying network is not observable. While many projects assume phone call logs are true social/friendship networks, others may use whatever network that is available as the underlying social network. On the other hand, smart phones can easily sense multiple networks using built-in sensors and software: Call logs, bluetooth proximity, GPS co-location, surveys, etc. In this work, our key idea is to infer an *optimal composite network*, the network that best describes diffusion, from multiple layers of different networks easily observed by modern smart phones.
2. Analysis for epidemics and Twitter networks (Yang and Leskovec 2010) is based on the fact that network is the only mechanism for adoption; i.e. the only way to get the flu is to catch the flu from someone else. For mobile app, this is, however, not true at all. Any user can buy an app online directly without peer influence.
3. The individual behavioral variance in app installation is so significant that any network effect might possibly be rendered unobservable from the data. We have both "grandma" type users and geek users in the network.
4. There are exogenous factors in the app installation behaviors. One particular factor is the popularity of apps. For instance, the Pandora Radio app is vastly popular and highly ranked in the app store, while most other apps are not.

Statistical analysis used by social scientists such as matched sample estimation (Aral, Muchnik, and Sundararajan 2009) are only for identifying network effects and mechanism. Recently works in computer science are only applicable to artificial simulation data on real networks (Gomez Rodriguez, Leskovec, and Krause 2010) (Myers and Leskovec 2010). However, by making inference from data rather than making assumptions, our model successfully addresses the above issues in practical social-based prediction. To our knowledge, we don't see other related works for similar problems.

This work is originally published in Proceeding of AAAI 2011. You can download the full version at http://arxiv.org/abs/1106.0359

Data

We collected our data from March to July 2010 with 55 participants, who are residents living in a married graduate student residency of a major US university. Each participant is given an Android-based cell phone with a built-in sensing software developed by us. The software runs in a passive manner, and it didn't interfere the normal usage of the phone. Please refer to Aharony et al.(Aharony et al. 2011) for more information about data collections.

Our software is able to capture most data from phones. We summarize all the networks obtained from both phones and surveys in Table 1. Notice that we collected the affiliation network and the friendship network by deploying a survey, which lists all the participants and ask each one to list their affiliations (i.e. the academic department), and rate their relationships. However we believe for app market makers the affiliation network can also be inferred simply by using phone GPS/cell tower information as shown by Farrahi et al(Farrahi and Gatica-Perez 2010). We refer to all networks in Table 1 as *candidate networks*, and all candidate networks will be used to compute the optimal composite network. It should be noted that all networks are non-directional in this work. Our built-in sensing platform is constantly monitoring the installation of mobile apps. Overall, we receive a total of 821 apps installed by all 55 users.

Prediction by Learning

In this section, we describe our novel learning model for capturing the app installation behaviors in networks. In the following content, $\mathbf{G}$ denotes the adjacency matrix for graph G. Each user is denoted by $u \in \{1, ..., U\}$. Each app is denoted by $a \in \{1, ..., A\}$. We define the binary random variable x_u^a to represent the status of adoption (i.e. app installation): $x_u^a = 1$ if a is adopted by user n, 0 if not.

As introduced in the previous section, the different social relationship networks that can be inferred by phones are de-

Network	Type	Source	Notation
Call Log Network	Undirected,Weighted	# of Calls	G^c
Bluetooth Proximity Network	Undirected,Weighted	# of Bluetooth Scan Hits	G^b
Friendship Network	Undirected,Binary	Survey Results (1: friend; 0: not friend)	G^f
Affiliation Network	Undirected,Binary	Survey Results (1: same; 0: different)	G^a

Table 1: Network data used in this study.

noted by $\mathbf{G}^1, ..., \mathbf{G}^M$. Our model aims at inferring an optimal composite network $\mathbf{G}^{\text{opt}}$ with the most predictive power from all the candidate social networks. The weight of edge $e_{i,j}$ in graph $\mathbf{G}^m$ is denoted by $w^m_{i,j}$. The weight of an edge in $\mathbf{G}^{\text{opt}}$ is simply denoted by $w_{i,j}$.

Adoption Mechanism: One base idea of our model is the non-negative accumulative assumption, which distinguishes our model from other linear mixture models. We define $\mathbf{G}^{\text{opt}}$ to be:

$$\mathbf{G}^{\text{opt}} = \sum_m \alpha_m \mathbf{G}^m, \text{where } \forall m, \alpha_m \geq 0. \quad (1)$$

The intuition behind this non-negative accumulative assumption is as follows: if two nodes are connected by a certain type of network, their app installation behaviors may or may not correlate with each other; On the other hand, if two nodes are not connected by a certain type of network, the absence of the link between them should lead to *neither positive or negative effect* on the correlation between their app installations. As shown in Table 2 in the experiment session, our non-negative assumption brings significant performance increase in prediction. Non-negative assumption also makes the model stochastic and theoretically sound. We treat binary graphs as weighted graphs as well. We later refer to the vector $(\alpha_1, ..., \alpha_M)$ as the optimal composite vector. We continue to define the network potential $p_a(i)$:

$$p_a(i) = \sum_{j \in \mathcal{N}(i)} w_{i,j} x^a_j, \quad (2)$$

where the neighbor of node i is defined by:

$$\mathcal{N}(i) = \{j | \exists m \text{ s.t. } w^m_{i,j} \geq 0\}. \quad (3)$$

The potential $p_a(i)$ can also be decomposed into potentials from different networks:

$$p_a(i) = \sum_m \alpha_m \underbrace{(\sum_{j \in \mathcal{N}(i)} w^m_{i,j} x^a_j)}_{p^m_a(i)}, \quad (4)$$

where $p^m_a(i)$ is the potential computed from one single candidate network. We can think of $p_a(i)$ as the potential of i installing app a based on the observations of its neighbors on the composite network. The definition here is also similar to incoming influence from adopted peers for many cascade models (Kempe, Kleinberg, and Tardos 2003).

Finally our conditional probability is defined as:

$$\text{Prob}(x^a_u = 1 | x^a_{u'} : u' \in \mathcal{N}(u)) = 1 - \exp(-s_u - p_a(u)), \quad (5)$$

where $\forall u, s_u \geq 0$. s_u captures the individual susceptibility of apps, regardless of which app. We use the exponential function for two reasons: a) The monotonic and concave properties of $f(x) = 1 - \exp(-x)$ matches with recent research (Centola 2010), which suggests that the probability of adoption increases at a decreasing rate with increasing external network signals. b) It forms a concave optimization problem during maximum likelihood estimation in model training. As shown in the experiment section and based on our experiences, this exponential model yields the best performance.

Model Training: We move on to discuss model training. During the training phase, we want to estimate the optimal values for the $\alpha_1, ..., \alpha_M$ and $s_1, ..., s_U$. We formalize it as an optimization problem by maximizing the sum of all conditional likelihood.

Given all candidate networks, a training set composed of a subset of apps TRAIN $\subset \{1, ..., A\}$, and $\{x^a_u : \forall a \in \text{TRAIN}, u \in \{1, ..., U\}\}$, we compute:

$$\arg \max_{s_1,...,s_U,\alpha_1,...,\alpha_M} f(s_1, ..., s_U, \alpha_1, ..., \alpha_M),$$
$$\text{Subject to:} \quad \forall u, s_u \geq 0, \forall m, \alpha_m \geq 0 \quad (6)$$

where:

$$\begin{aligned} & f(s_1, ..., s_U, \alpha_1, ..., \alpha_M) \\ = & \log \Big[\prod_{a \in \text{TRAIN}} \prod_{u: x^a_u = 1} \text{Prob}(x^a_u = 1 | x^a_{u'} : u' \in \mathcal{N}(u)) \\ & \prod_{u: x^a_u = 0} \big(1 - \text{Prob}(x^a_u = 1 | x^a_{u'} : u' \in \mathcal{N}(u))\big) \Big] \\ = & \sum_{a \in \text{TRAIN}} \Big[\sum_{u: x^a_u = 1} \log(1 - \exp(-s_u - p_a(u)) \\ & - \sum_{u: x^a_u = 0} (s_u + p_a(u)) \Big] \quad (7) \end{aligned}$$

(8)

This is a concave optimization problem. Therefore, global optimal is guaranteed, and there exist efficient algorithms scalable to larger datasets (Boyd and Vandenberghe 2004).

We emphasize that our algorithm doesn't distinguish the causality problem (Aral, Muchnik, and Sundararajan 2009) in network effects: i.e.,we don't attempt to understand the different reasons why network neighbors have similar app installation behaviors. It can either be diffusion (i.e. my neighbor tells me), or homophily (i.e. network neighbors share same interests and personality).

Virtual Network for Exogenous Factors: Obvious exogenous factors include the popularity and quality of an app, which affect the ranking and review of the app in the AppStore/AppMarket. We can model this by introducing a virtual graph G^p, which can be easily plugged into our composite network framework. G^p is constructed by adding a virtual node $U+1$ and one edge $e_{U+1,u}$ for each actual user u. The corresponding weight of each edge $w_{U+1,u}$ for computing $p_a(u)$ is C^a, where C^a is a positive number describing the popularity of an app. In our experiment, we use the number of installations of the app in this experimental community as C^a. In practice for app market makers, we argue that C^a can be easily obtained accurately by counting app downloads and app ranks.

The exogenous factors also increase accuracy in measuring network effects for a non-trivial reason: Considering a network of two nodes connected by one edge, and both nodes installed an app. If this app is very popular, then the fact that both nodes have this app may not imply a strong network effect. On the contrary, if this app is very uncommon, the fact that both nodes have this app implies a strong network effect. Therefore, introducing exogenous factors does help our algorithm better calibrate network weights.

Experiments[1]

Model Validation: Our algorithm predicts the probability of adoption (i.e. installing an app) given its neighbor's adoption status. $p_i \in [0,1]$ denotes the predicted probability of installation, while $x_i \in \{0,1\}$ denotes the actual outcome. The most common prediction measures is the Root Mean Square Error (RMSE $= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i - x_i)^2}$), Mean Precision at k (MP-k) and Optimal F_1-score (the largest F_1 Score on the Precision-Recall curve).

We test different configurations in Table 2 to check how three major factors in our approach (networks, individual variances and exogenous factors) affect prediction accuracy by testing our algorithm with some factors omitted. We also test our non-negative accumulation assumption here.

Conditions	RMSE	MP-5	F_1
Net.+ Ind. Var. + Exogenous Factor	**0.25**	**0.31**	**0.43**
Net. + Ind. Var. Only	0.26	0.29	0.42
Ind. Variance Only	0.29	0.097	0.24
Net. Only (non-negative)	0.26	0.24	0.37
Net. Only (allow negative)	0.30	0.12	0.12

Table 2: The performance of our approach under five different configurations. We observe that modeling both individual variance and exogenous factors are crucial in performance as well as enforcing non-negative composition for candidate networks as in Eq. 1.

In addition, we illustrate the prediction performance when our algorithm is only allowed to use one single network. The results are shown in Fig. 1. We discover that our composite network performs better than single network alone.

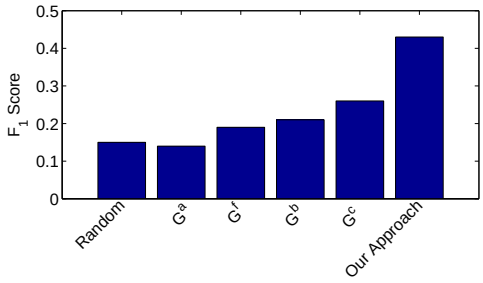


Figure 1: We demonstrate the prediction performances using each single network here. For comparison, we also show the result of random guess, and the result using our approach, which combines all potential evidence.

Comparison and Performance: We discover that our model performs much better than alternative approaches, and predicts almost half of all apps users install ($\sim 45\%$ precision at 45% recall). We demonstrate that our algorithm performs very well in both predicting future app installation and in predicting with significant missing data. We refer readers to our full paper for details.

References

Aharony, N.; Pan, W.; Ip, C.; Khayal, I.; and Pentland, A. 2011. The social fmri: Measuring, understanding and designing social mechanisms in the real world. In *Proceedings of the 13th ACM international conference on Ubiquitous computing*. ACM.

Aral, S.; Muchnik, L.; and Sundararajan, A. 2009. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences* 106(51):21544.

Boyd, S., and Vandenberghe, L. 2004. *Convex optimization*. Cambridge Univ Pr.

Centola, D. 2010. The Spread of Behavior in an Online Social Network Experiment. *science* 329(5996):1194.

Ellison, S. 2010. Worldwide and U.S. Mobile Applications, Storefronts, and Developer 20102014 Forecast and Year-End 2010 Vendor Shares: The "Appification" of Everything.

Farrahi, K., and Gatica-Perez, D. 2010. Probabilistic Mining of Socio-Geographic Routines From Mobile Phone Data. *Selected Topics in Signal Processing, IEEE Journal of* 4(4):746–755.

Gomez Rodriguez, M.; Leskovec, J.; and Krause, A. 2010. Inferring networks of diffusion and influence. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1019–1028. ACM.

Kempe, D.; Kleinberg, J.; and Tardos, É. 2003. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, 137–146. ACM.

Myers, S., and Leskovec, J. 2010. On the convexity of latent social network inference. *Arxiv preprint arXiv:1010.5504*.

Yang, J., and Leskovec, J. 2010. Modeling Information Diffusion in Implicit Networks.

[1]Due to space limitation, please refer to our full paper for detail on experiments: `http://arxiv.org/abs/1106.0359`.

Mobile phone and motorway traffic: a panel data perspective

Emmanouil Tranos, e.tranos@vu.nl

John Steenbruggen, john.steenbruggen@rws.nl

Peter Nijkamp, p.nijkamp@vu.nl

Henk Scholten, hscholten@feweb.vu.nl

Dept. of Spatial Economics, VU University

De Boelelaan 1105, 1081 HV Amsterdam

The Netherlands

The objective of our paper is to use mobile phone data as the main predictor of motorway traffic. Such a modelling exercise can provide a useful tool for transport engineers as it will enable the (near) real-time estimation of car traffic in specific segments of motorways avoiding the use of other more expensive and less efficient surveying techniques. Apart from mere academic interest in utilizing this new and rich data source in a social science domain, potential applications could include transport and incident management (Steenbruggen et al. 2011).

The case study for our research is the city of Amsterdam. The data we will utilize for this paper has been supplied by a major telecom operator in the Netherlands and provides aggregated information about mobile phone usage at the level of the GSM cell for the period 2007-2010. The spatial dimension is comprised by c. 1200 cells. The temporal dimension provides information on an hourly basis creating a very detailed pool of data. Indeed, such a rich dataset appears to be a 'luxury' for spatial analysts, but at the same time increases the complexity of the necessary analytical methods. The mobile phone dataset will be used to create a key variable indicating the mobile phone usage intensity. This variable will be the main explanatory variable of interest in the model we will build in our research. As our objective here is to predict car density in specific motorway segments, the dependent variable for this model will be based on motorway traffic measures based on a detection loop method. Such data has been made available to the research group and will be used here as the dependent variable. The main challenge is to spatially link these two heterogeneous datasets which are different. In order to do so, an area selection process will take place. Although we have data available for the city of Amsterdam, we will focus only on these cells that intersect with motorways for which there is traffic data available from the detection loop dataset.

Nonetheless, mobile phone data is not the only predictor for car density. Consequently, a set of control variables will be built. These variables will include both temporal and spatial information. In more detail, control – dummy – variables will be included in the model indicating times of interest during the day such as rush hours, working hours and night time hours, as well as days of interest

such as weekdays or weekends and bank holidays. Other temporal control variables of interest could be the time of year (winter/summer) and holiday periods. In addition, land use variables will also be incorporated in this model. Examples include the location of transport hubs or business areas inside the cell, as these types of land use will increase the intensity of mobile phone usage in a cell. On the contrary, land cover such as water-covered areas will be negatively related with phone usage.

From an econometric point of view, panel data will be the preferred specification for this model, as it will enable us to take advantage of the two very detailed dimensions of the data (space and time). Panel data improves the researchers ability to control for missing or unobserved variables (Hsiao 2003). Such an omitted-variable bias as a result of unobserved heterogeneity is a common problem in cross-section models. In addition, potential selection bias can also be addressed more efficiently with panel data. In total, panel data specification reduces the risk of obtaining biased estimators (Baltagi 2001).

Altogether, we 'will likely have around 100 cells of interest for a four year period (4years *x* 365days *x* 24hours) and without considering any potential data gaps we will achieve a panel data of c. 3.5 million observations. This will provide the basis for a robust modelling exercise. However, two econometric challenges might occur here: temporal autocorrelation due to the repeated over time measurements, and also spatial autocorrelation because of the potential dependence of mobile phone intensity between adjacent cells (Sevtsuk and Ratti 2010). In order to reveal these issues, relevant econometric tests will be performed and alternative to OLS specifications will also be utilized. In addition, the analysis of residuals will reveal information about cells and points in time where and when our prediction was not successful.

In a nutshell, we are aiming to build a generic model in order to see if and how mobile phone use intensity can be utilised as a predictor for car density and motorway traffic. Such a model will provide the basis for a real-time system for traffic and incident management.

Reference list

Baltagi BH (2001) Econometric analysis of panel data. 2nd edn. John Wilet & Sons, Chichester

Hsiao C (2003) Analysis of panel data. 2nd edn. Cambridge University Press, Cambridge

Sevtsuk A, Ratti C (2010) Does Urban Mobility Have a Daily Routine? Learning from the Aggregate Data of Mobile Networks. Journal of Urban Technology 17 (1):41-60

Steenbruggen J, Borzacchiello MT, 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 forthcoming

Churn Analysis in Mobile Telecom Data using Hybrid Paradigms

Yeshwanth V and Saravanan M
Ericsson R & D
Ericsson India Global Service Private Ltd, Chennai, India
{yeshwanth.v, m.saravanan}@ericsson.com

Abstract

Churn in the telecom domain refers to the movement of customers from one operator network to another. With the high number of operators competing in every market ranging from the developing to developed countries, the relevance to churn has increased dramatically. The key point being that the cost of bringing in a new customer is much greater than retaining and maintaining the value of an existing customer. While many churn prediction models are out in the market, when applied to the live environment many fail to live up to their advertised capabilities. Being a predominantly social phenomenon, it's a challenge to come up with a single good model. While prediction models have grown leaps and bounds, their application to the domain still remains the bottleneck. Our models and analysis have taken into the account the vital cogs in the mobile operator's network and also by a proper data preprocessing with respect to our needs. We have employed a split data preprocessing technique coupled with a hybrid methodology to perform churner prediction on a mobile network. The hybrid approach has been tried out with different classifier pairs to see the difference on two different mobile telecom datasets. We have compared the different pairs of classifiers with more preference given to the hit rates of churners while keeping the misclassification rate of non churners under an appreciable limit. We have focused on improving the classification/prediction model in this paper as an extension to an existing model [1] which has shown good results on a particular mobile telecom dataset. We have also addressed the issue of class imbalance problem, as a skewed distribution of churner percentage plays a major role in the quality of the results.

1. Introduction

Churn is the buzzword in today's telecom scenario. With emergence of high end technologies and a saturated market, the onus is heavily on retaining existing customers i.e. see to it that they don't churn out. The mobile network operators are struggling to comprehend their customer problems and in that respect they are facing difficulty in understanding them accurately and that timely knowledge translates into improved business performances. The aim of this study is to show an improvement on an existing methodology which involved a combination hybrid paradigms and influence analysis for the prediction of churners in a network. We will be showcasing the modification to the data mining phase of the model by evaluating it on a customer data of a South American (SA) operator.

By identifying the core set of possible churners we can deduce the factors common to their usage pattern. This turns out to be a more than a useful spin off. Due to the complexity in prediction and difficulty in building a good single model that shows high performances, many researchers have employed a hybrid model approach in solving the problem [2, 3, 4]. Our previous model used a combination of *tree induction* with *genetic algorithms* in the data mining phase, whose output was used to seed the community based churn propagation values [1]. The community churn value is described as *which describes how "influenced" or "liable to churn" the node is due to the effect of his/her neighbor's behavioral patterns*. We will not delve into the community churn part as it's covered in detail [1].

2. Data Preprocessing

In a typical billing system of a mobile operator, for every operation performed by the customer vary from Voice, SMS usage to GPRS usage and each individual events are recorded and stored. These records are referred to as Call Detail Records (CDRs). We have considered around 0.5 million subscriber's CDRs over a 6 month period from a SA telecom operator for our analysis. From the available users dataset 0.3 million is taken for training and the remaining as a test set. The required fields from CDRs are aggregated over time intervals (in our case month, week and day level) and possible churners are identified. In earlier churn prediction models the fields were aggregated over entire time period [5, 6]. The disadvantages of other approaches is that in case of new user who has joined in later part of the month may be classified as churner due to his low usage and also it's difficult to generate any patterns from those data, as differences in usage pattern cannot be derived.

These issues are analyzed thoroughly in our approach. For example, if we have data for a month, then we can split into 4 weekly intervals. Now we'll have number of calls split into 4 equal intervals. If he is a new user (who has joined in the end of month) the number of calls made by him in the first intervals might be zero. In other models he will be classified as a churner but in our model that user will not be considered as a churner. We have the data whether a person has churned out or not as classified by the operator over the time period. We have attributes related to usage, spend, refill and interconnection. We reduce our attribute list into minimal level with the help of derived gain score. In this case, we have used an in house Hadoop [7] based data processing engine to preprocess and analyze the bulk of the data for quicker turnaround time.

3. Hybrid Paradigm

A. Existing Hybrid Methodology

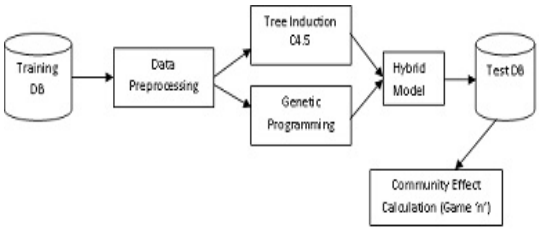


Fig.1 Hybrid Approach for Community effect of churn

In our earlier paper [1], we have used a combination of C4.5 and Genetic Algorithm to predict the churn score. Now we applied the same hybrid approach on the current split aggregated data. We will be comparing the various methodologies as we proceed in the paper. The churn score [1] is defined as '*the accuracy of the prediction that a person will churn out with*'.

$$\textbf{ChurnScore} = (OP_{C4.5} + OP_{GA})/2$$
$$OP_{C4.5} = \text{output(C4.5)} * \text{Accuracy of C4.5}$$
$$OP_{GA} = \text{output(GA)} * \text{Accuracy of GA}$$

The churn score is a function of the number of times an instance has been classified as a particular class and the hit-rate for that class in that individual model. The community churn phase used a Game Theoretic Centrality [8] concept to model the social spread of churn [1]. This paper does not focus on the community aspect instead our main intention is to present a strong data mining phase to the churn prediction issue, which we felt could enhance and fulfill the prediction phase of our approach. The observed statistics of this method is presented as follows by using WEKA implementations [9].

Results:

Correctly Classified Instances	*139545*	*83.7268 %*
Incorrectly Classified Instances	*20589*	*12.3534 %*
Mean absolute error	*0.1286*	

TP Rate	FP Rate	Precision	Recall	ROC Area	Class
0.999	0.167	0.855	0.999	0.916	0
0.087	0.004	0.632	0.087	0.541	1

We see that the same hybrid model which worked well in the previous African operator's dataset fails miserably with this data. Very few churners are detected which makes the usefulness of the model questionable to our purpose even though it picks the non churners properly.

B. Proposed Methodology

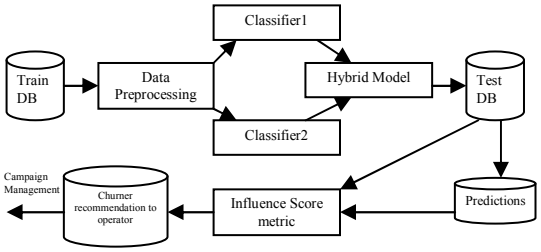


Fig 2. Proposed Framework for Churn Prediction

We now experiment with a different ensemble of classifiers along with the voting system of the hybrid mechanism. We take out the complexity of the community churn part from the base model and use a simpler influence score metric for simple priority based recommendation mechanism as described in Section 4. As it is observed from below, we keep the C4.5 algorithm (classifier 1) constant in our experiments. This is because while the results are important to the operator, the explanation of the results also plays a big part i.e. using the decision trees/if-then rules generated by the latter, we are able to derive some conclusions about the data and see which combination of factors plays a part in predicting customer churn. Classifier 2 has been changed to SVM [10] and Random Forest for comparison.

C4.5 and SVM

We now use the combination of SVM and C4.5.

Results:

Correctly Classified Instances	*139718*	*84.1371 %*
Incorrectly Classified Instances	*26341*	*15.8629*
Kappa statistic	*0.2528*	
Mean absolute error	*0.2358*	

TP Rate	FP Rate	Precision	Recall	ROC Area	Class
0.965	0.772	0.861	0.965	0.655	0
0.228	0.035	0.564	0.228	0.655	1

The above combination does not give us great results but is relatively better to the genetic programming approach. The time taken for execution for the SVM approach is also very high which is not favorable for production environments.

C4.5 and Random Forest

We now use a combination of Random Forest and C4.5.

Results
======

Correctly Classified Instances	*146162*	*87.697 %*
Incorrectly Classified Instances	*19898*	*11.9388 %*
Kappa statistic	*0.399*	
Mean absolute error	*0.1259*	

TP Rate	FP Rate	Precision	Recall	ROC Area	Class
0.962	0.628	0.905	0.962	0.714	0
0.372	0.038	0.609	0.372	0.712	1

We see the improvement in our approach significantly in this combination. The non churner's have been classified satisfactorily while getting a decent hit rate for the churners too. This performs consistently on other data sets too, namely the African operators dataset. For the further improvement of our model, we undertake the class imbalance problem in the hybrid approach.

Handling Class imbalance problem

The percentages we get with respect to the churn label seem to be less impressive mainly due to the skew in the class distribution. With an average monthly churn rate of 3%, the model is trained on a heavily skewed data, hence the results. We have around 46,000 churners out of the 0.3 million subscribers in the training set. We now split (not equal parts) the training set into 3 parts of each 0.11 million such that each and every one of the partitions includes the 46,000 churners in them. So now the percentage split in the training set which was 13% churners has been increased to around 39%. The final hybrid approach is applied on all these partitions to build the model and the results which are combined to give a prediction on the test set. We get vector of 6 predictions from the models of each partition (results based on two different classifiers). We use a voting mechanism based on the hit rates of the predicted class for every partition's model and decide the final class label for us to output as the prediction of the hybrid approach.

Results:

Correctly Classified Instances	*103467*	*97.2251 %*
Incorrectly Classified Instances	*2953*	*2.7749 %*
Kappa statistic	*0.9355*	
Mean absolute error	*0.0529*	

TP Rate	FP Rate	Precision	Recall	ROC Area	Class
0.998	0.082	0.962	0.998	0.961	0
0.918	0.002	0.996	0.918	0.961	1

We see that the hit rate with respect to the churners has shown a huge increase by handling the class imbalance in the dataset. Our main concern is the hit rate for the churners as we want to target as many possible churners as possible. But theoretically speaking if we extended the targeting list for retention campaigns as far as the entire network, we would have chance to understand every possible churner. This is not a sensible approach as such a prediction list would involve a huge proportion of misclassified non churners which is undesirable. Hence we would want to maintain a good ratio between the hit rate of churners and the misclassification of non churners. In terms of the ROC curve we would want a model that comes near the left hand side of the graph. The reason we stress on the misclassification rate of non churners is that, the operator would be giving an unnecessary discounted rate/offer to the misclassified customer which would lead to unnecessary drop in revenue with respect to those subscribers.

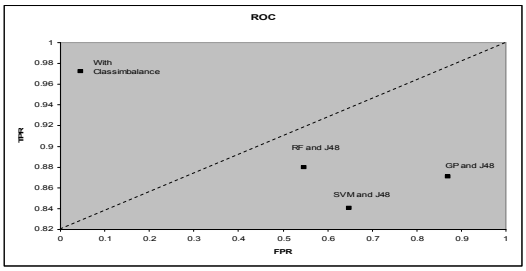


Fig. 3. ROC chart for FPR verses TPR

4. Influence Score

While we are not interested to go into community effects in churn as mentioned before, we do need a simple way of selecting subscribers from our prediction list for us to target using campaigns to try and retain them. While it is arguable that the churn score itself is a good parameter to use as a selection parameter, we need to select the subscribers in order of their influence or in other words target those who might affect those around them.

For that we devised a function (it is normalized)

Influence_score = ***f(no. of incoming calls, no. of outgoing calls, total no. of distinct connections)***

We take the top 69000 possible churners based on the churn score and distribute them based on the calculated influence score using the function above. We clustered the subscribers based on the 6 month usage statistics using K-Means clustering methodology [9] into High, Medium and Low usage clusters. Now we cross referenced the influence scores with the cluster segments as shown below. Depending upon the requirements of the operator's campaign management they can target specific risk prone users with appropriate campaigns. This is a less expensive priority/need based selection on a practical basis for possible churner prediction without having to go through the complex graph and centrality measure calculation in the previous paper[1]. But by no means is this a replacement to the effectiveness of the community churn methodology.

Average Influence Score based on Segments			
	High	**Medium**	**Low**
Max	12.39	10.01	12.97
Min	0.21	0.08	0.00034
Avg	2.25	1.24	0.54

Number of predicted Churners		
	Min - Avg	**Avg - Max**
High	2624	1914
Medium	12486	9178
Low	25041	17757

5. Conclusion

In this paper we have presented some extensions to our earlier work on churn prediction especially in the data mining aspect. The hybrid paradigm we had adopted in the previous study turned out to be ineffective when we used it for analysis on the new dataset. We explored new combinations and evaluated it on different datasets to see their effectiveness in predicting churners. While keeping the C4.5 decision tree algorithm constant is backed by the rule based output; the genetic programming classifier failed miserably on the current dataset. While the SVM classifier did relatively well, the Random Forest classifier performed exceedingly well with respect to predicting churners. In order to improve the churner prediction we took care of the class imbalance problem by splitting the data into smaller splits with an improved churner-non churner distribution and used the hybrid mechanisms on it. The results were very good and ended up predicting most of the churners while keeping the low misclassification rate. In the end we used an influence score metric which was calculated using the connection statistics of the subscribers to use as a selection criteria to pick the most important predictions to use for targeted campaigns. This is highly useful in production environments as it is a less expensive operation than using social network analysis etc.

References

[1] Yeshwanth V, Vimal Raj A, and Saravanan, M. "Evolutionary Churn Prediction in Mobile Networks using Hybrid Learning", in Proceedings of 24th International Florida Artificial Intelligence Research Society Conference (FLAIRS-24), May 18-20, 2011, Palm Beach, Florida, USA.

[2] Wai-Ho Au Chan, K.C.C and Xin Yao. 2003. A novel evolutionary data mining algorithm with applications to churn prediction. *IEEE Transactions on Evolutionary Computation*, 7, .532-545.

[3] Bala, J, Huang, H. Vafaie, K. DeJong and H.Wechsler. 1995. Hybrid Learning Using Genetic Algorithms and Decision Trees for Pattern Classification, *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), August 19-25, 1995, Montreal, Canada*

[4] Jae Sik Lee and Jin Chun Lee. 2006. Customer churn prediction by hybrid model, *Advanced Data Mining and Applications*, 4093, 959-966, Springer Berlin.

[5] Koustuv Dasgupta, Rahul Singh, Balaji Viswanathan, Dipanjan Chakraborty, Sougata Mukherjea, Amit Anil Nanavati, and Anupam Joshi. 2008. Social ties and their relevance to churn in mobile telecom networks. EDBT 2008: *Proceedings of the 11th international conference on Extending database technology*, 668-677, New York, USA. ACM

[6] Shin-Yuan Hung, David C. Yen, and Hsiu-Yu Wang. 2006. Applying data mining to telecom churn management. *Expert System Applications,* 31(3), 515-524.

[7] http://hadoop.apache.org/

[8] Aadithya, K. V. and Ravindran, B. 2010. Game Theoretic Network Centrality: Exact Formulas and Efficient Algorithms". *Proceedings of the Ninth International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2010).*

[9] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. *The WEKA Data Mining Software: An Update; SIGKDD Explorations*, 11. 1.

[10] Archaux C. And Laanaya H. and Martin A. and Khenchaf A. 2004. An SVM based Churn Detector in Prepaid Mobile Telephony", *Proceedings of the International Conference on Information & Communication Technologies: from Theory to Applications (ICTTA*), Damascus, Syria, 19-23 (April 2004).

Relational Learning for Customer Churn Prediction: The Complementarity of Networked and Non-Networked Classifiers

Wouter Verbeke[a], Thomas Verbraken[*,a], David Martens[b], Bart Baesens[a,c,d]

[a] *Department of Decision Sciences and Information Management, Katholieke Universiteit Leuven, Naamsestraat 69, B-3000 Leuven, Belgium*
[b] *Faculty of Applied Economics, University of Antwerp, Prinsstraat 13, 200 Antwerp, Belgium*
[c] *School of Management, University of Southampton, Highfield Southampton, SO17 1BJ, United Kingdom*
[d] *Vlerick, Leuven-Gent Management School, Reep 1, B-9000 Gent, Belgium*

Abstract

This study examines the applicability of relational classification algorithms for customer churn prediction in the telco industry, and the existence and usability of non-Markovian social network effects. A range of new and adapted techniques are proposed which are designed to handle the massive size of the call graph, the time dimension, and the skewed class distribution present in a customer churn prediction setting. The proposed techniques are experimentally tested on a large-scale, real life telco data set containing both networked (call detail records data) and non-networked (customer related) information about millions of subscribers. The results indicate the existence of a limited yet highly relevant impact of social network effects on customer churn behavior, including non-Markovian network effects. A parallel setup to combine the output of a relational and non-relational churn prediction model leads to substantially improved performance and boosts the generated profits.

1. Introduction

In recent years, vast amounts of networked data on a broad range of network processes and information flows between interlinked entities have become available, such as for instance calls and text messages linking telephone accounts (Dasgupta et al., 2008; Richter et al., 2010), money transfers connecting bank accounts (Martens and Provost, 2011), and messages relating email accounts. These massive, networked data logs open new perspectives for innovative business applications (Bonchi et al., 2011) and potentially hide information that is highly valuable to companies and organizations. This information is however extremely difficult to discover due to the size and the fragmentation of the data.

Networked data present both complications and opportunities for predictive data mining. The data are patently not independent and identically distributed, which introduces bias to learning and inference procedures (Jensen and Neville, 2002; Macskassy and Provost, 2007). Relational learning aims to exploit the information contained within the network structure of data instances, and to incorporate this information within a network classification or regression model (Džeroski and Lavrač, 2001; Getoor and Taskar, 2007). Relational classifiers learn directly from a graph or network, as opposed to non-relational classifiers which require an attribute-value representation of the data. An alternative approach to relational learning exists in aggregating the information contained within a network structure into variables or attributes, which can then be incorporated straightforward within a non-relational model. This approach is called featurization or propositionalization (Kramer et al., 2001).

2. Data description

Telco operators collect call detail records data, which contain detailed logs about all the transactions made by the customers of the operator. The basic information that is typically contained within CDR logs are the identity of the sender and the receiver as well as the telco operator of the sender and receiver, a time stamp, duration or size of the transaction, and the product type. Possible types of products are voice to voice calls (VVC), text messages (SMS), multimedia messages (MMS), and other specific telecommunication services that are offered by telco operators.

Given the information contained within CDR data, a graph can be constructed that represents the social network of the subscribers of the telco operator. This graph will be denoted the *call graph*. The nodes in the call graph represent the customers of a telco operator, and links between customers indicate they communicated. Optionally, the network may include customers of other telco operators as well, or a *competing operator vertex* bundling all the

*Corresponding author. Tel. +32.16.326.880; Fax +32.16.326.624
Email addresses: wouter.verbeke@econ.kuleuven.be (Wouter Verbeke), thomas.verbraken@econ.kuleuven.be (Thomas Verbraken), thomas.verbraken@econ.kuleuven.be (David Martens), bart.baesens@econ.kuleuven.be (Bart Baesens)

customers of competing telco operators in a single node.

Call detail record data was provided by an anonymous European telco operator, and covers a period of six months. Churn labels, indicating the exact date when customers churned, were available for this period and one month prior and after. The number of subscribers in the customer base is $1,925,427$, and the number of edges connecting customers in the social network derived from the CDR data equals $7,330,860$. The class distribution is very skewed, which is typically the case for customer churn data sets, with on average only 0.52% of the customers that churn each month. Furthermore, 58 local data attributes were provided for the each customer in the database.

3. Methodology

3.1. Relational classifiers, collective inference procedures, and propositionalization

Macskassy and Provost (2007) introduced a framework for classification in networked data. In this node-centric framework, a *full* relational classifier comprises a local classifier , a *pure* relational or network classifier , and a collective inference (CI) procedure.

A *local classifier* only considers attributes that are related to the entity that is to be classified. This type of classifiers has extensively been studied by the data mining community (an extensive literature overview is given by Verbeke et al. (2011)). A *relational classifier* makes use of the links between entities, as well as attribute values at linked entities, to estimate the label of a given node. However, in general, the labels of a multitude of nodes are unknown at the time of estimation and these nodes may be in each others neighborhood. Therefore, an iterative *collective inference procedure* is employed to simultaneously discover the labels for all unknown nodes in the network. In this study, the techniques developed by Macskassy and Provost (2007) and Dasgupta et al. (2008) have been implemented using sparse and parallel computation techniques in order to be applicable on massive networks consisting of millions of nodes.

An alternative approach to the relational classifiers and collective inference procedures exists in transforming the information that is contained within a graph into a set of network variables or attributes. These network variables can then be used as explanatory variables by traditional data mining techniques such as for instance logistic regression. Methods that transform a relational representation of a learning problem into a propositional, feature-based or attribute-value representation are known as *propositionalization* or featurization approaches (Kramer et al., 2001).

3.2. Combining collective inference, local and relational classifiers

A method to combine collective inference procedures and relational classifiers in a uniform setup follows directly from the definitions of the CI algorithms. In each iteration of the CI procedure the relational classifier is applied, and the result is used as input to the next iteration.

A first approach to combine a local classifier with both a relational classifier and a collective inference procedure in a unified setup follows as well from the definitions of the CI algorithms; the probabilities or scores resulting from a local classification model can be used as the initial labels of the nodes in the network. The local model is used in this setup to provide a first estimate of the class labels of the nodes in the network. Subsequently, the relational classifier and the collective inference procedure are added as a second model layer on top of the local classifier, with the intention to refine and improve the results of the local model and by using the information that is incorporated within the network data.

Conversely, the predicted probabilities by the network model can be included in the local model as an explanatory variable. In this approach the local model constitutes a second model layer on top of the network model. In fact, this can be regarded as an *automated* propositionalization method, leading to a network variable that is not explicitly defined but nonetheless aggregates network information.

An alternative approach exists in combining the output scores of the relational classifier, collective inference procedure, and local classifier by learning a *meta-model*. The meta-model uses the probabilities resulting from the local classification model and the relational classification model, whether or not in combination with a CI procedure, as input variables. In principle, any classification technique could be applied to induce a meta-model on top of the relational and non-relational classification models.

Finally, the network model and the local model can be applied in a parallel, non-integrated setup, by selecting customers indicated to have a high probability to churn either by the local model or by the network model (or by both models).

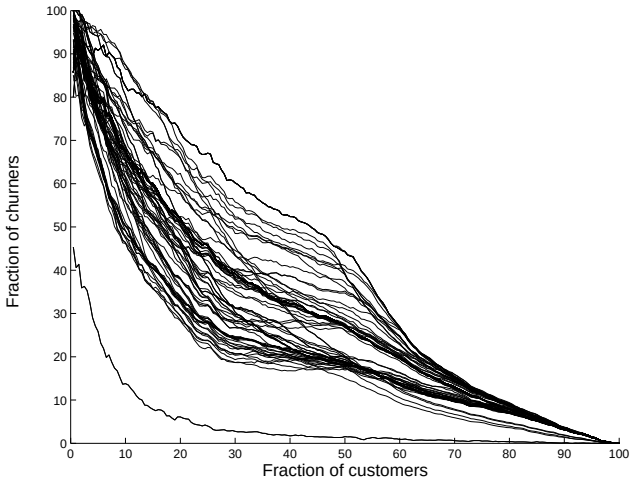
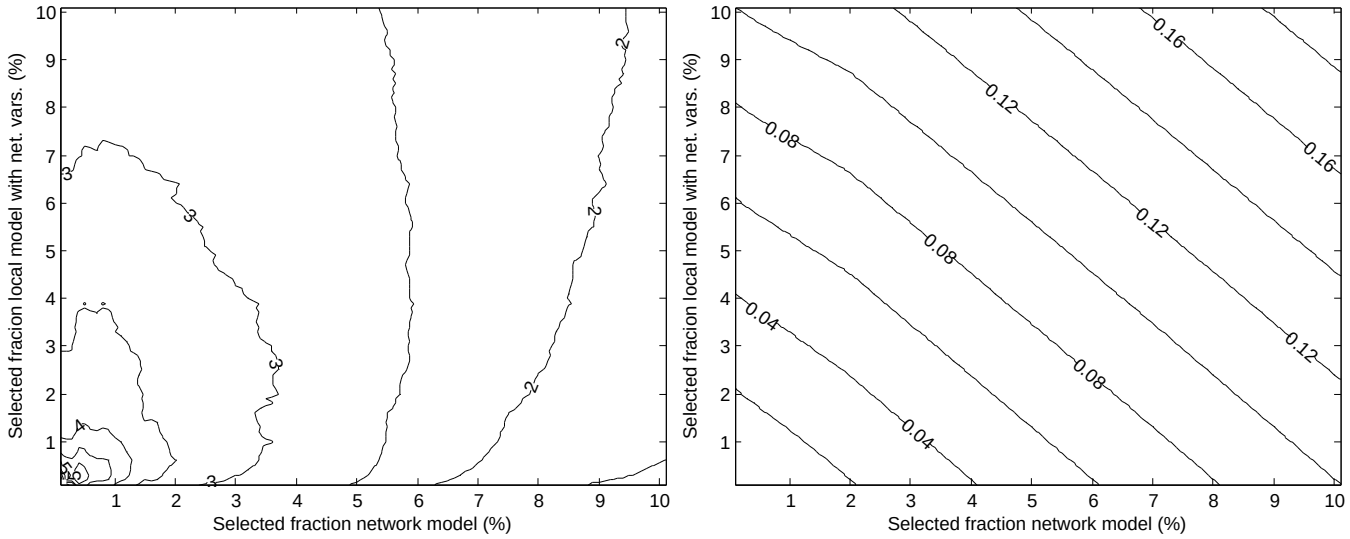


Figure 1: The fraction of the churners detected by the different network models that are not detected by the local model as a function of the selected fraction of customers with the highest predicted probability to churn of both models.



(a) Contour plot of lift curve.

(b) Contour plot total fraction of included subscribers.

Figure 2: Contour plot of a two dimensional lift curve with based on the local model with propositionalized variables (LM prop) and the network model (NM) (left), and contour plot of the total fraction of subscribers as a function of the selected fraction of the local model and network model respectively.

4. Results

In this study, four methods of combining different types of classifiers, described in the previous section, have been investigated. The results of this study indicate that the fourth approach, i.e. a non-integrated setup, proves to be promising. The reason lies in the fact that the local model and the relational classifier identify *different* customers as likely churners, which is illustrated in Figure 1. As a result, when combining both models in a unified framework yielding one single ranking of customers, a lower performance is reached as compared to a parallel non-integrated approach.

Figure 2(a) shows the contour plot of a three dimensional lift curve resulting from a parallel setup, for the top fraction of subscribers selected by the local model with network variables on the y-axis and selected by one of the network models on the x-axis. Some overlap exists between the two top fractions of selected subscribers, and hence combining the top 5% of subscribers selected by both models yields in total a fraction smaller than 10% of the customer base. Figure 2(b) plots the total fraction of the customer base that is selected as a function of the fractions selected by the local and the network model.

The contour plots in Figure 2 allow to interpret the three dimensional lift curve. Starting from the y-axis, representing the lift curve of the local model with network variables, the iso-lift curves in Figure 2(a) roughly consist of a short horizontal part followed by a curve downwards and a vertical part. The iso-fraction curves in Figure 2(b) on the other hand are slanted, almost straight, lines. Hence, on the parts of the iso-lift curves with a slope that is less steep than the slope of the iso-fraction curves, the lift remains constant for an increasing size of the selected top-fraction. This in fact means that in these areas of the three dimensional lift curve the parallel model obtains the same lift as the local model for a larger fraction of subscribers. And thus, depending on the fractions that are selected from each model, the parallel model yields a higher lift than the local model.

References

Bonchi, F., Castillo, C., Gionis, A., Jaimes, A., 2011. Social network analysis and mining for business applications. ACM Transactions on Intelligent Systems and Technology 2.

Dasgupta, K., Singh, R., Viswanathan, B., Chakraborty, D., Mukherjea, S., Nanavati, 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, EDBT '08. pp. 697–711.

Džeroski, S., Lavrač, N., 2001. Relational Data Mining. Kluwer, Berlin, Germany.

Getoor, L., Taskar, B., 2007. Statistical Relational Learning. MIT Press, Cambridge, MA, USA.

Jensen, D., Neville, J., 2002. Linkage and autocorrelation cause feature selection bias in relational learning. In: Proceedings of the 19th International Conference on Machine Learning. pp. 259–266.

Kramer, S., Lavrač, N., Flach, P., 2001. Relational Data Mining. Kluwer, Berlin, Germany, Ch. Propositionalization approaches to relational data mining, pp. 262–286.

Macskassy, S., Provost, F., 2007. Classification in networked data. Journal of Machine Learning Research 8, 935–983.

Martens, D., Provost, F., 2011. Construction and inference of networked data in a bank setting. Working paper, to be submitted for publication.

Richter, Y., Yom-Tov, E., Slonim, N., 2010. Predicting customer churn in mobile networks through the analysis of social groups. In: Proceedings of the Tenth SIAM International Conference on Data Mining. pp. 732–741.

Verbeke, W., Dejaeger, K., Martens, D., Hur, J., Baesens, B., 2011. New insights into churn prediction in the telco sector: a profit driven data mining approach. Submitted for publication in European Journal of Operations Research.

Network neighbor effects on customer churn in cell phone networks

Pavel N. Krivitsky Pedro M. A. Ferreira Rahul Telang

iLab, H. John Heinz III College, Carnegie Mellon University

1 Introduction

In today's deregulated mobile communication markets, carriers experience churn rates of as high as 25% per year, and because attracting a new customer tends to be substantially more costly than retaining an existing one, being able to predict and head off churn is eminently important to a carrier's profitability. An important development in churn prediction in mobile communication networks has been to take into account the inherently social nature of these networks: the potential of those with whom one communicates to affect one's propensity to churn in both causal and predictive sense. In particular, churn has been modeled as a diffusion process over the network [2]; by partitioning subscribers into groups and predicting churn based on group structure and their most central members [7]; and using Markov Logic Networks to take into account the effect of network neighbors on churn [3]. Econometric models have previously been applied to model the actions of mobile phone network customers [5] and survival analysis has been used to model customer churn [6, 1].

In this paper, we provide preliminary evidence of contagious churn, that is, the propensity of a subscriber to churn increases after her neighbors do so. When the neighbors of a subscriber churn the subscriber's cost of service is likely to increase because calling outside the network is typically more expensive across all tariff plans, which might, per si, increase the tendency to churn. Yet, we show that when the neighbors of a subscriber churn the subscriber's propensity to churn increases beyond the impact of her changing social circle on costs, an effect we associate with influence and word of mouth across subscribers. To show this, we use survival analysis, with instrumental variables to alleviate endogeneity problems, over a large dataset obtained from a European carrier covering eleven months worth of detailed data on call and SMS records.

2 Data

Our dataset comprises records of all calls and all SMS messages between August 2008 and June 2009 involving all subscribers in a large European mobile carrier, which we will call OurNet. OurNet has roughly 4 million active subscribers with prepaid consumer grade plans during this period. For each call, the anonymized phone numbers of the caller and the callee, the day and time of the call, its duration in seconds, and the initial cell towers used by OurNet to service the call are available. For each SMS message, the anonymized phone numbers of the sender and the receiver and the day and time of when the SMS was sent are. In addition, although the numbers are anonymized, their country and area/carrier code are available. For most subscribers, postal code of their billing address has been recorded, and for some, age and gender as well. For all subscribers, we have the history of their tariff plans and add-on purchases going back to 2003. For the most used tariff plans we have information on the price of both the first minute and of subsequent minutes to both other OurNet subscribers and subscribers of other carriers. Add-ons include a number of products, such as caller ring back tones, free calls at night or during the weekend and packs of SMS.

Over the 334 days for which we have data, 3.2 billion calls were placed for a combined 13.1 thousand years of airtime. Roughly 14.8 billion SMS were observed during this time period. OurNet subscribers have initiated 2.5 billion calls, 1.9 billion of them to other OurNet subscribers. OurNet subscribers have sent approximately 14.2 billion SMS, 13.6 billion to other OurNet subscribers. The disparity between on-net and

off-net traffic is clear and staggering. If subscribers are called "neighbors" if they hold a call or exchange an SMS within the same calendar month, then, based on the same sample of 100,000 prepaid subscribers, an OurNet subscriber has, on average, 15.23 neighbors. If two subscribers are only considered neighbors if they exchange at least one SMS in each direction within the same calendar month, then this number reduces to 3.99; and if they are neighbors only if they call each other at least 3 (5) times within the same calendar month, then this number reduces to 3.41 (2.11). These later definitions of neighbors capture better the idea of a persistent relationship. The numbers of neighbors within OurNet and outside are statistically different, the former being always higher.

The only indication that a prepaid subscriber had churned is prolonged absence of activity. OutNet considers a subscriber churned when he or she is inactive (not placing calls or sending SMS) for at least 3 months in a row, and we follow this definition in our study (the number of prepaid subscribers who reactivate after 3 months of inactivity is negligible. Churn is significant and appears to increase over time for OurNet. Between September 2008 and April 2009, about 17% of OurNet subscribers left the network. The monthly probability of churn for a subscriber that spends more time talking to other OurNet subscribers is 0.016. This increases to 0.021 for the subscribers that spend more time talking to subscribers in other networks.

3 Methodology

Although our data only contain voice and SMS records for an eleven-month period, they also contain pricing plan records for a much longer period. In particular, for each OurNet subscriber active in the period of interest, the initial subscription time is known, but the churn time is only known for those subscribers who had churned during the period of observation and is not known for those who did not. Our information about time from subscription to churn is thus right-censored. This makes survival analysis a natural approach to modeling churn. In our application, the predictors of interest, such as the fraction of neighbors who had churned, vary over time, and because there is likely to be inhomogeneity among the subscribers, we use a frailty model with both time-constant and time-varying covariates. [4, pp. 296–308] In short, the hazard of subscriber i churning t months after subscribing is modeled as

$$h(t, x_i, x_i^{t-1}, \alpha, \beta) = z_i h(t) e^{x_i \cdot \alpha + x_i^{t-1} \cdot \beta}$$

where z_i are Gamma-distributed individual subscriber effects, x_i are the non-time-varying covariates for subscriber i, x_i^t are i's time-varying covariates t months after subscriber i's initial subscription, α and β are parameters. The main focus of the analysis in this paper is the effect of network neighborhood on the propensity to churn. Thus, our main predictors of interest include the fraction of airtime a subscriber spends speaking with subscribers outside OurNet, the fraction of SMS exchanged with subscribers outside OurNet, and the fraction of a subscriber's neighbors, weighted by airtime and number of SMS, who had churned in the previous month. To control for other factors likely to affect a subscriber's propensity to churn — such the overall level of usage — we also include tariff plan information, linear and quadratic effects of total airtime and of number of SMS sent and received, and a rough estimate of the total cost of service, derived from usage and the information available on tariff plans. Finally, note that those subscribers who have churned before the period of observation are, effectively, left-truncated. [4, p. 228] This does not preclude our analysis, but it does make it more sensitive to the proportionality of hazard assumptions.

Yet, endogeneity concerns may lead to inconsistent estimates for α and β in the model above. Consider two neighbors, subscribers i and j that churn from OurNet. We do not know who influenced whom to churn and the fact that subscriber j might have churned before subscriber i is just weak evidence that it was him who induced subscriber i to churn. In other words, we need to instrument subscriber j's decision to churn. We postulate that subscriber j churned because she was influenced to do so by her neighbors, call them subscribers k. Therefore, we relate the decision of subscribers k to churn to subscriber j's decision to do so, but we would like to avoid the former to have an effect on subscriber i's decision to churn. To this end, we use only those subscribers k that are not neighbors of subscriber i. Clearly, subscribers k and subscriber i can still influence each other through common neighbors other than subscriber j, but these would certainly be second other effects (which we intent to control for in future versions of our work) with less impact.

4 Results and Discussion

The table below reports preliminary results from our 2SLS estimation. Columns 1 and 2 show the first stage results. The instruments work as expected and are highly statistically significant. An increase in the number of the neighbor's neighbors (subscribers k) who churn, weighted by either airtime or SMS, leads to more churn of the neighbors (subscribers j), as indicated by the first two rows in these columns. Column 3 in the same table shows the results of the second stage. An increase in the number of neighbors who churn (subscribers j)in the previous month, weighted by airtime, leads to an increase in the propensity to churn, as indicated by the last two rows in column 3. This is highly statistically significant. Likewise for weighting by number of SMS, though in this case only at the 5% level.

Subscribers without airtime or SMS exhibit a higher propensity to churn. More airtime or SMS to outside OurNet increase the propensity to churn, though the former is not statistically significant. The cost of service is not statistically significant in our regression, maybe due to multi-colinearity with usage or just because of inaccuracy in estimating costs. The linear and quadratic effects on airtime and number of SMS show that, on average, subscribers involved in more calls tend to churn less, whereas subscribers involved in more SMS tend to churn more. This difference is worth analyzing further, namely the extent it might be confounding with age given that younger subscribers use more SMS and are more price sensitive thus more likely to churn. Finally, having a pricing plan without a mandatory monthly contribution to the prepaid account significantly increases the tendency to churn. In these regressions, we added monthly dummies to control for seasonal effects and in the first stage we used a polynomial on the time with the company to capture the effect of the baseline churning hazard.

In sum, in this paper we observe evidence of contagious churn, that is, when a subscriber leaves OurNet her neighbors exhibit an increased tendency to churn too. We show that this is the case after controlling for the fraction of airtime and SMS with subscribers outside OurNet, so our result holds beyond the fact that when a subscriber's neighbors churn the percentage of communication to outside the network also changes, which already affects, per si, the subscriber's decision to churn. Leveraging our 2SLS approach, which is based on the latent structure of the social network across subscribers, we associate this effect to influence and word of mouth across users. We are currently working on a structural model whereby subscribers choose networks from time to time to maximize their utility and conditional on carrier choice they then call and send messages to other subscribers. We hope to integrate soon this structural model with the reduced form estimations presented in this paper into a unified coherent framework able to explain the neighboring effects on churn over cell phone networks.

		1st Stage Airtime	1st Stage SMS		2nd Stage Prob Churn
% airtime-weighted neighbors' neighbors churned		0.2220***	0.0750***		
		(113.52)	(41.05)		
% SMS-weighted neighbors' neighbors churned		0.0648***	0.1710***		
		(29.73)	(83.91)		
% airtime outside OurNet		0.0025***	0.0028***		0.0312
		(8.37)	(10.16)		(1.10)
% SMS outside OurNet		0.0039***	-0.0032***		0.1204***
		(13.18)	(-11.50)		(4.07)
Airtime (linear)		-0.4855***	-0.1922***		-81.4432***
		(-6.89)	(-2.92)		(-7.38)
Airtime (quadratic)		0.6190***	0.4343***		32.9145***
		(10.04)	(7.54)		(11.26)
SMS (linear)		0.4403***	0.0468		21.9607***
		(6.55)	(0.75)		(2.58)
SMS (quadratic)		-0.3746***	-0.0970		13.2095**
		(-5.84)	(-1.62)		(2.19)
No airtime		-0.0072***	0.0028***		0.6025***
		(-23.39)	(10.16)		(25.40)
No SMS		0.0021***	-0.0080		0.6609***
		(8.22)	(-33.00)		(27.40)
Costs		0.0000***	0.0000***		0
		(8.68)	-7.03		(0.01)
No mandatory topup		0.0021***	0.0026***		0.6439
		(10.83)	(14.55)		(28.82)
Intercept		0.0044***	0.0040***		
		(16.88)	(16.64)		
Poly months subscribed		yes	yes		
Monthly dummies		yes	yes		yes
% airtime-weighted neighbors churned					3.974***
					(4.98)
% SMS-weighted neighbors churned					2.1047**
					(1.92)

Results from Survival Analysis with IV. Significance Levels: *** 1% ** 5%

References

[1] R. N. Bolton. A dynamic model of the duration of the customer's relationship with a continuous service provider: The role of satisfaction. *Marketing Science*, 17(1):45–65, 1998.

[2] K. Dasgupta, R. Singh, B. Viswanathan, D. Chakraborty, S. Mukherjea, A. A. Nanavati, and A. Joshi. Social ties and their relevance to churn in mobile telecom networks. In *EDBT '08: Proceedings of the 11th international conference on Extending database technology*, pages 668–677, New York, NY, USA, 2008. ACM.

[3] T. Dierkes, M. Bichler, and R. Krishnan. Estimating the effect of word of mouth on churn and cross-buying in the mobile phone market with markov logic networks. *Decision Support Systems*, In Press, Corrected Proof, 2011.

[4] D. W. Hosmer, S. Lemeshow, and S. May. *Applied Survival Analysis: Regression Modeling of Time-to-Event Data*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., 2nd edition, 2008.

[5] Y. Kim, R. Telang, W. B. Vogt, and R. Krishnan. An empirical analysis of mobile voice service and sms: A structural model. *Management Science*, 56:234–252, Feb. 2010.

[6] B. Larivière and D. Van den Poel. Investigating the role of product features in preventing customer churn, by using survival analysis and choice modeling: The case of financial services. *Expert Systems with Applications*, 27(2):277–285, Aug. 2004.

[7] Y. Richter, E. Yom-Tov, and N. Slonim. Predicting customer churn in mobile networks through analysis of social groups. In *Proceedings of the 2010 SIAM International Conference on Data Mining (SDM 2010)*, 2010.

Subscriber Behaviour in a Cellular Network implementing Dynamic Pricing for Voice Calls

Han Wang, Liam Kilmartin

Electrical & Electronic Engineering,
National University of Ireland, Galway, Ireland
h.wang1@nuigalway.ie, liam.kilmartin@nuigalway.ie

***Abstract*— In this paper we present an analysis of a large Call Detail Record (CDR) data set gathered from a cellular network implementing a novel Dynamic Pricing Service (DPS). In the DPS, the tariff applied to voice calls is discounted depending on time of day and cell location in a directed manner to typically increase or decrease subscriber calls for revenue or traffic load reasons. We have examined the CDRs to determine the subscriber calling and mobility characteristics with a view towards modelling this behaviour and in order to examine whether there is significant evidence of subscriber's attempt to exploit the discounting algorithm by means of a behaviour we have termed *discount chasing*.**

I. INTRODUCTION

Recent years have seen significant research interest in the analysis of records relating to the activity of subscribers of cellular networks as a means of providing insight into subscribers' daily behaviour, their mobility patterns and their social interactions. A typical mobile phone network will process several million users' phone calls every day and, with such high volumes of phone calls and other subscriber activity such as SMS, multimedia messaging and data usage, the daily datasets generated by these individual subscriber interactions with the network are extremely large. In this study we are specifically focussed on the analysis of Call Detail Records (CDRs) generated by a specific entity in a mobile network in an African country which is responsible for implementing an opt-in real-time *Dynamic Pricing Service* [1-3] for voice calls made by prepaid subscribers in that network.

In the Dynamic Pricing Service (DPS), subscribers who opt into the service are offered a discount on the tariff applied to each call. This discount varies continuously during the day and it is also dependent upon in which cell in the network the subscriber is currently located. All subscribers in a cell are aware of the discount currently on offer to users of the service by means of a Cell Broadcast service notification which is used to advertise the current discount rate in the cell. The system from which our CDRs are taken processes all call attempts made by subscribers using the service and confirms the discount being applied to each individual call by means of a USSD notification sent to the caller's handset during the call set-up. The rationale from a network operator's perspective of using such a real time DPS is to maximise cell utilisation or revenue by offering discount in order to encourage larger volumes of calls from the subscriber base. Hence, it is common, as in this case, that the discount on offer will be varied throughout the day in each individual cell based on current traffic load being experienced in that cell. Another significant motivator for the utilisation of a DPS is as a flexible admission control strategy which attempts to discourage users from accessing the network under heavy load conditions by offering little or no discount and encouraging them to utilise the network services during light load periods by offering significant discounts at those times.

The data set which has been analysed in this paper consists of a record of every call attempt made in the network by "opted in" prepaid subscribers on a given week day (i.e. Wednesday) over a period of several months during which time the DPS was active. On any given day there would typically be several million such call attempts in the data set which was analysed. There are several aims to this initial study:

1. To obtain a general understanding of the nature of the subscriber behaviour in the network during the study time period.
2. To analyse whether there is significantly different caller behaviour depending on the level of discount on offer to the subscriber.
3. To determine whether there is a significant level of "*discount chasing*" visible in the network. Anecdotal reports from the network operator

suggested the growth of "micro-businesses" which involved individuals moving from cell to cell during the day "chasing" the largest discount on offer and then selling on calls from their handsets to others in their physical vicinity at that time.

II. DATASET DESCRIPTION

The raw data (CDRs) analysed in this research were downloaded from the DPS platform for 19 weeks. The records for each Wednesday were downloaded starting from 28/04/2010 and ending on 22/09/2010. Typically the daily CDRs generated by the system represented approximately 6.5 million call attempts involving 2 million unique participants. A subset of the CDRs were analysed namely those which were call attempts to other prepaid subscribers of the same network (i.e. *on-net* calls). This typically consisted of CDRs representing an average of 3.5 million call attempts between approximately 800,000 unique participants each day. An example of the format of an annonymised CDRs is shown in Table 1.

Time Stamp	Caller ID	Called ID	Cell ID	Discount	Cell Utilization
734263	1	2	265	70	0.2365

Table 1: Example of individual CDRs

III. RESULTS

A. *Traffic Periods in Mobile Network*

Firstly, in order to get an overview of the general traffic patterns in the network, we examined the distribution of number of calls within 10 discrete discount ranges (i.e. 0-10%, 10-20%, etc.) over the day. Figure 1 is the call attempt intensity map for date 05/05/2010 where the number of calls was calculated during each 30 minutes time interval during the day.

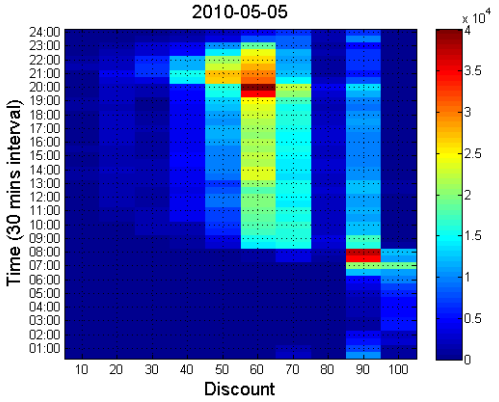


Figure 1: Example of the daily call attempt distribution

The same type of analysis was completed for the remaining 18 days in the dataset and the results show that the daily modal discount range did vary slightly each day. However, the periods of high traffic intensity were consistent and they happened in the morning between 7am and 8am and at night between 8pm and 9pm.

B. *Call and Mobility Distributions*

As has been reported in other studies [4-5] with such large dataset, the probability distribution function (PDF) of call attempts in this data set does appear to exhibit characteristics of a "power-law like" distribution. In addition, the mobility pattern of subscribers (which we have estimated by computing the probability that subscriber will make call attempts for multiple cell sites during the 24 hour period) also shows a similar "power-law like" distribution. Figure 2 shows the resultant distributions of call attempts and visited cells on a log-log scale.

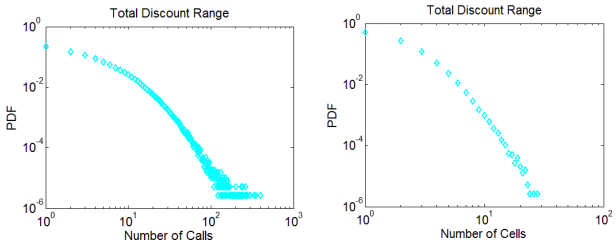


Figure 2: Distribution of call attempts (Left) and visited cells (Right) on a log-log scale

The CDRs analysed in this research come from a dynamic pricing mobile network which means that the subscriber who makes a phone call will be assigned a dynamic discount rate for each call, this discount being calculated in real time based on the cell utilization at that time. The inclusion of this "pricing" information gives us the opportunity to obtain the average discount for each subscriber and then to utilize this average discount to separate the total subscriber base into 10 sub-groups (based on the average discount which they obtained during the day). Figure 3 is an example plot showing the distributions of call attempts and visited cells for subscribers with an average discount of between 50% and 60%.

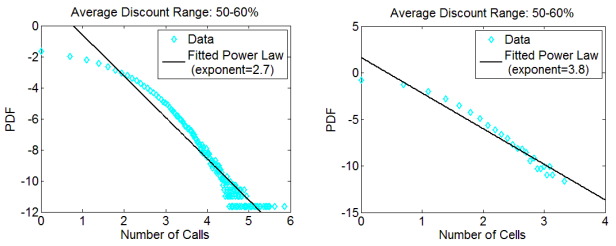


Figure 3: Power Law fit to the distribution of call attempts (Left) and visited cells (Right) in log-log scale (Discount Range: 50-60%)

As we have seen, the power law does not fit the distribution very well over its complete range. Therefore, a lognormal distribution was examined as a possible better fit for the above two distributions, as such an approach was also successfully applied in [6]. The optimal lognormal fit to the distribution of call attempts is shown in Figure 4 (Left). This appears to offer a substantially better fit across the complete range compared to the power-law fit. Similarly, Figure 4 (Right) is the lognormal fit applied to the distribution of number of visited cells.

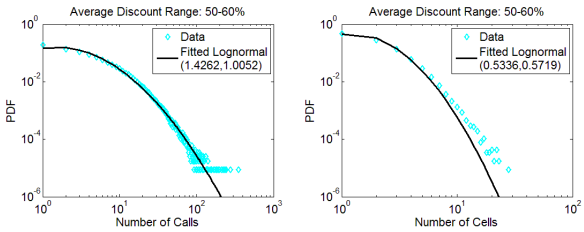


Figure 4: Lognormal fit to the distribution of call attempts (Left) and visited cells (Right) in log-log scale (Discount Range: 50-60%)

With the new fitting method, the fitted lognormal parameters, namely mean and standard deviation are examined based on different average discount range across all 19 days. Figure 5 illustrates the variation of these parameters distributions across the discount range as determined by estimating these two parameters for each discount range on each of the 19 days under investigation. This analysis does show that there appears to be a relatively well behaved and predictable variation in the lognormal mean and standard deviation versus the average dynamic pricing discount.

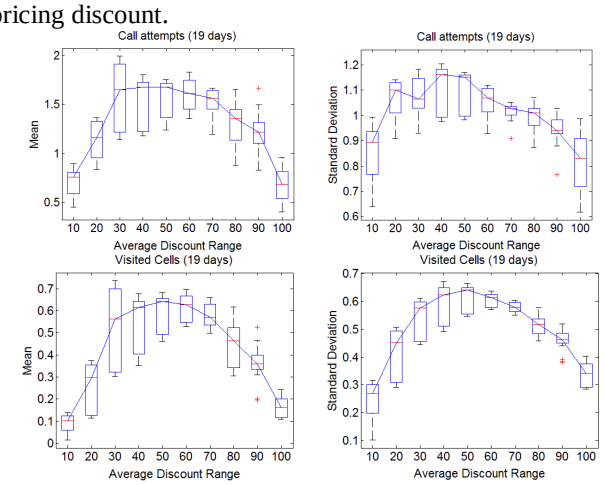


Figure 5: Variation of mean and standard deviation of log-normal fit for call attempts and visited cells distribution

In order to investigate the prevalence of *discount chasing* in the network, a subscriber mobility analysis was undertaken. In this analysis, the joint calling and mobility pattern was examined (i.e. an investigation of whether highly mobile subscribers make more or fewer calls than static subscribers). Figure 6 shows a sample density contour map which reflects the number of subscribers making X calls from Y cells (locations) during one specific day.

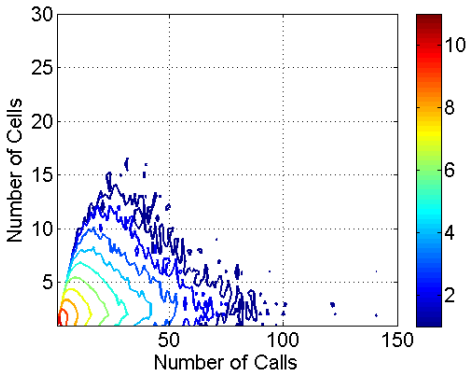


Figure 6: Joint calling and mobility pattern density contour map

Figure 6 illustrates that most of the subscribers visited fewer than 6 cells and that they usually made fewer than 6 calls per day. The higher mobility subscribers tended to make more phone calls and it reached the peak at about 32 calls with 17 cells visited. Above that level of usage, the subscribers tended to be more static with for example the subscribers who made 140 calls per day typically only visiting fewer than 6 cells. This is not surprising since highly mobile subscribers tend to have occupations requiring contacts with multiple individuals every day (e.g. business people). Overall, on examining this behaviour over the 19 days under study, we found that this joint mobility\calling pattern analysis produced consistent results and within these there was little evidence to support any widespread *discount chasing* behaviour in significant numbers in the subscriber base of the service.

C. Subscriber Calling Behaviour

As the data we analysed consists of 19 different days of CDRs, this can be used to categorise the subscribers into 19 groups based on the number of days they appears in the total of 19 days. This measure provides a surrogate for the *regularity* with which a subscriber uses the network's voice services. Figure 7 (Left) shows the average discount that the subscribers obtained versus the number of days during which they made calls. The average discount obtained by subscribers decreases as the subscribers' *regularity of access* increases. In other words, subscribers who make phone calls every day appear to be less concerned with the discount offered by the operator. On a related note, figure 7 (Right) shows that the subscribers who regularly make phone calls (i.e. make one or more calls on a given day) also tend to make more calls per day, with a very distinctive change in behaviour in both plots once subscribers access the network on 10 or more of the 19

days (i.e. subscribers who access the network on more than ≈53% of days).

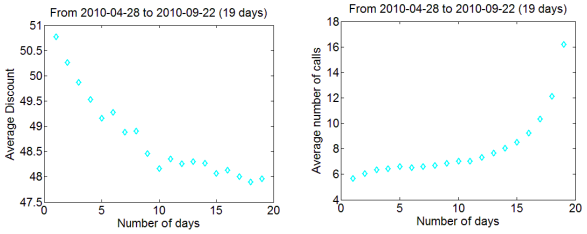


Figure 7: Distribution of average discount versus access regularity (Left) and distribution of average number of calls versus access regularity (Right)

IV. CONCLUSIONS AND FUTURE WORK

In this paper, we have provided a summary of, to our knowledge, a unique dataset containing call detail records for subscriber voice calls made in a network implementing real time dynamic pricing. We analysed the subscriber calling and mobility behaviours from a general view based on different discount ranges. The results appear to show that there is variation in subscriber behaviour based on the average discount which they obtain but that there does not appear to be any systematic abuse of the service through the behaviour of *discount chasing*. Further analysis of this data set is ongoing and focussed on determining parameters which would facilitate the development of a statistical model for subscriber behaviour in this environment. Also, we have recently commenced work on a graph theoretic analysis of the dataset with the aims of identifying a means of "individualising" the dynamic pricing strategy in a manner which will maximise its influence on the complete subscriber base behaviour.

REFERENCES

[1] Olivré, A., Call Admission Control and Dynamic Pricing in a GSM/GPRS Cellular Network, in Computer Science. 2004, University of Dublin.

[2] Khanifar, E.D.F.-N.a.A., Dynamic pricing in mobile communication systems, in First International Conference on 3G Mobile Communication Technologies. 2000. p. 416-420.

[3] Fitkov-Norris, E.D. and A. Khanifar. Dynamic pricing in cellular networks, a mobility model with a provider-oriented approach. in 3G Mobile Communication Technologies, 2001. Second International Conference on (Conf. Publ. No. 477). 2001.

[4] Gonzalez, M.C., C.A. Hidalgo, and A.-L. Barabasi, Understanding individual human mobility patterns. Nature, 2008. 453(7196): p. 779-782.

[5] Onnela, J.P., et al., Structure and tie strengths in mobile communication networks. Proceedings of the National Academy of Sciences, 2007. 104(18): p. 7332-7336.

[6] Seshadri, M., et al., Mobile call graphs: beyond power-law and lognormal distributions, in Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. 2008, ACM: Las Vegas, Nevada, USA. p. 596-604.

What is the Economic Value of Cell Phone Location Data?

Francois Baccelli (ENS/INRIA)
Francois.Baccelli@ens.fr
http://www.di.ens.fr/~baccelli/

Jean Bolot (Technicolor)
Jean.Bolot@technicolor.com
http://jeanbolot.com/

***Abstract*—The defining characteristic of mobile and cell phone users is how they change location over time. Information about location is becoming critical, and therefore valuable, for an increasingly larger number of location-based or location-aware services. One key open question, however, is how valuable exactly this information is.**

Our goal in this work is to develop an analytic framework, namely models and the techniques to solve them, to help quantify the economics of location information. We consider in particular the fundamental problem of quantifying the value of different granularities of location information, for example how much more valuable is it to know the GPS location of a mobile user compared to only knowing the access point, or the cell tower, that the user is associated with. We illustrate our approach by considering what is arguably the quintessential location-based service, namely proximity-based advertising or mobile coupons.

We make three main contributions. First, we find that determining the economic value of location is inseparable from determining the value of user preference or user profile data. The question then is not "what is the value of location data" but "what is the relative value of location vs preference data" or equivalently: what is more valuable - know where you are or to know what you like? We then develop several novel models, based on stochastic geometry, which capture the location-based economic activity of mobile users with diverse sets of preferences or interests and we derive closed-form analytic solutions for the economic value generated by those users. Third, we augment the models to consider uncertainty about the users' location, and derive expressions for the economic value generated with different granularities of location information.

To our knowledge, our work is the first one to present and analyze economic models which can help understand the economic value generated by mobile users with location based services, for different granularities of location information in wireless networks. A detailed description of the work is available in reference [1].

I. Introduction

The analysis of cell phone data is a very exciting and rapidly growing area of research because the data reveals interesting and previously hidden aspects of the behavior of cell phone users, and by extension of human behavior in general. Because it reveals personal and aggregate aspects of human behavior, the data is valuable to both the users (who would like to protect the more personal aspects of it) and to a plethora of service providers such as operators of location-based software service providers (who would like to present or advertise well-targeted or personalized information and offers to users). Therefore, two key questions related to cell phone data are those of **privacy** and **economics**, namely: how revealing is cell phone data and how can it be shared in a privacy-preserving manner on one hand, and how can it be used to personalize and improve user experiences and what is the value that it brings in the process on the other hand? We consider the second question in this work. Specifically, since location data enables new services and new economic activities, it is therefore naturally seen as economically valuable. This raises the fundamental question then of how valuable it is, and how to quantify that value.

Developing models to study and quantify the economic value of location data is a very challenging task, because the models need to capture a number of different factors such as i) spatially distributed users, ii) spatially distributed businesses and/or entities that trigger economic transactions, iii) transactions dependent on user location and, likely iv) transactions dependent on specific user preferences and interests. Furthermore, the models need to capture the granularity at which location information is available. We do not present in this work models that capture all the properties of mobile users involved in and responding to proximity-based advertising (for example we do not consider non-Poisson or non-homogeneous user distributions). Deriving and especially solving such a general model is still an extremely challenging problem. This is a first step, which presents what we believe is a useful and very promising approach to tackle and formalize the general problem.

We consider in this paper what is arguably the quintessential location-based service, namely proximity-based advertising: a mobile user, with a given set of preferences or interests, is offered an advertisement (in the form of an SMS coupon, or a discount of some kind) when getting close to a business or a store which offers products and services that match (at least partially) with the interests of the user - in practice, a hungry user who likes Italian food would get a coupon from a pizzeria for discounted pizza when getting within some range of that pizzeria. Mobile coupon usage is expected to triple by 2014, exceeding 300 million people; almost 3 billion coupons are expected to be issued by 2011 with coupon redemption value (the amount of discounts redeemed) expected to approach $7 billions globally [2]. An economic model of this situation helps answer the following questions

- What is the overall economic value generated by such transactions?
- What is the additional value created by fine-granularity location data (such as GPS data) compared to coarse-

granularity data (such as cell tower or access point-level data)?
- The location-aware advertisement situation described above typically involves three parties: 1) the mobile operator which provides or exposes the location data, 2) an entity which provides or exposes the interests or preferences of the user (for example Google, based on past search queries of the user), and 3) the source of the ad (the restaurant). Given that the user responds to the ad and walks into the restaurant to buy a pizza, how should the mobile operator, Google, and the restaurant (all of which provided component of the transaction) share the revenues created by the transaction?

II. Description of the model

We model the spatial distribution of businesses (such as the pizzeria above) which offer services that might be of interest to the mobile users (and which therefore might lead them to send mobile ads or coupons). We assume that there is a denumerable number of businesses or services which are randomly deployed in an infinite plane. We assume that services of type n (e.g. pizzeria versus movie theater vs coffee shop) are deployed according to a homogeneous Poisson point process Φ_n of intensity λ_n and that all these Poisson processes are independent (refer here and throughout the rest of the paper to [3] for background and details on stochastic geometry). Let us stress that the Poisson assumption is adopted for tractability of the analysis and that it is beyond the scope of the present paper to validate it in statistical terms. The homogeneity assumption should also be challenged as it is clear that urban densities vary from city centers to suburbian environments. There are two ways of addressing this last question. The simplest way consists in considering the model to represent a large but homogeneous area. The second way consists in extending the analysis to non-homogeneous Poisson point processes, which seems feasible. This last line of thought is left for future work.

Users are characterized by a random preference list which is a list of services, namely a subset of the integers. Users of class $(k, i_1, \ldots, i_k)$ have k elements in their list and these elements are the services $(i_1, \ldots, i_k) = \mathbf{i}$ they are interested in. The following notation is used:

- The probability to have a user of this class is denoted by $\pi(k, \mathbf{i})$.
- The radius of the ball defining the vicinity of a user is denoted by R.
- The spatial density of users is denoted by ν.

In the basic model described here, we adopt the following assumptions:

- The propensity for users to stop and check out the available services depends on the *total number of services* m matching their list and located in their vicinity; this is quantified by a function $f(m)$; it makes sens to assume that f is non-decreasing.
- Given that a user stops, the revenue generated is proportional to the *number of different services* located in his/her vicinity.

The f function is central in the model in that it captures the essence of the physical/psychological process in action: because mobile users are localized (by the wireless network operator), and because their interests are known (thanks to Google), they can be informed on how well their current location matches their interest through the variable m (we have also developed more elaborate scenarios based on ratings) and this triggers (through some psychological process) a decision of the user to check the services with the probability $f(m)$ and eventually to purchase.

Now that the bases for the model are in place, we can examine how to quantify the economic value generated by the users responding to the location-aware advertisements in this case.

Pick a typical user. Given this user is of the $(k, \mathbf{i})$ type, what matters for his/her is the Poisson point process of intensity $\lambda(k, \mathbf{i}) = \sum_{j=1}^{k} \lambda_{i_j}$, which is that of services present in his/her preference list.

It is easy to see that in case of a perfect localization, a user of type $(k, \mathbf{i})$ could be sent a proactive message informing him/her of the presence of m services matching his/her preferences iff his/her location belongs to the *m-coverage region* of the Boolean, or germ-grain model with germs (the points of the Poisson point process) of intensity $\lambda(k, \mathbf{i})$ and with grains equal to balls of radius R, centered on these points. By definition, a location belongs to this m-coverage region if exactly m balls cover it.

Because the number of grains that cover a given point follows a Poisson law on the integers of parameter $\lambda(k, \mathbf{i})\pi R^2$, the probability that a typical location is m-covered is

$$p(m, k, \mathbf{i}) = e^{-\lambda(k,\mathbf{i})\pi R^2} \frac{\left(\lambda(k, \mathbf{i})\pi R^2\right)^m}{m!}.$$

Hence the *mean revenue generated per unit of space* under Model 1 is

$$\rho = \nu \sum_k \sum_{\mathbf{i}} \pi(k, \mathbf{i}) \sum_m f(m) g(m, k, \mathbf{i}) \, e^{-\lambda(k,\mathbf{i})\pi R^2} \frac{\left(\lambda(k, \mathbf{i})\pi R^2\right)^m}{m!} \quad (1)$$

where $g(m, k, \mathbf{i})$ denotes the mean number of different services among the m for a user of type $(k, \mathbf{i})$. Using the fact that the probability that there is no service of type p among the m is $\left(1 - \lambda_{i_p}/\lambda(k, \mathbf{i})\right)^m$, with $\lambda(k, \mathbf{i}) = \sum_{q=1}^{k} \lambda_{i_q}$, we get that

$$g(m, k, \mathbf{i}) = k - \sum_{p=1}^{k} \left(1 - \frac{\lambda_{i_p}}{\lambda(k, \mathbf{i})}\right)^m$$

To get (1), we used: (i) the assumption that the probability of stopping given an m-match depends only on m through a function that we denote by $f(m)$; (ii) the assumption that the mean revenue given an m-match is proportional to the number of different services among the m (we take the mean revenue

per service given that this service is visited by the user equal to 1).

In the special case when:

- There are N types of services.
- $f(m) = 1 - \alpha^m$ with $0 < \alpha < 1$, where the constant α^{-1} is the propensity to react to proactive information.
- $\pi(k, \mathbf{i}) = \beta^k(1-\beta)\binom{N}{k}^{-1}$, i.e. we have a geometric size for the list of preferences and a uniform law on the services; the parameter β determines the mean size of the list of preferences as provided by e.g. Google.
- $\lambda_n = \lambda$ for all n; λ is the spatial density of services.

then $g(m, k, \mathbf{i}) = k\left(1 - \left(1 - \frac{1}{k}\right)^m\right)$, so that

$$\rho = \nu \sum_k \beta^k (1-\beta) \sum_m (1-\alpha^m)\left(k - k(1-1/k)^m\right) e^{-\lambda k \pi R^2} \frac{\left(\lambda k \pi R^2\right)^m}{m!}$$

or

$$\rho \approx \nu\beta \left(\frac{1-e^{-\lambda\pi R^2}}{1-\beta} - \frac{(1-\beta)\left(e^{-\lambda(1-\alpha)\pi R^2} - e^{-\lambda\pi R^2}\right)}{\left(1-\beta e^{-\lambda(1-\alpha)\pi R^2}\right)^2} \right)$$

We also compute the mean revenue per unit space $\rho'$when the user location is not known accurately (specifically when the user is mistakenly located at distance r from its true location) and the mean revenue per unit space ρ_0 when neither the user location nor its preference lists are known accurately.

Comparing ρ_0 to ρ' lets us quantify the value brought by a knowledge of the user preference list, and comparing ρ' to ρ lets us quantify the value brought by a knowledge of the user location. With closed form expressions for ρ, ρ' and ρ_0, we are in a position to answer the question we raised at the beginning of the paper, namely: "What is more valuable: to know where you are or to know what you like?"

The answer is shown in Figure 1 below. We plot the mean revenues per unit space ρ (in red), ρ' (in green) and ρ_0 (in yellow) for comparison for three different values of λ, i.e. for three different densities of services - low (left graph) corresponding to a rural or sparsely service environment, medium (middle graph) and high (right graph) corresponding to a dense urban core or a locally densely service area.

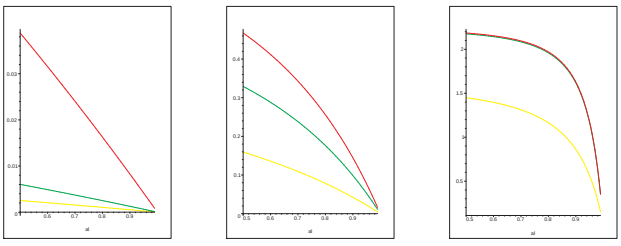


Fig. 1. Mean revenues per unit of space for three different values of $\lambda = .01, .1$ and 1

Not surprisingly, ρ is always higher than ρ' or ρ_0, meaning that knowing both the location and the preferences of users creates the largest revenue per unit space, primarily by providing the best *positive correlation* between the users' decision of stopping and their interests.

We also observe that the impact of the parameter α, the propensities to stop, is moderate.

Finally, and most interestingly, we observe that accurate information about location is the source of most of the revenue in rural or sparsely serviced areas (green curve is almost similar to yellow curve meaning profile information does not add much value, compared to red curve meaning location adds most of the value), but that profile or preference data adds most of the value in densely serviced areas or urban core. Therefore the answer to the question above is: **It is better to know where you are in rural areas, but better to know what you like in urban areas.**

Of course, the last figures do not have universal validity and only hold for the chosen set of parameters. However, they suggest a methodology to answer the question of the sharing of the revenues created by the transactions between the mobile operator and e.g. Google: it makes sense to propose that this revenue be shared proportionally to the relative gains brought by exact localization and by the list information respectively, which in turn means that revenues should preferentially go to the provider of user profiles rather than location in urban environments, and vice versa in rural environments.

III. Conclusion and Future Work

We have presented and analyzed a novel model allowing one to capture how to jointly leverage the combination of three types of basic information: 1) that of the geographic location and mobility of users of a cellular phone networ,k 2) that of their needs and interests as obtained from their preference lists and 3) that of the services available at all locations of space and their rating by the users. We have shown how the model could be used to quantify the revenue of this combination of informations and thus to quantify the economic value of cell phone location and user preference data.

The model does not capture all the properties of "real-life" mobile users involved in and responding to proximity-based advertising. Deriving and especially solving such a general model is clearly a challenging area for future research, but also an extremely important area since much of the activity and interest around cell phone data mining is supported or motivated by economic considerations, ranging from marketing opportunities (as in mobile proximity advertising) to societal improvements (as in location-based traffic optimization for example).

References

[1] F. Baccelli and J. Bolot, "Modeling the economic value of location and preference data for mobile users," in *Proc. IEEE Infocom'11*, Shanghai, China, April 2011.

[2] H. Wilcox, *Mobile Coupons and NFC Smart Posters: Strategies, Applications and Forecasts 2009-2014*. Juniper Research, Dec. 2009.

[3] D. Stoyan, W. Kendall, and J. Mecke, *Stochastic Geometry and its Applications*. Wiley, 1995.

MIT Campus Map

Welcome to MIT

All MIT buildings are designated by numbers. Under this numbering system, a single room number serves to completely identify any location on the campus. In a typical room number, such as 7-121, the figure(s) preceding the hyphen gives the building number, the first number following the hyphen, the floor, and the last two numbers, the room.

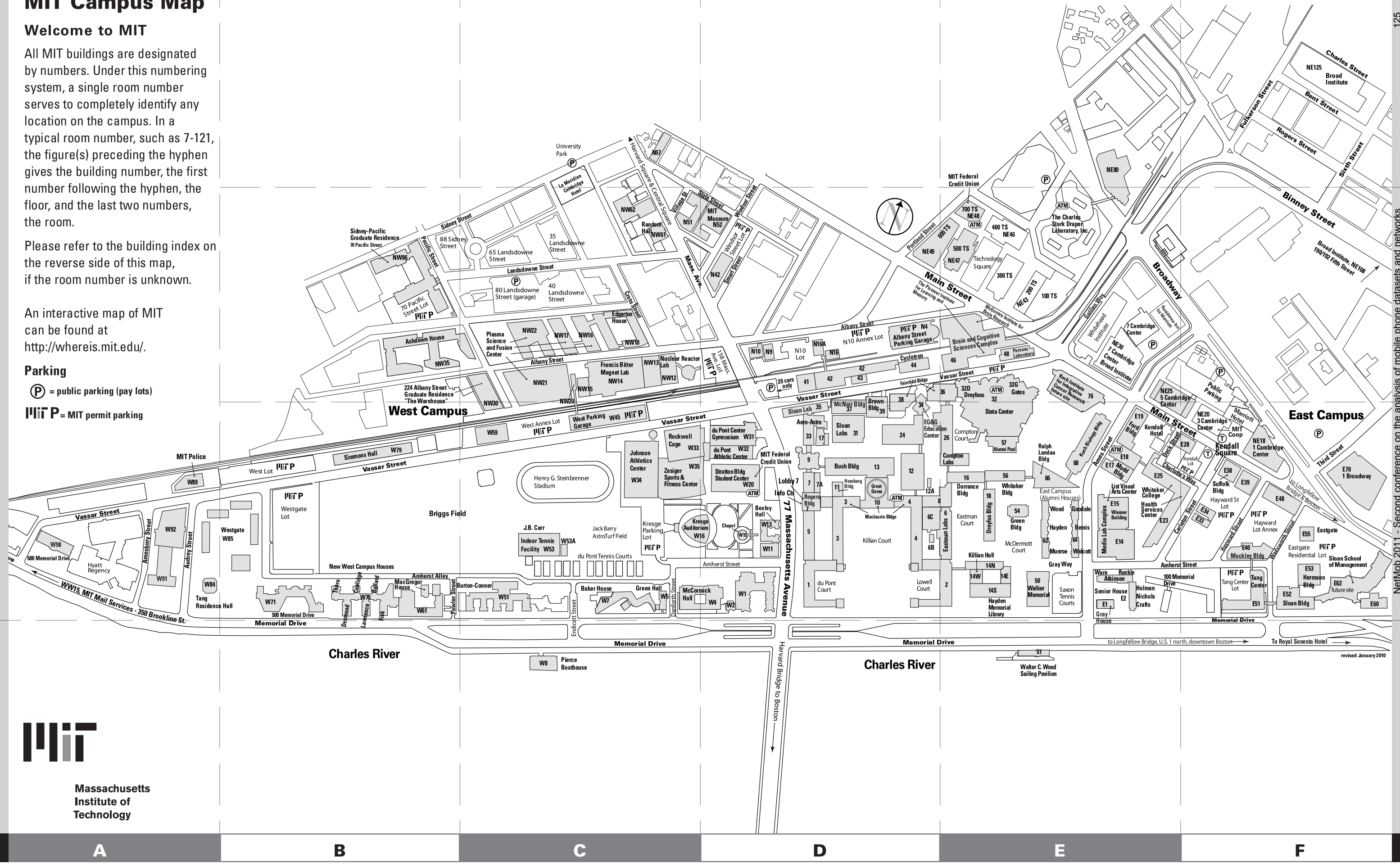
Please refer to the building index on the reverse side of this map, if the room number is unknown.

An interactive map of MIT can be found at http://whereis.mit.edu/.

Parking

Ⓟ = public parking (pay lots)

MIT P = MIT permit parking



revised January 2010

Massachusetts Institute of Technology

NetMob 2011 - Second conference on the analysis of mobile phone datasets and networks

MIT campus map index (building number follows name)

Visitor information

As parking is limited and regulations are aggressively enforced, visitors are strongly encouraged to use public transportation.

By public transportation, MBTA ("The T") Ⓣ

A parking garage at Alewife Station (Routes 2 and 16 at the Cambridge/Belmont/Arlington line) allows access to the Red Line subway. Take the Red Line from Alewife Station in Cambridge or from Park Street Station in Boston to Kendall/MIT Station or to the Central Square Station, both of which are within walking distance of MIT. There is regular MBTA bus (#1) service along Massachusetts Avenue from MIT to Boston and Harvard Square.

From Logan Airport by MBTA Ⓣ

Taxi fare is about $20–$30. To travel by subway, at Logan Airport take the airport shuttle bus that runs to the Blue Line Subway Station. Take the Blue Line to the Government Center Station and go upstairs to board the Green Line Trolley to Park Street. At Park Street Station go downstairs to board the Red Line Subway going outbound towards Alewife and exit at Kendall/MIT or Central Square Station.

By car from Logan Airport

Leaving the airport follow the signs to the Summer Tunnel. Entering the tunnel keep in the right lane for Storrow Drive North. Continue in the right lane following the Storrow Drive West signs. Exiting the tunnel you will see the Massachusetts Eye and Ear Infirmary on your left and the Charles River on your right. Keep to the right and continue to follow the Storrow Drive West signs. After passing under the Arthur Fiedler pedestrian walk bridge, change to the left lane and take exit for Massachusetts Avenue/Cambridge (2AN). Bear right and cross the Harvard Bridge. MIT's main entrance at 77 Massachusetts Avenue will be on your right, at the third set of traffic lights.

By car from Route I-93 or Southeast Expressway

Take exit 26. Keeping in the right lane, follow the Storrow Drive West signs. After passing under the Arthur Fiedler pedestrian walk bridge, change to the left lane and take exit for Massachusetts Avenue/Cambridge (2AN). Bear right and cross the Harvard Bridge. MIT's main entrance at 77 Massachusetts Avenue will be on your right, at the third set of traffic lights.

By car from Route I-90

Take the Cambridge/Allston Exit (18) on the left off of the Massachusetts Turnpike (Interstate 90). Bear right at the end of the ramp and go through two sets of traffic lights. Follow the signs to Cambridge. Cross the River Street Bridge and continue straight ahead for about one mile to Central Square. Turn right onto Massachusetts Avenue. MIT's main entrance, 77 Massachusetts Avenue, is about one-half mile down on the left.

By car from Route I-95 low the I-93 directions to MIT as given. From the south take I-90 off of I-95 and follow the I-90 directions to MIT as given.

Call a taxi

Yellow Cab; 617-492-0500

Parking

Ⓟ = public parking (pay lots) MIT P = MIT permit parking

Campus telephones

MIT house telephones are located in many of the campus lobbies including the Student Center (map section D). To reach an office extension from a house telephone dial the last 5 digits (i.e. 3-4795) of the number. There are pay telephones in the Student Center (map section D), as well as other locations around the campus. To contact a person, department, or residence at MIT, dial 0 from a house phone or 617-253-1000 from a pay telephone.

Campus tours

Tours of campus: 11:00 am and 3:00 pm weekdays except holidays. Tours leave from 77 Massachusetts Avenue, Lobby 7 (map section D).

Admissions Office undergraduate information sessions: 10:00 am and 2:00 pm (preceding tour) from mid-March through mid-December except holidays. Confirm at http://web.mit.edu/admissions/.

Dining on campus

Snacks and meals are available in the Student Center (map section D) and in the Stata Center (map section E). There are restaurants and small eating places in the Kendall Square area of the campus and in the local hotels adjacent to the campus.

The MIT Press

One of the country's largest university presses, the MIT Press publishes books and journals circulated throughout the world. Its titles include professional, reference, and scholarly books; graduate and undergraduate texts; and books for general audiences. The MIT Press Bookstore is located at 292 Main Street (map section F).

MIT events and exhibits

The **MIT Events Calendar** is available online at **http://events.mit.edu**.

A map giving locations of the **public art** in MIT's Permanent Collection, overseen by the List Visual Arts Center, may be found at http://web.mit.edu/lvac.

The following 24-hour numbers are available for recorded information on current arts events:

Concerts 617-253-9800

List Visual Arts Center 617-253-4680

MIT Museum 617-253-4444

Theater Arts 617-253-4720

Student Center facilities
W20 - 84 Massachusetts Avenue (map section D)

Bank, 1st floor
Cafeteria, 2nd floor
Campus Police/event registration detail, basement
Cleaners, basement
Conor Moran Lounge, 5th floor
Copy Technology Center, 1st floor
Food Market/convenience store, 1st floor
Game Room, 1st floor
Hair salons, basement
Manager, Campus Activities Complex, 5th floor
MIT Card Services, basement
Optical Store, basement
Parking and Transportation Office, basement
Police, Detail Office, W20-020B
Post Office (U.S.), basement
Restaurants, 1st and 2nd floors
Stratton Lounge, Catherine N., 2nd and 3rd floors
Tech Coop, 1st floor (no textbooks)
Wiesner Student Art Gallery, 2nd floor

For more information

Massachusetts Institute of Technology
Information Center
Room 7-121
Telephone 617-253-4795
http://web.mit.edu
77 Massachusetts Avenue
Cambridge, MA 02139-4307

Massachusetts Institute of Technology

January 2010

Table of contents