Third International Conference on the Analysis of Mobile Phone Datasets

UCL
Université
catholique
de Louvain

Massachusetts
Institute of
Technology

MIT Media Lab, Cambridge, MA

May 1–3, 2013

www.netmob.org

Book of abstracts

Editors
Vincent Blondel
Adeline Decuyper
Pierre Deville
Yves-Alexandre De Montjoye
Jameson Toole
Vincent Traag
Dashun Wang

Sponsored by
orange™
Real Impact Analytics

Contents

Session 5

Mobility patterns

Frequencies, temporal patterns, and spatial regularity of mobile-phone data

Philipp Hövel,[1, 2, 3, *] Filippo Simini,[1, 4, 5] Chaoming Song,[1, 6] and Albert-László Barabási[1, 6, 7]

[1] *Center for Complex Network Research, Northeastern University, Boston, USA*
[2] *Institut für Theoretische Physik, Technische Universität Berlin, Germany*
[3] *Bernstein Center for Computational Neuroscience, Humboldt-Universität zu Berlin, Germany*
[4] *Dipartimento di Fisica "G. Galilei", Università di Padova, Padova, Italy*
[5] *Institute of Physics, Budapest University of Technology and Economics, Budapest, Hungary*
[6] *Center for Cancer Systems Biology, Dana-Farber Cancer Institute, Boston, USA*
[7] *Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, USA*

In recent years more and more data have been generated contributing to the rise of network science. In particular, social networks can be studied in greater details as an increasing number of datasets become readily available in the digital age. The analysis of what is generally termed *big data* allows for the inference of human behavior such as mobility or statistics about the social environment. One type of datasets that is especially suited to serve as a proxy for human behavior arises from the usage of mobile-phones [1–5]. The mobile-phone dataset used in this contribution, for instance, consists of anonymized call data records (CDRs) of 10 million customers from a single phone company.

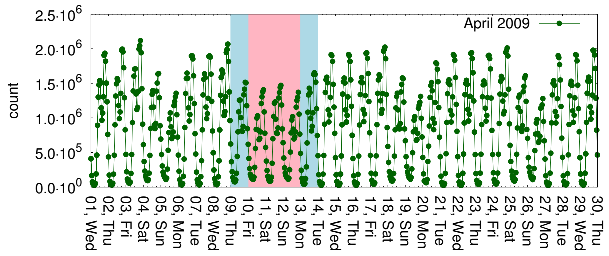


FIG. 1: Call activity during one month with an hourly resolution. The colored days mark Easter holidays.

We present a data analysis of CDRs with the purpose to extract various patterns of synchronization and to identify different rhythms of daily life. Our analysis spans multiple time periods, which allows to discover rhythms on a daily, weekly, and even longer time frames. For example, we study routines during weekdays and the deviations from these temporal patterns during weekend activities. Furthermore time-resolved evaluations of the total number of calls provides a simple insight into the heartbeat of the society as a whole. See Fig. 1, which shows a repeating call activity profile during weekdays interrupted by weekends or holidays (marked by color). Systematic sorting of users into pre-defined groups, that is, the use of additional meta data such as the age of the mobile-phone users, enables further investigation of the social behavior of different subpopulations.

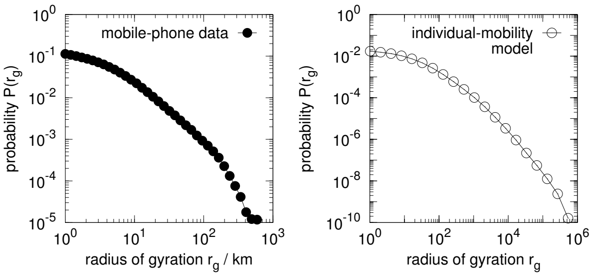


FIG. 2: Normalized distribution $P(r_g)$ of the radius of gyration r_g derived from the mobile-phone data (left, observation period: 1000 hours) and the individual-mobility model (right, arbitrary spatial units). Reproduction of Figs. 4(a) and (c) of Ref. [6].

Next to the social network (who calls whom), the dataset at hand also provides information about the individual mobile-phone user's whereabouts (where does a call originates from). The position of the caller is approximated by the location of the nearest mobile-phone tower that handles the call. We demonstrate that the study of this spatial layer of information offers an insightful perspective on human mobility. In particular, it will lead to scaling laws of human travel calculated, for instance, for the radius gyration [6]. The radius of gyration r_g for each user, who is recorded during L events at positions $\vec{r}_1, \ldots, \vec{r}_L$, is defined as

$$r_g = \sqrt{\frac{1}{L}\sum_{i=1}^{L}\left(\vec{r}_i - \vec{r}_{cm}\right)^2}$$

with the center of mass $\vec{r}_{cm} = L^{-1}\sum_{i=1}^{L}\vec{r}_i$. The distribution of the radius of gyration is depicted in Fig. 2.

In addition, we will consider limits of predictability [7]. For this, we use a measure of regularity given by the ratio R_i of the number of recordings at the most frequently visited tower to all calls,

$$R_i = \frac{\text{number of appearances at primary location}}{\text{total number of appearances}},$$

[*] Electronic address: `phoevel@physik.tu-berlin.de`

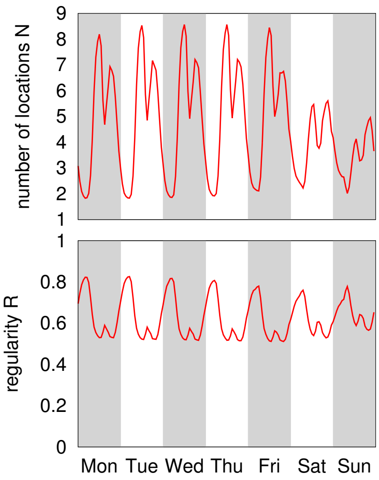


FIG. 3: Regularity of mobility: (bottom) ratio R to find a user at his/her most frequent location and (top) number of different locations N with an hourly resolution for each day of the week. Data: May 2009 for a class of users between 12 and 100 events on the daily average.

for each hour of the week, $i = 1, \ldots, 168$. Figure 3 depicts the averaged regularity R (bottom) and the average number of locations N (top). That figure confirms an intuitive expectation: The average person is most regular (largest R) with the least number of locations during early-morning hours and least regular (smallest R) with the largest number of locations during commuting times between home and workplace. Similar to Fig. 1, one can see that this effect repeats for each weekday, but is less pronounced on the weekend.

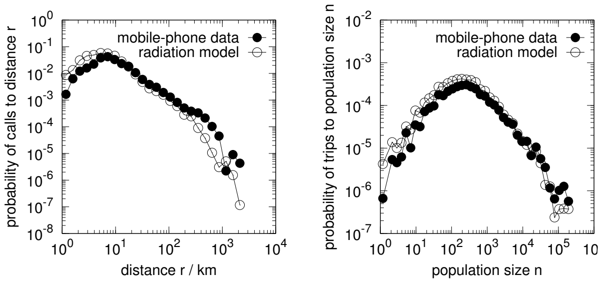


FIG. 4: (Left) probability distribution of a call over a distance r and (right) to a municipality with population size n. Data: number of phone calls between users living in different municipalities during a period of 4 weeks with a total number of 38,649,153 calls placed by 4,336,217 users. The data was aggregated to obtain the total number of calls between every pair of municipalities. Reproduction of Figs. 3(g) to (h) of Ref. [8].

Finally, we will briefly review simple models like the *individual-mobility model* [6] or the *radiation model* [8] that reproduce the empirical findings to large extent and provide mechanisms for the discovered scaling laws and scales of human travel. For a comparison of these models to the empirical findings see Figs. 2 and 4.

Acknowledgments

PH acknowledges support of the German Academic Exchange Service (DAAD) via a postdoctoral fellowship.

[1] M. C. Gonzalez and A.-L. Barabási: *From data to models*, Nature Physics **3**, 22 (2007).

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

[3] D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A.-L. Barabási: *Human mobility, social ties, and link prediction*, ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD) (2011).

[4] J. P. Bagrow, D. Wang, and A.-L. Barabási: *Collective response of human populations to large-scale emergencies*, PLoS ONE **6**, e17680 (2011).

[5] J. P. Bagrow and Y.-R. Lin: *Mesoscopic structure and social aspects of human mobility*, PLoS ONE **7**, e37676 (2012).

[6] C. Song, T. Koren, P. Wang, and A.-L. Barabási: *Modelling the scaling properties of human mobility*, Nature Physics **6**, 818 (2010).

[7] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási: *Limits of Predictability in Human Mobility*, Science **327**, 1018 (2010).

[8] F. Simini, M. C. Gonzalez, A. Maritan, and A.-L. Barabási: *A universal model for mobility and migration patterns*, Nature **484**, 96 (2012).

Discovering urban and country dynamics from mobile phone data with spatial correlation patterns

Roberto Trasarti[a], Ana-Maria Olteanu-Raimond[b], Mirco Nanni[a], Thomas Couronné[b], Barbara Furletti[a], Fosca Giannotti[a], Zbigniew Smoreda[b], Cezary Ziemlicki[b]

a. *KDD lab, Istituto di Scienza e Tecnologie dell'Informazione, CNR Pisa, Italy*
b. *Sociology and Economics of Networks and Services dept., Orange Labs, Paris, France*

Abstract

Based on data coming from mobile communication infrastructures, this paper proposes an analytical process aimed at extracting interconnections between different areas of the city that emerge from highly correlated temporal variations of population local densities. The proposed methods are experimented on two real scenarios of different spatial scale: the Paris Region and the whole France.

1 Introduction and objectives

In recent years massive mobile phone location data have been studied and shown to have great potential to model human mobility (González et al., 2008; Song et al., 2010). In particular, studies on long-term mobility data proved how this kind of data could be important for urban planning and transportation studies (Reades et al., 2007; Calabrese et al., 2011). The difficulty to conduct classical travel surveys using self-report records (diaries or questionnaires) as well as the growing need to collect longitudinal data have drawn attention to automatic mobile phone data collection systems (Wang et al., 2010). However, one of the main difficulties with mobile phone data is the incompleteness of the users' traces. In fact, people are localized only when they are using their phone (calling or sending SMS); it leads to several problems in finding mobility patterns by means of classical data mining algorithms. Recent experiences using mobile phone data have shown that they can provide a good understanding of how the density of population changes during the day in various regions of a given (urban or larger) area, the existing body of research appears to be mostly focused on the discovery of local phenomena, such as increases of population, or simple flows of population between pairs of regions. On the contrary, the dynamics of a city naturally create links of several different natures between regions of the city, sometimes even very far apart: some regions tend to get congested together as response to some external event (e.g. intense precipitations); others might be connected through a cause-effect chain, where the population of a region flows from one to the other in exceptional measure when the former exceeds some levels of saturation. In this paper, we propose a new pattern definition, *correlation pattern*, which tackles the problem of finding correlations between areas using the presence of cellphone users. The proposed algorithm and tools are integrated in an existing mobility data analysis platform: M-Atlas (Giannotti et al., 2011), thus allowing the analyst to take advantage of all the pre-existing features. Our approach is based on the observation that the density distribution of population tends to be regular (periodic) in almost all regions. That is the result of the sum of several routine human activities such as going to work, going to school, etc. which are constantly generated by the same residents, as well as more random activities due to tourism, use of services (e.g., shopping, leisure activities), generated by different people yet yielding overall stable densities. Therefore, finding evidence of connections between areas by simply looking at raw densities would essentially lead to link everything to everything. A more promising approach, therefore, consists in looking for exceptions to regular behaviors. Following this idea, we start by searching *events* that represent significant deviations from regular trends, and locate them in space and time. Then, we try to detect recurrent combinations of events, therefore extracting frequent *patterns* of deviations. The types of patterns we look for essentially have the form of sequential patterns and are defined as follows:

DEFINITION 1. ***Correlation patterns****. A correlation pattern (C-pattern) is a sequence* D = <$D_1,\dots,D_n$> *of sets of events, where* $D_i = \{d_1,\dots,d_m\}$ i*s a set of events, each defined as* $d_j = (s_j, w_j)$: s_j *is a spatial region and w_j is the weight associated to the event.*

A *C-pattern* describes a set of regions that often experience a (significant) deviation from their common behavior, and do that either at the same time (in the case the events belong to the same event set) or at different times (if they belong to different event sets) but always in the specific order described by the pattern.

2 Methodology

In order to support the detection of the events and patterns mentioned above, in this section we introduce methods and algorithms to achieve three basic objectives: first, detect the locations and times where relevant variations of population take place; second, infer from them more complex patterns that link regions where variations tend to appear together or in some constant sequence; finally, navigate the discovered patters along the spatial and temporal dimensions, and enrich patterns with additional information (derived from raw data) to help their interpretation.

Stop Detection. The main concept behind our work is the population presence in a spatial region. Such presence can be estimated through mobile phone data in various ways, mainly depending on whether the presence measure for a regior (within a time window) should include only users that stopped in the region or it should also count users that simply crossed it while moving towards a different destination. In particular, we propose a stop detection criterion inspired by Palma et al. (2008), where spatial positions and time intervals are used instead of the speed: a *stop* for user *uid* is any ordered pair (p_k, p_m) of mobile phone points such that their location is the same ((p_i.x, p_i.y)=(p_k.x, p_k.y)), between them there are no points in different locations, and the temporal distance between them is longer than a given minimum duration threshold.

Density estimation and events detection. The basic elements of correlation patterns are the single events, which represent all the relevant variations of population for a given region. In this work, in particular, events are computed by comparing the density of population within a region in a given moment against the expected density for that area at that hour of the day. After partitioning the space into *regions* and time into *time-slots* of appropriate size (given by the user), the input data is divided into a training and a test dataset. For both datasets the spatiotemporal grid of densities is computed. The first is used to compute the expected densities of a typical period for each region. The second dataset is then compared against such typical period in order to detect significant deviations. For instance, we might obtain an expected density for each pair *(region, hour of the day)*, i.e., 24 values for each region, assuming 24 one-hour time-slots. Then, for each region and each time-slot, the corresponding density is compared against its expected value: if the difference is significant (another parameter of the method), an event of form *(region, weight, time slot)* is produced, representing its spatiotemporal slot and a discretized measure (*weight*) of how strong was the deviation.

C-patterns extraction. The extraction of *C-patterns* focuses on those patterns that appear frequently, i.e. they occur with some given minimum frequency. In particular, *C-patterns* are computed as sequential patterns over the dataset of events obtained in the previous step. To do that, we employ a simple extension of the standard SPAM (Agrawal & Srikant, 1995) that integrates spatial and temporal constraints, as well as the extraction of maximal and closed patterns. Finally, only the *C-patterns* that show a high correlation are retained, the latter being evaluated by an ad hoc variant of the standard *lift* index defined for item sets, expressing the ratio between the actual frequency of the pattern and its expected frequency, computed under the assumption of complete independence between events.

Spatial and temporal navigation of C-patterns. Adopting a hierarchical clustering method we are able to reorganize the patterns w.r.t. their temporal or spatial distributions in a tree structure. The result is a dendogram where each leaf represents a pattern and the intermediate nodes represent groups of similar patterns in terms of their temporal distribution. An example is shown in Figure 1(center), as produced by the tool that we implemented to cluster the temporal distributions. The tool can also visualize the temporal distribution of each intermediate node, as shown in Fig.1(bottom). It represents the typical distribution in its sub-tree (bold red line) with additional information about the minimum and maximum values of the group in each time interval (pink shadow). Practically, the analyst can study the dendogram at different levels in order to interpret group of patterns by their common temporal distribution, e.g., *morning patterns* as the patterns having all the occurrences in the morning. Analogously to the temporal navigation, it is possible to organize the patterns using their spatial component.

3 Case studies

In this section, we briefly summarize some of the results obtained by applying the methodology to extract patterns from two different CDR datasets: at a city level, covering Paris, and at a national level, covering all the French territory.

Urban level: Paris. Figure 1 shows an example of results obtained, focused on the C-patterns that had the Charles De Gaulle Airport (CDG) as starting area. The lines connect the areas belonging to the same pattern. This result shows that all the major train stations are influenced by the airport, e.g., an increasing of +10% in the presences at CDG airport impacts on Gare de l'Est (train station) by a +10% in 2 hours at maximum. For better understanding of the phenomena, we analyze the temporal dimension of patterns (right part of the figure), since this chain of events may happen in different periods of a day. The temporal distribution shows that the patterns cover several parts of the day, suggesting a further drill-down navigation of the dendogram (possible through the GUI by a simple click on a node) to better separate patterns based on their temporal profile. On the right, a focus on two subtrees of the dendogram are shown, corresponding to the blue and green nodes. Beside a first differentiation on the temporal profile, the two subtrees correspond to significantly different spatial area.

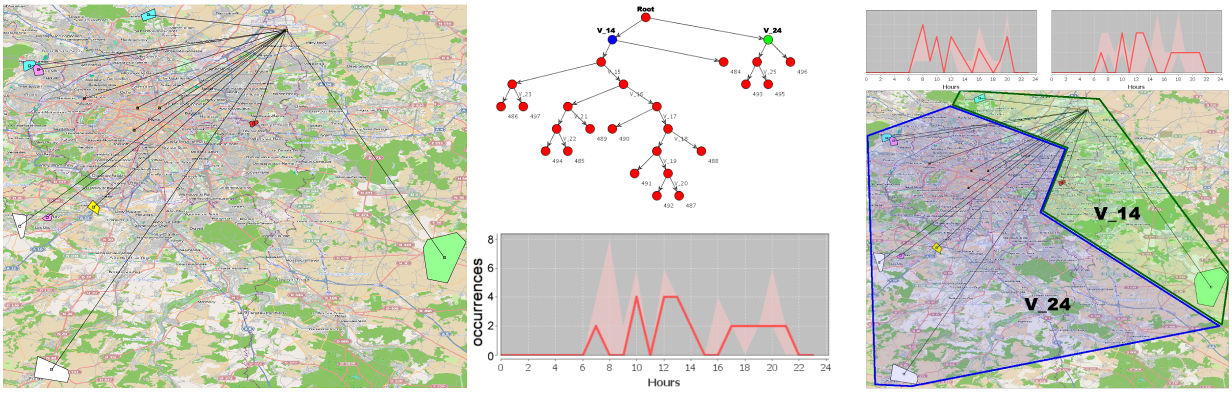


Figure 1: A selection of C-patterns obtained on the Parisian area, starting from CDG airport. On the left, a spatial representation of all patterns selected; on the center, the dendogram (top) and hourly frequencies (bottom) of the temporal distribution of their occurrences; on the right, the temporal and spatial distribution of two subtrees of the dendogram.

National level: France. A second analysis was performed at the national level, using *departments* as spatial units. The C-patterns obtained are shown in Figure 2(left), with a focus on the *Seine-Saint-Denis* department (center), and the temporal distributions of the corresponding patterns (right). More in detail, we can notice that these patterns can be divided into in- and out-coming ones: the groups of patterns *a,b,c,d,e* move from the neighborhood departments to *Seine-Saint-Denis*, while patterns *f,g,h* move from *Seine-Saint-Denis* outwards. Also, we can notice that the *incoming* patterns are more present during mid-day or early afternoon, while the *outcoming* one are more present during the evening, with exceptions in *e* and *f*, since they have a two-peeks distribution which most likely follows the systematic movements of the area.

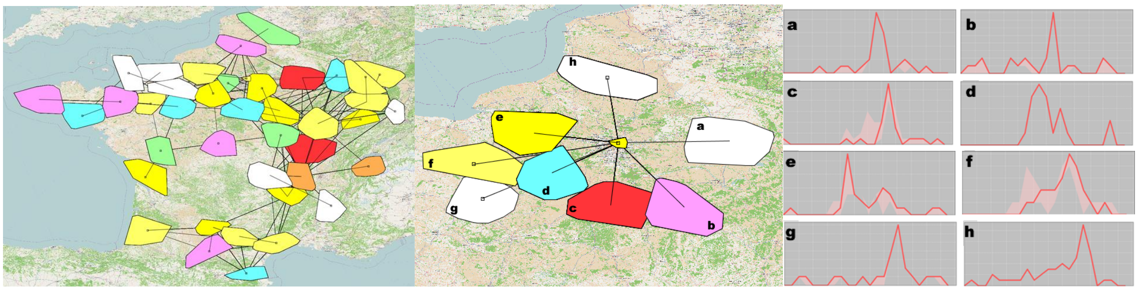


Figure 2: A set of patterns extracted on France at the level of departments (left), a selection of patterns focused on Seine-Saint-Denis and the neighborhood departments (center), and the temporal profiles of the patterns (right).

4 Conclusion

We presented a new kind of pattern called *C-pattern*, aiming to discover hidden logic of connections between regions of a city (or other kinds of areas, at different scales), by analyzing frequently co-occurring changes in population densities. We have developed an extraction process to discover these patterns, and tested it on real cases studies at two different granularities (urban vs. national), showing examples of results and their temporal and spatial navigation.

References

Agrawal, R. & Srikant, R. (1995). Mining sequential patterns. *Proc. 11th International Conference on Data Engineering,* Taipei, pp. 3-14.

Calabrese, F., Di Lorenzo, G., Liu, L., & Ratti, C. (2011). Estimating origin-destination flows using mobile phone location data. *Pervasive Computing*, 10, pp. 36-44.

Giannotti, F., Nanni, M., Pedreschi, D., Pinelli, F., Renso, C., Rinzivillo, S., & Trasarti R. (2011). Unveiling the complexity of human mobility by querying and mining massive trajectory data. *The VLDB Journal*, 20, pp. 695-719.

González, M., Hidalgo, C., & Barabási, A.L. (2008). Understanding individual human mobility patterns. *Nature,* 453(7196), pp. 779–782.

Palma, A. T., Bogorny, V., Kuijpers, B., & Alvares, L.O (2008). A clustering-based approach for discovering interesting places in trajectories. *ACMSAC,* ACM Press, pp. 863-868.

Reades, J., Calabrese, F., Sevtsuk, A., & Ratti, C. (2007). Cellular Census: Explorations in urban data collection. *IEEE Pervasive Computing,* 6, pp. 30-38.

Song, C., Qu, Z., Blumm, N., & Barabasi, A.L. (2010). Limits of predictability in human mobility. *Science,* 327(5968), pp. 1018–1021.

Wang, H., Calabrese, F., Di Lorenzo, G., & Ratti, C. (2010). Transportation mode inference from anonymized and aggregated mobile phone Call Detail Record. *13th International IEEE Annual Conference on Intelligent Transportation Systems*, Madeira Island, Sept. 2010.

Location Patterns of Mobile Users : A Large-Scale Study

Ashwin Sridharan
AT&T Labs,
asridharan@research.att.com

Jean Bolot
Technicolor,
bolot@technicolor.com

Mobile devices have become a common accessory for a large fraction of the population who carry and utilize it as they move. The very nature of the wireless service allows the network to record location of users on an almost continuous basis via active and passive means. This has provided a convenient source of location information for large populations, which in turn has spurred a large interest in the study human mobility from various different perspectives.

In this abstract, our perspective is one of characterizing the entire *footprint* or a *location pattern* of a user. Broadly speaking, we define a *location pattern* of a user to be the set of *all* locations (cell towers) touched by users when making calls over an observation period. Location patterns possess significant useful information and hence are interesting and useful to study. For example, information regarding the range covered by a user as well as characteristics of *how* a user moves within the range, are important inputs for modeling and sizing the paging 'zones' to optimize paging load. Knowledge about the size and shape of the footprint of the *ensemble* of users based within a certain locale, e.g a city, allows us to determine how far they range and in which direction, which major routes they take, all of which are useful inputs for urban and traffic planning. Last but not least, clearly such information can be utilized in building mobility models, e.g. to determine parameters related to the area covered by users, the directions they choose, length of major routes *etc.,*.

In this work, we develop a systematic methodology that utilizes geometric constructs to capture salient features of location patterns such as their size, shape, or internal structure. We apply these techniques to study the location patterns of several million users extracted from a large nation-wide data set comprising of several billion call data records, which record the location of users over a period of 1 month with a spatial granularity of a cell tower (meaning that the location of the user is identified by the cell tower the user is associated with). As explained in detail below, one of our main results is the discovery that **the distribution patterns of users in general, at both nationwide and city levels, follow a statistical distribution with a simple generative process, namely the *Double Pareto Log Normal* distribution**($DPLN$). Specifically, the $DPLN$ is an excellent fit to model the distribution of several features of a location pattern, including its size, shape and structure. In addition, we also show how these features can be used to discern and compare human footprints across locales.

We characterize the location pattern of a user from two aspects : a) the coverage spread of the pattern which relates to its size and shape and b) the arrangement of points within the pattern which relates to the movement of the user as well as radio coverage.

A. Size, Shape and Structure of Location Patterns

The size and shape of the area covered by a location pattern is one of its most basic features. In order to study this aspect, we *circumscribe* the 'hull' or boundary of each location pattern with a simple geometric shape : a rectangle. We then focus on studying properties of the rectangle, which are much more amenable to analysis. The specific rectangle we choose is the *Minimum Area Bounding Rectangle (MABR)* [6]. The *MABR* covers the *convex hull* of *all* the points comprising the cell-towers visited by the user (this includes air-travel)[1] with the least amount of area, which allows us to minimize the error induced by this approximation in estimating the coverage area. Once we circumscribe the pattern in a *MABR* , we study three basic properties of the rectangle (and hence the pattern) : 1) the area of the rectangle, 2) the *skew* defined as ratio of width to length and 3) the *orientation* of the rectangle (with reference to true north).

B. Clusters and Trajectories

In order to study the internal arrangement of points within a location pattern, we classify points into structures that aim to answer two intuitive questions with regards to movement :

- Does a user's calling/movement pattern occur in a *contiguous* area or is the movement disparate, *i.e.,* the user appears at distinctly apart locations?
- Does a user's pattern contain trajectories, e.g.a highway route?

The first question, apart from providing an intuitive sense of movement of the user, also has the benefit of characterizing the radio coverage patterns in an area and can provide important information to design paging 'zones' since it provides information about the area in which a user is most likely to make calls. To address this, we extract 'dense' clusters of points that would identify preferred 'areas' of a user using a modified version of the well-known dbScan [1] algorithm that can identify clusters of arbitrary shape and size in the presence of noise.

The second question is related to major routes that a user may follow. Our interest is in identifying major routes followed

[1] The cell-towers are mapped to a $2-D$ space which is reasonably accurate for small distances. In practice they actually reside on a 3D ellipsoid.

a user over the entire duration of observation rather than a specific call. Since such movement would normally be along roadways that (at least in the continental US) are generally straight, we approximate the paths a user follows in the real world as *segments* in 2D space and extract segments from the set of points by applying a version of the *Iterative-End-Point-Fit* ($IEPF$) segment extraction algorithm [4] that includes a pre-partitioning phase presented in [2]. We visually verified the effectiveness of both the cluster and segment extraction algorithms over several examples.

Having defined the geometric constructs that characterize a location pattern, our objectives in applying them to our large data-set are two-fold : First, study statistical properties of actual location patterns and in particular determine good models that can characterize them. Second, utilize these constructs to understand how user location patterns compare across various geographical areas. For the latter, we computed and studied location patterns of users at the nationwide level, as well as for eleven cities : New York, San Francisco, Los Angeles, Chicago, Boston, Seattle, Kansas City, Denver, Albuquerque, Tulsa and Dallas.

C. Results

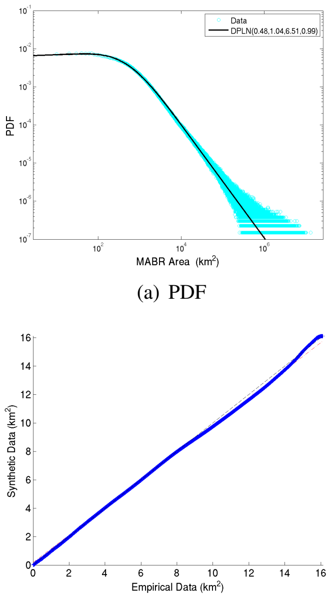
With respect to the first objective, we find rather surprisingly that a *single* statistical distribution, the *Double Pareto LogNormal Distribution* (or $DPLN$) provides an excellent fit for the area of the location patterns and internal clusters as well as trajectory lengths of location patterns at both national and city level.

The Double Pareto Log Normal Distribution is a four parameter distribution $(\alpha, \beta, \nu, \tau)$ introduced by Reed ([3]). The complete $DPLN$ distribution is given by :

$$f(x) = \frac{\alpha\beta}{\alpha+\beta}\Big[e^{\alpha\nu\ +\alpha^2+\tau^2} x^{-\alpha-1}\Phi(\frac{\log x-\nu-\alpha\tau^2}{\tau}) + e^{-\beta\tau+\beta^2\tau^2/2} x^{\beta-1}\Phi^c(\frac{\log x-\nu+\beta^2\tau^2}{\tau})\Big] \quad (1)$$

where Φ is the Normal CDF of $N(0,1)$. An attractive property of the $DPLN$ distribution is that it arises via exponential sampling of a generative process that was originally lognormally distributed and then evolves as a Brownian process. Consequently, this process can also be used as a basis to model evolution of the empirical process being observed. Qualitatively speaking, on a log-log plot, the $DPLN$ distribution is characterized by two linear curves corresponding to the head and tail of the distribution with slopes $\beta - 1$ and $-(\alpha + 1)$ respectively and a log-hyperbolic middle section.

Our analysis shows that $dPLN$ provides a very good fit to empirical distributions of the size of location pattern (approximated by $MABR$ area), size of clusters as well as trajectory lengths. Furthermore, this is observed to hold at both the nationwide level as well as city-level.



(a) PDF

(b) QQ Plot

Fig. 1. *MABR* Area : Nation

Fig. 1 plots the empirical density function (1(a)) for the area of the *MABR* of location patterns for users nationwide[2]. In the same graph, we also plot the MLE fit of $DPLN$ distribution computed using the EM algorithm outlined in [3] along with the associated parameters in the legend. Visually, one can see that the $DPLN$ is an excellent fit to the empirical distribution. To further strengthen this assertion, Figure 1(b) shows a *quantile-quantile* (QQ) plot between the empirical data and samples generated from a $DPLN$ distribution with parameters from the MLE fit (the graph is on a log-scale) . A clear linear relation is evident in the QQ plot between the empirical data and random samples from the fitted $DPLN$ distribution up to about $\approx$ 1 million km^2 thereby verifying the goodness of the fit.

Fig. 2 shows the empirical distribution as well as the $DPLN$ fit and associated QQ plot for the size of location pattern for users in New York. Again, we see the distribution provides an excellent fit, albeit with different parameters. We have observed that the $DPLN$ distribution provides similarly good fits for other features such as cluster size and trajectory length at both nationwide *and* city scale. Further results are available in [5].

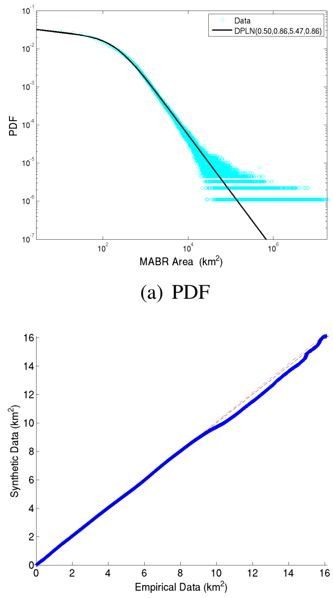
The second aspect of our work involved comparing and studying location patterns across different cities. For brevity, we present a sample of our results in Fig. 3.

Fig. 3(a) compares the average *MABR* area[3] per user at each city. The graph gives a sense of the range that users travel around each . For example, it shows that users associated with dense urban areas such as New York, Boston, Chicago, and Los Angeles have a smaller footprint than areas such as Dallas, Denver, Kansas City, Tulsa or Albuquerque. Overall, we observe that the typical average *MABR* area is $\approx 20,000\ km^2$ which translates to user range of about 150 km in each direction (assuming equal length and breadth of the *MABR*).

The difference in *orientation* of the $MABR$ as a function of geography is explored in Fig. 3(b). We plot the empirical

[2] We truncate the range of the x-axis such that y-axis values $\approx 10e-6$.

[3] To ensure the measure stays city-specific, we computed the mean over 95% of the CDF mass since the tail values typically correspond to air travel.



(a) PDF

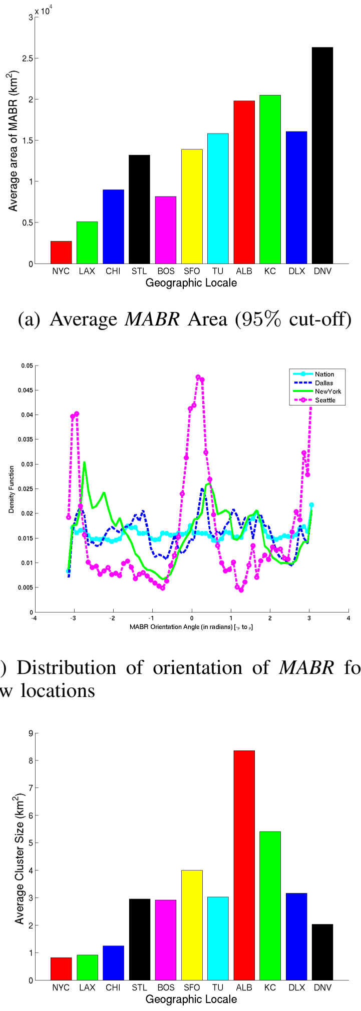
(b) QQ Plot

Fig. 2. *MABR* Area : NewYork

density function of this metric for three representative locales and compare it against the nationwide orientation. Note that the x-axis in the figure ranges from $-\pi$ to π and the y-axis is the PDF. From the figure, the distribution for the 'Nationwide' data set is close to *uniform* indicating that at a global scale there is no preferred direction, which is as expected: user footprints are essentially oriented in independent random directions. Next, the distribution for Dallas shows a few peaks indicating a slight preference to certain directions of travel. The peaks get more pronounced with New York and are most evident for Seattle, suggesting that as an ensemble, the movement pattern in this city is highly correlated (*eg.,* everybody uses the same set of few roads).

Finally, Fig. 3(c) provides a sample result from comparison of the *internal* structure of the location patterns in the form of average cluster size. Interestingly, we note that some cities, specifically, Albuquerque, Tulsa, Kansas City, Dallas and San Francisco have very large clusters which indicates that the popular calling 'areas' of users have a large range. In contrast, New York and Los Angeles have very small clusters pointing to a user calling pattern that is more concentrated. This differentiation in behaviour can be useful to optimize network paging [5].

In summary, we have developed a simple methodology to analyze the location patterns of users and applied it to a large nationwide data set of call data records. Our observation that many salient features of location patterns can be modeled by a *single* distribution, the $DPLN$, is one of the main contributions, of particular interest given the simple generative process of that distribution. In addition, we have also provided a sample overview of results that show how our constructs can be used to compare and understand the difference in user footprints across different cities.



(a) Average *MABR* Area (95% cut-off)

(b) Distribution of orientation of *MABR* for a few locations

(c) Cluster Size

Fig. 3. Comparison of Location Patterns

References

[1] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of ACM KDD*, 1996.

[2] V. Nguyen, S. Gächter, A. Martinelli, N. Tomatis, and R. Siegwart. A comparison of line extraction algorithms using 2d laser rangefinder for indoor mobile robotics. *Journal of Autonomous Robots*, 23, August 2007.

[3] W. Reed and M. Jorgensen. The double pareto-lognormal distribution - a new parametric model for size distribution. *Communications in Statistics -Theory and Methods*, 33(8):1733–1753, 2004.

[4] R.O.Duda and P.E.Hart. *Pattern Classification and Scene Analysis*. John Wiley & Sons, 1973.

[5] A. Sridharan and J. Bolot. Location patterns of mobile users : A large case study. *IEEE INFOCOM*, 2013.

[6] G. T. Toussiant. Solving geometric problems with rotating calipers. In *MELECON*, 1983.

A Multi-Scale Multi-Cultural Study of Commuting Patterns Incorporating Digital Traces

Yingxiang Yang, Marta C. Gonzalez

Department of Civil and Environmental Engineering, Massachusetts Institute of Technology,
Cambridge, MA 02139, USA

yxyang@mit.edu

Abstract

In this paper, we propose an extended radiation model to predict commuting flow OD matrix by performing a multi-scale study. The extended radiation model overcomes the shortcomings of both the gravity model and the radiation model. Unlike the gravity model, it has only one parameter which can be determined largely just by the study region size, so that the new model doesn't necessarily need empirical OD matrices for parameter calibration. The added one parameter makes the new model more flexible than the original radiation model and thus can be applied to study regions of different scales. For countries without detailed census data but with rich cell phone data, we propose a cell phone user OD matrix expansion method so that we could gain insight of these regions' commuting patterns from cell phone records. This method is validated and then tested on regions in three different continents.

The results show that as a combination of the radiation model and the gravity model, the extended radiation model overcomes the shortcomings of each. We show that the radiation model is applicable to some certain scales. But it's not flexible enough to adjust to the homogeneity of opportunities and varying scales. On the other hand, because of the large number of parameters the doubly constrained gravity model is flexible enough to fit to most given datasets, but the fitted parameters cannot be applied elsewhere. In the extended radiation model there is only one parameter α, but it's enough to take into account the effect of the scale and the opportunity heterogeneity. Unlike the doubly constrained gravity model in which the values of parameters are totally unpredictable, the parameter is to a large extent predictable given the size of the study region. The results are validated on countries from four different continents.

We use the Bay Area as an example to demonstrate that cell phone records is an ideal alternative for movement pattern analysis when traditional data sources are not available. Home and work locations are inferred at individual level and then aggregated to show its equivalence to the census data. The three different models' prediction results for cell phone users are compared at different countries. These results show not only the applicability of each model, but also the unique commuting patterns in each country.

The automatically collected geographical information data, the point of interest, is verified to be a suitable proxy for commuting generation rates. The potential of both the cell phone records and the point of interests can be further exploited in future studies. From cell phone records we can

reconstruct the activity chains and regularities in daily activity patterns at individual level, which are very hard to acquire by traditional survey methods. Together with digital footprints such as the point of interest or Foursquare records, non-commuting trips, which have higher flexibility, can also be traced.

Session 6

Mobility modelling

Scaling Theory of Human Mobility and Spatial Network[a]

Pierre Deville,[1] Dashun Wang,[2,3] Chaoming Song,[2,3] Nathan Eagle,[4] Vincent Blondel,[1] and Albert-László Barabási[2,3,5]

[1]*Department of Applied Mathematics, University of Louvain, Belgium*

[2]*Center for Complex Network Research, Department of Physics,*
Biology and Computer Science, Northeastern University, Boston, Massachusetts 02115, USA

[3]*Center for Cancer Systems Biology, Dana Farber Cancer Institute, Boston, Massachusetts 02115, USA*

[4]*College of Computer Science, Northeastern University, Boston, Massachusetts 02115, USA*

[5]*Department of Medicine, Brigham and Women's Hospital,*
Harvard Medical School, Boston, Massachusetts 02115, USA

Abstract

With the increasing availability of large-scale datasets that simultaneously capture human movements and social interactions, advances in human mobility and spatial networks have rapidly proliferated during the past years [1, 2], impacting in a meaningful fashion a wide range of areas, from epidemic prevention and emergency response to urban planning and traffic forecasting [3, 4]. As human mobility and spatial networks have developed in parallel, being pursued as separate lines of inquiry, we lack any known relationships between the quantities explored by them, despite the fact that they often study the same systems and datasets. Here, by exploiting three different cell phone datasets, we find a set of scaling relationships, mediated by a universal flux distribution, that link the quantities characterizing human mobility and spatial networks, showing that the widely studied scaling laws uncovered in the two areas represent two facets of the same underlying phenomena.

[a] Corresponding authors: pierre.deville@uclouvain.be, dashunwang@gmail.com

INTRODUCTION

Our knowledge of the interplay between individual mobility and social network is limited, partly due to the difficulty in collecting large-scale data that record, simultaneously, dynamical traces of individual movements and social interactions. This situation is changing rapidly, however, thanks to the pervasive use of mobile phones. Indeed, the records of mobile communications provide extensive proxy of mobility patterns and social ties, by keeping track of each phone call between any two parties and the localization in space and time of the party that initiates the call. These data, collected by telecommunication carriers in a truly objective manner, serve as an unprecedented social microscope helping us scrutinize the mobility patterns together with social structures.

DATA

To demonstrate the practical relevance and universality of our results, we compiled three large-scale mobile phone datasets from three different countries in two continents:*D*1, that contains 1.3 Million users in a western European country and covers a period of one month; *D*2 is the dataset from another European country that covers one year long period of around 6 Million users; emphD3 is collected by the largest mobile phone carrier in Africa, covering a period of four years. These three datasets, of same level of details yet with different demographics and scales, allow us to assess the universality of previously reported scaling laws on individual mobility and spatial networks in a systematic manner, and most importantly, lead us to uncover the scaling relationship between these scaling laws, establishing the first formal link between the two fields.

RESULTS

The main result is the discovery of a scaling relationship between the exponent characterizing spatial networks (β) and the exponent characterizing human movements (α) given by

$$\beta = \alpha\theta - \delta. \tag{1}$$

where θ and δ are exponents characterizing the relationship between social and mobility fluxes and are constant across the datasets.

The values of the exponents in Eq. 1 can be summarized in a table for all datasets and demonstrate a good agreement between empirical measurements and our theoretical results.

Taken together, Eq. 1 offers an explicit link between human movements and social communications, showing that the social exponents characterizing the distance distribution (β_r) and rank distribution (β_s) can be expressed in terms of the mobility exponents characterizing individual movements (α_r and α_s). The relationship between these classes of exponents is mediated by a universal flux distribution, which is independent of geography. This scaling relationship directly bridges two fields that were perceived as distinct, showing that they represent different facets of a deeper underlying reality, and offers us a powerful framework to derive the characteristics of one field from those of the other.

CONCLUSION

The unexpected duality between human mobility and social communications opens a new avenue in many areas previously improbable, making it possible, for instance, predicting traffic flows and transportation patterns by using communication volumes alone. Indeed, as technology continues to inundate us with increasingly detailed data on individual activities, the results presented here are expected to have an increasing value, impacting all phenomena driven by human behavior, from epidemic spreading and emergency response to traffic forecasting, urban planning and more.

[1] Gonzalez, M. C., Hidalgo, C. A., and Barabasi, A.-L. (2008) *Nature* **453(7196)**, 779–782.

[2] Candia, J., González, M. C., Wang, P., Schoenharl, T., Madey, G., and Barabási, A.-L. (2008) *Journal of Physics A: Mathematical and Theoretical* **41(22)**, 224015.

[3] Colizza, V., Barrat, A., Barthelemy, M., Valleron, A.-J., and Vespignani, A. (2007) *PLoS Medicine* **4(1)**, e13.

[4] Helbing, D., Farkas, I., and Vicsek, T. (2000) *Nature* **407(6803)**, 487–490.

Proxy networks for human mobility in Europe: the impact on epidemic modeling

Michele Tizzoni[1], Paolo Bajardi[2], Adeline Decuyper[3], Guillaume Kon Kam King[4], Christian Schneider[5], Vincent Blondel[3], Zbigniew Smoreda[6], Marta C. Gonzalez[5], Vittoria Colizza[7,8,9]
1 Computational Epidemiology Laboratory, Institute for Scientific Interchange (ISI), Torino, Italy
2 GECO - Computational Epidemiology Group, Department of Veterinary Sciences, University of Torino
3 ICTEAM Institute, Universit Catholique de Louvain, Belgium
4 CNRS, UMR5558, F-69622 Villeurbanne, France
5 MIT - Department of Civil and Environmental Engineering, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA
6 Sociology and Economics of Networks and Services Department, Orange Labs, France
7 INSERM, U707, Paris, France
8 UPMC Universit Paris 06, Facult de Mdecine Pierre et Marie Curie, UMR S 707, Paris, France
9 Institute for Scientific Interchange (ISI), Torino, Italy
Email: MT: michele.tizzoni@isi.it, PB: paolo.bajardi@unito.it, GKKK: guillaume.konkamking@gmail.com, AD: adeline.decuyper@uclouvain.be, CS: schnechr@mit.edu, VB: vincent.blondel@uclouvain.be, ZB: zbigniew.smoreda@orange.com, MC: martag@mit.edu, VC: vittoria.colizza@inserm.fr

Abstract

Introduction. The modeling of human mobility plays an essential role in the development of realistic mathematical and computational models for the spatial spread of infectious diseases [1]. When available, empirical data extracted from official census surveys have been successfully used to integrate human movements into epidemic models [2, 3]. On the other hand, a range of mobility models, such as gravity models [4] and radiation models [5], have been developed to predict population movements in case of missing empirical data and have been used to fill this gap in epidemic models [2, 6]. More recently, the use of spatially explicit mobile phone data as a tool to investigate human mobility patterns has gained great popularity, leading to the discovery of universal characteristics of individual mobility patterns [7, 8]. Similarly to other approaches, mobile phone data can be used as a proxy for people movements and then integrated into epidemic models, once proper scales of time and space are defined. For instance, recent epidemiological studies on malaria have used mobile phone data to estimate human movements in countries where there is limited availability of official data on human mobility [9, 10]. However, despite the variety of approaches, the impact of using different proxies for human movements in epidemic models is still poorly understood. In particular, a comprehensive comparison between different methods is urgently needed in order to assess the reliability of mobile phone data as a proxy for human mobility.

Methods. In this paper we use a metapopulation network approach to address two main issues:

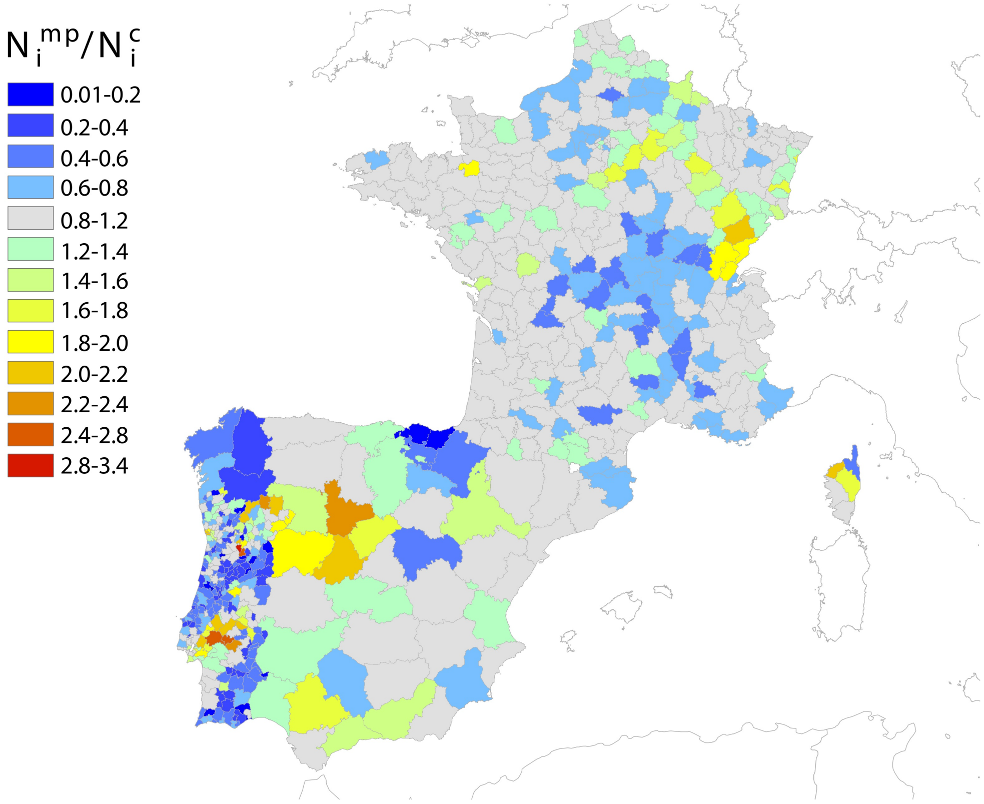


Figure 1: **Mobile phone coverage.** Map of the ratio between the population estimated by mobile phone records (N_i^{mp}) and the census population (N_i^{c}) in the subdivisions of the European countries under study. Blue regions are undersampled by the mobile phone dataset, while regions that appear in green to red are oversampled.

(i) evaluating the adequacy of mobile phone data as a description of commuting patterns in Europe; (ii) evaluating the impact of using mobile phone data as a proxy for human movements in epidemic models. To this aim, we compare the commuting networks extracted from the official census surveys of three European countries to the corresponding proxy networks extracted from three high resolution mobile phone datasets tracking the daily movements of millions of users. For the first time we are in the position of examining, through a detailed statistical analysis, the ability of mobile phone data to match the empirical commuting patterns reported by census surveys, at different geographic scales in a set of European countries. Then, we directly compare the outcomes of stochastic epidemics simulated on a metapopulation model that is based either on the empirical commuting networks or on the mobile phone commuting networks. In this analysis, we focus on the differences between the modeling approaches in terms of arrival times (i.e. time of first infection) in a subpopulation and invasion paths from the source of the infection to the rest of the network. In addition, we perform the same analyses on a set of synthetic networks generated using the radiation model and compare the results to those obtained using the mobile phone networks.

Results. The statistical analysis reveals that mobile phone data can well predict the total traffic, both incoming and outgoing, of a given node. Commuting flows estimated from mobile phones are able to reproduce the topology of the census commuting network to a high level of detail. Commuters' flows are generally found to be in good agreement between the two sources of data, however the agreement is not statistically significant because of the discrepancies emerging in the total number of commuters estimated by mobile phones tracks, due to the presence of sampling biases in the mobile phone dataset. Indeed, mobile phone data tend to overestimate the magnitude of commuting flows between residence and workplace, especially in those regions that are undersampled; hence, epidemics on mobile phone networks spread usually faster than on census networks, leading to earlier arrival times for all the subpopulations. Discounting this effect, the general epidemic behavior of the mobile phone commuting networks and the census commuting networks shows a good and statistically significant agreement.

Conclusions. Statistical differences between empirical census data and commuting flows inferred from mobile phone users can be relevant, but they do not essentially alter the outcomes of simulated epidemics, especially if we consider those observables that are important for evaluating strategies for disease control. Epidemic results are comparable to those obtained using synthetic commuting networks generated by the radiation model. Our results confirm the valuable role of mobile phone data to estimate population movements which can be integrated into spatial epidemic models to provide support to public health policies.

References

[1] S. Riley, Science **316**, 1298 (2007).
[2] D. Balcan *et al.*, Proc. Natl. Acad. Sci. USA **106**, 21484 (2009).
[3] M. L. Ciofi degli Atti *et al.*, PLoS ONE **3**, e1790 (2008).
[4] J. Ortúzar and L. Willumsen, *Modelling transport* (Wiley, Chichester, United Kingdom, 2001).
[5] F. Simini, M. González, A. Maritan, and A.-L. Barabási, Nature **484**, 96 (2012).
[6] S. Merler and M. Ajelli, Proc. R. Soc. B **277**, 557 (2009).
[7] M. C. González, C. A. Hidalgo, and A.-L. Barabási, Nature **453**, 779 (2008).
[8] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, Science **327**, 1018 (2010).
[9] A. Le Menach *et al.*, Sci. Rep. **1**, 93 (2011).
[10] A. Wesolowski *et al.*, Science **338**, 267 (2012).

Human Mobility Modeling at Metropolitan Scales

Sibren Isaacman*, Richard Becker†, Ramón Cáceres†, Margaret Martonosi‡,
James Rowland†, Alexander Varshavsky†, Walter Willinger†
*Loyola University †AT&T Labs ‡Princeton University
isaacman@cs.loyola.edu {rab,ramon,jrr,varshavsky,walter}@research.att.com mrm@princeton.edu

Human mobility models have myriad uses in mobile computing research and other fields of study. Models that faithfully reproduce the movements of real people can help answer questions in areas as varied as mobile sensing, opportunistic networking, urban planning, ecology, and epidemiology. For example, a model of how people move around a city can help evaluate whether a sensing application running on mobile phones would be able to attain the desired geographic coverage.

Our work aims to produce accurate models of how large populations move within different metropolitan areas. In pursuit of this general aim, we have a number of more specific goals. Our first goal is to generate sequences of locations and associated times that capture how individuals move between important places in their lives, such as home and work. Previous work has shown that people spend most of their time at a few such places [3, 4, 12]. Our second goal is to aggregate the movements of many such individuals to reproduce human densities over time at the geographic scale of metropolitan areas. A model that operates at these scales can help address important societal issues such as the environmental impact of home-to-work commutes. Our third goal is to take into account how different metropolitan areas exhibit distinct mobility patterns due to differences in geographic distributions of homes and jobs, transportation infrastructures, and other factors. Previous work has shown significant differences between cities along metrics such as commute distances [4, 5, 6, 11].

Many human mobility models that fall short on one or more of these goals have been proposed in the past. Some models produce random motion that does not correspond to actual mobility patterns, e.g., [8, 10]. Their lack of memory about recurring movement patterns and of spatiotemporal realism about population densities results in unrealistic motion of modeled individuals. Some models are tailored to a small geographic area such as a university campus, e.g., [9]. They do not apply to larger geographic areas with more diverse populations. Some models aim to be universal, e.g., [3], and thus do not adapt to different geographic areas. There remains a need for a realistic model that matches empirical observations for large and distinct geographic areas.

This paper introduces a modeling approach that takes as input certain spatial and temporal probability distributions drawn from large populations of real people living across wide geographic areas. An especially good source of these distributions are the Call Detail Records (CDRs) maintained by cellular network operators. Billions of cellphone users worldwide keep their phones near them most of the time, and the networks need to know the rough location of all active phones to provide them with voice and data services. CDRs contain 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. When joined with information about the locations of those towers, CDRs can serve as sporadic samples of the approximate locations of the phone's owner. A growing body of work has shown that information derived from anonymized CDRs can accurately characterize many aspects of human mobility [1, 2, 3, 4, 5, 6, 12].

With cellular network data becoming more available, it is tempting to think that creating human mobility models from such data should be easy. However, this is not the case. For example, while CDRs readily yield insights into aggregate population densities, they do not convey whether their associated locations correspond to home, work, or other important places for particular cellphone users. Without such semantic information, it is difficult to abstract CDRs into models applicable to scenarios, regions, or populations that vary from those for which the real-life CDR data was collected. Furthermore, both the spatial and temporal granularity of CDR data is quite coarse. Spatially, CDRs are only accurate to the granularity of celltower spacings. Temporally, CDRs are only generated when phones are actively involved in a voice call or text message. Our work makes key contributions in overcoming the challenges stemming from lack of semantic information and coarse granularity, to produce usefully accurate models for arbitrary metropolitan regions.

Our modeling approach intelligently samples the spatial and temporal probability distributions from CDRs, or other population data, to generate sequences of locations and times for any number of synthetic people in any region for which the required distributions can be obtained. A generative model derived from CDRs has flexibility, compactness, and availability advantages over using CDRs directly. First, our mod-

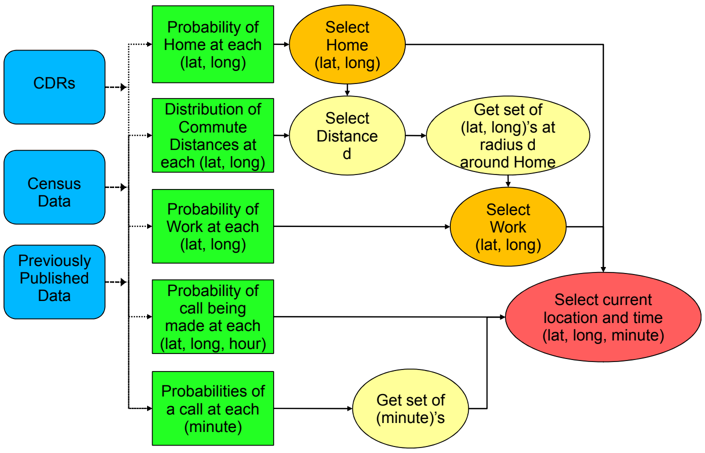

CDRs
Census Data
Previously Published Data
Probability of Home at each (lat, long)
Distribution of Commute Distances at each (lat, long)
Probability of Work at each (lat, long)
Probability of call being made at each (lat, long, hour)
Probabilities of a call at each (minute)
Select Home (lat, long)
Select Distance d
Get set of (lat, long)'s at radius d around Home
Select Work (lat, long)
Select current location and time (lat, long, minute)
Get set of (minute)'s

Figure 1: Overview of the WHERE modeling approach.

els offer the option of perturbing the input distributions to evaluate what-if scenarios, for example to consider how the addition of a new residential or employment area might change traffic patterns. In contrast, the original CDRs are difficult to manipulate in meaningful ways. Second, our model for a metropolitan area with a 50-mile radius can be stored as a set of histograms that fit within 2 gigabytes. In contrast, an anonymized CDR dataset for the same area occupied approximately 100 gigabytes. Finally, our models can be made available to a larger research community because they do not to reproduce the mobility pattern of any individual real person. They thus avoid many of the privacy concerns associated with source CDRs.

The final stage of our modeling approach produces locations and times in the form of synthetic CDRs. These synthetic CDRs have the same format and call/text frequency characteristics of real CDRs. They are modeled to approximate the actual movement patterns of users. Increased model complexity results in more accurate movement patterns, which in turn produces higher-fidelity synthetic CDRs. We chose the CDR output format for several pragmatic reasons. One, we can compare this output directly against real CDRs, our best source of location information for large populations and regions. Two, this output can plug in directly into the growing body of analysis software that uses CDRs as input.

In our full paper [7], we propose and evaluate WHERE ("Work and Home Extracted REgions"), a region-scale modeling approach. First, we identify the key properties of human movement, such as important locations and commute distances, that need to be represented as probability distributions. Then, we describe how these probability distributions can be used to generate synthetic CDRs for an arbitrary number of synthetic people. Figure 1 summarizes the overall flow of our approach.

We validate our approach by comparing the spatiotemporal dynamics of synthetic populations generated by WHERE to those of real populations. In particular, we use Earth Mover's Distance (EMD) as a metric to compare the spatial population densities on an hourly basis for synthetic and real CDR sequences. Our validation begins with stylized examples that confirm our models' fidelity both quantitatively and visually. We validate both at the aggregate level, where simpler models may perform well, as well as at a finer granularity, which exposes the advantages of WHERE compared to other models considered. We then scale up our validation to large datasets containing real anonymized CDRs for the Los Angeles (LA) and New York City (NY) metropolitan areas. Our LA and NY datasets each span three months of activity for hundreds of thousands of phones, yielding billions of location samples.

Recognizing that real CDRs are not available to all researchers, we also evaluate models in which the same input distributions are derived from publicly available US Census data [13]. We show that models based on real CDRs closely approximate the real populations and movements of these cities. Models based on census data are also viable, but at a loss of significant accuracy.

Finally, we present example applications of our modeling approach. We create models for the LA and NY metropolitan areas and use the resulting synthetic CDRs to perform calculations that one may wish to perform on real CDRs. We show that calculations performed on the WHERE model produce far more accurate results than those performed on more naive models. For example, we can calculate daily ranges of travel that agree with real ranges, as well as perform more complex tasks such as investigating opportunistic message propagation in large urban environments.

The overall contributions of our work are:

- We introduce an approach to modeling human mobility patterns by generating fully synthetic CDRs from real-world probability distributions.
- Our approach works at the scale of large metropolitan areas and accounts for mobility differences between metropolitan areas.
- We show that our technique is extensible to greater lev-

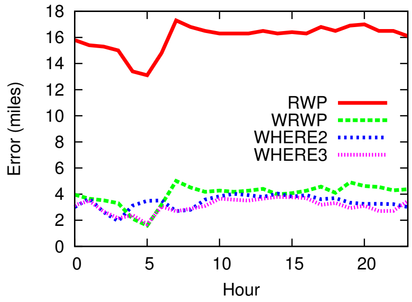


Figure 2: EMD error over time for different models of human movement in the NY area. Our WHERE3 model shows average errors smaller than 3 miles. WHERE2 fits between WHERE3 and WRWP.

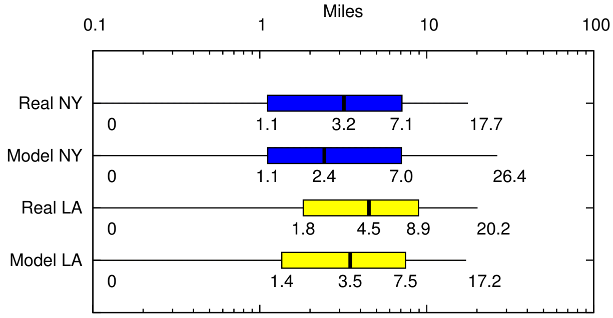


Figure 3: Daily range statistics for the NY and LA regions as calculated from real data and the output of our WHERE2 model. WHERE2 accurately recreates the characteristics of NY and LA mobility.

els of precision by providing it more complete input probability distributions (at the cost of increased model complexity). For example, we can create WHERE2 models with only home and work locations, and WHERE3 models that include an additional important location.

- We validate our approach against large-scale location datasets drawn from two major US metropolitan areas. We compare our generated CDRs against real CDRs, and show that our location distributions achieve more than 4 times error reduction compared to a Random Waypoint (RWP) model. Figure 2 shows these results for the NY region.
- As an example of how our models can help answer concrete questions about human mobility, we use our synthetic CDRs to compute daily ranges of travel. Our synthetic CDRs exhibit error at the median of less than 0.8 and 1 mile for NY and LA residents, respectively. This accuracy constitutes a more than 14 times improvement over a Weighted Random Waypoint (WRWP) model. Figure 3 shows these results for the NY region.

Please see our full paper [7] for additional details.

1. References

[1] M. A. Bayir, M. Demirbas, and N. Eagle. Discovering spatiotemporal mobility profiles of cellphone users. *World of Wireless, Mobile and Multimedia Networks and Workshops*, 2009.

[2] F. Girardin, A. Vaccari, A. Gerber, A. Biderman, and C. Ratti. Towards estimating the presence of visitors from the aggregate mobile phone network activity they generate. In *Intl. Conference on Computers in Urban Planning and Urban Management*, 2009.

[3] M. C. González, C. A. Hidalgo, and A.-L. Barabási. Understanding individual human mobility patterns. *Nature*, 453, 2008.

[4] S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, M. Martonosi, J. Rowland, and A. Varshavsky. Identifying important places in people's lives from cellular network data. In *9th International Conf. on Pervasive Computing*, 2011.

[5] S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, M. Martonosi, J. Rowland, and A. Varshavsky. Ranges of human mobility in los angeles and new york. In *Eighth IEEE Workshop on Managing Ubiquitous Communications and Services*, 2011.

[6] S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, J. Rowland, and A. Varshavsky. A tale of two cities. In *Workshop on Mobile Computing Systems and Applications (HotMobile)*, 2010.

[7] S. Isaacman, R. Becker, R. Caceres, M. Martonosi, J. Rowland, A. Varshavsky, and W. Willinger. Human Mobility Modeling at Metropolitan Scales. In *Proceedings of the Tenth Annual International Conference on Mobile Systems, Applications, and Services (MobiSys 2012)*, June 2012.

[8] D. B. Johnson and D. A. Maltz. Dynamic source routing in ad hoc wireless networks. In *Mobile Computing*, pages 153–181. Kluwer Academic Publishers, 1996.

[9] M. Kim, D. Kotz, and S. Kim. Extracting a mobility model from real user traces. In *IEEE INFOCOM*, 2006.

[10] W. Navidi and T. Camp. Stationary distributions for random waypoint models. *IEEE Transactions on Mobile Computing*, 3(1):99–108, 2004.

[11] A. Noulas, S. Scellato, R. Lambiotte, M. Pontil, and C. Mascolo. A tale of many cities: universal patterns in human urban mobility. *PLoS ONE*, 2012.

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

[13] US Census Bureau. `http://www.census.gov`.

Geographic Similarity Within Social Networks

Jameson L. Toole
Engineering Systems Division, MIT, Cambridge, MA 02139

Carlos Herrera
Departamento de Matemática Aplicada, Universidad Politécnica de Madrid, Madrid, Spain

Christian M. Schneider and Marta C. González
Civil and Environmental Engineering, MIT, Cambridge, MA 02139

I. INTRODUCTION

Each day over 3.5 billion people wake up in a city. They commute to work, schedule meetings, and convene with friends for dinner. Mobile phones are increasingly used to coordinate these behaviors and they are collecting and storing massive amounts of data along the way. Originally designed as communication devices, mobile phones are exceptionally well-suited to measure social behaviors such as who a person talks to, how often, and when. Now innovations in hardware and software make it possible for phones to provide high-resolution data on the geographic location of their users. To better understand how cities function and the behavior of those living in them, this work analyzes social and geographic data collected by roughly 800,000 mobile phones in two large cities in a European country over a 15 month period. The goal is to shed light on a simple, but important question: How does a person's social network affect how much and to where they travel?

Much research has approached social behavior and mobility separately. Studies have mapped the local and global structure of large call networks [1, 2] and used this structure to predict characteristics and activity patterns of individuals [3]. Geographic data on its own has revealed universal patterns of human mobility and a surprisingly high amount of predictability to a person's movement [4–6]. Recent efforts have combined the two data types. Correlations between the proximity of two individuals and the likelihood they are friends [7, 8] have proven especially important in the diffusion of information [9, 10]. Moreover, predictions of where an individual will be in the future are improved by incorporating information on the whereabouts of friends [11, 12]. Results confirm humans as a social species, willing to travel to be with friends and family [13].

However, these studies have been limited in their approach, seeking to predict movement to and from only a small subset of places using only a few good friends. This work addresses builds upon previous studies to more fully understand the complexities of social behavior and mobility.

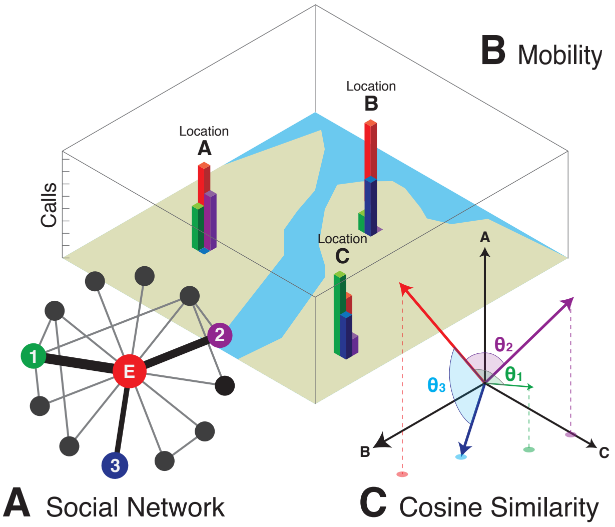


FIG. 1: **A**) We construct a social network where nodes are users and links are weighted by the number of calls between two individuals. For each user, E, we construct an ego-network and rank social contacts based on the number of calls with the ego. **B**) A location vector is constructed for each user by counting the number of calls made at each tower in the region. In this diagram, four uses (including the ego) record calls at 3 locations. **C**) Finally, we measure the similarity between the movement of two users by calculating the cosine similarity of their location vectors. Users who visit the same locations with the same relative frequency will have locations vectors which point in the same direction, thus giving a cosine similarity value of 1 while users that do not share any visited locations in common have a cosine similarity of 0.

II. METHODS

Figure 1 diagrams our metrology. For each city, we begin by constructing a social network where links between two users are weighted by the number of calls between them. Locally, we construct an ego network consisting of all users called by an individual and rank each of these contacts by the number of calls between them (Figure 1A). Globally, we assess how the geodesic distance between two individuals relates to the similarity of two user's movement. To describe the mobility of a person in a city, we construct it location vectors for each individual. Given a set of T locations (in this case mobile phone towers) in a city, construct a T-dimensional vector, $\vec{u}$ for each user where element u_i is given by the number of times a user calls from location i(Figure 1B). To assess the similarity between the mobility patterns of two individuals, we compute the *cosine similarity* between their vectors in this location space (Figure 1C).

III. RESULTS

We measure how the similarity of users changes with the social relationship between them. Figure 2A shows the average cosine similarity between two users separated by a given geodesic distance d in the call network. The similarity of location vectors for direct neighbors is 10 times that of two randomly chosen individuals. Significantly higher similarity values are only measured between individuals separated by up to three degrees ($d = 3$) in the network.

Interpretation of these results is complicated by the fact that individuals who live in an area are likely to have access to similar places as well as being far more likely to be friends. In order to measure the contributions of social and geographic proximity to similarity, we compare four combinations of two variables: sharing or not-sharing the same *home* and *work* locations and are being or not-being direct neighbors in the call graph. We find that sharing the same home and work locations with a user contributes only slightly more to the similarity of visits to other places than being direct contacts (Table I).

TABLE I: Four groups are constructed from combinations of (not) sharing the same home and work locations (+/-HW) and (not) making calls between them (+/-F)

Similarity	+HW	-HW
+F	0.44	0.19
-F	0.25	0.04

Next, we explore how similarity changes with tie strength and relative location importance. The insert in Figure 2B shows that tie strength is highly correlated with location similarity. An individual is nearly twice as similar to the contact they call the most than to the 30th ranked social contact. Furthermore, by randomizing location data we find that having one-half to two-thirds of the similarly between contacts is related to placing the same relative importance on locations rather than simply visiting the same places (regardless of how often).

The relative importance of a place is also correlated with the social rank of a contact. We begin by calculating the average similarity between a user and each of her 30 people she makes the most calls to. We then successively remove locations from each user's location vectors based on importance and recalculate similarity. For example, the first step sets the number of visits to each user's most visited location to 0, before recalculating similarity. The second step sets visits to the top two most visited locations of all users to 0, and so on. Figure 2B shows that removing the most visited location of all users dramatically decreases the similarity between a person and their top 4 social contacts, but increases their similarity to contacts 5-30 (and to a random user). This suggests that individuals share their most visited location with their top 4 friends. Moreover, we see non-monotonic changes in similarity over successive removals. We hypothesis that regimes featuring increasing similarity reflect locations that are not shared with social contacts (thus removing them makes a user more similar to a contact), while the reverse is true for regimes featuring decreasing similarity. Moreover, these dynamics appear robust to changes in the size of the space of possible locations to visit.

Finally, having explored the relationship between an individual's social behavior and where he or she moves, we quantify the correlation between social behavior and how much movement takes place. We focus on three mobility metrics, the radius of gyration, number of towers visited, and the entropy of those visits. As the degree of an individual in the call network increases from 0 to 70, we find that individuals travel farther, to more places, and with in less predictability. However, this affect is inverted for users with degree beyond 70. Moreover, we find that, unlike other forms of social contagion [14], it is only the number of connections that correlates with mobility and not the structural diversity of them.

IV. FUTURE WORK

For future work, we hope to relate these findings to the predictability of users in order to develop an analytic model that reproduces these patterns. We will also explore the sensitivity of our measurements to the dimension of the location space and the length of the time periods used.

[1] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A.-L. Barabási, Proceedings of the National Academy of Sciences of the United States of America **104**, 7332 (2007).

[2] A.-L. Barabási, Nature **435**, 207 (2005).

[3] N. Eagle and A. S. Pentland, Behavioral Ecology and Sociobiology **63**, 1057 (2009).

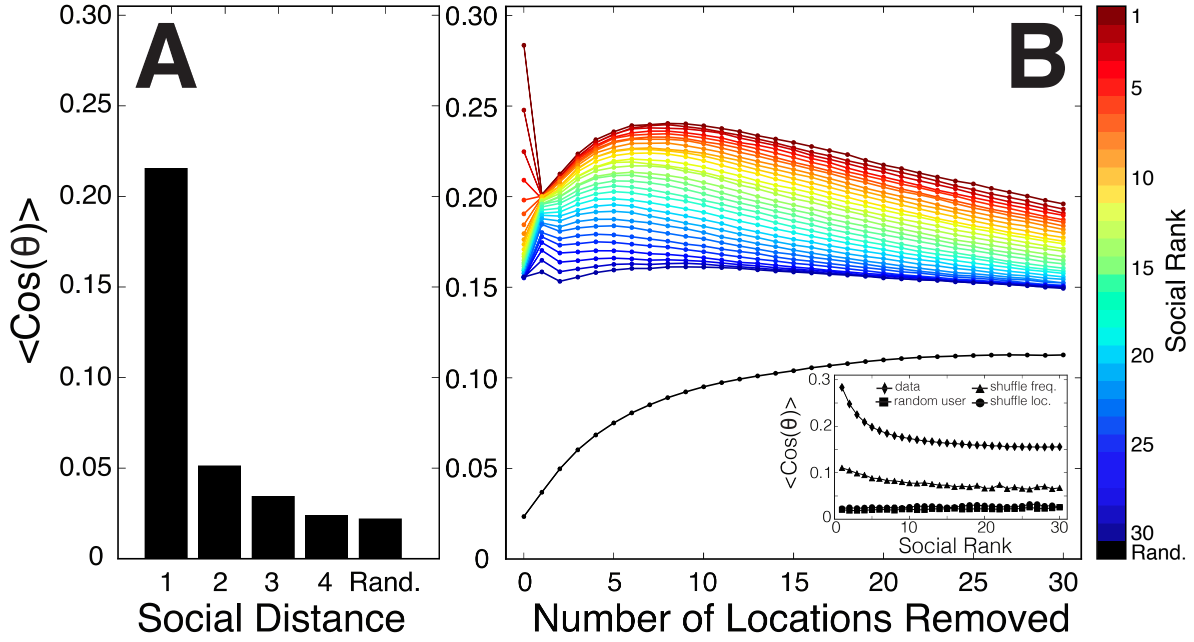


FIG. 2: **A**) We show the average cosine similarity of visited locations between users separated by a geodesic distance d in the call graph. The movement of users who are direct neighbors are 10 times more similar than two users selected at random. Moreover, we find significant increases in similarity up to 3 degrees removed from an individual. **B**) We show how similarity depends on the strength of a relationship (ranked by number of calls) and how this similarity changes when important locations are systematically removed. The stronger the tie, the more similar the locations and frequency of visits. Moreover, successively removing the k most visited locations from each user's vector then recalculating similarity reveals which locations are shared between friends of a given rank.

[4] F. Simini, M. C. González, A. Maritan, and A.-L. Barabási, Nature **484**, 8 (2012), arXiv:1111.0586.

[5] M. C. González, C. A. Hidalgo, and A.-L. Barabási, Nature **453**, 779 (2008).

[6] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, Science **327**, 1018 (2010).

[7] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins, Proceedings of the National Academy of Sciences of the United States of America **102**, 11623 (2005).

[8] L. Backstrom, E. Sun, and C. Marlow, North WWW '10, 61 (2010).

[9] J. L. Toole, M. Cha, and M. C. González, PLoS ONE **7**, e29528 (2012), arXiv:1110.0535.

[10] J.-P. Onnela, S. Arbesman, M. C. González, A.-L. Barabási, and N. A. Christakis, PLoS ONE **6**, 7 (2011).

[11] M. D. Domenico, csbhamacuk **2012** (2012).

[12] V. Etter, M. Kafsi, and E. Kazemi, in *Mobile Data Challenge by Nokia Workshop in conjunction with Int Conf on Pervasive Computing* (2012).

[13] E. Cho, S. A. Myers, and J. Leskovec, in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining KDD 11*, KDD '11 (ACM Press, 2011) p. 1082.

[14] J. Ugander, L. Backstrom, C. Marlow, and J. Kleinberg, Proceedings of the National Academy of Sciences **2012**, 1 (2012).

Session 7

Trajectories and regularities

Late for Good

Vsevolod Salnikov and Renaud Lambiotte
naXys, University of Namur, Belgium

In this work, we perform a large-scale experiment on human mobility by using smartphones as sensors. To do so, we have developed an app continuously tracking significant displacements of phones and providing in return to their user an attractive service. Contrary to standard tracking procedures based on Call Detail Records, our methods offers the advantage of collecting data in real-time, in an anonymous way, and without the constraints usually imposed by mobile phone operators.

The last few years have witnessed the increasing use of mobile phone devices as a way to collect social data at a large-scale [1, 2]. Each of us carries a mobile phone, almost continuously, and its sensors and online services provide a more and more complete view on our life and our environment. Our phone knows our whereabouts, our friends, where we meet, our taste in music, etc. The integration and monetization of these data by corporations raises alarming privacy issues. Yet, when treated ethically and used for the common good, they also have the potential to provide innovative ways to solve problems in areas such as public health, ecology or urban planning [3, 4]. Examples include the automatic detection of dysfunctions in the transport infrastructure, and the participatory report of ecological or epidemiological data.

In the particular case of mobility tracking, two different types of approach have been developed to extract individual trajectories from mobile phone data. The first approach exploits data already routinely collected by cellular network or by online services. Examples include Foursquare check-in data [5], or Call Detail Records of a phone connecting to a cell tower [6, 7]. These data have the advantage of being acquired for free and of being analyzed at virtually no cost, but they have the disadvantage of being proprietary, often with strict confidentiality agreements, and of variable quality. For instance, CDRs are known to provide a sparse and heterogeneous sampling of the trajectories, with a fairly poor spatial resolution.

The second approach aims at tracking mobility in a more controlled way, either by handing mobile devices to a limited number of users, or by distributing software, typically downloadable apps, tracking motion. However, this approach either relies on ad-hoc infrastructure or on the active participation of users, and it has, as a consequence, been limited to small sample sizes so far [8]. The aim of our work is to address this limitation and to explore the possibility to attract a large number of individuals in a mobility tracking experiment. Because the deployment of infrastructure is prohibitively expensive at this scale, we focus on the development of apps and their adoption by the broader public [9, 10]. To be installed, an app needs to compete against thousands of others and, typically, to be run continuously in the background of the phone OS despite its energy cost. To be successful, the deployment of a tracking experiment thus requires proper incentives for a user to participate. For this reason, a majority of experiments have failed to reach a large part of the population, and they have typically been limited to circles of *geeks*, students and researchers.

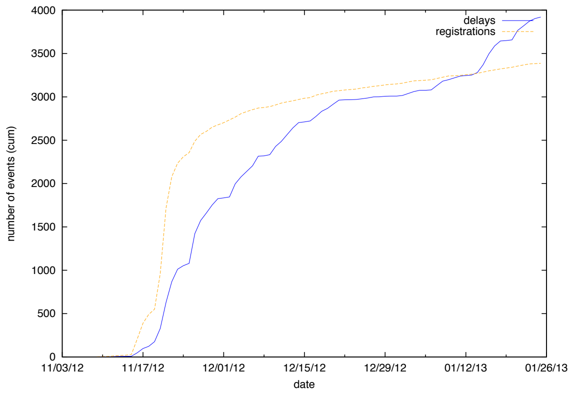


FIG. 1: Time evolution of the total number of users and submitted delays. After an initial rapid growth, we are in a steady regime, where the number of new users and delays is roughly constant.

In this work, we have developed an app for iPhones and Android devices, called *lateforgood* [13], where we offer an attractive service to the user, and not only the prospect of participating in a research experiment. The app is dedicated to Belgian train users, and helps them keep track of their train delays and submit a form for financial compensation to the national train company [14]. Incentives for the user are thus: money (paid by the train company for delays), time and convenience (the otherwise tedious form is automatically filled on our servers) but also the important feeling of being heard by the train company as a collective voice (the punctuality of train delays has been the subject of animated debates in Belgium in the last years).

In practice, the app works as follows. When it is first opened, it starts tracking the motion of the user, in order to detect closeby train stations. The app runs in

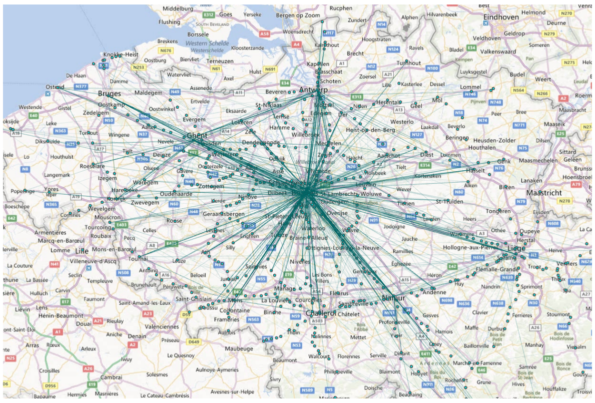


FIG. 2: Geographical representation of train delays. Two train stations are connected when a delay has been notified. Darker lines correspond to more delays.

the background on the mobile device, using a technology similar to the one used by *openpath* [11] to minimize impact on battery life: the app does not record the position continuously via GPS, but tracks only significant changes in position determined by the device API, via GPS, WI-FI access and triangulation of cell phone towers. When the user re-opens the app to declare a train delay, a list of likely official train schedules is proposed, based on recent mobility patterns, and a delay can be declared and uploaded to the user profile. When a sufficient number of delays has been accumulated, a compensation form is pre-filled and ready for download. Let us also stress that the app uploads data to our servers only when the app is used by its user. Moreover, the issue of privacy has been considered carefully as no personal information is hold: on our side, it is as if mobility patterns of anonymous mobile phones were collected.

The experiment is still in its acquisition phase. Since the launch of our service in November 2012, around 3500 users have downloaded it, among which 40% use it actively (Fig. 1). During this time period around 4000 delays have been submitted, leading to 600000 data points for the position of the users, and providing us with information on the main train lines in Belgium (Fig. 2). We are currently developing algorithms to distinguish between car and train mobility, to properly assign sampled trajectories to train routes [12], including connections between train lines, and to uncover a list of train lines and/or stations particularly affected by delays.

Acknowledgment

We acknowledge financial support from FNRS.This paper presents research results of the Belgian Network DYSCO (Dynamical Systems, Control, and Optimization), funded by the Interuniversity Attraction Poles Programme, initiated by the Belgian State, Science Policy Office.

[1] R. Kwok, “Phoning in data”, Nature **458**, 959-961 (2009)
[2] J. Reades, F. Calabrese, A. Sevtsuk and C. Ratti, “Cellular Census: Explorations in Urban Data Collection”, IEEE Pervasive Computing **6**, 30–38 (2007)
[3] A. Pentland, D. Lazer, D. Brewer and T. Heibeck, “Using reality mining to improve public health and medicine”, Stud. Health Technol. Inform. **149**, 93–102 (2009)
[4] V.D. Blondel, M. Esch, C. Chan, F. Clerot, P. Deville, E. Huens, F. Morlot, Z. Smoreda and C. Ziemlicki, “Data for development: the D4D challenge on mobile phone data”, arXiv:1210.0137 (2012)
[5] A. Noulas, S. Scellato, R. Lambiotte, M. Pontil and C. Mascolo, “A Tale of Many Cities: Universal Patterns in Human Urban Mobility”, PLoS ONE **7**, e37027 (201
[6] M.C. González, C.A. Hidalgo, A.-L. Barabási, “U standing individual human mobility patterns”, N **453**, 779-782 (2008)
[7] S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, Rowland and A. Varshavsky, “A tale of two cities”, Proceedings of the Eleventh Workshop on Mobile Computing Systems (HotMobile ’10), 19–24 (2010)
[8] N. Eagle and A. Pentland, “Reality Mining: Sensing Complex Social Systems”, Personal and Ubiquitous Computing **10**, 255–268 (2006)
[9] http://www.funf.org/
[10] M. Wirz, T. Franke, D. Roggen, E. Mitleton-Kelly, P. Lukowicz and G. Troster, “Inferring crowd conditions from pedestrians’ location traces for real-time crowd Monitoring during city-scale mass gatherings”, Proceedings of the IEEE 21st international workshop on enabling technologies, 367–372 (2012)
[11] https://openpaths.cc/
[12] R.A. Becker, R. Cáceres, K. Hanson, J. Meng Loh, S. Urbanek, A. Varshavsky and C. Volinsky, “Route classification using cellular handoff patterns”, Proceeding of Ubiquitous Computing (UbiComp 2011)
[13] More info can be found on http://www.sci-app.com
[14] The compensation is equal to 100% of the value of the ticket if the delay is of at least 60 minutes. In case of multiple delays within a 6 month period, a compensation of 25% (50%) can be claimed if, over a period of 6 months, at least 20 (10) delays of at least 15 (30) minutes are suffered. The compensation is received in the form of vouchers.

Do Mobile Phone Data Allow Estimating Real Human Trajectory?

Sahar Hoteit[†][§], Stefano Secci[†], Stanislav Sobolevsky[§], Guy Pujolle[†], Carlo Ratti[§]
[†] LIP6/UPMC - University of Paris VI; 4 Place Jussieu, 75005 Paris, France
[§] MIT Senseable City Laboratory, 77 Massachusetts Avenue Cambridge, MA 02139 USA.
E-mail: sahar.hoteit@lip6.fr, stefano.secci@lip6.fr, stanly@mit.edu, guy.pujolle@lip6.fr, ratti@mit.edu

Nowadays, the huge worldwide mobile-phone penetration is increasingly turning the mobile network into a gigantic ubiquitous sensing platform, enabling large-scale analysis and applications. In recent years, mobile data-based research reaches important conclusions about various aspects of human mobility patterns and trajectories. But how accurately do these conclusions reflect the reality?

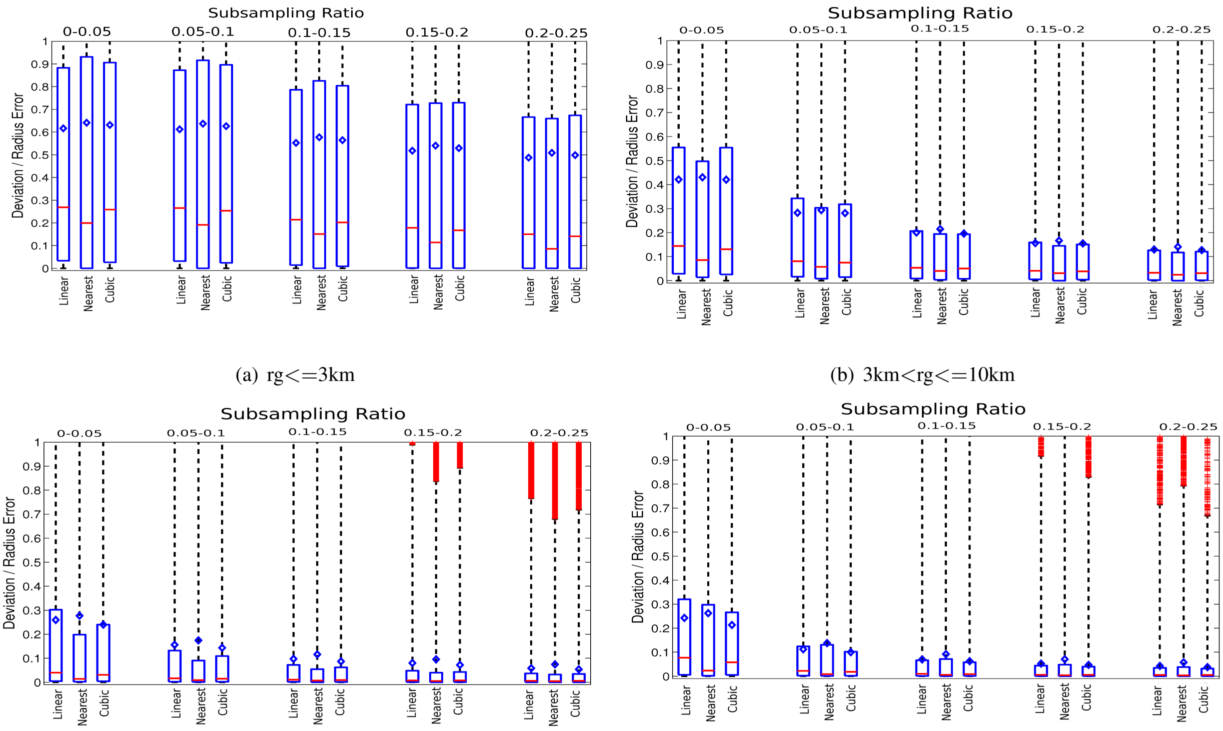
In order to evaluate the difference between the reality and the approximation methods, we study the error between real human trajectory and the one obtained through mobile phone data using different interpolation methods (linear, cubic, nearest, spline interpolations) and taking into consideration mobility parameters.

We use for this aim a dataset consisting of anonymous cellular phone signaling data, it consists of location estimations for about one million devices in the Boston metropolitan area.

To evaluate the error between real human trajectories and the estimated ones, we fine-select data of those smartphones holders with a lot of samplings, typically those data-plan users with persistent Internet connectivity due to applications such as e-mail synch. Then, in order to reproduce *artificial* “normal user” sampling, we subsample *real* data-plan smartphone quasi-continuous traces according to an experimental inter-event statistical distribution. Therefore, we extract, from the real trajectory, a first random position then the corresponding next positions are extracted according to the inter-event time distribution values.

Hence, given a real trajectory with a high number of positions, and its subsampling that reproduces normal user’s activity, we apply an interpolation method to estimate the trajectory across the given points. Given the real trajectory points P_i, we estimate its corresponding position in time in the estimated trajectory: P_i'. Then we determine the deviation between the two points]as the distance separating the exact position P_i to the estimated position P_i' in the interpolating curve joining the samples. To take into account mobility habits, we categorise the users depending on their “radius of gyration” defined by the deviation of user positions from the user centroid position.

From extensive evaluations based on real cellular network data of the Boston metropolitan area, we show that the linear interpolation offers the best estimation for sedentary people (with a small radius of gyration) and the cubic one for commuters (having a big radius of gyration). Moreover, the nearest interpolation appears as the best



(a) rg<=3km

(b) 3km<rg<=10km

(c) 10km<rg<=32km

(d) rg>32km

Fig. 1: Boxplots of trajectory error

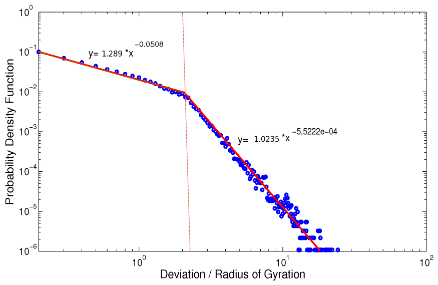


Fig. 2: Probability density function of error

one for “ordinary people” doing regular stops and standard displacements (Figure 1).

Another important experimental finding is that trajectory estimation methods show different error regimes whether used within or outside the “territory” of the user defined by the radius of gyration. The distribution of errors over all users’ positions is approximated by a combination of two power law distributions joined by a breakpoint (approximately equal to 2.2) for the different interpolation methods (Figure 2).

As a future work we aim to estimate the positions of hotspots in a region knowing only the mobility characteristics of its residents.

Depict Urban Activities from Real Movement with Auto-GPS

Teerayut Horanont
Institute of Industrial Science
The University of Tokyo
Komaba, Tokyo 153-8505, JAPAN
teerayut@iis.u-tokyo.ac.jp

Apichon Witayangkurn
Department of Civil Engineering
The University of Tokyo
Komaba, Tokyo 153-8505, JAPAN
apichon@iis.u-tokyo.ac.jp

Ryosuke Shibasaki
Center for Spatial Information Science
The University of Tokyo
Kashiwa-shi, Chiba 277-8568, JAPAN
shiba@csis.u-tokyo.ac.jp

Abstract

Today, the urban computing scenario is emerging as a concept where human can be used as a component to probe city dynamics. The urban activities can be described by the close integration of ICT devices and humans. In the quest for creating sustainable livable cities, the deep understanding of urban mobility and space syntax is crucial importance. This research aims to explore and demonstrate the vast potential of using large mobile GPS dataset for the analysis of human activity and urban connectivity. The new type of mobile sensing data called "Auto-GPS" has been anonymously collected from 1.5 million people for a period of one year in Japan. The analysis delivers some insights on interim evolution of population density, urban connectivity and commuting choice. The results enable planner to better understanding of urban organism with more complete inclusion of urban activities and their evolution through space and time.

1 Introduction

How new technology can help cities manage and deliver a sustainable future. In the past few years, it has become possible to explicitly represent and account for time-space evolution of the entire city organism. Information and communication technology (ICT) has the unique capability of being able to capture the ever-increasing amounts of information generated in the world around us, especially the longitudinal information that enables us to investigate patterns of human mobility over time. Thus, the use of real-time information to manage and operate the city is no longer just an interesting experience but a viable alternative for future urban development.

In this research, the analysis of mobile phone location, namely "Auto-GPS", has been used to serve as frameworks for the variety of measures of effective city planning. More specifically, we explore the use of location information from Auto-GPS to characterize human mobility in two major aspects. First is the commuting statistics and second is the city activity, how the change of activities in part of urban space can be detected over times.

2 Dataset

There were two datasets used in this study. The main dataset were collected from approximately 1.5 million mobile Auto-GPS users of a certain mobile phone service provided by a leading mobile phone operator in Japan. Under this service, handsets provide a regular stream of highly accurate location data, and thereby enable support services that are closely linked with the user's behavior. Technically, an Auto-GPS-enabled handset position is measured within 5 min and sent through a network of registered services. (Fig.1) The data was recorded from August 2010 to October 2011. In order to preserve user privacy, Auto-GPS data is provided in a completely anonymous form to ensure privacy of personal information.

Fig. 1 shows a graph of the average number of GPS points per day in this dataset. A small sample of the raw data is shown in Fig 3.

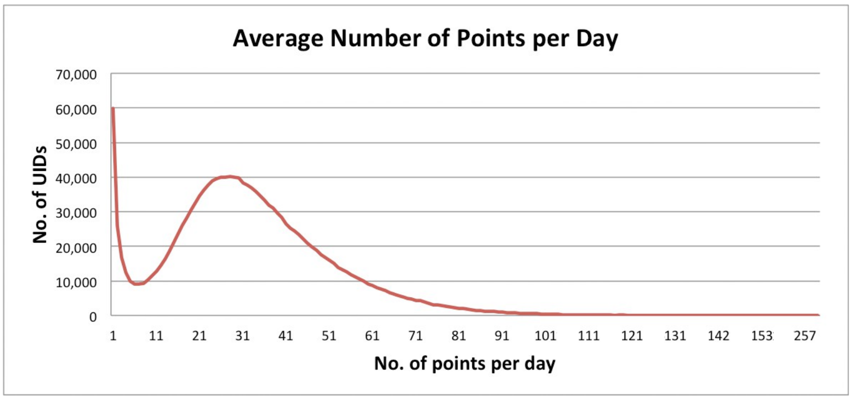


Fig. 1. The average number of GPS points per day is 37, indicating that Japanese users spent approximately 3 hours traveling with their handset each day.

3 Results and Discussions

Finding the urban descriptive knowledge such as where/how/when/why of the people who use the area is one of the most important information for urban planners. Our first result attempted to explain where the people come from. We constructed multiple criteria to define visitors in the area. We used the minimum stay of 30 minutes and excluded people who have home and work location in the area. The maximum annual visit is set to 8 times as it is the third quartile of the entire dataset. The total annual visitors to Odaiba area was estimated at 80,463 people from 1.5 million total samples or 5.36% of the population. Fig. 2 shows the choropleth map of estimate annual visitor. As expected, the closer prefecture the most likely people come from. There are some exception for the big cities such as Nagoya, Osaka, Fukuoka and Hokkaido where air transport services are operated more often.

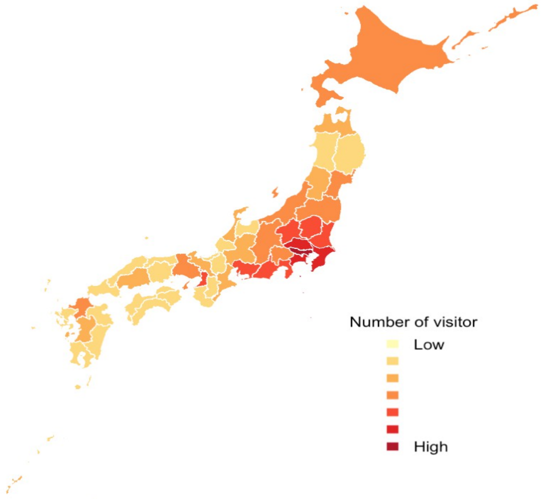


Fig. 2. Estimate annual visitor to the Odaiba area by prefecture.

We continue observed the number of visitors in Odaiba per day, it appears that the area are more popular during summer. This is because of the big event arranged by Fuji TV. The highest visit to Odaiba is on August 14 when the Tokyo Bay Grand Fireworks Festival was held. The second most visit is on the Christmas Eve as the area is known as couple-y place. We notice a significant distinct drop in the number of visitors suddenly in March 11. It was the day of the earthquake magnitude 9.0 hit Japan in 2011, followed with the "radiation leakage" of 2 nuclear power plants in Fukushima prefecture. We could captured 3 weeks of anomaly reduction in visiting the area before it returned to the normal situation.

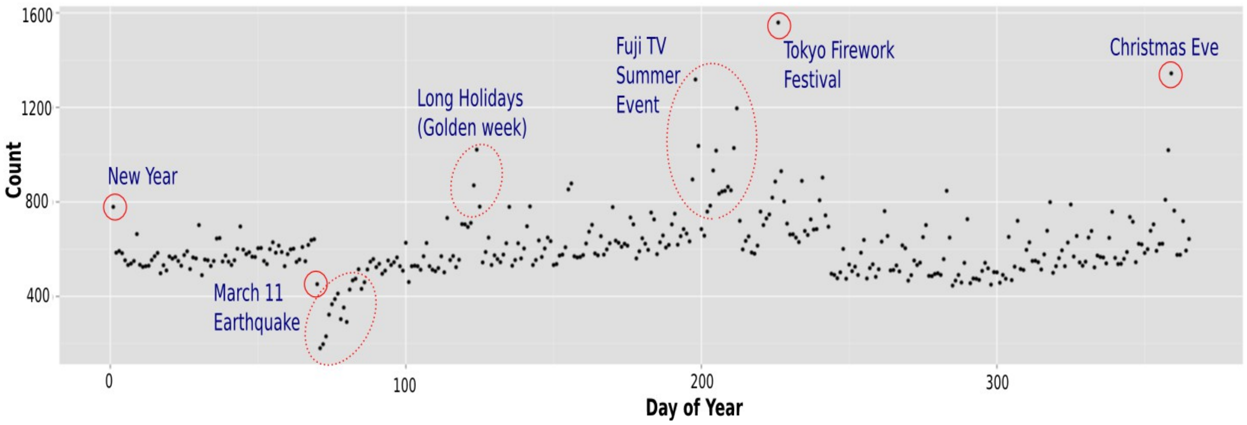


Fig. 3. Estimate daily visit to Odaiba area. The magnitude of anomalies can vary greatly between events, and this could lead to composite dominated by a few major events.

Next, we visualized how different the activities between weekday and weekend by overlaid weekday stay points over the weekend. (Fig. 4) Surprisingly, there are several clusters that highly dominate over others at particular location. By incorporating prior knowledges of the area and collection of news, it revealed clear evidence how the patterns created. The location marked with “a” are complex buildings where shopping malls, hotels and restaurants are located in. This yield the similar distribution of both weekday and weekend. The “b” mark is an open space where we can see half of the area are more active during the weekend. This is because of the special events are usually held only on the weekend. The area in upper part are served as outdoor parking space and is the main area of Fuji TV summer event. This event is usually held for 3 months in summer regardless weekend or weekday. The “c” areas are event spaces that mainly occupied during weekend. The “d” areas are office building that is more dominant in weekend. Please note that the d1 area is a construction area during our data collection period.

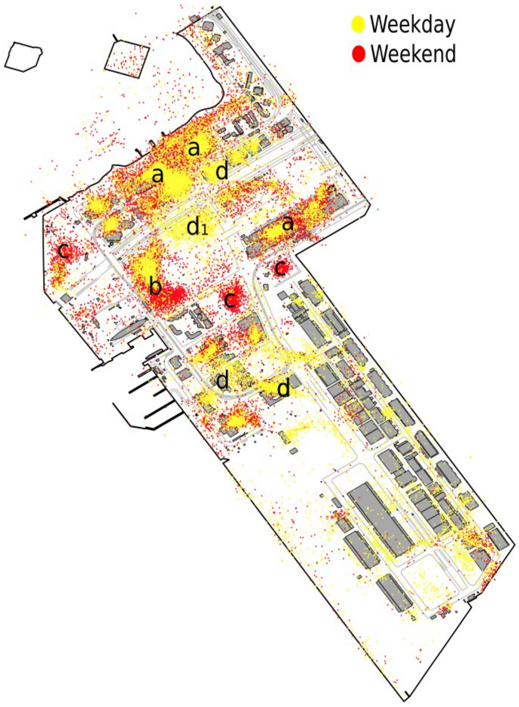


Fig. 4. Comparison of estimated people activities between weekday (yellow) and weekend (red).

Fig. 5 provides valuable population count information in each building of the entire year. The hight of 3D building corresponded to the number of visitors. It is clear that Shopping malls, restaurant and complex type buildings are the most destination place in Odaiba area. All of them are 10 times more visitors than the office area.

6 Conclusions

This study explores the potential of using mobile Auto-GPS enables in the new context and broader advances towards the understanding of today’s excessive mobility. The finding of this remarkable dataset is to capture the urban evolution from the real movement of people. The results display summarizes the findings of the comprehensive and creative process using Auto-GPS data. Finally, the importance of this research ultimately lies on how it can be practically applied and utilized massive amount of high accurate GPS data for future sustainable urban development.

Acknowledgements

This research carried out under the support of DIAS (Data Integration and Analysis System) by MEXT (Ministry of Education, Culture, Sports, Science and Technology). We would like to thanks ZENRIN Data Com (ZDC) CO., LTD. for archiving the dataset used in this research.

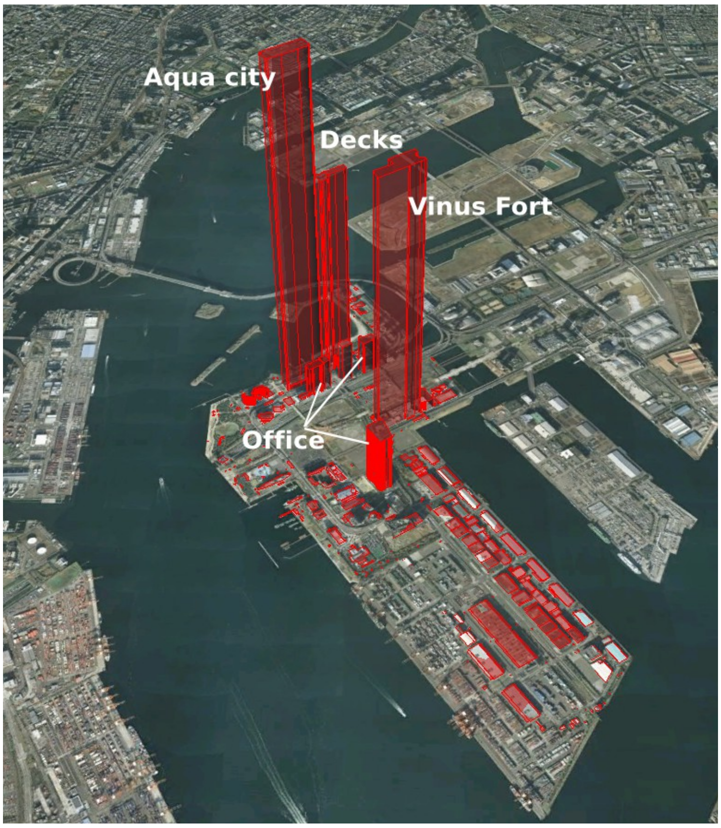


Fig. 5 The yearly visit counted in every buildings in Odaiba area. The hight of building represented the number of visitors who stop by

References

[1] M. C. Gonza ́lez, C. A. Hidalgo, and A.-L. Baraba ́si. (2009) Understanding individual human mobility patterns, Nature, vol. 458, pp. 238–238

[2] C. Song, Z. Qu, N. Blumm, and A.-L. Baraba ́si. (2010) Limits of predictability in human mobility, Science, vol. 327, no. 5968, pp. 1018–1021.

[3] Isaacman, S., Becker, R., Cáceres, R., Kobourov, S., Rowland, J., & Varshavsky, A. (2010). A tale of two cities. Proceedings of the Eleventh Workshop on Mobile Computing Systems & Applications - HotMobile ’10, 19.

[4] Robert J. and Wytse M. (2003). Wearable GPS device as a data collection method for travel research. Working Paper, ITS-WP-03-02, Institute of Transport Studies, University of Sydney

[5] Hongmian G., Cynthia C., Evan B., and Catherine L. (2011). A GPS/GIS method for travel mode detection in New York City. Computers, Environment and Urban Systems

[6] Ashbrook D., Starner T. (2003) Using GPS to learn significant locations and predict movement across multiple users. Personal and Ubiquitous Computing 7:275–286.

[7] Li, Z., Wang, J., & Han, J. (2012). Mining event periodicity from incomplete observations. Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’12, 444.

[8] Fontana, D., & Zambonelli, F. (2012). Towards an Infrastructure for Urban Superorganisms: Challenges and Architecture. The IEEE International Conference on Cyber, Physical and Social Computing.

[9] Ministry of Internal Affairs and Communications. Retrieved January 27, 2013, from http://www.soumu.go.jp/english/index.html

Netmob 2013 “Third conference on the Analysis of Mobile Phone Datasets”
May 2-3, 2013, MIT

Aggregated OD tracks of mobile phone data for the recognition of daily mobility spaces: an application to Lombardia region

Paolo Tagliolato, Fabio Manfredini, Paola Pucci
Dipartimento di Architettura e Studi Urbani – Politecnico di Milano
paolo.tagliolato@polimi.it, fabio.manfredini@polimi.it, paola.pucci@polimi.it

Introduction

Interpretative tools for the identification of mobility actices in the contemporary metropolis are needed, t only for the some known limitations of traditional ata sources but also because new forms of mobility are emerging, describing new city dynamics and time-variations in the use of urban spaces by temporary populations. In Italy, the traditional data sources for urban and mobility investigations (ie surveys, census) have some known limitations, including the high cost of surveys, the difficulty of data updating, the difficulty of describing city dynamics and time dependent variations in intensity of urban spaces usages by temporary populations at different scales.

New forms of mobility are changing the way in which urban spaces are used. They are characterized both by being based on the use of transportation system, and by the efficient appropriation of information technologies (internet, mobile phones). As underlined by some authors (Ehrenberg, 1995; Urry, 2000; Kaufmann, 2002; Ascher, 2004; Bourdin, 2005; Scheller et al. 2006), changes in management of mobility in contemporary cities are a useful key for understanding the transformations of times, places and modes of social life and work programs, structuring the metropolitan areas.

Considering the role of mobility practices in social and spatial differentiation, it becomes important to formulate pertinent analytical approaches, aimed at describing the different densities of use of the city as a new challenge and a prerequisite for understanding the city and its dynamics.

Hence, from an analytical point of view, it becomes important to accompany the traditional quantitative approaches referred to a geographic displacement that tends to focus on movement in space and time, in an aggregate way and for limited periods, with data sources able to describing fine grain over-time variation in urban movements.

In this direction, an interesting contribution may come from mobile phone network data as a potential tool for the development of real-time monitoring, useful to describe the urban dynamics as it has been tested in several experimental studies (Ahas et al., 2005; Ratti et al., 2006; Gonzales et al., 2008).

In this general context, we used mobile phone data provided by Telecom Italia, the main Italian operator, in form of aggregated mobility traces in order to test the potentialities of this information to identify temporary populations and different forms of mobility that structure the relationships in the contemporary city and to propose diversified management policies and mobility services that city users require, increasing the efficiency of the supply of public services.

Methodology

Milan is placed in an urban region which goes far beyond its administrative boundaries (fig.1). The core city and the whole urban area have been affected in the last 20 years by changes in their spatial structures and have generated new relationships between the centre and suburbs. At the moment, the urban region of Milan is a densely populated, integrated area where 4.000.000 inhabitants live, where there are 370.000 firms and large flows of people moving daily in this wide area (Balducci et al., 2010).

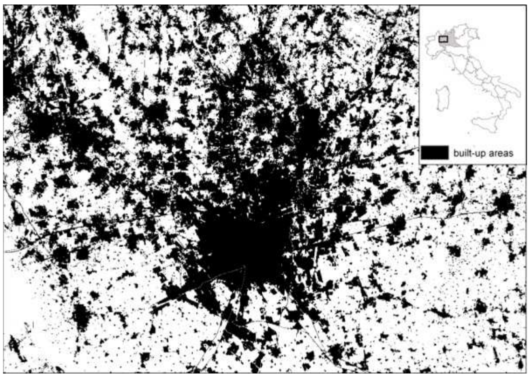


Figure 1 - Map of built-up areas in the Milan urban region (2007).

We used localized and aggregated tracks of anonymized mobile phone users. The data set was collected in different working days (five Wednesday in July, August, September, October and November 2012). In this case the available information was the geolocation of users’ mobile phone activity in time and in space. With mobile phone activity we intend each interaction of the device with the mobile phone network (i.e. calls received or made, SMSs sent or received, internet connections, etc..). This information was available at the level of the antenna which handled the activity.

We performed an aggregation of the information related to individual cells (antennas) in order to obtain useful polygon elements which could gave us the possibility to map and to interpret main spatial patterns of mobile phone users' mobility. The aggregation was determined by means of the application of an algorithm of hierarchical clustering of the location of the antennas, resulting in 526 polygons. The final zoning has been obtained by calibrating the algorithm in order to reach sufficiently balanced clusters (i.e. with an homogeneous number of antennas per polygon). Through a process of tessellation, we defined a set of polygons and it was therefore possible to map the direction and the intensity of mobile phone users' movements at an hourly basis. Using this data, we performed an analysis aimed at evaluating the overall mobility of cell phone users in the Lombardia region.

Aggregated tracks of mobile phone users

The analysis of the activity of mobile phone users permitted to put in evidence the main hourly distribution of origin – destination movements of a huge sample of people (more than one million per day).

We started from an hourly origin destination matrix (OD) of mobile phone users among the 526 zones obtained through the automatic clustering of antennas. For each zone it was available a set of directed connections towards the other zones of the aggregation and for each connection it was available the number of traced users.

Our goal was to display prevalent fluxes of mobility at different hour of a typical working day through a visualisation of the sum vector moving from each zone. The sum vector is the single vector resulting from the sum of all the single connections between each zone and the others and is characterized by two dimensions: the magnitude, which is function of the magnitudes of the original vectors and the angle which expresses the direction of the flux. The sum vectors have been finally applied to each zone of the automatic clustering of antennas.

A set of maps of the sum vector moving from each zone at different hours has been produced in order to highlight the main patterns of mobility during a typical working day. In the maps the dimension of the arrow is proportional to the absolute value of the sum vector and is therefore related to the prevalent fluxes of mobile phone users at specific hours.

The convergence of travels toward the main centres during the morning, the more complex direction of movements during the afternoon, are some interesting phenomena emerging from our analysis.

A broad use of the territory and an articulation of daily moves are visible every hour. The maps can be used as meaningful tools for monitoring the use of the infrastructural networks and of the urban spaces. On the one hand, the morning map (9pm; fig 2) confirm a polarization of movements towards the main centres offering job opportunities and highlights also the most commonly used infrastructures. On the other hand, the aggregated flows of mobile phone users in the afternoon (at 5 pm; fig 3, fig. 4) allow to recognize significant places for shopping and leisure, that are attended after work. This type of information is difficult, if not impossible, to monitor through conventional data at a comparable spatial and temporal resolution.

Our analysis shows, in synthesis, a wide and dense of the territories of the Milan urban region, a region which the attractiveness of new places emerges from the mobility practices. The multi-directional mobility intensifies and describes a complex network of relationships such as the growing transversal travels that define a not hierarchical system of relationships in the most dynamic territories of the Milan urban region.

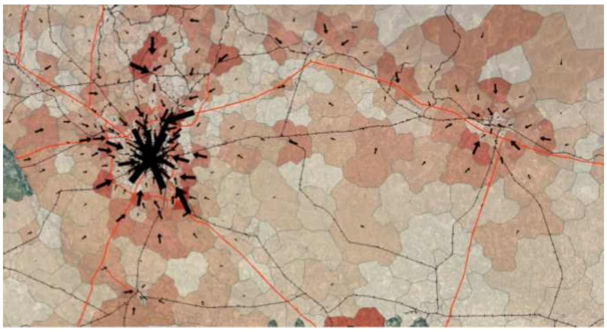


Figure 2 - Aggregated flows of mobile phone users: 9 am – 2011-10-19

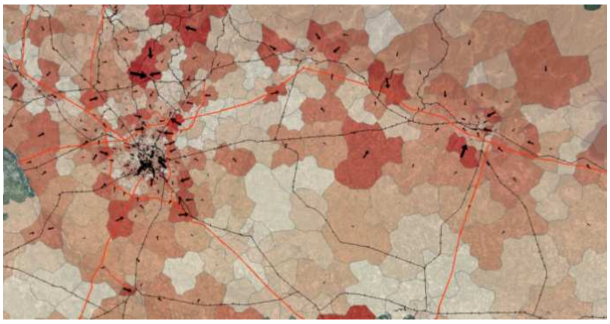
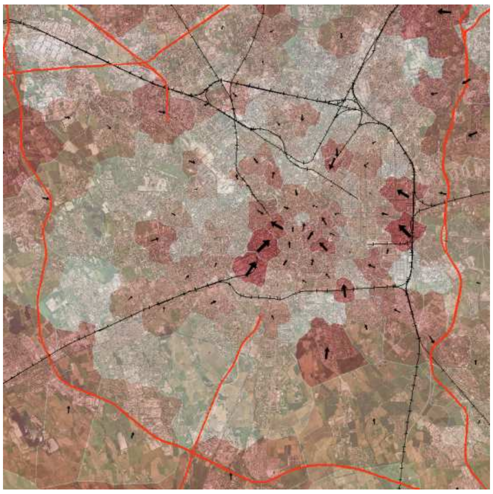


Figure 3 - Aggregated flows of mobile phone users: 5 pm – 2011-10-19



ɡure 4 – Milan city aggregated flows of mobile phone users: 5 pm – 2011-10-19

The interested reader can visit our interactive web version of the map showing prevalent fluxes of mobility in a working day at the following URL: http://www.ladec.polimi.it/maps/od/fluxes.html .

Implications for Policies

The research allowed us to test the potential of mobile phone data in explaining relevant urban usage and mobility patterns at the Milan urban region scale that can be hardly intercepted through traditional data source. This opens new implications for the urban research community which needs to elaborate new strategies to integrate traditional data with user generated data, such as mobile phone activity, in order to achieve a better comprehension of urban usages, in time and in space. Describing the trends of use of the urban spaces, the maps of mobile phone data give important information for mobility policies: the lack of coincidence between the mobility practices in the peak hours in the morning and in the afternoon when the chains of displacements are very articulate and complex, allows to recognize not only the variability in mobility practices, but also the places where these practices are occurring.

The commuters between 8 am and 9 am, become city users between 5 pm and 7 pm. This phenomenon strictly affects land use and can pose new questions and indications also for transport policy.

Indeed, if we overlaid the boundary of the institutional management of local public transport in the Milan area with the areas of mobility practices, taken from the mobile phone data, we could observe the "deep structural effects of the mobility of people on urban policies" and the obvious disconnection between fixed jurisdictions and “mobile factors" (Estèbe, 2008).

The same data helps us to question some interpretations in the literature on the erratic behaviors of metropolitan populations and on the nomadism that characterizes the contemporary practices, that surveys on mobile phone data have already undertaken (Gonzales et al., 2008).

References

Ahas, R., Aasa, A., Silm, S. and Tiru, M. (2010). Daily rhythms of suburban commuters’ movements in the Tallin metropolitan area: case study with mobile positioning data. Transportation Research Part C: Emerging Technologies 18(1), 45–54.

Ahas, R., Mark, Ü. (2005). Location based services–new challenges for planning and public administration?. Futures 37(6), 547–561.

Ascher F. (2004). Les sens du mouvement : modernités et mobilités. in Allemand S., Ascher F., Lévy J. (eds). Le sens du mouvement, Belin Bourdin, 21-34.

Balducci, A., Fedeli, V. and Pasqui, G. (2010) Strategic planning for contemporary urban regions: city of cities : a project for Milan, Ashgate, Burlington, VT.

Bourdin A. (2005). Les mobilités et le programme de la sociologie . Cahiers internationaux de sociologie, 1 (118), 5-21.

Ehrenberg, A. (1995). L’individu incertain. Paris, Calmann-Lévy.

Estèbe, P. (2008). Gouverner la ville mobile, PUF, Paris 2008.

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

Kaufmann, V. 2002. Re-thinking mobility. Aldershot: Ashgate.

Manfredini F., Pucci P., Tagliolato P. (2012). Mobile phone network data. New sources for urban studies? In Borruso G., Bertazzon S., Favretto A., Murgante B. and Torre Cm (eds), Geographic Information Analysis for Sustainable Development and Economic Planning: New Technologies, IGI Global.

Nuvolati G.(2003). Resident and Non-resident Populations: Quality of Life, Mobility and Time Policies. The Journal of Regional Analysis and Policy, 33 (2), 67-83.

Ratti, C., Pulselli, R. M., Williams, S. and Frenchman, D. (2006). Mobile landscapes: using location data from cell phones for urban analysis. Environment and Planning B: Planning and Design 33(5), 727–748.

Reades, J., Calabrese, F., Sevtsuk, A. and Ratti, C. (2007). Cellular census: Explorations in urban data collection. IEEE Pervasive Computing 6(3), 30– 38.

Scheller M., Urry J. (2006). The new mobilities paradigm. Environment and Planning A, 38, 207-226.

Urry, J. (2000). Sociology Beyond Societies. London: Routledge.

Understanding Human Mobility Due to Large-Scale Events

Faber Henrique Z. Xavier
Pontifical Catholic University of Minas Gerais (PUC MINAS)
Brazil – 30.535-901
faber.xavier@sga.pucminas.br

Lucas M. Silveira
Pontifical Catholic University of Minas Gerais (PUC MINAS)
Brazil – 30.535-901
lmsilveira@sga.pucminas.br

Jussara M. Almeida
Universidade Federal de Minas Gerais (UFMG)
Brazil – 31.270-010
jussara@dcc.ufmg.br

Carlos Henrique S. Malab
Oi Telecom
Brazil – 20.230-070
malab@oi.net.br

Artur Ziviani
National Laboratory for Scientific Computing (LNCC/MCTI)
Brazil – 25.651-075
ziviani@lncc.br

Humberto T. Marques-Neto
Pontifical Catholic University of Minas Gerais (PUC MINAS)
Brazil – 30.535-901
humberto@pucminas.br

I. Introduction

Analyzing the mobility patterns of cellphone users is a challenging task, but also a great opportunity to better understand the human dynamics in a covered area. Although recent studies show that human mobility in urban areas can be predictable considering daily routines [1], cellphone carriers still have difficulties for planning the necessary communication infrastructure to support the unusual workload that arises during large-scale events [2]. Such events typically involve a large number of people within an urban area, such as the final match of a soccer championship, a major rock concert (e.g., Rock in Rio), New Year's Eve celebrations, a religious pilgrimage, political manifestations, or the Olympics.

We here consider a large-scale event to be characterized by a huge number of people with similar interests directly related to the event's main subject who move towards/from a specific place in order to participate on a set of collective activities during a period of time. Even though many of these large-scale events are scheduled and planned in advance, and are expected to cause collective changes in the mobile phone workload [3], it remains common to notice the congestion of the carrier's resources during them.

In this paper, we present our on-going work towards understanding the human mobility and the workload dynamics of mobile phone networks due to large-scale events. To that end, we analyze the impact of some types of large-scale events on the workload of a mobile phone network. We use our recently proposed methodology [4], applying it to real anonymized mobile phone datasets provided by a major mobile operator in Brazil. Whereas in [4] the methodology was applied to datasets collected during major Brazilian soccer matches, we here apply it to a different type of large-scale event: New Year's Eve celebrations in three large Brazilian cities. Our results could be used to improve the understanding of human mobility in urban areas due to large-scale events, thus contributing to the network management of the mobile phone operators and also to the development of new applications and devices.

II. Related Work

Better understanding the dynamics of the workload imposed on mobile phone networks is increasingly gaining attention from different research efforts [5]. The prediction of user mobility patterns can help, for instance, detecting routine patterns [6], urban planning efforts (e.g., urban traffic planning [7]), and public health management (e.g., disease spread control [8]). Other authors argue that characteristics of the network workload can be exploited to identify the kind of events users are experiencing (e.g., an emergency or a concert) [3].

Two previous studies are particularly related to our project. Batty et al. [9] analyze human dynamics and social interactions of mobile phone users during large-scale events and Calabrese et al. [10] investigate the relationship between different types of events and the home area of the attendees.

We have recently developed a methodology to analyze the workload dynamics of a mobile phone network during large-scale events, with an initial focus on soccer matches [4]. In that work, we characterized how mobile phone users move during part of the day, around the time and location where major soccer matches took place. We analyzed soccer matches in the same location (one soccer stadium) on different days, aiming at comparing the workloads imposed on the carrier's infrastructure (the same set of antennas near the stadium) on days with and without matches. The analysis from a single point of interest also makes it easier to determine how people move towards the stadium before the match starts and where they go afterwards, following an approach similar to [10].

In this paper, we apply our proposed methodology to analyze another kind of large-scale event, i.e. New Year's Eve celebrations, similar to what has been done by [9]. We also include in our study the analysis of three new metrics: (i) call durations, (ii) call inter-arrival, and (iii) call inter-departure times, extending what was done in our previous work [4].

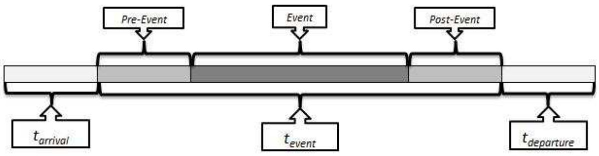


Fig. 1. Timeline adopted in our methodology.

III. Proposed Methodology

Our proposed methodology [4] aims at providing insights r better management and capacity planning of the carrier's frastructure to support the demands of cellphone users due o their mobility during large-scale events. Basically, our methodology is designed with the purpose of answering three main questions: (i) *who moved towards the surroundings of the large-scale event when it took place?* (ii) *where did they come from?* and (iii) *where did they go after the event?*

Towards answering such questions, we restrict our analysis to cellphone calls made during a period of time around the event time. Specifically, we adopt the *timeline* notation shown in Figure 1, which is discussed next.

The first step of our methodology is to identify the antennas that cover the region where the event was held. We then select the users who made at least one cellphone call in one of the selected antennas before and after the event. To that end, we define the *Pre-Event* and *Post-Event* periods as the time intervals that cover the k minutes preceding the beginning of the event and following its end, respectively. We refer to the period from the beginning of *Pre-Event* until the end of *Post-Event*, including the event's total duration, as t_{event}. Users who made at least one call during t_{event} in one of the selected antennas are considered *attendees*. The durations of these time intervals, i.e., determining k, somewhat depends on local aspects and on the event characteristics.

In the second step, we identify the subset of event attendees who also made calls before the beginning (i.e., before *Pre-Event*) or after the end (i.e., after *Post-Event*) of the event. These calls allow us to track the movements of these users before or after the event duration. Thus, our analysis of mobility patterns focuses on these selected users. To identify these users, we define two other periods ($t_{arrival}$ and $t_{departure}$) corresponding to the time intervals before and after the t_{event} period, respectively, during which people are moving towards and arriving at the event place as well as departing from the location.

At this point, the third and last step of our methodology, we use the geolocation of the antennas to locate where each call started and ended, thus determining from where the event attendees (identified in the previous step) came and to where they moved after the event. Ultimately, this enables us to analyze the workload dynamics of the antennas located along the main routes to and from where the large-scale event takes place. To visualize such dynamics, we use *heat maps*[1] to represent the intensity of activity in the mobile phone network (the darker the color in the heat map, the larger the number of calls received by the antenna in this area).

TABLE I. Overview of the Datasets (event days in bold).

City	Day	# Calls done by Attendees	# Attendees	Average Calls per Attendees
BH	**Dec 31, 2011**	**5187**	**1938**	**2.7**
BH	Jan 03, 2012	779	365	2.1
Recife	**Dec 31, 2011**	**9951**	**3566**	**2.8**
Recife	Jan 03, 2012	924	444	2.1
Salvador	**Dec 31, 2011**	**12826**	**7458**	**1.7**
Salvador	Jan 03, 2012	1019	689	1.5
Rio	**Dec 04, 2011**	**4284**	**1754**	**2.4**
Rio	Oct 30, 2011	1270	691	1.8

IV. Results

Our methodology was applied to datasets containing mobile phone calls made during the 2012 New Year's Eve celebrations in three large Brazilian cities: Belo Horizonte (BH), Recife, and Salvador. The datasets contain for each call a unique user identifier[2], the geographical locations (latitude and longitude) of the antennas where the call started and ended, as well as the time instants when it started and ended.

To define the event location and associated antennas, we considered large-scale New Year's Eve celebrations organized in each city, such as a celebration on a beach in Salvador which received, reportedly, one million attendees. Six antennas covered this area. The celebrations in BH and Recife, hosted in an area covered by 3 antennas each, received around 100,000 and 10,000 people, respectively. Regarding the proposed timeline, we considered the total period from 9:45PM to 2:30AM, with the event starting at 11:15PM and lasting for 105 minutes. We set the durations of $t_{arrival}$, $t_{departure}$, *Pre-Event* and *Post-Event* to 45 minutes each. For comparison purposes, we also analyzed other datasets from the same cities, collected on January 3rd, 2012 (a day without event) using the same antennas and timeline. Similarly, we compared these results with those obtained for another type of large-scale event - a soccer match in Rio on December 4th 2011 - reported in [4]. The antennas covering the location (stadium) of the match and the timeline were defined using the same methodology, taking the time and soccer match stadium into account. As basis for comparison, we also analyzed data collected on October 30th corresponding to the same antennas and timeline of the match.

Table I summarizes the analyzed datasets, presenting the total numbers of calls and attendees. Note that the numbers of calls are increased by a factor between 6.7 and 12.6 during the New Year's Eve celebrations (BH, Recife, and Salvador), comparing with the numbers of a day without event at each city. This could be expected if we consider the time (late at night), the event nature, and the huge number of attendees. Note also that the growth factor of the number of calls during the soccer match, comparing with a day with no events in Rio (two last rows of Table I), is lower (factor of 3.4). This indicates the importance of the event nature for better understanding the human mobility and the workload dynamics of the mobile phone network.

To further understand the characteristics of the workload imposed on the selected antennas during each event, Figure 2 shows the numbers of calls made by attendees of the

[1] Heat maps are generated using the Google Maps Javascript API V3.

[2] This id is generated by the cellphone carrier. It is completely anonymized, and thus, cannot be used for identifying the user; although, it allows us to identify multiple calls made by the same user in different points in time.

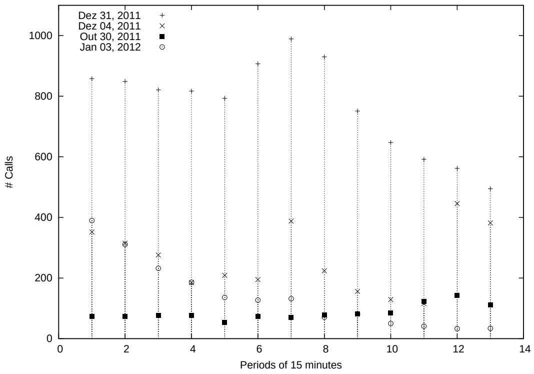


Fig. 2. Calls made by attendees of Recife's New Year's Eve (Dec 31, 2011), at one soccer match in Rio (Dec 04, 2011), and on two days with no large-scale event in Rio (Oct 30, 2011) or in Recife (Jan 03, 2012).

celebration in Recife, in successive 15-minute time bins during t_{event}. For comparison purposes, corresponding results for a day with a soccer match (Dec 04, 2011) and two days without event (Oct 30, 2011 in Rio and Jan 03, 2012 in Recife) are also shown in Figure 2. This figure shows that, for the day of Recife's New Year's Eve, the number of calls peaks at around midnight (bins 6–8), dropping quickly after the celebration finishes. For a soccer match and regular days, instead, the number of calls shows different patterns, decreasing sharply in the ending of t_{event}. Results for the other cities are very similar, being thus omitted.

Call durations, inter-arrival times (IAT), and inter-departure times (IDTs) are key metrics for capacity management and planning as they allow us to assess whether an antenna throughput is meeting its imposed load. We computed all these metrics over 15-minute intervals for each New Year's Eve celebration and soccer matches. In general, all events are typically composed of a large volume of short calls. We also observed that the average of IATs and IDTs are similar for each celebration and slightly different of days with soccer matches. This also points out the difference of the workload imposed on a mobile phone network during events of distinct nature.

Finally, in our methodology, we use heat maps to analyze the mobility of event attendees before, during and after the analyzed events, which allows us to infer the most used access routes to/from the event's location. Figures 3 and 4 show the heat maps produced for the $t_{arrival}$ and t_{event} periods of the New Year's Eve celebration in Salvador. The observation of these heat maps in sequence illustrates human mobility towards the location of the large-scale event.

V. Conclusions and Future Work

In this paper, we extended our recently proposed methodology [4], applying it to analyze human mobility and the workload dynamics of a mobile phone network during large-scale events. Our results show that user behavior patterns that arise during different kinds of events can be used to analyze phenomena related to human mobility due to such events. As future work, we intend to expand our analysis to include other types of large-scale events, like major concerts. We believe that identifying mobility patterns based on cellphone calls during large-scale events can drive the development of target applications and gadgets tailored for such events.

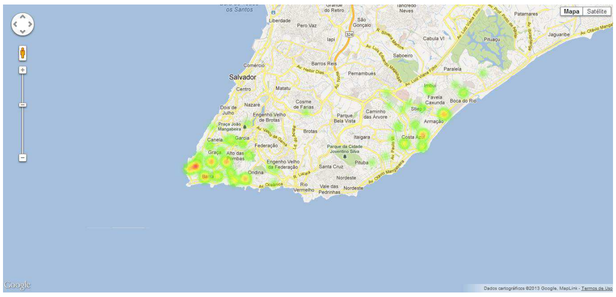


Fig. 3. Heat map of New Year's Eve at Salvador during its $t_{arrival}$.

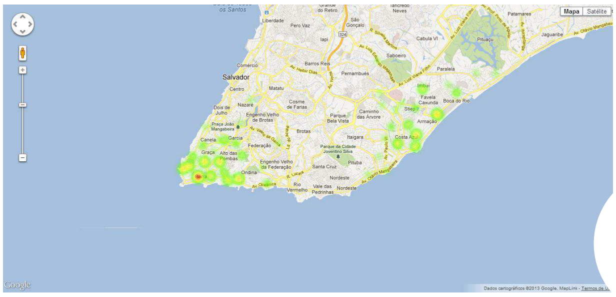


Fig. 4. Heat map of New Year's Eve at Salvador during its t_{event}.

Acknowledgment

This work is supported by FIP-PUC MINAS, InWeb (MCT/CNPq 573871/2008-6), CNPq, CAPES, FAPEMIG, and FAPERJ.

References

[1] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010.

[2] A. Bleicher, "The on-demand olympics," *IEEE Spectrum*, vol. 49, no. 7, pp. 9–10, jul 2012.

[3] J. P. Bagrow, D. Wang, and A.-L. Barabási, "Collective response of human populations to large-scale emergencies," *PLoS ONE*, vol. 6, no. 3, p. e17680, 03 2011.

[4] F. H. Z. Xavier, L. M. Silveira, J. M. Almeida, A. Ziviani, C. H. S. Malab, and H. Marques-Neto, "Analyzing the workload dynamics of a mobile phone network in large scale events," in *Proceedings of the UrbaNe Workshop – ACM CoNEXT 2012*, 2012.

[5] N. Eagle, A. Pentland, and D. Lazer, "Inferring social network structure using mobile phone data," *Proceedings of the National Academy of Sciences (PNAS)*, vol. 106, no. 36, pp. 15 274–15 278, 2009.

[6] M. C. González, C. A. Hidalgo, and A.-L. Barabási, "Understanding individual human mobility patterns," *Nature*, vol. 453, pp. 779–782, 2008.

[7] J. Yuan, Y. Zheng, and X. Xie, "Discovering regions of different functions in a city using human mobility and POIs," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, ser. KDD '12. New York, NY, USA: ACM, 2012, pp. 186–194.

[8] D. Balcan, V. Colizza, B. Gonçalves, H. Hu, J. J. Ramasco, and A. Vespignani, "Multiscale mobility networks and the spatial spreading of infectious diseases," *Proceedings of the National Academy of Sciences*, vol. 106, no. 51, pp. 21 484–21 489, 2009.

[9] M. Batty, J. Desyllas, and E. Duxbury, "The discrete dynamics of small-scale spatial events: agent-based models of mobility in carnivals and street parades," *International Journal of Geographical Information Science*, vol. 17, no. 7, pp. 673–697, 2003.

[10] F. Calabrese, F. Pereira, G. Di Lorenzo, L. Liu, and C. Ratti, "The geography of taste: Analyzing cell-phone mobility and social events," in *Pervasive Computing*, ser. Lecture Notes in Computer Science, P. Floréen, A. Krüger, and M. Spasojevic, Eds. Springer Berlin / Heidelberg, 2010, vol. 6030, pp. 22–37.

The Social Amplifier – Reaction of Human Communities to Emergencies

Yaniv Altshuler, Michael Fire, Erez Shmueli, Yuval Elovici,
Alfred Bruckstein, Alex (Sandy) Pentland and David Lazer

I. Introduction

Imagine a scenario where some set of individuals witness an extraordinary event which impels them to communicate regarding that event to other individuals, who in turn will communicate with yet others. In this scenario, it is possible for an external observer to witness the fact of communication, but not the content. How might that observer effectively make the inference that an extraordinary event has occurred?

This is in fact a plausible scenario, with the existence of communication systems (most notably phones) where timing and volume of traffic is observed, but (typically) not content. Mobile phones are particularly notable in this regard, because of how pervasive they are. Here we build on work examining detection of anomalous events in networks [2], but with the focus on how to aggregate those signals in a computationally efficient fashion. That is, if one cannot observe all nodes and edges, how best to sample the network?

Analyzing the spreading of information has long been the central focus in the study of social networks for the last decade [4], [5]. One of the main challenges associated with modeling of behavioral dynamics in social communities with respect to anomalous external events stems from the fact that it often involves stochastic generative processes. A further challenge is the trade off that exists between coverage and prediction accuracy [1]. While simulations on realizations from these models can help explore the properties of networks [3], a theoretical analysis is much more appealing and robust. The results presented in this work are based on a pure theoretical analysis, validated both by extensive simulations as well as by real world data derived from a unique dataset.

Contribution: In this work we present an innovative approach for studying the network dimension of the changes that take place in social communities in the presence of emergencies. We do so using a mechanism we call a "*Social Amplifier*" – a method for analyzing local sub-networks spanning certain high-volume network nodes. The innovation in our proposed approach is twofold: (a) using a non-uniform sampling of the network (namely, focusing on activity in the social vicinity of network hubs), and (b) projecting the network activity into a multi-dimensional feature space spanned around a multitude of topological network properties. We show using both simulation and real world data that starting certain coverage level of the network, our method outperforms the use of either random sampling, as well as single signal analysis.

II. The Social Amplifier

The proposed method is comprised of three stages as follows.

In the initial stage, we track the traffic volume in the network's nodes, looking for hubs – nodes with high traffic (either incoming or outgoing). The rationale behind the use of hubs is that hubs are highly likely to be exposed to new information, due to their high degree.

Given available resources ϵ, we select network nodes, $v_1, \ldots, v_n$ such that those nodes have the highest degrees in the network and the set $S_M = \bigcup_{1 \le i \le n} E^{1.5}(v_i)$ does not contain more than ϵ portion of the edges, where $E^{1.5}(v)$ denote the 1.5 ego-network around node v, that is – the edges between v and all of v's neighbors, as well as the edges between v's neighbors and themselves.

The use of the 1.5 ego-network is required in order to analyze not only the overall number of calls in the network (sampled by the hubs), as done in works such as [2], but rather to generate the actual networks around the hubs, in order to enable their in-depth analysis. More specifically, analyzing only the overall number of calls, can only detect massive global events, but not local ones (unless the local events are known in advance, and the local data is analyzed in retrospective).

In the second stage, for each day during the test period, and each phone social network, we extract a set of 21 topological features, such as the In Degree, Out Degree, Number of Strong Connected Components, Subgraph Density, etc.

In the third stage, we detect anomalies in the dynamics of the social network around the network hubs, using the Local-Outlier-Factor (LOF) algorithm. Applying the LOF algorithm on each hub, detects days which anomaly features occurred. Then, by using ensemble of all the hubs, we detect which dates have the highest probability for anomaly.

We do so by ranking each day according to the number of hubs that reported it as anomalous. Then, for each day we look at the 29 days that preceded it, and calculate the final score of the day by its relative position in terms of anomaly-score within those 30 days. Namely, a day would be reported

Y. Altshuler, E. Shmueli and A. Pentland are with MIT Media Lab. E-mail: {yanival,shmueli,sandy}@media.mit.edu.

M. Fire and Y. Elovici are with Deutsche Telekom Lab & Department of Information Systems Engineering, Ben-Gurion University. E-mail: {mickyfi,elovici}@bgu.ac.il

A.M. Bruckstein is with Computer Science Department, Technion. E-mail: freddy@cs.technion.ac.il

D. Lazer is with College of Computer and Information Science & Department of Political Science, Northeastern University. E-mail: d.lazer@neu.edu

as anomalous (e.g., likely to contain some emergency) if it is "more anomalous" compared to the past month, in terms of the number of hubs-centered social networks influenced during it. Each day is given a score between 0 and 1, stating its relative "anomaly location" within its preceding 30 days.

III. Validation

A. Analytic Evaluation

Alongside its increased sensing capability, our proposed mechanism has also an additional overhead, in terms of additional edges that should be monitored, compared to the standard approach of "number of calls analysis". This is the result of the following two reasons:

- **Hubs**: Due to their high degree, whenever the edges associated with an additional hub are added to the monitored edges set they increase its size substantially (unlike the addition of a randomly selected node, that is expected to be of a much lower degree).
- **1.5 Ego-Network**: For some node v, although the number of nodes in its 1 ego-network equals exactly the number of nodes in its 1.5 ego-network, the latter is usually expected to have substantially larger amount of edges.

We therefore write the utilization of the Social Amplifier mechanism as follows :

$$E = E_{INITIAL} + E_{AMPLIFIER} + E_{DETECT} \qquad (1)$$

whereas E is the "energy" supplied to the system for monitoring some k edges, $E_{INITIAL}$ is the overhead spent on monitoring the first few hubs until we achieve good topographical coverage of the network, $E_{AMPLIFIER}$ is the energy spent on maintaining a 1.5 ego-network closure (that is, the number of edges of the 1.5 ego-network minus the number of edges at the 1 ego-network), and E_{DETECT} denotes the resources spent on the actual detection of the signal.

We note that $E_{INITIAL}$ decreases with the time it takes the detection process to complete. In other words, as the event to be detected is more explicit and broadly observed, it will be detected using a shorter time, which implicitly increases the relative portion of $E_{INITIAL}$. We can therefore write :

$$E_{INITIAL} \approx \alpha \cdot E$$

for $\alpha \in [0, 1]$ the *exposure coefficient* of the event.

Notice that as the exposure coefficient of an event decreases, it means that additional edges (and nodes) are required in order to detect the event. For extreme low values of the exposure coefficient there is no longer much difference between adding "hubs" and adding random nodes (in terms of their degrees) to the monitored set of nodes. This means that the ratio between the number of edges between hubs' neighbors and the edges to and from the hubs increases, resulting in an increase in $E_{AMPLIFIER}$.

Namely, for high exposure coefficient values the ratio between $E_{AMPLIFIER}$ and E_{DETECT} is proportional to the ratio between the average aggregate degrees of hubs' neighbors and the average degree of the hubs themselves. For low exposure coefficient values this ratio converges to $\frac{1}{<k>}$ (denoting by $< k >$ the average degree of the network) :

$$\frac{\lambda}{k_{MAX}} \leq \frac{E_{AMPLIFIER}}{E_{DETECT}} \leq \frac{\lambda}{< k >}$$

denoting by k_{MAX} the maximal degree, and for $\lambda \geq 1$ being the *Social Amplification Constant* of the network.

The same effect is obtained when the portion of the edges being monitored ϵ changes, as low values for ϵ cause the ratio $\frac{E_{AMPLIFIER}}{E_{DETECT}}$ to decrease, and very high values of it cause it to converge to $\frac{\lambda}{<k>}$. We can therefore write :

$$E_{AMPLIFIER} \approx \frac{\lambda \cdot E_{DETECT}}{< k > + \alpha\epsilon(k_{MAX} - < k >)} \approx$$

$$\approx \frac{\lambda \cdot E_{DETECT}}{< k > (1 - \alpha\epsilon) + \alpha\epsilon k_{MAX}}$$

We shall therefore rewrite Equation 1 as follows :

$$E_{DETECT} = \frac{E \cdot (1 - \alpha)}{1 + \frac{\lambda}{<k>(1-\alpha\epsilon)+\alpha\epsilon k_{MAX}}} \qquad ($$

Figure 1 illustrates the behavior of E_{DETECT} as a function of the changes in the exposure coefficient α and in the portion of edges being monitored ϵ. Notice how E_{DETECT} has a non-monotonous dependency on α, obtaining a global maximum for intermediate values.

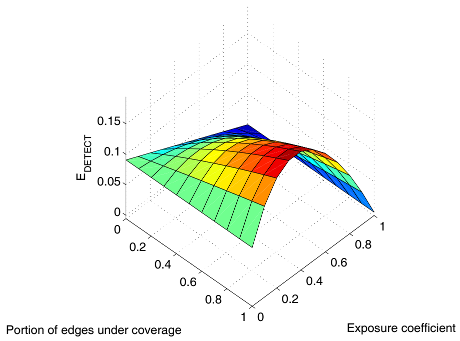


Fig. 1. The dependency of E_{DETECT} on the exposure coefficient α and on the number of edges being monitored. The illustration assumed $k_{MAX} = 10 \cdot < k >$.

B. Simulation

The goal of our simulation was to check how the two methods for selecting the subset of monitored edges (i.e. Social Amplifier vs. Random) influence the time required to detect an event. In order to achieve this goal we simulated the spreading of events in generated scale-free graphs and measured the time taken to detect those events when using the two different methods for selecting the subset of monitored edges.

Our simulation included a tremendously large number of executions ($\approx 10^6$) and used different parameters:

- cp - the coverage percentage of the mobile operator.
- w - the number of initial witnesses to the event.
- c - the confidence level, i.e., the minimum number of "spreading edges" that need to be sensed in order to be confident that an event has occurred.

Following the analysis in Section III-A, we defined the exposure coefficient, denoted by α, as $\alpha = log_2(c)/w$.

Fig. 2 shows the influence of α and cp on $\Delta(\alpha, cp)$ where $\Delta(\alpha, cp)$ is the mean difference in detection time (between the two methods) over all executions with the given cp and $log_2(c)/w = \alpha$. (Note that the original results were smoothened with $R = 0.804$ and $R^2 = 0.646$.) As shown in the figure, for medium α values, the Social Amplifier method outperforms the Random method. In addition, we observe that ı this range of medium α values, the advantage of the Social nplifier method increases with larger cp values.

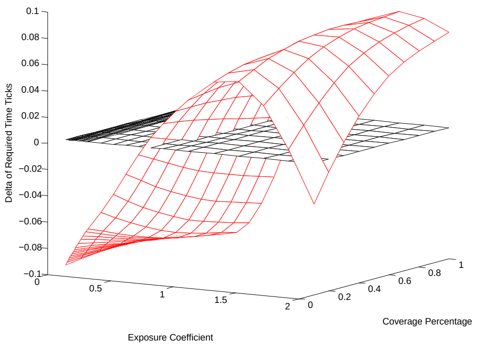


Fig. 2. The influence of α and cp on $\Delta(\alpha, cp)$ as evaluated using simulative environment. $\Delta(\alpha, cp)$ is represented by the red area in the figure. The dark grid represents the fixed $z = 0$ plane. Positive values of $\Delta(\alpha, cp)$ mean an advantage for the Hubs method.

Note that the efficiency of our method as illustrated in Figure 2 closely resembles that of our analytic model, as discussed in Section III-A (Equation 2 and Figure 1).

C. Real World Data

We also validated our method using a comprehensive dataset, containing the entire internal calls as well as many of the incoming and outgoing calls within a major mobile carrier in a west European country, for a period of roughly 3 years. During this period that mobile users have made approximately 12 billion phone calls. We used the company's log files, providing all phone calls (initiator, recipient, duration, and timing) and SMS/MMS messages that the users exchange within and outside the company's network. All personal details have been anonymized, and we have obtained IRB approval to perform research on it.

For evaluating the Social Amplifier technique as an enhanced method for anomalies detection we have used a series of anomalous events that took place in the mobile network country, during the time where the call logs data was recorded.

We have divided the anomalies into the following three groups : (1) "Concerts and Festivals" Events that are anomalous, but whose existence is known in advance to a large enough group of people; (2) "Small exposure events" Anomalous events whose existence is unforeseen, and that were limited in their effect; and (3) "Large exposure events" Anomalous events whose existence is unforeseen, that affected a large population.

For each of the events we used the method described in Section II in order to rank each day between 0 and 1, according to its "anomalousness". This was done for increasingly growing number of monitored edges, in order to track the evolution of the detection accuracy. The result of this process was a series of numeric vectors pairs: $(\mathcal{V}_{BASE}, \mathcal{V}_{AMPLIFIED})_{|E|}$, corresponding to the two networks used (e.g. the random network sampling for $\mathcal{V}_{BASE}$ and the social-amplified hubs-sampling for $\mathcal{V}_{AMPLIFIED}$), for $|E|$ edges which were monitored. In addition, we created a binary vector $\hat{\mathcal{V}}$ having '1' for anomalous days and '0' otherwise.

For $|E|$ edges which were monitored we denote by $\delta_{|E|}$ the difference between the correlation coefficient of $\mathcal{V}_{AMPLIFIED}$ and $\hat{\mathcal{V}}$, and the correlation coefficient of $\mathcal{V}_{BASE}$ and $\hat{\mathcal{V}}$, namely :

$$\delta_{|E|} = CORR(\mathcal{V}_{AMPLIFIED}, \hat{\mathcal{V}}) - CORR(\mathcal{V}_{BASE}, \hat{\mathcal{V}})$$

for $(\mathcal{V}_{BASE}, \mathcal{V}_{AMPLIFIED})_{|E|}$, and for $CORR(x, y)$ the correlation coefficient function.

Notice that whereas $\delta_{|E|}$ measures the delta in detection accuracy, it has somewhat similar meaning to $\Delta(\alpha, cp)$, which measures delta in detection speed.

Figure 3 presents the values of $\delta_{|E|}$ for number of monitored edges between 300 and 800, for the three types of events. Notice how the results strongly coincide with the analytic model as is illustrated in Figure 1, as concerts and events have the highest exposure value a, and the small exposure events have the lowest value.

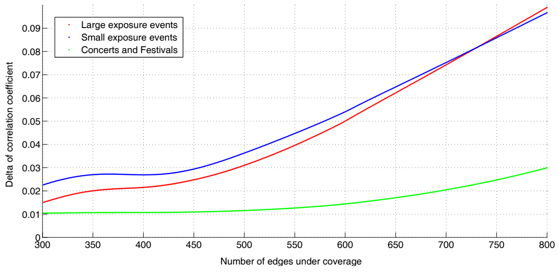


Fig. 3. The changes in the value of $\delta_{|E|}$ for growing numbers of edges being analysed, segregated by the type of event detected. Notice how concerts and festivals that have high exposure value a generate relatively lower values of $\delta_{|E|}$ (but still monotonously increase with $|E|$), while the small exposure events are characterized by the highest values of $\delta_{|E|}$, specifically for low values of $|E|$. It is important to note that a low value of $\delta_{|E|}$ does not imply that the accuracy of the detection itself is low, but rather that the difference in accuracy is small.

References

[1] Altshuler, Y., Aharony, N., Fire, M., Elovici, Y., Pentland, A.: Incremental learning with accuracy prediction of social and individual properties from mobile-phone data. CoRR (2011)

[2] Bagrow, J., Wang, D., Barabási, A.: Collective response of human populations to large-scale emergencies. PloS one **6**(3), e17,680 (2011)

[3] Herrero, C.: Ising model in scale-free networks: A monte carlo simulation. Physical Review E **69**(6), 067,109 (2004)

[4] Huberman, B., Romero, D., Wu, F.: Social networks that matter: Twitter under the microscope. First Monday **14**(1), 8 (2009)

[5] Leskovec, J., Backstrom, L., Kleinberg, J.: Meme-tracking and the dynamics of the news cycle. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 497–506. Citeseer (2009)

Session 8

Social data collection

Detecting Face-to-face Meetings using Smartphone Sensors

Piotr Sapiezynski
s091728@student.dtu.dk

Arkadiusz Stopczynski
arks@dtu.dk

Sune Lehmann
sljo@dtu.dk

Techical University of Denmark

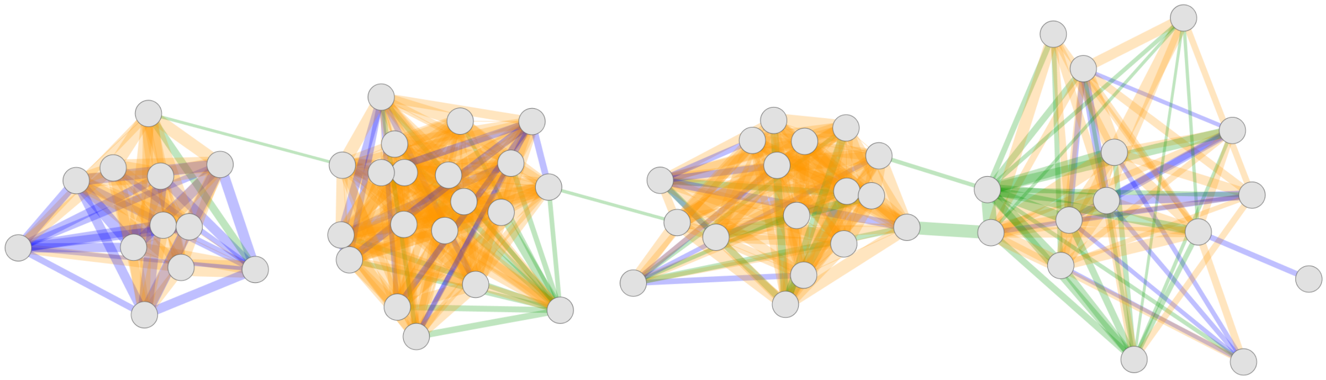


Figure 1: The network composed of 466 most active dyads. 366 of them are among most active both on and outside of the campus (orange edges), while 50 are only significantly active on (blue edges) and 50 outside of the campus (green edges). The campus-centered subset is a good approximation of the whole network, but note that the bridging edges between distinct communities occur only outside of the campus

ABSTRACT

Face-to-face meetings are used in computational social science as one of the most prominent signals for discovering social ties. Different methods have been used for recording such meetings, for example video analysis, infrared sensors, or more recently Bluetooth scanning using mobile phones. In this paper we examine data collected during large computational social science study deployment (N=130 people). The data is recorded using Funf framework[1] running on participants' smartphones. We perform Bluetooth and WiFi scanning, location readings, and cell tower tracing, as well as data collection from a campus-wide WiFi system.

Using a curated Bluetooth signal as a ground truth for face-to-face meetings, we examine how this signal can be recovered from other channels. Our main result is that interactions, which occur on campus (and can thus be discovered using the system WiFi), constitue a relevant approximation of all the interactions among the participants. These findings may be beneficial for 1) discovering co-location networks in contexts, where a WiFi system is already deployed, and using additional mobile devices is not feasible, such as university and company campuses, schools, and other institutions, 2) planning more energy-efficient deployments of experiments involving mobile devices, and 3) understanding how different signals may introduce bias into analysis and conclusions.

Further, this contribution reports on a thorough examination of the data in search of quality deficits, biases, and interplay between the information channels and argue this to be crucial step, before the data can be used to generate and verify scientific hypotheses.

RESULTS OVERVIEW

We use our real world data to show the following:

- Interactions between the students, which can be observed through a campus-only system WiFi network, can be successfully used to estimate the proximity networks emerging between the participants also outside of campus, see Figure 1. This finding indicates a possibility of deploying co-location based studies in environments featuring a WiFi network, without a need for buying additional hardware.
- User perspective WiFi is a rich source of information, because of inherent oversampling (Android forces often WiFi scans whenever WiFi is enabled), giving comparable results accross participants (contrary to GSM towers where the visibility dependents on the provider, users see the same WiFi access points in the same location), and easy recognition of dynamic and static contexts.
- GPS based location estimation has a low recall and precision in face-to-face meeting detection, due to transient nature of contexts where the GPS signal is available.

- GSM towers are a poor medium to study social interaction, due to lack of data about their geographical position, inconsistent identifiers across providers, and very low precision.
- In an experiment where three phones located close to each other periodically scan Bluetooth environment, a successful diadic discovery between them occurs with a rate of approx. 75%. Thus, Bluetooth scan results cannot be directly treated as ground truth without sensible preprocessing.

REFERENCES

1. Aharony, N., Pan, W., Ip, C., Khayal, I., and Pentland, A. The Social fMRI: Measuring, Understanding, and Designing Social Mechanisms in the Real World. In *Proceedings of the 13th International Conference on Ubiquitous Computing*, ACM (2011), 445–454.

Vehicular Traffic Estimation Leveraging Location Area Updates of Mobile Phones

Andreas Janecek
Research Group Entertainment Computing
University of Vienna, Austria
andreas.janecek@univie.ac.at

Danilo Valerio
Telecommunications Research Center (FTW)
Vienna, Austria
valerio@FTW.at

Karin A. Hummel
Communication Systems Group
ETH Zurich, Switzerland
karin.hummel@tik.ee.ethz.ch

Fabio Ricciato
Telecommunications Research Center (FTW)
Vienna, Austria
fabio.ricciato@FTW.at

Helmut Hlavacs
Research Group Entertainment Computing
University of Vienna, Austria
helmut.hlavacs@univie.ac.at

1. INTRODUCTION

Being a large-scale ubiquitous mobility sensor, the mobile cellular network provides valuable data for human mobility modeling [3] and vehicular traffic analysis. So far, road traffic has been mainly monitored by means of static sensors or derived from floating car data. These approaches are either costly or significantly suffer from the low fraction of vehicles captured.

In our study, we analyze the mobile cellular network signaling data to infer road status and, specifically, congestion events in real-time. One major novelty compared to most previous studies on the topic [5] that considered only mobile phone data related to 'active' terminals (Call Detail Records), is the use of the more complete signaling data provided by the network links near the Radio Access Network (RAN) of the cellular network. This way, also the position of 'idle' terminals can be observed on Location Area (LA) level. On the downside, the LA level provides only coarse grained location information. However, the overwhelming majority of the mobile terminal population can be included leading to a much better coverage of moving vehicles. This property is crucial, as we have shown in [2, Section 4] that the exclusive observation of 'active' terminals produces a biased picture of the overall human mobility.

By validating our approach against four different data monitoring datasets on a sample highway in Vienna, Austria (see Figure 1), over one month, we show that our method can indeed detect congestions very accurately and in a timely manner. The cellular dataset used comprises anonymized mobile cellular data of the signaling traffic from 2G/3G cells of a real operational network over 31 days. For details of our work, refer also to [4].

2. METHODOLOGY

Our method builds mainly on leveraging location area and routing area (for simplicity, we use LA to refer to both) updates to detect congestion events; this part is termed 'stage 1'. Additionally, we introduced also a second stage ('stage 2') to include cell handovers of 'active' terminals in order to localize congestions events more precisely [4]. This second stage has the potential to differentiate between congestions types as identified in road traffic research [6].

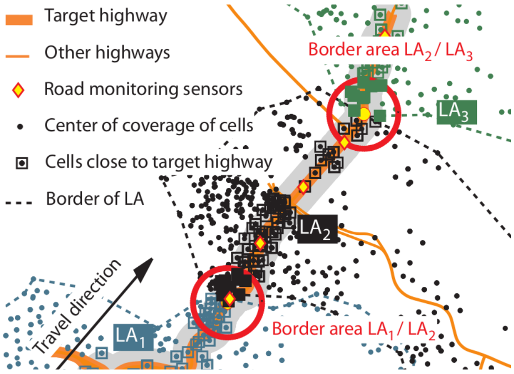

Target highway
Other highways
Road monitoring sensors
Center of coverage of cells
Cells close to target highway
Border of LA
Border area LA$_2$ / LA$_3$
LA$_3$
LA$_2$
Travel direction
LA$_1$
Border area LA$_1$ / LA$_2$

Figure 1: Excerpt of the sample highway, Vienna, Austria: topography, cellular network, and traditional road sensors. Devices are tracked across LA$_2$ based on events triggered in the entry cells (south) of LA$_2$ (start border area LA$_1$/LA$_2$) and the entry cells (south) of LA$_3$ (arrival border area LA$_2$/LA$_3$).

The stages differ in type and amount of cellular network signaling events that are considered, and in terms of length and granularity of the considered road section. Stage 1 offers a continuous estimation of the travel times at large-scale with good terminal coverage but with spatial accuracy limited by the large radii of LAs (up to several kms), in urban areas comparatively small (5-10 km on our target highway). Stage 2 offers higher spatial accuracy (up to few hundred meters) but less terminal coverage. We remark that, in general, it is not sufficient to rely exclusively on active terminals as the coverage is too low. Stage 2 can be used to further localize the area of congestion, but only after the congestion has been detected.

We will now focus on stage 1. The calculation of the travel time t of one mobile terminal is processed as follows: Every time the mobile terminal attaches to a cell belonging to a new LA, it emits an LA update event. In order to identify the terminals traveling along a target road segment, a suitable set of cells is pre-selected meeting the following criteria: (i) the cells are located at the border of two LAs and (ii) they are in close proximity of the target highway. We

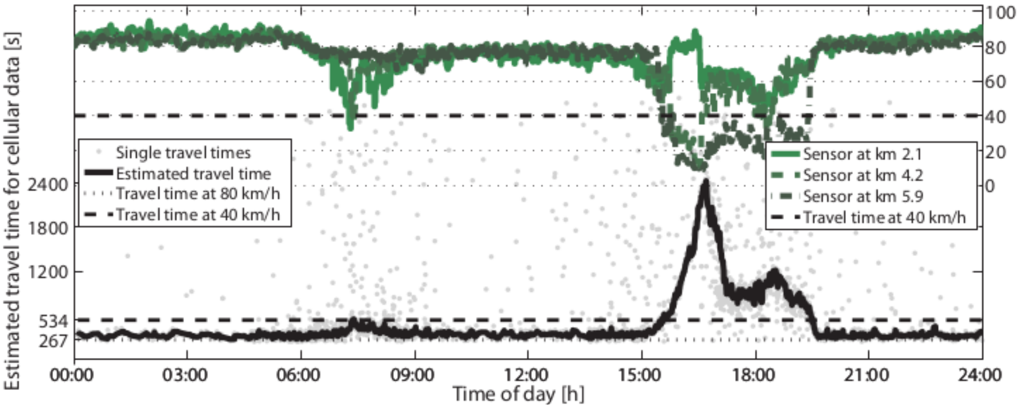


Figure 2: Visualization of one day (June 30th, 2011): comparison of estimated travel times for mobile cellular data and speed measured through fixed sensors during the day.

term these cells *entry cells*. Let now LA_1, LA_2, and LA_3 be three adjacent LAs crossed in sequence by a mobile terminal. The set of entry cells of LA_2 indicating a change from LA_1 to LA_2 is termed *border area* LA_1/LA_2while the set of entry cells of LA_3 indicating a change from LA_2 to LA_3 is termed LA_2/LA_3.The mobile terminal (vehicle) traveling from LA_1 to LA_3 will generate at least one event in LA_1/LA_2 and one in LA_2/LA_3. The travel time estimate of the terminal is now simply calculated as $t = t_a - t_s$, where t_a is the time of the first event in LA_2/LA_3 and t_s is the time of the last event in LA_1/LA_2. Hereby, we assumed that the highway is the fastest connection between the start and arrival area, i.e., the fastest mobile device users are all traveling on the target highway (see Figure 1).

A congestion is detected by evaluating the travel time of fastest users. Due to the nature of the cellular network and to the heterogeneity of mobile terminals, one cannot predict exactly at which position a cell change occurs, and thus, the length of the segment under investigation. As a consequence, it is difficult to determine the minimum travel time. Instead of using a possibly imprecise static estimate for the minimal travel time, we adapt the set of fastest users by introducing an adjustable parameter κ: The travel time of the κ-quantile of all users is used to estimate the minimum travel time.

3. DATASETS

We apply or methods to anonymized traces of 2G/3G signaling traffic observed in an operational cellular network of a major Austrian mobile operator. We evaluate our results by comparing them to results achieved by major traditional sources for traffic estimation.

Cellular data

The anonymized traces consist of 400–500 million events on average per day, and contain signaling messages for both the packet switched and circuit switched domains. To preserve user privacy, all sensitive identifiers are removed from the traces. Distinct users are discriminated only by means of pseudonyms computed via a one-way hash function. The pseudonyms are changed every 24 hours. The event-based tickets used contain information such as: `anonymous_ID` of the user generating the event, `timestamp` of the event, cell information including the `LAC` (Location Area Code) and `Cell_ID`, the `coordinates`, `type`, etc. of the base station antenna, information about the event such as `type_of_event`, etc.

Validation data

We validate the results of our approach against various datasets originating from traditional road monitoring sources. We use road *sensors* as point-based road monitoring data (measuring speed of passing vehicles, in our investigation nine road sensors covered the highway and the measurement frequency is once per 60 seconds) and *toll* gantries monitoring trucks via RFID-based transponders to estimate the average travel time of trucks between two toll gantries (every 15 minutes). Further, *taxi* floating car data provide GPS positions of taxis and, thus, average taxi speed. Additionally, we use *radio* events, that are radio broadcasts about road incidents reported by registered drivers.

4. ANALYSIS RESULTS

The sample period stretches over 31 days. While the mobile cellular data are available for all days, the validation data are only partially available due to the nature of the sensors or temporary faults: *toll* data are partly not available on weekends (due to ban of trucks), *sensor* data are missing for eight days (due to a problem in the recording system), and concerning the *taxi* data, only half of period is provided. *Radio* data are available throughout the whole period. The target highway stretches over 32 km from a rural area into the center of Vienna, Austria. It is covered by four different LAs.

Detecting congestions

As there is no general agreement in the literature about the definition of a traffic congestion [1], we adopt the following definition: a road segment is marked as congested if the estimated speed of the fastest vehicles considered falls below half the speed limit (i.e., the travel time is doubled). We mark a congestion event in our dataset if *at least one data source* out of *toll*, *sensors*, or *taxi* triggers the above condition. In this way, 74 congestion events could be identified over the sample period, and 58 of them were sent as broadcast also via radio. All events in the *radio* dataset could be detected by at least one of the other three validation sources. We refer to our approach with the term *mobile*. We analyze the quality of congestion estimation based on cellular data vs. other data sources in terms of the *number of detected congestions* and the *estimation delay*.

Let us first exemplify the analysis with a comparison related to a single congestion. Figure 2 shows the estimated travel times through a given area on our target highway over a whole day for cellular data vs. the speed estimates for fixed sensor data. A severe congestion is visible in the afternoon and a less severe in the morning. The morning and afternoon incidents were also visible in toll and taxi data, the incident in the afternoon was broad-casted on radio at 15:33 h ('heavy traffic').

Our results of the full month period are summarized in Table 1. The average advance ("-") or delay ("+") of *mobile* against each validation data source is reported. κ determines the fraction of users considered as being in the set of fastest users (upper κ-quantile of all users). Here, smaller values of κ (0.03 or 0.04) make the approach aggressive as it identifies congestions fast, but with an increase in false positive classifications. Setting $\kappa = 0.05$ is a good trade-off between FPs, FNs and advance over other data sources.

Table 1: Delay of *mobile* vs. validation data [in s]: number of comparable events (i.e., identified by both data sources) is given n brackets. Correct, false positive (FP), and false negative (FN) lassifications are shown at the bottom of the table.

κ	0.03	0.04	**0.05**	0.06	0.07	0.08	0.09
Toll:	-533 (66)	-282 (66)	**-152** (65)	0 (62)	+72 (62)	+175 (61)	+196 (61)
Sensors:	-586 (25)	-377 (25)	**-259** (24)	-93 (22)	-15 (22)	+37 (22)	+79 (22)
Taxi:	-583 (25)	-383 (25)	**-311** (25)	-182 (24)	-113 (24)	-11 (23)	+7 (23)
Radio:	-451 (58)	-195 (58)	**-177** (57)	+5 (57)	+77 (57)	+197 (56)	+217 (56)
Numbers of identified congestions for *mobile* data							
Correct:	74	74	**73**	70	70	68	68
FN / FP:	0 / 65	0 / 19	**1 / 3**	4 / 2	4 / 1	6 / 1	6 / 0

5. CONCLUSION

We presented a novel approach for estimating vehicular travel times based on anonymous location area updates of mobile phones collected from an operational mobile cellular network. Although spatially coarse, mobility data from all terminals (most of which in idle state) can be exploited to detect speed deviations at long road sections and congestion events. Experiments on a major traffic route in Austria showed that our method yields higher detection success rates and at least a similar detection delay when compared to traditional road monitoring sensory technologies. Thus, we conclude that cellular network data are a valuable source for accurate traffic monitoring. At the same time, our approach does not require investments in a new infrastructure but leverages the mobile cellular network – as a large-scale mobility sensor.

Acknowledgments

We acknowledge the support of the FTW project RoadCell by the partners ASFiNAG, A1 Telekom Austria, Kapsch CarrierCom, and Nast Consulting. Further, we acknowledge the support of the Austrian Government and the City of Vienna under the COMET program and the Commission of the European Union under the FP 7 Marie Curie project MOVE-R, contract PIEF-GA-2010-276336.

6. REFERENCES

[1] R. Bertini. You are the traffic jam: an examination of congestion measures. Technical report, Department of Civil and Environmental Engineering, Portland State University (2005), 2005.

[2] P. Fiadino, D. Valerio, F. Ricciato, and K. A. Hummel. Steps towards the extraction of vehicular mobility patterns from 3G signaling data. In *Traffic Monitoring and Analysis (TMA) Workshop*, 2012.

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

[4] A. Janecek, K. A. Hummel, D. Valerio, F. Ricciato, and H. Hlavacs. Cellular data meet vehicular traffic theory: Location area updates and cell transitions for travel time estimation. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp '12, pages 361–370, 2012.

[5] G. Rose. Mobile phones as traffic probes: Practices, prospects and issues. *Transport Reviews*, 26(3):275–291, 2006.

[6] M. Treiber, A. Kesting, and D. Helbing. Three-phase traffic theory and two-phase models with a fundamental diagram in the light of empirical stylized facts. *Transportation Report Part B*, (44):983–1000, 2010.

Emergence of Congestion In Road Networks based on Realistic Demand obtained from Mobile Phone Data

by Serdar Çolak[1], Christian M. Schneider[2], Pu Wang[3], Marta C. González[4]
Emails: serdarc@mit.edu, puwang@mit.edu, schnechr@mit.edu, martag@mit.edu
[1,2,4]Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, USA
[3]School of Traffic and Transportation Engineering, Central South University, China

The first requirement a road network needs to fulfill is overall connectivity: there must be an adequately functioning path between any two places. Roads have limited capacity and queues of vehicles accumulate; therefore it is of great interest to study the effect of both network topology and travel demand on traffic flow leading to the emergence of congestion. Upon a certain threshold in the number of vehicles, the capacity of the roads is exceeded, retarding the efficient functioning of the whole network. In analyzing congestion, estimation of the demand is of prime importance.

Transition to congestion emerges for the rate of trips starting per time step, R, exceeding a certain threshold R_c. A network is considered to remain functional if an equilibrium in the number of vehicles travelling at any time is reached. First, we study the influence of the network topology on the emergence of congestion. Inspired by models investigating network resilience in the context of information packets and of the Internet, we propose a model to analyze the resilience of urban road networks described as follows: At each time step R vehicles with assigned sources and destinations enter the system. Roads have different capacities for delivering vehicles, at each time step every segment can deliver at most C vehicles one step towards their destinations following a fixed routing table. A vehicle, upon reaching its destination, is removed from the system.

We are interested in the critical value R_c, measured by the number of trips beginning at each time step, at which a phase transition takes place from free flow to congested traffic. This critical value reflects the network's capability of handling its traffic demand. Particularly, for $R < R_\mathrm{c}$, the numbers of starting and completed trips are balanced, leading to a free traffic flow. For $R > R_c$, traffic congestion occurs as the number of accumulated vehicles increases with time simply because the capacity of the roads is exceeded.

The travel time dimension is incorporated into this model using a point-queue (PQ) macroscopic link model. Vehicles move along the road at the speed limit before they reach its end where a point queue is formed if the traffic arriving is greater than the capacity at the exit. The PQ model is enhanced to a spatial PQ model (SPQ) by limiting the number of vehicles a road segment can contain at once.

We analyze both theoretical and real road networks by performing simulations on periodic and non-periodic lattices, random networks as well as on the real San Francisco road network. Our analytical and numerical findings indicate that the critical point of the transition is determined by the ratio of the capacity to a modified maximum betweenness centrality. Moreover, the network response can be analytically obtained by iteratively solving a set of coupled equations. We also show that at R_c, the timespans during which the congested road operates at its flow capacity exhibit no characteristic time scale, as the timespan distribution follows a power law with an exponent around -0.5.

To estimate realistic demand, we use a dataset of mobile phone activity for the Bay Area, California. Using the tower scale and the timestamps we generate trajectories, which are then matched onto the road network. The trajectories are scaled using census data and aggregated to form realistic origin-destination matrices. The real demand distribution is applied on the San Francisco road network to analyze congestion and its emergence. It is found that the emergence of congestion based on demand distributions obtained from mobile phone users arises not from lack of outflow capacities of the road segments but from lack of volume capacities, namely, the number of vehicles that can simultaneously use the segment. This suggests that congestion is inherent in downtown areas where road segments are short and dense.

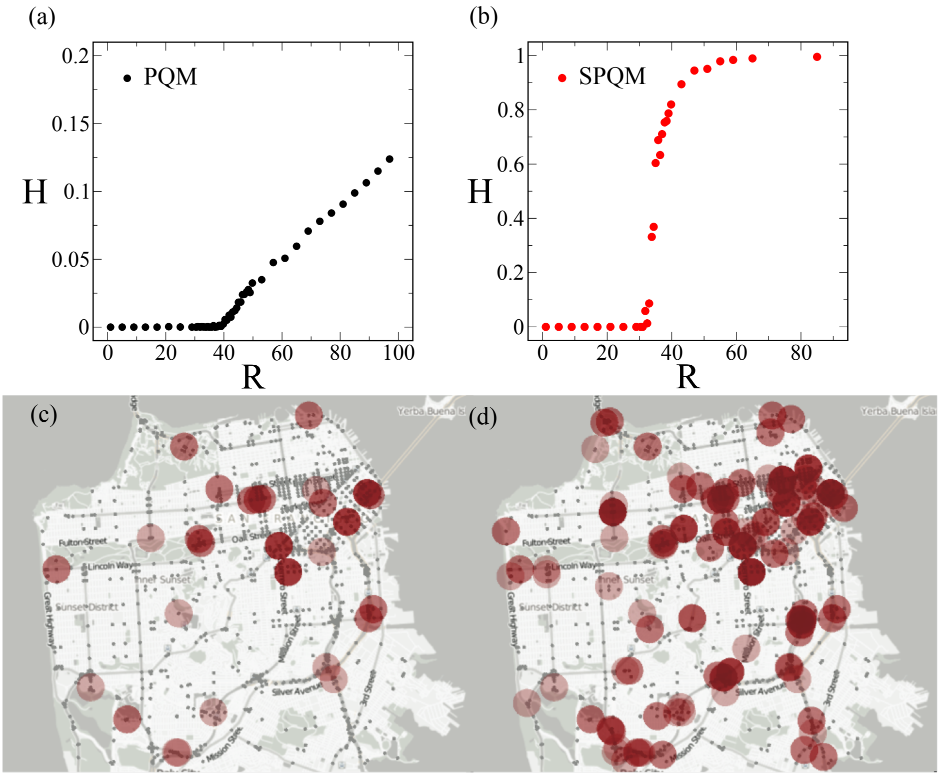


FIGURE. The transitions for the San Francisco road network for **(a)** PQM with R_c=40 (14400 vehicles/hr) and **(b)** SPQM with R_c=32 (11520 vehicles/hr). **(c)** Network response as red points representing the locations of the queues of links that are above 90% of their volume capacities at the time of this snapshot, for R=0.98R_c and **(d)** R=1.02R_c, respectively. A slight increase in travel demand results in an explosion of queues in the downtown area, and the spillback of congestion to the arteries that enter the downtown area can be observed.

Indicators of wealth, economic diversity and segregation in Côte d'Ivoire using Mobile Phone datasets

Thoralf Gutierrez[1,*], Gautier Krings[1,2] and Vincent D. Blondel[1,†]

[1] Institute for Information and Communication Technologies, Electronics and Applied Mathematics (ICTEAM), Université catholique de Louvain, Avenue Georges Lemaître, 4, 1348 Louvain-la-Neuve, Belgium
[2] Real Impact Analytics, 30 Grand Rue, 1660 Luxembourg, Luxembourg
[*] thoralf.gutierrez@student.uclouvain.be,
[†] vincent.blondel@uclouvain.be

In this study, we demonstrate how mobile phone datasets can be used to generate indicators of wealth, economic diversity and economic segregation in the African country of Côte d'Ivoire. We are missing statistical information for many African countries. We illustrate in this paper how mobile phone data sets can be used to infer rough economic parameters without having to carry large and expensive surveys.

We use anonymized Call Detail Records (CDRs) and top-up history from a major cellular network company for a period of seven months[‡]. Most mobile phone subscriptions in Côte d'Ivoire are pre-paid subscriptions. The top-up history of a user gives a complete description of every credit addition on the user's subscription. The top-up amounts and frequency may be used as an indicator of a user wealth. The CDRs include, for each call or text message, the caller and callee number, the location of the tower where the call originated, the date and hour of the call/text message and, if it is a call, its length in seconds. These CDRs allows us to use the method described in [1] to localize a user's home cell and to construct a social graph of users.

We notice that a user's top-up history is often quite stable; users tend to buy credit in chunks of the same amount. This stable behavior allows us to assign an average top-up amount to every user without losing much information. We hypothesize that the wealth of a user is correlated with his or her recharging habit: someone richer will be able to buy larger amounts of phone credit at a time. Further analysis would be needed to assess the quality of this indicator.

We represent on Figure 1 the average user top-up in the different regions of the country. We may also quantify the wealth diversity in the regions by computing the variance of the average top-up (Figure 2). As expected, the region of Abidjan is the wealthiest and most diverse. Some border regions also stand out as wealthier than the rest of the country but with a lower diversity. The administrative capital Yamoussoukro stands out in terms of its economic diversity but not by its wealth. Finally, the typical rural regions are poor and have limited diversity.

After constructing a social graph based on the CDRs, we identify communities in the social graph using the Louvain Method described in [2]. We then look at how many people have an average top-up close to their community's average top-up. This gives an indication of how economically diverse

[‡] Although the dataset is for Côte d'Ivoire, this is not the dataset used for the D4D challenge organized by Orange. The data comes from another mobile phone operator in Côte d'Ivoire.

communities are. In order to give a representation of this diversity in the country, we consider only those communities whose members all live in the same region and we compute for every region the average of diversity of all communities in the region. The result is represented on Figure 3. We find that regions with high variability in top-up behavior often show higher levels of diversity within communities. In other regions, high variability in top-up behavior is accompanied by low diversity within communities. This may be seen as an indication of segregation between social classes in the corresponding regions.

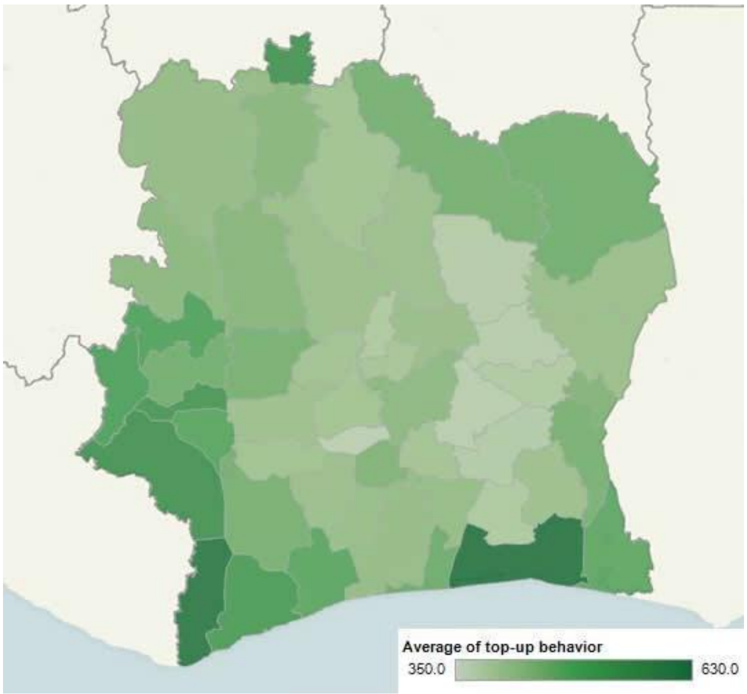


Figure 1 – Average of top-up behavior for each region, which we interpret as an indicator of wealth for each region.

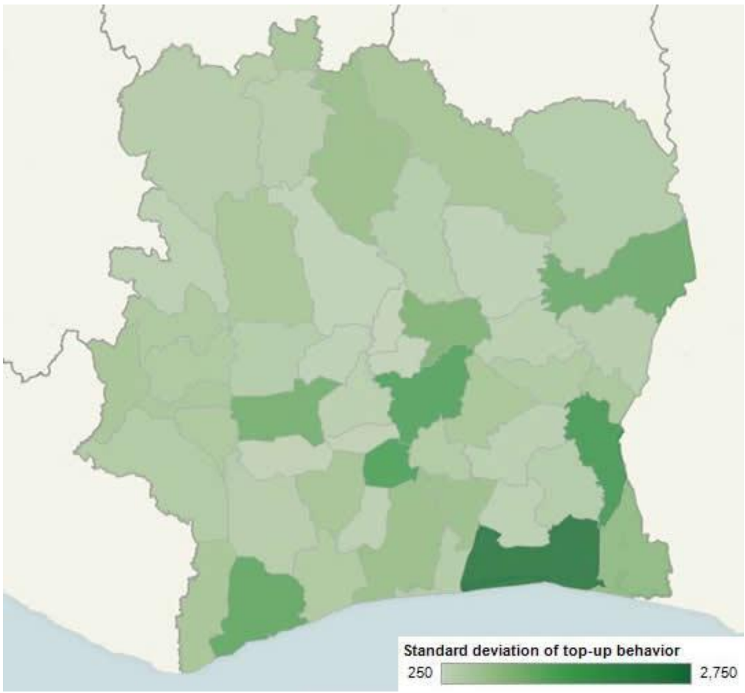


Figure 2 – Standard deviation of top-up behavior for each region, which we interpret as an indicator of wealth diversity for each region.

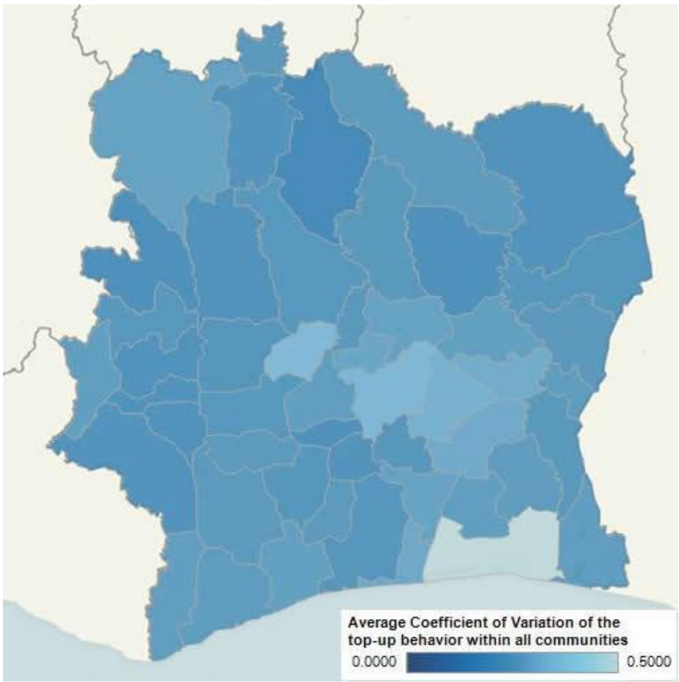

Average Coefficient of Variation of the
top-up behavior within all communities
0.0000
0.5000

Figure 3 – Average Coefficient of Variation (CV) of the top-up behavior within all communities in each region, darker means that communities are less economically diverse.

References

[1] Sibren Isaacman, Richard Becker, Ramón Cáceres, Stephen Kobourov, Margaret Martonosi, James Rowland, Alexander Varshavsky "Identifying important places in people’s lives from cellular network data." *Pervasive Computing* (2011): 133-151.

[2] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte and Etienne Lefebvre "Fast unfolding of communities in large networks."*Journal of Statistical Mechanics: Theory and Experiment* 2008.10 (2008): P10008.

[3] Jukka-Pekka Onnela, Samuel Arbesman, Marta C. González, Albert-Lászlό Barabási, Nicholas A. Christakis "Geographic constraints on social network groups." *PLoS one* 6.4 (2011): e16939.

[4] Ott Toomet, Siiri Silm, Erki Saluveer, Tiit Tammaru, Rein Ahas, "Ethnic Segregation in Residence, Work, and Free-time Evidence from Mobile Communication." (2011).

Can Cell Phone Traces Measure Social Development?

Vanessa Frias-Martinez *, Victor Soto *†, Jesus Virseda *‡, Enrique Frias-Martinez *
* Telefonica Research, Spain † Columbia University, USA ‡ Carlos III University, Spain
vanessa,vsoto,jvjerez,efm@tid.es

Abstract—Census maps contain important socio-economic information regarding the population of a country. Computing these maps is critical given that policy makers often times make important decisions based upon such information. However, the compilation of census maps requires extensive resources and becomes highly expensive, especially for emerging economies with limited budgets. On the other hand, the ubiquitous presence of cell phones, both in developed and emerging economies, is generating large amounts of digital footprints. These footprints can reveal human behavioral traits related to specific socio-economic characteristics. In this paper we propose a new tool, *CenCell*, to approximate census information from behavioral patterns collected through cell phone call records. The tool provides affordable census information by accurately classifying socio-economic levels from cell phone call records with classification rates of up to 70%.

I. Introduction

Census maps gather large amounts of information regarding the socio-economic status of households at a national scale. These maps contain information that characterizes various social and economic aspects like the educational level of the citizens or the access to electricity. Such information is aggregated and reported at various granularity levels, from a national scale, to states, all the way down to urban geographic areas of a few square kilometers. The accuracy of these maps is critical given that many policy decisions made by governments and international organizations are based upon variables measured through census maps. National Statistical Institutes compute such maps every five to ten years, and typically require a large number of enumerators that carry out interviews gathering information pertaining the main socio-economic characteristics of each household. All these prerequisites make the computation of census maps highly expensive, especially for budget-constraint emerging economies. To reduce costs, countries have made cuts both in the number of interview questions and in the number of citizens interviewed, which unfortunately impacts the quality of the final census information.

On the other hand, the ubiquitous presence of cell phones in emerging economies is generating large datasets of digital footprints. Data mining techniques applied to such datasets can be used to reveal cell phone usage patterns specific to socio-economic levels. Previous research has already shown that cell phone-based behavioral patterns might be correlated to specific socio-economic characteristics [1], [2]. For example, Eagle *et al.* showed correlations between the size of a cell phone social network and the socio-economic level of a person, and Frias *et al.* observed strong relationships between mobility and socio-economic indices [3].

In this paper, we propose a new tool for governments and policy makers that allows to compute affordable census maps by decreasing the number of geographical areas that need to be interviewed by the enumerators. The tool, called *CenCell*, is designed to allow institutions to approximate the census information of areas not covered by the enumerators using anonymized cell phone call records gathered by telecommunication companies. At its core, *CenCell* consists of a classification algorithm that determines the socio-economic level of a region based on the aggregated cell phone behavioral patterns of its citizens. Thus, *CenCell* significantly decreases the workload of the enumerators that carry out the interviews and as such, allows to reduce the budget allocated for the computation of census maps.

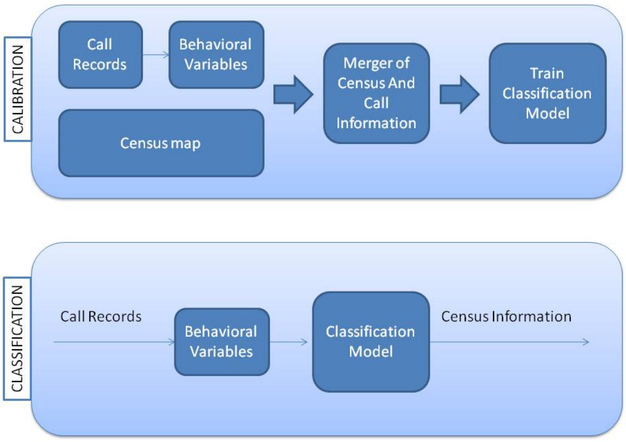


Fig. 1. *CenCell* Architecture.

II. CenCell: General Architecture

Figure 1 shows the general architecture of the tool. It consists of two main components: (1) the *calibration* phase, which needs to be executed only once to set up the system for a region; and (2) the *classification* phase, which is executed every time census information is required for a specific geographical area in the region that was not covered by the enumerators through household surveys.

The *calibration* phase needs two datasets, one containing anonymized cell phone call records for the region under study and another one containing the regional socio-economic levels computed by the local National Statistical Institute through household surveys. This phase first computes a set of cell phone usage behavioral patterns from call records. Next, it combines both datasets to obtain a map that associates to each cellular tower in the region under study a set of cell phone behavioral variables and a socio-economic level. This map is used to train a classification model that will output the socio-economic levels of the areas not covered by the enumerators based on the cell phone behavioral patterns of its citizens (*classification* phase). It has to be noted that *CenCell* uses anonymized aggregated patterns of behavior so no individual information is used to build the classification models.

During the *calibration* phase *CenCell* tests the classification accuracy of a battery of supervised and unsupervised algorithms and selects the one with the best results, which will be used in the *classification* phase. Previous work explored the use of supervised techniques (SVMs and Random Forests) to forecast socio-economic levels from cell phone records [2]. However, given that *CenCell* computes the socio-economic levels (SELs) for each cellular tower based on a weighted average of overlay maps, the final values might be smoothed or blurred depending on the information distribution in the original maps. In an attempt to overcome this problem, *CenCell* also explores unsupervised techniques to identify groups of cell phone behaviors without prior knowledge of their socio-economic values. Sections IV and V cover details about the different techniques explored by *CenCell* and its results. Additionally *CenCell's calibration* phase, computes classification models for different socio-economic level granularities. Although the SEL is a continuous variable, it is often times expressed as a discrete value through a letter (A, B, C, etc.). The granularity of the SELs *i.e.,* the number of SEL classes in which the continuous values are divided into, varies a lot across studies. Some researchers differentiate three socio-economic levels while others prefer to use a larger range of values in their analyses. To account for this need, *CenCell* outputs the best predictive technique for each granularity value in the *calibration* phase. As such, *CenCell* provides a *knob* that allows researchers to select a specific granularity depending on the classification error they are willing to accept. In Section V, we delve more into results and implications of this approach. Finally, it is important to highlight that in order to build accurate classification models the tool needs that: (i) the area selected for the *calibration* phase is representative of the different socio-economic levels of the country and (ii) both call records and socio-economic variables correspond to a similar period in time.

On the other hand, the *classification* phase uses the models generated by the *calibration* phase to compute the socio-economic level of the geographical areas in the region that were not covered by the enumerators. This phase, which only requires access to anonymized aggregated calling records, can be executed as many times as needed and allows policy makers to compute affordable census maps without the need to held household surveys across all the region under study but rather in a few areas. Next, Sections III and IV describe the *calibration* phase in detail.

III. Datasets and Information Merging

Call Detail Records

Cell phone networks are built using a set of base transceiver station (BTS) towers that are responsible for communicating cell phone devices within the network. Each BTS or cellular tower is identified by the latitude and longitude of its geographical location. The area of coverage of a BTS can be approximated using Voronoi tesselation. Call Detail Records (CDRs) are generated whenever a cell phone connected to the network makes or receives a phone call or uses a service (e.g., SMS, MMS). In the process, the BTS details are logged, which gives an indication of the geographical position of the user at the time of the call. From all the information contained in a CDR, *CenCell* only considers the encrypted originating number, the encrypted destination number, the time and date of the call, the duration of the call, and the BTS that the cell phone was connected to when the call took place. Using call detail records, *CenCell* computes three sets of variables per subscriber so as to model cell phone usage: (1) consumption variables; (2) social network variables and (3) mobility variables. The *consumption variables* characterize the general cell phone use statistics, measuring, among others, the number of input or output calls, the duration of the calls or the expenses. The *social network variables* compute measurements relative to the social network of each subscriber. These variables include the input and output degree of the social network, the social distance between contacts (*diameter of the social network*) or the strength of the communication ties. Finally, the *mobility variables* characterize the geographic areas where a person typically spends most of his/her work and leisure time as well as the spatio-temporal mobility patterns. In total, *CenCell* computes 69 consumption variables, 192 social network variables and 18 mobility variables.

Census Data

CenCell uses the census maps collected by National Statistical Institutes (NSI) to gather socio-economic information about the population under study. NSIs carry out individual and household surveys at a national level every five to ten years. These surveys employ a large staff of enumerators that are responsible for interviewing every household head within their assigned geographical area. The enumerators have been especially trained to be able to gather all the required information in a proper manner. Although in some cities in emerging economies the census information is collected with laptops, in general, paper survey forms are still very common, which makes the collection process even more expensive and time consuming. Given the private nature of the individual census information, NSIs only make public average values over specific geographical areas. These areas might represent states, cities, neighborhoods or *geographical units (GUs)*, the smallest geographical division which divides cities into small areas of up to a few squa kilometers (approximately blocks). The census variables gather by NSIs are usually divided into three groups: *education variabl* *demographic variables* and *goods' ownership variables*. With th information NSIs compute the socio-economic level (SEL) of a region as a weighted average of all the census variables. As mentioned earlier, SEL values are typically represented as a set of discrete values typified through letters.

Merging Call Records with Census Data

The objective of the *calibration* phase is to build a socio-economic classification model from cell phone records. To do so, we need to compute a training set that associates to each BTS cell tower: (1) aggregated cell phone behavioral variables for the citizens that live within the cell tower coverage area and (2) the corresponding socio-economic level (SEL) for that same area. However, given that call records are gathered per BTS area and that the census information is reported per geographical unit (GU), the tool uses a three step protocol to merge the two sources of information [4]: (Step 1) Associate the residential location of each subscriber to a BTS area; (Step 2) Compute the overlapping between Voronoi diagrams and Geographic Units (GUs). This mapping allows us to merge census and BTS maps so as to associate socio-economic levels to each BTS coverage area; and finally, (Step 3) For each BTS, compute the aggregated consumption, social and mobility variables of all the subscribers whose residential location is within that area. These three steps produce a map that associates to each BTS a socio-economic level as well as a set of variables (features) characterizing the average cell phone usage for that area. Next, we explain the machine learning techniques used in the *calibration* phase.

IV. Classification Model Generation

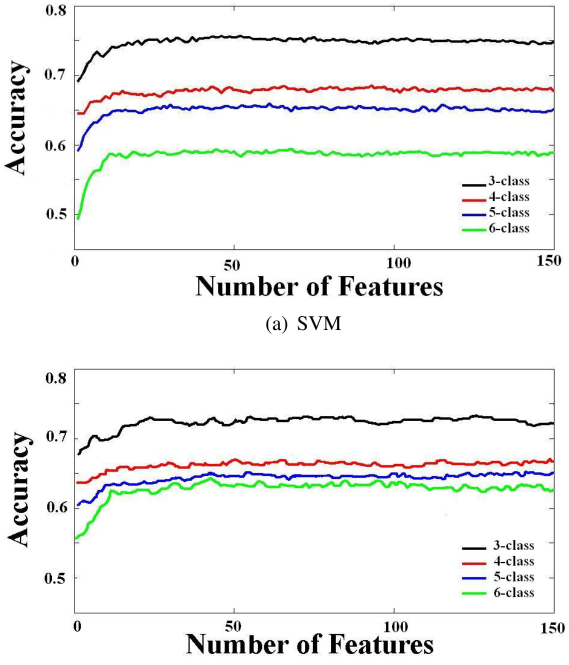
The generation of the Classification Model represents the last step in the *calibration* phase. It builds a model that will allow the *classification* phase to approximate the socio-economic level of geographical areas that have not been covered by the evaluators to save budget. Specifically, it takes as input the $(SEL, features)$ dataset for a specific region and identifies the best classification technique for each socio-economic granularity. *CenCell* considers four different SEL granularities: three, four, five and six ranges (classes). For six SEL classes A covers range $[100-83)$, B $[83-66.4)$, C $[66.4-49.8)$, D $[49.8-33.2)$, E $[33.2-16.6)$ and F $[16.6-0]$. Smaller granularities follow a similar distribution in ranges. In terms of the features (cell phone variables), having a large number might boost classification, but it can also generate a lot of noise (*the curse of high dimensional datasets*). For that reason, *CenCell* first identifies the significance of each feature and applies classification techniques on the features ordered by their relevance. Specifically, it evaluates two different feature selection techniques: maxrel and mRMR (as difference mRMR-MID or quotient mRMR-MIQ). Once the features have been ordered according to their significance, *CenCell* tests supervised SVMs as well as unsupervised EM Clustering. We selected SVMs since they have been successfully used in similar classification problems [5]. On the other hand, unsupervised EM clustering was selected among all clustering unsupervised techniques, since populations have been previously shown to follow Gaussian distributions in terms of socio-economic variables [6]. *CenCell* evaluates each combination of technique and granularity and outputs the one that has the best predictive power for each SEL granularity. As a result, the tool provides policy makers with the possibility of selecting a granularity and classification quality according to their own interests.

This step first partitions the BTS dataset with the ordered features and SELs for training and testing, containing 2/3 and 1/3 respectively. Using each supervised and unsupervised technique, *CenCell* computes the classification rate for each SEL granularity, from three classes (A, B and C) to six (A, B, C, D, E and F), and for each subset of ordered features in $n = \{1, \ldots, 279\}$ produced by Maxrel, mRMR-MIQ and mRMR-MID. *CenCell* implements the SVM using a Gaussian RBF kernel and identifying its two parameters (C and γ) through a 5-fold cross-validation over the training set for each combination of technique, granularity, and subset of features. As for EM clustering, *CenCell* computes a mixture Gaussian models for each socio-economic level, granularity and subset of features until the log-likelihood values are maximized. During testing, each final cluster is labeled with the dominating socio-economic level.

V. Experimental Evaluation

The objective of this section is to evaluate the classification power of *CenCell* to determine the socio-economic level of regions that are not covered by household surveys to save budget. Our CDR dataset contains 6 months of cell phone calls, SMSs and MMSs from over $500,000$ pre-paid and contract subscribers from a large city in an emerging economy in Latin America. The city has a total of 920 BTS cellular towers and the subscribers represent a 20% of the total population. On the other hand, the census information was acquired from the local NSI and contained a total of 1200 GUs with their SEL expressed as a continuous value between 0 and 100. The city was selected because it covered all ranges of SELs. Our final dataset consists of 920 pairs $(SEL, features)$ that we divide into training (552 pairs) and testing sets (368 pairs).

The classification accuracies for each pair of technique and socio-economic granularity explored by *CenCell* are presented in Figure 2. We observe that SVMs achieve classification rates of up to 76% when differentiating three SEL classes and the top 38 ordered features selected by mRMR-MIQ. We also notice that as we increase the granularity of the SEL, the classification accuracy decreases reaching a value of 57% for six SEL classes and the top 19 features. On the contrary, EM clustering achieves worse classification rates than SVM for granularities three, four and five. However, EM clustering yields better results when six socio-economic levels are differentiated. It shows accuracies of 63% compared to the 57% reached by SVMs, with 6 clusters and the top 20 features. We hypothesize that as the socio-economic granularity increases, the map-overlay technique in (Step 2) generates more blurred SEL levels thus making unsupervised techniques a more adequate approach. At the end of the *calibration* phase, *CenCell* would select SVMs as classification tools for granularities three, four and five; and would select EM clustering when six socio-economic levels are differentiated. To understand better the nature of the classification errors, we analyzed the confusion matrices for the best approaches selected by *CenCell*. These confusion matrices revealed that when incorrectly predicted, the output SEL class *tends to be* adjacent to the correct one. Such errors reflect an implicit order between the SELs which limits the impact of the errors on the computation of the maps.



(a) SVM

(b) EM Clustering

Fig. 2. Accuracy for (a) SVM and (b) EM clustering with mRMR.

References

[1] N. Eagle. Behavioral inference across cultures: Using telephones as a cultural lens. *IEEE Intelligent Systems*, 23:4:62–64, 2008.

[2] V. Soto, V.Frias-Martinez, J. Virseda, and E. Frias-Martinez. Prediction of socioeconomic levels using cell phone records. In *User Modelling, Adapt. and Pers., UMAP*, 2011.

[3] V. Frias-Martinez and J. Virseda. On the relationship between socio-economic factors and cell phone usage. *ICTD, Atlanta, USA*, 2012.

[4] V. Frias-Martinez, J. Virseda, A.Rubio, and E. Frias. Towards large scale technology impact analyses: Automatic residential localization from mobile phone-call data. *ICTD*, 2010.

[5] Robert Burbidge and Bernard Buxton. An introduction to support vector machines for data mining. Technical report, UCL, 2001.

[6] M. Yadollahi. Factors affecting family economic status. *European Journal of Scientific Research*, 37:94–109, 2009.

Session 9

Social structure

Mobile communication in business networks: structure and leadership

Gautier Krings[1,*], Didier Baclin[1], Loic Jacobs van Merlen[1], Marcelo Lobato Pimenta[2], Felipe De Abreu Galli[2]

[1] Real Impact Analytics, 30 Grand Rue, 1660 Luxembourg, Luxembourg
[2] Telefonica/Vivo, Av. Eng. Luís Carlos Berrini, 1376 - Brooklin, Itaim Bibi, São Paulo, 04571-000, Brazil
[*] gautier.krings@realimpactanalytics.com

Mobile Call Data Records (CDRs) have been widely used for the last 15 years for the purpose of Social Network Analysis (SNA). Typical questions are about the structural organization of these networks [1], the detection of leaders among the nodes [2] or the spreading of information [3]. In a majority of the research published in this topic, the analysed CDRs are related to private accounts and the analyses are mostly focused on social interactions such as friendships or families. Little research has been conducted so far on the structure of such communication networks in corporate environments (see related work for studies of face-to-face communications [4], email networks [5] or in political environments [6]).

In business networks, we state the principle of leadership in a different way, which is bound to the way companies tend to work. We will highlight the differences between our B2B decision makers' discovery method and the existing leadership algorithms. We use this method to score the entire network of B2B customers. This allows the operator to prioritize some lines in order to offer to decision makers an optimal experience, which is especially relevant in emerging markets where the growth of the customer base has often outpaced the investments in the underlying network that is often the cause of poor experience, call quality, slow data transfer speed.

We study a network built on two months of call data records of approximately six million business customers from one large Brazilian operator. For each customer, the company ID is known allowing us to segment the network into subnetworks corresponding to the internal communication inside the company. The initial communication network is hence splitted into 334,000 networks representing each one company, with sizes ranging from 1 to 58,000 nodes.

The original CDRs consist of about 2.3 billions calls, SMS and MMS (in 2 months), which result into an undirected weighted network where links are weighted by the number of calls, SMS and MMS that have been exchanged by both numbers. It has often been pointed out that a preliminary filtering of the network allows to remove fake calls and non-social interactions [7], however we decided not to apply such a filtering, given that calls between two employees of the same company is a priori expected to be a meaningful interaction.

Previous research about networks in a corporate environment has been mostly focused on searching information in the network structure about the company's organization. Here, we go a step further into the analysis by searching how to find to infer leadership from the network structure. Leadership in social networks has been a long-studied topic and can be defined in various ways. In social networks, leaders may be individuals being in the centre of dense parts of the network [2], individuals being referenced by others, or hubs instead, that are sources of useful references [8]. In business networks, we state the principle of leadership in a different way, which is bound to the way companies tend to work.

While communication is crucial for an organization to work correctly, it is observed that most companies work as silos: a large fraction of the communication stays inside departments and only few calls are passed across different departments. In such a configuration, managers and team leaders are located in a central position of the network, and act as bridges of information across departments. We propose a centrality measure that integrates these hypotheses and works similarly to Google's PageRank measure. We first estimate the structure of departments inside each company communication network using the Louvain method with standard modularity. The choice of using the standard modularity instead of a scaled variant of this quality measure [9,10] is made on purpose; the standard modularity suffers of the so-called resolution limit [11] which makes the sizes of the communities dependent of the size of the network. While this limit is often seen as a problem, in our case we use it to our advantage: in small companies, we will detect small departments, while in large companies, the Louvain method returns large communities, hence large departments.

Using the communities as approximation of the departmental structure of the company, when calculate the leadership measure of each node using the following two assumptions:

- Leaders communicate often with leaders of other departments
- Leaders communicate with most of the departments of the company

Based on these assumptions, we calculate recursively the leadership measure $l(i)$ of each node with

$$l(i) = \frac{1}{|C| - 1} \sum_{k \in C\{c_i\}} l\big(N_k(i)\big),$$

where C is the set of all communities in the network, c_i is the community index of node i, and $N_k(i)$ returns the node of community k that has the strongest connection with node i. In other words, the leadership of a node is the average of the leadership of his closest connection in each community different of his own community. If a node does not communicate with one or several communities, his leadership level is hence penalized in comparison to nodes that communicate to all other communities. We apply this computation recursively and normalize the leadership values at the end of each iteration such that the highest leader in each community has a leadership measure of 1. After the convergence of the iterative process, we recover the leadership measure for all nodes of the network.

We have tested this measure on the internal communication network of a part of the operator's employees, consisting of approximately 1,000 individuals for which the department and the position in the company have been made available. As a first result, we observe that the detected communities and the exact departments have a very significant overlap. Moreover, we used the leadership measure in conjunction with several other network variables such as the core number in a data mining model, and managed to predict correctly over 70% of the leaders in the company.

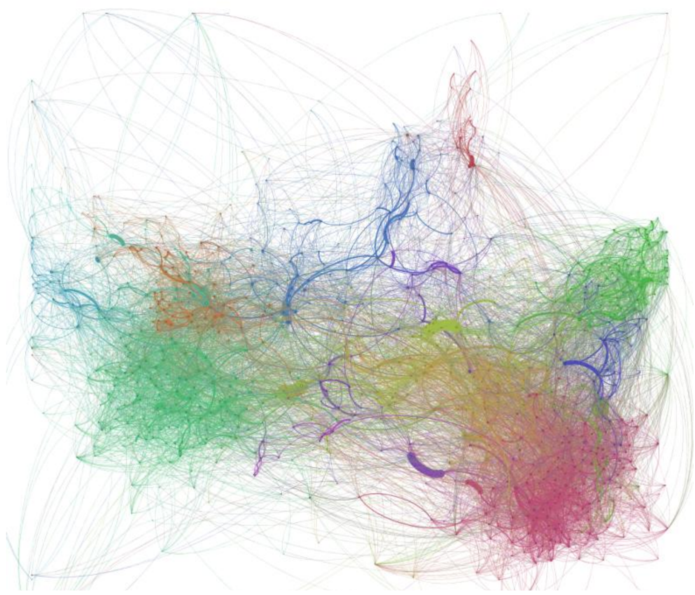


Figure 1: network of the operator's internal communications, colors correspond to communities and overlap strongly with the actual company's departments.

Although the leadership measure is based on a community detection algorithm that has a non-deterministic component, the obtained scores remain consistent when the communities returned by the detection method differ one from another.

We score the entire network of B2B customers, and observe that decision makers are a useful source of information about the structure of the network. Their communication patterns provide information on the company's way of working. On a commercial side, we show how leaders act as drivers of adoption for several products offered by the operator: if the decision maker of a community is adopting a product, the members of his department are more likely to adopt this product as well. Finally, decision makers help as well the operator to rank lines that are out of his network based on the leadership score of his own customers.

References:

[1] Blondel, V., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), P10008.
[2] Blondel, V., De Kerchove, C., Huens, E., & Van Dooren, P. (2006). Social leaders in graphs. *Lecture notes in control and information sciences*, *341*, 231.
[3] Tabourier, L., Stoica, A., & Peruani, F. (2012). How to detect causality effects on large dynamical communication networks: a case study. In *Communication Systems and Networks (COMSNETS), 2012 Fourth International Conference on* (pp. 1-7). IEEE.
[4] Olguín, D. O., Waber, B. N., Kim, T., Mohan, A., Ara, K., & Pentland, A. (2009). Sensible organizations: Technology and methodology for automatically measuring organizational behavior. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, *39*(1), 43-55.
[5] McCallum, A., Wang, X., & Corrada-Emmanuel, A. (2007). Topic and role discovery in social networks with experiments on enron and academic email.*Journal of Artificial Intelligence Research*, *30*(1), 249-272.
[6] Porter, M. A., Mucha, P. J., Newman, M. E., & Friend, A. J. (2007). Community structure in the united states house of representatives. *Physica A: Statistical Mechanics and its Applications*, *386*(1), 414-438.
[7] Onnela, J. P., Saramäki, J., Hyvönen, J., Szabó, G., De Menezes, M. A., Kaski, K., ... & Kertész, J. (2007). Analysis of a large-scale weighted network of one-to-one human communication. *New Journal of Physics*, *9*(6), 179.
[8] Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment.*Journal of the ACM (JACM)*, *46*(5), 604-632.
[9] Lambiotte, R., Delvenne, J. C., & Barahona, M. (2008). Laplacian dynamics and multiscale modular structure in networks. *arXiv preprint arXiv:0812.1770*.
[10] Traag, V. A., Van Dooren, P., & Nesterov, Y. (2011). Narrow scope for resolution-limit-free community detection. *Physical Review E*, *84*(1), 016114.
[11] Fortunato, S., & Barthelemy, M. (2007). Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, *104*(1), 36-41.

A Place-focused Model for Social Network Formation in Cities

Chloë Brown, Anastasios Noulas, Cecilia Mascolo
Computer Laboratory
University of Cambridge
email: firstname.lastname@cl.cam.ac.uk

Vincent Blondel
Department of Mathematical Engineering
Université Catholique de Louvain
email: vincent.blondel@uclouvain.be

1 Introduction

It is an intuitive idea that social relationships between people, arise out of meetings and shared activities in common spaces. Scott Feld's theory of the *focused organization* of social ties posits that friendships form between individuals whose interactions are organized around extra-network foci, which can include physical places. In the paper outlining this theory [Feld1981], Feld discusses the implications of this theory for the structure of social networks, and shows how commonly observed structural properties could arise from the formation of social ties in the way he describes.

Empirical investigation of Feld's theory has traditionally been difficult and time-consuming, requiring interviews with and observation of necessarily small groups of people, and these difficulties have meant that it has been impossible to test on a large scale. However, the recent widespread adoption of location-based online social services has provided us with a huge volume of data both about the structure of people's social networks, as described by social ties explicitly declared by users of these services, and about their activities and meetings at places in their local environment, thanks to the location-sharing dimension. We now therefore have an unprecedented opportunity to investigate the role that places may play in the formation of the structure of social networks on a scale not previously possible. Furthermore, the semantic information about places available in location-based online social services allows us to investigate for the first time the relationship between the categories of places where people meet and the likelihood that those people are friends. This is a great advantage offered by these services, where user position is becoming available when they use the respective service similarly to cellular datasets, yet compared to the latter, the combination of multiple layers of data opens new avenues for addressing previously unanswered research questions.

In this work, we study a large dataset from Foursquare, the most popular location-based online social network, which has over 35 million users as of January 2013, analyzing the social and spatial properties of social networks in cities. We then use our observations to present a model for social network formation based on Feld's focused organization theory, and show that the model produces networks with the structural properties expected of social networks. Our work makes the following contributions:

- We first define and analyze place-based social networks at the city scale, to answer the question: *what do intra-city social networks look like, and do they have common structural characteristics*? We show that the social networks in the Foursquare dataset have the structural properties observed by sociologists studying real-world social networks, namely a power-law degree distribution, small-world properties (high clustering and small diameter), and strong community structure. While the global properties of online and offline social networks have previously been analyzed, we believe that our work is the first to examine and compare the structures of place-based social networks *within different cities* and to show these common structural properties.

- We then address the question: *does this large location-based social dataset support Feld's theory of focused tie organization*? Exploiting the combination of social information and specific semantic information available in the Foursquare dataset, we are able to investigate for the first time the relationship between the category of a place where people meet and the probability of friendship. We find that the type of a place where people meet has a strong influence on the likelihood that they are friends, which provides support for Feld's theory of focused organization at a scale not previously possible.
- Inspired by Feld's theory, we present a model for the formation of a social network in a city based around meetings at places, and show that this model is able to produce networks with the structural properties observed in real social networks. The fact that this model is able to reproduce empirically observed social network features represents a computational validation of the focused organization theory, and supports its suggested mechanisms for the formation of friendship between individuals.

We believe that our work has intrinsic interest, as we investigate an area largely unexplored, namely, that of the structural similarities between social networks at the city scale within different cities, and demonstrate that the networks in different cities show striking similarities. Furthermore, our model demonstrates computationally that Feld's theory of the focused organisation of social ties, with places as foci, results in networks with the structural features commonly observed in social networks.

From a practical perspective, our observation that the type of a place where people meet strongly affects the probability of friendship could be useful to online location-based social services such as Foursquare. For example, one important application in location-based social networks is the recommendation to users of venues they might want to visit [Berjani and Strufe2011, Long, Jin, and Joshi2012, Noulas et al.2012]. Our finding that the type of place where people meet has a strong influence on friendship suggests that different recommendations would be appropriate depending on the other users present.

This has potential applications in the development of smarter privacy controls in location-based online social networks: Page et al. found that people's concerns about privacy in location-sharing services center around the desire to preserve one's existing offline relationship boundaries [Page, Kobsa, and Knijnenburg2012], and our observations suggest that these boundaries might be reflected in the types of places where friends meet (for example, closer friends at homes and nightlife spots, less close acquaintances only at professional venues or transport spots). Use of this information could enable services such as Foursquare to adjust the default audience of a check-in, for example, based on relationship semantics inferred using meeting places.

References

[Berjani and Strufe2011] Berjani, B., and Strufe, T. 2011. A recommendation system for spots in location-based online social networks. In *Proc. of the 4th Workshop on Social Network Systems*. ACM.

[Feld1981] Feld, S. L. 1981. The focused organization of social ties. *American Journal of Sociology* 86(5).

[Long, Jin, and Joshi2012] Long, X.; Jin, L.; and Joshi, J. 2012. Exploring trajectory-driven local geographic topics in foursquare. In *LBSN'12*.

[Noulas et al.2012] Noulas, A.; Scellato, S.; Lathia, N.; and Mascolo, C. 2012. A random walk around the city: New venue recommendation in location-based social networks. In *SocialCom '12*.

[Page, Kobsa, and Knijnenburg2012] Page, X.; Kobsa, A.; and Knijnenburg, B. 2012. Dont disturb my circles! boundary preservation is at the center of location-sharing concerns. In *ICWSM '12*.

A Comparative Study of Decentralized Routing in Social Network Based on Mobile Phone Data

Carlos Herrera,[1, *] Christian M. Schneider,[1, †] Thomas Couronne,[2]
Zbigniew Smoreda,[2] Rosa M. Benito,[1, ‡] and Marta C. González[1, §]
[1] *Department of Civil and Environmental Engineering,*
Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America
[2] *Sociology and Economics of Networks and Services department,*
Orange Labs, 38 rue du Général Leclerc, 92794 Issy les Moulineaux, France

In this work we focus on the study of social networks extracted from mobile phone data. The data has been filtered in such a way that the nodes are a subset of the mobile phone users and the links represent a social interaction that indicates an acquaintance or a friendship as opposed to just an occasional phone call. This is possible because our current data spans over a period of at least six months in each of the countries we study (displayed in Fig. 1).

The main idea behind this work is based on the famous small-world experiment performed by the social psychologist Stanley Milgram [1] in the 1960s. Milgram's experiment led to two striking discoveries, of which the existence of short paths was only the first. The second was that people in society were collectively able to forward a letter to a distant unknown target surprisingly fast with knowledge of only their own personal acquaintances (local information) and only the name, location and the profession of the target.

Many interesting questions remain open in relation to Milgram's, experiment, such as why should a social network contain such short paths and how people are able to select among hundreds of acquaintances the correct person to form the next link in the chain. To answer these questions several works have been carried out, both empirically and mathematically. Dodds et al. [2] repeated Milgram‘s experiment at large scale with e-mails, providing confirmation that geographical position of the nodes has a crucial role in the possibility of the network to be searchable. On the other hand, Lieben-Nowell et al. [3] proved a simple geo-greedy algorithm is able to efficiently route a message between different cities using the social network, with data from an online blog community where users declare the city where they live. Watts and Strogatz [4, 5] proposed a hierarchical network model that present the small world effect and high clustering. Kleinberg [6] has proposed several analytical treatable network models in which actors are capable of finding short paths with high probability just by using local information.

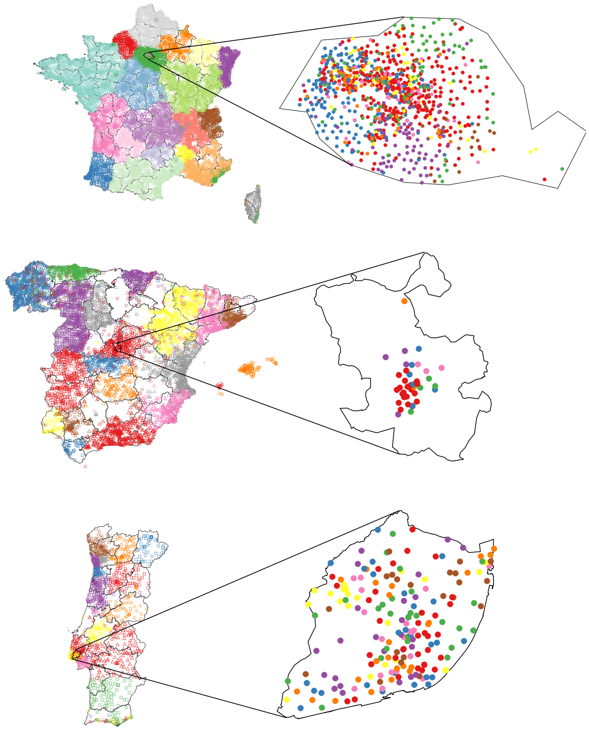


Figure 1. The social communities in the three studied countries and their capitols. While on the country level the communities divide different regions, within cities the geographical separation collapse and it reveal social groups.

The goal of this research is to understand how people are able to deliver a message within a small number of hops from a source user to a target user whose location is given in terms of his/her coordinates in a city of the country under study. Therefore, we study different strategies for the case of limited information (without global knowledge of the social network), just using local information such as the identities and connectivity of a node's neighbors based on the structure of the real social networks. This study is performed in different countries and both

[*] Email: carloshy@mit.edu
Permanent address: Universidad Politécnica de Madrid, 28040 Madrid, Spain

[†] Email: schnechr@mit.edu

[‡] Permanent address: Grupo de Sistemas Complejos, and Departamento de Física y Mecánica, Escuela Técnica Superior de Ingenieros Agrónomos, Universidad Politécnica de Madrid, 28040 Madrid, Spain

[§] Email: martag@mit.edu

at the inter-city and at the intra-city levels demonstrating that decentralized search algorithm can work well on real friendship networks on both levels. We use three different large data sets corresponding to six months of anonymized mobile phone calls of three European countries (France, Portugal, and Spain) with over 7 billion of calls. The data sets include information of the most used tower coordinates (France and Portugal), or billing zip codes (Spain) for each of the users present in the call data. From these data sets we construct the friendship networks based on reciprocal communications taking into account the user location.

First we show the existence of short paths with an average length of 7.75, 7.44, and 9.2 for France, Portugal and Spain, respectively. We further simulate delivering a message from a source user to a target user whose location is given in terms of his/her coordinates in a city of the country using only information of the source's neighbors. We report the success of three different decentralized routing strategies based on: random information, geographical information, and community information. Finally, we connect the results with the underlying social network structure and conclude that both information Milgram provided (geography and profession) are required to successfully route to an unknown person.

[1] Milgram, S., The small world problem. Psychology Today **1**, 62-67 (1967).

[2] Dodds, P.S., Muhamad, R., Watts, D.J., An Experimental Study of Search in Global Social Networks. Science **301**, 827-829 (2003).

3] Lieben-Nowell, D., Novak, J., Kumar, R., Raghavan, P., Tomkins, A., Geographic routing in social networks. PNAS **102**, 11623 (2005).

[4] Watts, D.J., Strogatz S.H., Collective Dynamics of Small World Networks. Nature **393**, 440 (1998)

[5] Watts, D.J., Doods, P.S., Newman, M.E.J., Identity and Search in Social Networks. Science **296**, 1302-1305 (2002).

[6] Kleinberg, J., Complex networks and decentralized search algorithms. Proceedings of the International Congress of Mathematicians: Madrid, August 22-30, 2006: invited lectures, 1019-1044 (2006).

Migration and Ethnic Segregation: Evidence from Estonia's Mobile Phone Logs

Joshua Blumenstock
University of Washington
joshblum@uw.edu

Ott Toomet
Tartu University
otoomet@gmail.com

DRAFT VERSION

February 2013*

Abstract:
What is the effect of migration on ethnic segregation in urban areas? The manner in which individuals segregate along ethnic lines has important economic and social consequences, affecting employment and education decisions, processes of information creation and diffusion, and several other aspects of society. Yet, due in part to data constraints, little is known about the extent and dynamics of migrant integration within and across ethnicities. We exploit a novel source of communications data that allows us to observe the ethnicity of hundreds of thousands of mobile phone subscribers in Estonia, a country with a history of tense ethnic relations. In addition to containing detailed information about inter- and intra-ethnic network structure, we are able to observe the migration and movement of each individual using mobile positioning data. We can thereby infer, for each individual, whether a migration takes place, the extent to which the migrant communicates with coethnics and non-coethnics pre- and post-migration, and whether the migrant is physically proximate to coethnics and non-coethnics. Our study contains three related sets of results: (i) Whether the average migrant integrates within or across ethnicities, (ii) How quickly integration with coethnics and non-coethnics occurs for different types of migrants; and (iii) Aggregate effects of migration and urbanization on city- and country-wide levels of ethnic segregation.

Keywords: Segregation, Migration, Homophily, Urbanization and Cities, Mobile Phones; Estonia.

*The authors are grateful for thoughtful comments from Rein Ahas, Matthew Jackson, Frederico Finan, Tyler McCormick, and to seminar participants at the University of Washington for helpful comments. We gratefully acknowledge financial support from CERGE-EI and the National Science Foundation. All errors are our own.

1 Introduction

Ethnic segregation is a common phenomenon in many of the world's cities, and a prominent feature of most developing economies (Taeuber and Taeuber, 1965; Massey and Denton, 1993). The consequences of segregation are far-reaching, affecting investment in human capital, the structure of labor markets, levels of violence and corruption, inequality, and patterns of prejudice and discrimination (Easterly and Levine, 1997; Miguel and Gugerty, 2005; Cutler and Glaeser, 2007; Bayard, Hellerstein, Neumark, and Troske, 1999).

Internal migration plays a critical role in both short- and long-term dynamics of segregration and integration. Migrants often choose to migrate to areas where their networks are stronger (Munshi, 2003), existing residents often strive to keep out dissimilar immigrants (Schelling, 1971), and political and institutional forces often prevent integration of migrants across ethnic lines (Yinger, 1986; Clark, 1986). However, due largely to shortcomings in the data used in the analysis of migration and segregation, very little is known about the extent to which migrants integrate into communities in the location of destination, and whether recent migrants affect aggregate levels of segregation. The vast majority of empirical studies on segregation rely on census or household survey data that is sampled infrequently, and which is notoriously bad at tracking domestic migration. These studies almost invariably focus on segregation at the place of residence, ignoring segregation in other aspects of life, such as place of work and leisure time (van Ham and Manley, 2010).

2 Context and Data

Estonia is an ethnically fractured society that provides an ideal context for studying patterns of ethnic segregation through mobile phone datasets. Besides native Estonians, the country houses a large and relatively homogeneous Russian-speaking population. The relationship between these two ethnic groups is a fragile one, with the tensions culminating in "Bronze Soldier" riots in 2007. In previous work, we have used the call data to show that ethnic Estonians and ethnic Russians remain highly segregated in major urban areas (Toomet, Silm, Saluveer, Tammaru, and Ahas, 2012) and in their respective communication network (Toomet, van der Leij, and Rolfe, 2012).

A unique feature of the data at our disposal is the fact that we observe the language spoken by each subscriber in the dataset. In Estonia, a bilingual economy where native Estonians speak Estonian and native Russians speak Russian, this allows us to very accurately infer the ethnicity of each subscriber. As with other Call Detail Records used in recent research (cf. Gonzalez et al. 2008, Song et al. 2010), we additionally observe the complete communication records of each individual, as well as a long history of passive positioning data, which indicates for each phone call and text message, the time and location of the subscriber involved in the event. The location is specified by the cell tower, which gives us a spatial resolution ranging between many kilometers in rural areas to a few hundred meters in dense urban settings.

3 Measuring segregation and migration from mobile call data

Our approach allows us to measure several types of segregation simultaneously, over a period of multiple years.[1] For the long-run data, we rely on the call records to infer the location of each individual throughout

[1] Census and survey data typically only permit the measurement of segregation at the place of residence. However, segregation is more complex, as important parts of our life are spent at work, school, and in places where we pursue various leisure activities. These various types of segregation are correlated (Schnell and Yoav, 2001), as employment opportunities are affected by local ethnic networks (Edin, Fredriksson, and Åslund, 2003), and by the fact that similar people tend to cluster into similar jobs (Sørensen, 2004). Still, Åslund and Skans (2010) find that, in Sweden, workplaces are less segregated than residences.

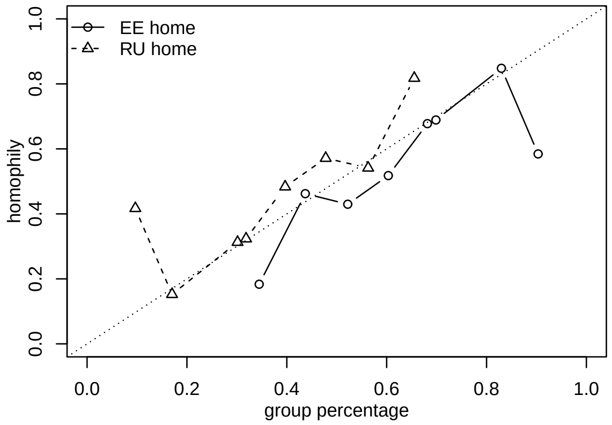


Figure 1: Relationship between the homophily at the home neighborhood as a function of the corresponding ethnic composition (Toomet, Silm, Saluveer, Tammaru, and Ahas, 2012). Group percentage is calculated from 2000 census. Points represent average values for residents in the corresponding Tallinn municipality districts. Circles represents ethnic Estonians; triangles ethnic Russians.

the day. This allows us to observe at which times of the day individuals of similar and dissimilar ethnicities are close to each other. Formally, we measure *copresence* from the passive positioning data. The "same" place and "same" time must be operationalized to *timeframes*, where we specify what is the required spatial proximity (for instance, a city tract) and time interval (for instance, 1 hour). We define copresence between individuals i and j, p_{ij}, as $p_{ij} = \sum_k \mathbb{1}(c_{jk} \in C_i)$, where $\mathbb{1}(\cdot)$ is the indicator function, c_{jk} is timeframe of call k of individual j, and C_i is the set of all timeframes of individual i. Copresence can be understood as the meeting potential between given two individuals and hence treated as a (proxy for a) social tie.

In addition to the positioning data, we observe the actual call graph, as determined by the calls made between individuals in the dataset. Knowing the ethnicity of the caller and recipient allows us to measure the social segregation in the call network, which we take as a proxy for underlying patterns of ethnic segregation. We can further infer the strength of each network edge using the the volume and length of calls between each dyad. Formally, homophily for individual i is defined as $h_i = \frac{s_i}{s_i + d_i}$, where s denotes the measure of contacts with the same-language individuals, and d that with the different-language ones. s and d are defined through copresence as

$$s_i = \sum_{\substack{j \in \mathbb{1} \\ j \neq i}} \mathbb{1}(\lambda_j = \lambda_i) \cdot p_{ij} \qquad \text{and} \qquad d_i = \sum_{\substack{j \in \mathbb{1} \\ j \neq i}} \mathbb{1}(\lambda_j \neq \lambda_i) \cdot p_{ij}, \tag{1}$$

where λ_i denotes the preferred language of individual i. As we have shown in prior work, the homophily in the location of residence follows the population percentage rather closely (see Figure 1), while at workplace and during free-time the relationship is much weaker (Toomet, Silm, Saluveer, Tammaru, and Ahas, 2012).

Given a continuous sequence of locations for all individuals over several years, we can similarly identify the home and work (geographic) locations for cellphone users using the anchor-point methodology of Ahas, Silm, Järv, Saluveer, and Tiru (2010). This permits us to identify both short and long-distance movers, and those who change their workplace while remaining settled in the same area. We are then able to classify these moves as temporary or permanent *migrations*, and thus construct a binary indicator variable M_{it} that indicates for each individual i whether he/she migrated at time t.

Using the above measure of homophily h_{it} as an indicator of the extent to which individual i prefers to associate with coethnics over non-coethnics at time t, and taking M_{it} to be a binary indicator of whether i migrates at time t, a basic fixed-effects model can be used to study the effect of the migration on i's preference for coethnics.

$$h_{it} = \alpha_1 + \sum_{s=0}^{T} \beta_s M_{i(t-s)} + \mu_i + \pi_t + \epsilon_{it} \tag{2}$$

where μ_i is an individual-specific fixed effect that accounts for the fact that each individual may have a different preference for segregation at baseline, and π_t is a time-specific fixed effect to reduce bias caused by common trends across all individuals over time. The coefficients on the β_s thus indicate the extent to which migrations in the past T periods (for s = 0,1,...,T) lead to changes in observed homophily for the average migrant. It is these estimated β_s coefficients that reveal whether migrants tend to be more segregated in the place of destination or of origin.

Adding heterogeneous effects to Model 2 permits us to identify which types of migrants are more or less likely to be integrated into communities of coethnics and non-coethnics. Letting X_i denote a vector individual characteristics that vary between individuals but which are constant over time, we estimate

$$h_{it} = \alpha_2 + \sum_{s=0}^{T} \beta_s M_{i(t-s)} + \sum_{s=0}^{T} \gamma_s (M_{i(t-s)} * X_i) + \mu_i + \pi_t + \epsilon_{it} \tag{3}$$

As before, the β_s indicate the extent to which past migrations affect current homophily, and now the γ_s estimate the extent to which individuals of different X_i are more or less likely to be affected through the migration. As an example, if X_i includes a measure of the homophily of i prior to migration, then γ_s indicates whether migrants who are more integrated across ethnic lines pre-migration are more or less likely to integrate post-migration.

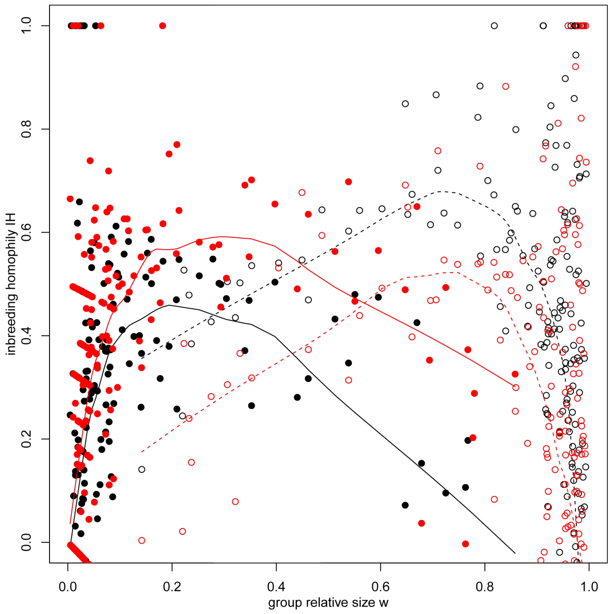


Figure 2: Inbreeding homophily, a measure of co-ethnic bias, as a function of the relative group size, for several hundred municipalities in Estonia. Red dots represent Russian speakers; black dots represent Estonian speakers. Solid dots correspond to outgoing calls; hollow dots to incoming calls.

References

Ahas, R., S. Silm, O. Järv, E. Saluveer, and M. Tiru (2010): "Using Mobile Positioning Data to Model Locations Meaningful to Users of Mobile Phones," *Journal of Urban Technology*, 17(1), 3–27.

Bayard, K., J. Hellerstein, D. Neumark, and K. Troske (1999): "Why are Racial and Ethnic Wage Gaps Larger for Men than for Women? Exploring the Role of Segregation," Working Paper 6997, National Bureau of Economic Research.

Clark, W. A. V. (1986): "Residential Segregation in American Cities; A Review and Interpretation," *Population Research and Policy Review*, 5, 95–117.

Cutler, D. M., and E. L. Glaeser (2007): "Social interactions and smoking," Working Paper 13477, NBER, 1050 Massachusetts Avenue, Cambridge, MA 02138, USA.

Easterly, W., and R. Levine (1997): "Africa's Growth Tragedy: Policies and Ethnic Divisions," *The Quarterly Journal of Economics*, 112(4), 1203–1250.

Edin, P.-A., P. Fredriksson, and O. Åslund (2003): "Ethnic Enclaves and the Economic Success of Immigrants—Evidence from a Natural Experiment," *The Quarterly Journal of Economics*, 118(1), 329–357.

Massey, D. S., and N. A. Denton (1993): *American Apartheid: Segregation and the Making of the Underclass.* Harvard University Press, Cambridge, MA.

Miguel, E., and M. K. Gugerty (2005): "Ethnic diversity, social sanctions, and public goods in Kenya," *Journal of Public Economics*, 89(11–12), 2325 – 2368.

Munshi, K. (2003): "Networks in the Modern Economy: Mexican Migrants in the U. S. Labor Market," *The Quarterly Journal of Economics*, 118(2), 549–599.

Schelling, T. C. (1971): "Dynamic models of segregation," *Journal of Mathematical Sociology*, 1, 143–186.

Schnell, I., and B. Yoav (2001): "The Sociospatial Isolation of Agents in Everyday Life Spaces as an Aspect of Segregation," *Annals of Association of American Geographers*, 91(4), 622–636.

Åslund, O., and O. N. Skans (2010): "Will I See You at Work? Ethnic Workplace Segregation in Sweden, 1985-2002," *Industrial and Labor Relations Review*, 63(3), 471–493.

Sørensen, J. B. (2004): "The Organizational Demography of Racial Employment Segregation," *American Journal of Sociology*, 110(3), pp. 626–671.

Taeuber, K. E., and A. F. Taeuber (1965): *Negroes in cities: Residential segregation and neighborhood change.* Aldine Pub. Co., Chicago, IL.

Toomet, O., S. Silm, E. Saluveer, T. Tammaru, and R. Ahas (2012): "Where do Ethnic Groups Meet? Copresence at Residence, Work, and Free-time," Tartu University.

Toomet, O., M. J. van der Leij, and M. Rolfe (2012): "Social Networks and Labor Market Inequality between Ethnicities and Races," Tinbergen Institute Discussion Papers 12-120/II, Tinbergen Institute.

van Ham, M., and D. Manley (2010): "The effect of neighborhood housing tenure mix on labour market outcomes: a longitudinal investigation of neighbourhood effects," *Journal of Economic Geography*, 10, 257–282.

Yinger, J. (1986): "Measuring Racial Discrimination with Fair Housing Audits: Caught in the Act.," *American Economic Review*, 76(5), 881–893.

Why does my phone always ring when I am about to make a call?

Jeppe Juul[1] and Albert-László Barabási[2,3,4]
[1] *University of Copenhagen, Niels Bohr Institute, Blegdamsvej 17, Copenhagen, Denmark, Electronic address: jeppesj@nbi.dk*
[2] *Center for Complex Network Research, Northeastern University, Boston, USA*
[3] *Center for Cancer Systems Biology, Dana-Farber Cancer Institute, Boston, USA*
[4] *Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, USA*

In large social structures, a tendency for individuals to form links with people similar to themselves has been observed. This is known as the *homophily principle* [1, 2].

Recently, there has been a lot of focus on the structural and dynamic properties of social networks. The increased availability of large sets of data on communication patterns has made it possible to study human interaction on a scale and at a level of detail not previously possible [3]. In particular, mobile phone call records have been used to investigate human mobility, responses to emergencies, and the formation of new social ties [4–6]. The number of people an individual has links to and the strength of these links characterize the topology of a social network und are important for understanding the formation of communities and the spread of information through the network [7, 8].

We studied the communication habits of 2.5 million mobile phone users over a year. We used a large dataset of 3 billion calls to examine their call profiles, defined as the fraction of outgoing calls made during each hour of the week. By grouping users according to gender, age, and geographical region we were able to identify characteristic call profiles for different demographics. We introduced a distance measure, based on the area of non-overlap between individual call profiles (see Fig. 1), and examined its distribution for different groups of users.

Interestingly, we find that the distance distribution for friends, defined as users with reciprocal calls, is centered at lower values than can be explained by gender, age, or regional similarity. Furthermore, weighting the distance measure according to the frequency of reciprocal calls shifts the distribution to even lower values (see Fig. 2). That is, close friends will have a tendency to make calls at the same times during the week. This could indicate the existence of a new of type of communication homophily. We explore the alternative hypothesis that friends establish mutual norms for phone usage and that this can account for the alignment effect observed for the call profiles of users who phone each other frequently.

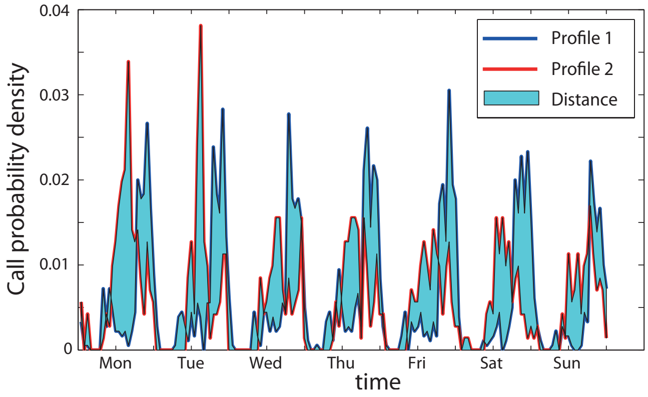


Figure 1: Mobile phone users can be characterized by their call profiles, defined as the fraction of calls made during each hour of the week. We introduce a distance measure between users as the non-overlap between individual call profiles.

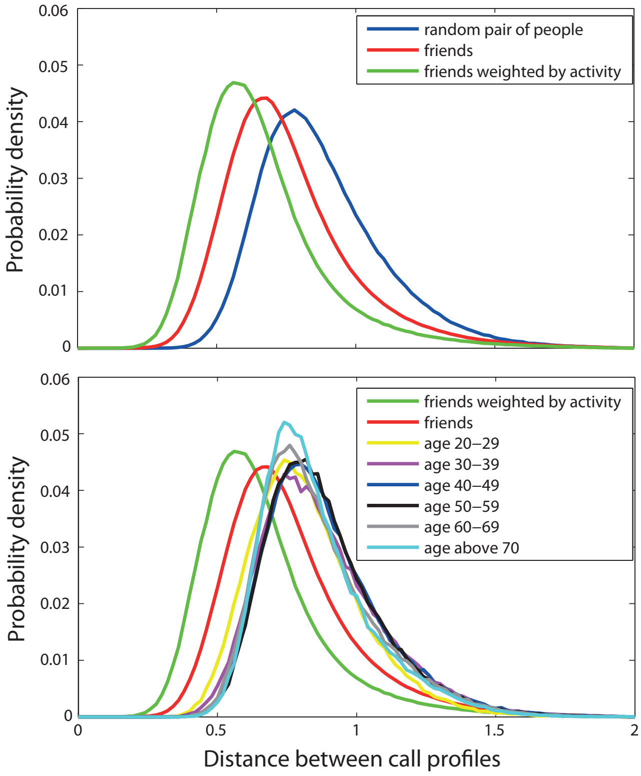


Figure 2: The distance between call profiles of mobile phone users with reciprocal calls (friends) is significantly smaller than between random pairs of user (top) or users of the same age (bottom), gender, or geographical area.

[1] M. McPherson, L. Smith-Lovin, and J. M. Cook, Annual review of sociology pp. 415–444 (2001).

[2] G. Kossinets and D. J. Watts, American Journal of Sociology **115**, 405 (2009).

[3] G. Miritello, E. Moro, R. Lara, R. Martínez-López, J. Belchamber, S. G. Roberts, and R. I. Dunbar, Social Networks (2013).

[4] D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A.-L. Barabasi, in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (ACM, 2011), pp. 1100–1108.

[5] J. P. Bagrow, D. Wang, and A.-L. Barabási, PloS one **6**, e17680 (2011).

[6] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, Nature **453**, 779 (2008).

[7] G. Palla, A.-L. Barabasi, and T. Vicsek, Nature **446**, 664 (2007).

[8] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, M. A. De Menezes, K. Kaski, A.-L. Barabási, and J. Kertész, New Journal of Physics **9**, 179 (2007).

Session 10

Privacy

Differentially Private Modeling of Human Mobility at Metropolitan Scales

Darakhshan J. Mir
Rutgers University
mir@cs.rutgers.edu[*]

Ramón Cáceres
AT&T Labs
ramon@research.att.com

Sibren Isaacman
Loyola University Maryland
isaacman@cs.loyola.edu

Margaret Martonosi
Princeton University
mrm@princeton.edu

Rebecca N. Wright
Rutgers University
rebecca.wright@rutgers.edu[*]

Models of human mobility have wide applicability in fields such as infrastructure and resource planning, analysis of infectious disease dynamics, and ecology. The abundance of spatio-temporal data from cellular telephone networks affords opportunities to construct such models. However, this data consists of individuals' locations and cellular phone activities, thus raising privacy concerns that have prevented the release and wider use of such models. In response to such privacy concerns, our work seeks to to adapt the WHERE [9] approach for modeling human mobility in metropolitan areas to be *differentially private.*

Differential privacy [4, 5] is a notion of privacy receiving increasing attention. It makes privacy a mathematical requirement on the results of interactions with data. It captures the intuition that an individual's risk of being identified should be almost the same whether or not they participate in a database. This is a strong notion of privacy that makes no assumptions about the power or background knowledge of a potential adversary.

Starting with Call Detail Records (CDRs) from a cellular telephone network that have gone through a straightforward anonymization procedure, WHERE [9] produces synthetic CDRs for a synthetic population. WHERE has been experimentally validated against billions of location samples for hundreds of thousands of cell phones in the New York and Los Angeles metropolitan areas. To ensure that the resulting synthetic CDR's are provably private, we propose to modify WHERE to be differentially private. The aim is to enable the creation and possible release of synthetic CDRs that capture the mobility patterns of real metropolitan populations while preserving individual privacy.

To the best of our knowledge, the problem of making human mobility models (based on sensitive spatio-temporal data) differentially private has not been studied before. Differential privacy has been examined in other contexts of spatio-temporal data. Chen et al. [3] study the problem of publishing a differentially private version of the trajectory data of commuters in Montreal. They then evaluate the utility of published private data in terms of count queries and frequent sequential pattern mining. WHERE does not directly model the sequentiality of the spatio-temporal data at the level of an individual. However, it would be interesting to compare the two approaches. Similarly, Ho and Ruan [7] consider the problem of location pattern mining with differential privacy for spatial databases. Andrés et al. [1] introduce the notion of *geo-indistinguishability* in location-based systems, which protects the exact location of a user while allowing release of information needed to gain access to a service.

1. BACKGROUND

Consider a database D of m simplified CDRs containing n distinct users. Each row of the database corresponds to a voice call or text message (hereafter both referred to interchangeably as calls) and includes the following attributes:

`user-id`	`date`	`time`	`lat`	`long`

Each user is indexed by a unique user ID in the set $[n] = \{1, 2, \ldots n\}$. For each user, `home` and `work` locations are estimated according to earlier techniques [8]. We assume these two locations are appended to each row of D.

1.1 WHERE

WHERE works with a CDR database D representing calls in a given metropolitan area that has been divided into smaller geographic areas by imposing a rectangular grid of granularity $d \times d$. To model the users in D, WHERE computes six types of spatial and temporal distributions:

Home and *Work*. For each grid cell, all users who have `home` and `work` locations in that area are counted. Next, the number of `home` and `work` locations in that area is divided by the total number of users in the database, thus giving a probability for the `home` and `work` locations for each area of interest. To construct a synthetic user, WHERE creates a cumulative distribution function (CDF) of the `home` or `work` probability for every possible area, randomly picks a number uniformly between 0 and 1, and selects the corresponding area as the synthetic `home` or `work`.

CommuteDistance. For each grid cell, we compute a distribution on home-to-work commuting distances of people who live in that cell. There are a total of d^2 of these *CommuteDistance* distributions.

CallsPerDay. For each user in the CDR, we calculate the mean (μ) and standard deviation (σ) for the number of calls made per day. Now, a two-dimensional histogram is constructed for the various values of μ and σ. Each user's μ

[*]Partially supported by NSF award CCF-1018445.

and σ is rounded to the tenths place. A two-dimensional matrix M is maintained, in which an appropriate cell indexed by a user's μ and σ is incremented. Then, dividing each cell by the total number of users gives a 2-D probability distribution function (that is, the sum of all the cells adds to 1) called *CallsPerDay*. We can traverse the matrix in either row or column major order to construct a CDF which can be easily sampled from and converted back into the μ and σ that index to the random cell. When determining how many calls a synthetic user makes on a particular day, WHERE samples a number uniformly at random between 0 and 1 and locate the (rounded) μ and σ that this number corresponds to in the CDF of *CallsPerDay*. A number sampled from a normal distribution determined by this μ and σ gives the number of calls that the user makes.

CallTime. For each user in the CDR database, WHERE computes the distribution of calls over the day. However, using X-Means clustering, user call-behavior can be clustered into two classes [9]. Subsequently, using the CDR database, per-minute call probability distributions are computed separately for each user class. From this, a probability distribution *CallTime*, for each hour of the day, for each user class is constructed. Previous work [9] showed that while both classes exhibit diurnal patterns in the calling behavior, one class favors evening calls while the other favors afternoon.

During the creation of a synthetic call by a synthetic user, WHERE first determines the number of calls made per day using *CallsPerDay* as described above. After assigning a user to one of the two user-classes, it uses the relevant *CallTime* distribution to synthesize call times.

Hourly. For every hour, WHERE computes a distribution of calls made over every latitude-longitude area (grid cell) in the grid. There are 24 distributions over the metropolitan area, one for each hour. Each distribution reflects the probability of people being at a specific location at a specific time. It is not explicitly tied to either a user or a `home` or `work` location.

1.2 Differential Privacy

Differential privacy rests on the guarantee that an individual's risk of being identified is almost the same whether or not the individual participates in a database, even for individuals with unique or outlier behaviors. To formalize this we need the notion of neighborhood [5] of two CDR databases. We introduce two notions of neighborhood: *call-level* and *user-level.*

Def. 1 (Neighbors). *Two CDR databases D and D' are* call-level neighbors *if $|D \oplus D'| = 1$. D and D' are* user-level *neighbors if for all CDRs, $c \in D \oplus D'$,* `user-id` *$(c) = k$, for some $k \in [n]$.*

In other words, neighboring CDR databases D and D' are call-level neighbors if they differ in exactly one CDR. They are user-level neighbors if they differ in the records of exactly one user (who may have made many calls).

We will also make use of the concept of the *global sensitivity* of a function of the database [5]. This is the maximum change in the function over all neighboring databases.

Def. 2 ([5]). *The* global sensitivity *of a function of a database D, $f : D \to \mathbb{R}^{\ell}$ is*

$$\mathrm{GS}_f := \max_{D,D'} \|f(D) - f(D')\|_1$$

where D and D' are (user or call-level) neighbors.

Known results show that differential privacy can be achieved by adding noise to the outcome of f that is proportional to the global sensitivity of f [5].

2. DIFFERENTIALLY PRIVATE WHERE

We discuss modifications to the distributions described in Section 1.1 to make them differentially private.

2.1 Probability Distributions

Home and *Work* . We need to compute a differentially private CDF for the *Home* and *Work* distributions. Let ϵ be the privacy level we want to achieve. (Values like 0.1 or 1 are typically used in practice.) Let $\mathrm{Lap}(0, \lambda)$ denote a Laplacian distribution with mean 0 and standard deviation λ. Assume a row or column major ordering of the d^2 grid cells.

Let $\mathrm{CountNum}(\texttt{home}, i)$ be a function that returns the number of distinct people in the database D with homes in grid cell i in this ordering. Then, applying standard noise adding techniques [5], using Algorithm 2.1 provides an ϵ-differentially private approximation of the CDF. It has an error proportional to $\sqrt{d^2}$.

```
Algorithm 2.1: DPhomeCDF($D, \epsilon$)

  Count ← 0
  for $i$ ← 1 to $d^2$
        Count ← Count +
        CountNum($home, i$)+
    do  Lap($0, \frac{\sqrt{2}}{\epsilon}$).
        CDF[$i$] ← Count .
  return (CDF)
```

The error can be improved by McSherry and Mahajan's method [11] to publish a CDF with error proportional to $(\log d^2)^{3/2}$. Moreover, the noisy CDF does not correspond to a legitimate probability distribution since the noisy counts are not necessarily non-decreasing. We can use Hay et al.'s post-processing techniques [6] to "clean-up" this noise and create a legitimate CDF.

CommuteDistance. Assume a min and max for the commuting distance of people living in each grid cell. Let $\mathrm{Range} = \max - \min + 1$. Let $\mathrm{CountCommute}(i, \mathrm{dist})$ count the number of people living in grid cell i who commute a distance of dist. Then Algorithm 2.2 provides an ϵ-differentially private computation of the original *CommuteDistance* distribution.

For each grid cell, the *CommuteDistance* distribution can be computed independent of the distribution for other cells. Applying the parallel composition theorem [10], the d^2 invocations of Algorithm 2.2, are still ϵ-differentially private.

Algorithm 2.2: COMMUTECDF(D, cor, ϵ)

```
Count ← 0
for $i \leftarrow 1$ to $Range$
    do { Count ← Count +
         CountCommute$(cor, i)$+
         Lap$(0, \frac{\sqrt{2}}{\epsilon})$.
         CDF$[i]$ ← Count .
return (CDF)
```

CallsPerDay. To make the CDF of *CallsPerDay* differentially private, we assume that the mean number of calls per day for any user falls in the range $[\mu_{\min}, \mu_{\max}]$. Similarly the standard deviation of the number of calls per day is in the range $[\sigma_{\min}\sigma_{\max}]$. Using this we have a fixed size two-dimensional matrix M as described above and a noise matrix N_ϵ, such that $\forall,\ i, j, N_\epsilon(i,j) = \text{Lap}(0, \frac{\sqrt{2}}{\epsilon})$. However, after every user's μ and σ is used to increment the appropriate matrix cell of M, M is replaced by $M + N_\epsilon$.

Now, we proceed as before with the modified M giving an ϵ-differentially private CDF of *CallsPerDay.*

CallTime. We cluster the users into one of the two classes using $\epsilon-$differentially private k-means clustering [2]. Subsequently, we compute a probability distribution *CallTime*, for each hour of the day, for each user class, separately by using a variant of randomized-response [12].

The stochastic creation of synthetic calls by synthetic users proceeds as before. WHERE first determines the number of calls made per day using the private *CallsPerDay* as described above. After assigning a user to one of the two user-classes, it uses the relevant private *CallTime* distribution to synthesize call times.

Hourly. For every hour, the global sensitivity of the query that counts the number of calls made over all grid cell is $\max_{calls}$. Add noise proportional to this quantity by adding a noise matrix N_ϵ, where

$$N_\epsilon(i,j) \sim \text{Lap}\left(0, \frac{\max_{calls}\sqrt{2}}{\epsilon}\right),$$

yields an ϵ-differentially private approximation of the matrix of calls made over the entire metropolitan over an hour. This can then yield a differentially private *Hourly* probability distribution. When a synthetic call is made using the *CallTime* and CallsPerDay distributions, its location is determined to be either home or work according to the probability of a person being in the location at that time of day using the *Hourly* distribution.

2.2 Summary and Evaluation Plan

In summary, the modified WHERE approach we have outlined above achieves differential privacy. This privacy may come at some accuracy cost, however, because noise is introduced in each probability distribution. We plan to compare the accuracy of the models produced by the differentially private and original versions of WHERE. The ϵ parameter offers us a "knob" by which to trade off privacy and accuracy. Our evaluations will compare the differentially private distributions to their original counterparts, using several metrics. As in [9], we will use Earth Mover's Distance (EMD) to compare the two sets of distributions. In addition, other metrics such as *daily range* can also be used. Overall, once evaluated for accuracy, our work can represent a significant step towards making large-scale well-validated mobility models provably private and therefore easier to distribute and build from.

3. REFERENCES

[1] M. E. Andrés, N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi. Geo-indistinguishability: Differential privacy for location-based systems. *CoRR*, abs/1212.1984, 2012.

[2] A. Blum, C. Dwork, F. McSherry, and K. Nissim. Practical privacy: the sulq framework. In *Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, PODS '05, pages 128–138, New York, NY, USA, 2005. ACM.

[3] R. Chen, B. C. M. Fung, B. C. Desai, and N. M. Sossou. Differentially private transit data publication: a case study on the montreal transportation system. In *KDD*, pages 213–221, 2012.

[4] C. Dwork. Differential privacy. In *ICALP '06: Proceedings of the 33rd International Colloquium on Automata, Languages and Programming (2)*, pages 1–12, 2006.

[5] C. Dwork, F. Mcsherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *TCC '06: In Proceedings of the 3rd Theory of Cryptography Conference*, pages 265–284, 2006.

[6] M. Hay, V. Rastogi, G. Miklau, and D. Suciu. Boosting the accuracy of differentially private histograms through consistency. *PVLDB*, 3(1):1021–1032, 2010.

[7] S.-S. Ho and S. Ruan. Differential privacy for location pattern mining. In *Proceedings of the 4th ACM SIGSPATIAL International Workshop on Security and Privacy in GIS and LBS*, SPRINGL '11, pages 17–24, New York, NY, USA, 2011. ACM.

[8] S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, M. Martonosi, J. Rowland, and A. Varshavsky. Identifying important places in people's lives from cellular network data. In *Proceedings of the 9th international conference on Pervasive computing*, Pervasive'11, pages 133–151, Berlin, Heidelberg, 2011. Springer-Verlag.

[9] S. Isaacman, R. A. Becker, R. Cáceres, M. Martonosi, J. Rowland, A. Varshavsky, and W. Willinger. Human mobility modeling at metropolitan scales. In *MobiSys*, pages 239–252, 2012.

[10] F. McSherry. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. *Commun. ACM*, 53(9):89–97, Sept. 2010.

[11] F. McSherry and R. Mahajan. Differentially-private network trace analysis. In *SIGCOMM*, pages 123–134, 2010.

[12] S. L. Warner. Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias. *Journal of the American Statistical Association*, 60(309):63+, Mar. 1965.

Unique in the Crowd: The privacy bounds of human mobility

Yves-Alexandre de Montjoye*[1,2], César A. Hidalgo[1,3,4], Michel Verleysen[2] & Vincent D. Blondel[2,5]

[1] *Massachusetts Institute of Technology, Media Lab, 20 Ames Street, Cambridge, MA 02139 USA*

[2] *Université catholique de Louvain, Institute for Information and Communication Technologies, Electronics and Applied Mathematics, Avenue Georges Lemaître 4, B-1348 Louvain-la-Neuve, Belgium*

[3] *Harvard University, Center for International Development, 79 JFK Street, Cambridge, MA 02138, USA*

[4] *Instituto de Sistemas Complejos de Valparaíso, Paseo 21 de Mayo, Valparaíso, Chile*

[5] *Massachusetts Institute of Technology, Laboratory for Information and Decision Systems, 77 Massachusetts Avenue, Cambridge, MA 02139, USA*

We study fifteen months of human mobility data for one and a half million individuals and find that human mobility traces are highly unique. In fact, in a dataset where the location of an individual is specified hourly, and with a spatial resolution equal to that given by the carrier's antennas, four spatio-temporal points are enough to uniquely identify 95% of the individuals. We coarsen the data spatially and temporally to find a formula for the uniqueness of human mobility traces given their resolution and the available outside information. This formula shows that the uniqueness of mobility traces decays approximately as the 1/10 power of their resolution. Hence, even coarse datasets provide little anonymity. These findings represent fundamental constraints to an individual's privacy and have important implications for the design of frameworks and institutions dedicated to protect the privacy of individuals.

Despite its importance, privacy has mainly relied on informal protection mechanisms. For instance, tracking individuals' movements has been historically difficult, making them de-facto pri-

*Correspondence should be addressed to Yves-Alexandre de Montjoye (email: yva@mit.edu)

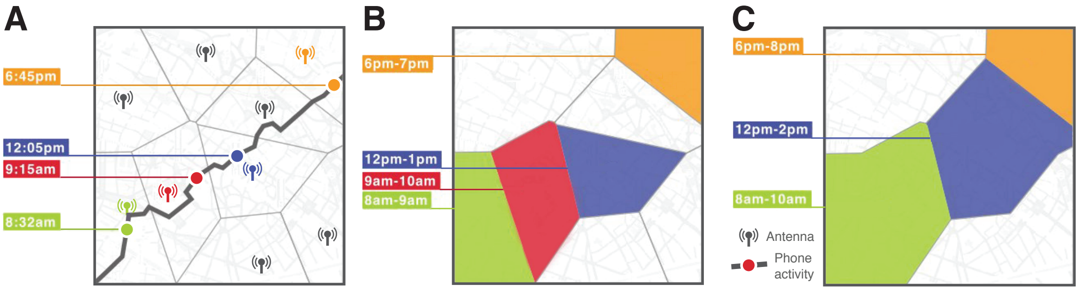


Figure 1: **(A)** Trace of an anonymized mobile phone user during a day. **(B)** The same user's trace as recorded in a mobility database. **(C)** The same individual's trace when we lower the resolution of our dataset through spatial and temporal aggregation.

vate. Modern information technologies such as the Internet and mobile phones, however, magnify the uniqueness of individuals, further enhancing the traditional challenges to privacy. While in the past, mobility traces were only available to mobile phone carriers, the advent of smartphones and other means of data collection has made these broadly available. Mobility data is among the most sensitive data currently being collected

A simply anonymized dataset does not contain name, home address, phone number or other obvious identifier. Yet, if individual's patterns are unique enough, outside information can be used to link the data back to an individual. We study the unicity of mobility traces in a mobile phone dataset containing 15 months of mobility data for 1.5M people, a significant and representative part of the population of a small European country. We show that four randomly chosen points are enough to uniquely characterize 95% of the users ($\mathcal{E} > .95$), whereas two randomly chosen points still uniquely characterize more than 50% of the users ($\mathcal{E} > .5$). This shows that mobility traces are highly unique, and can therefore be re-identified using little outside information.

Nonetheless, $\mathcal{E}$ depends on the spatial and temporal resolution of the dataset. Here, we determine this dependence by lowering the resolution of our dataset through spatial and temporal aggregation. We show that it is possible to find one formula to estimate the uniqueness of traces given both, the spatial and temporal resolution of the data, and the number of points available to an outside observer. We show that the uniqueness of a trace decreases as the power function $\mathcal{E} = \alpha - x^{\beta}$, for decreases in both the spatial and temporal resolution (x), and for all considered p = 4, 6, 8 and 10. The power-law dependency of $\mathcal{E}$ means that, on average, each time the spatial or temporal resolution of the traces is divided by two, their uniqueness decreases by a constant factor $\sim (2)^{-\beta}$. This implies that privacy is increasingly hard to gain by lowering the resolution of a dataset.

We also shows that $\mathcal{E}$ increases with p. The mitigating effect of p on $\mathcal{E}$ is mediated by the exponent β which decays linearly with p: $\beta = 0.157 - 0.007 * p$. The dependence of β on p implies that a few additional points might be all that is needed to identify an individual in a dataset with a lower resolution. Because of the functional of $\mathcal{E}$ on p through the exponent β, mobility datasets are likely to be re-identifiable using information on only a few outside locations.

We showed that the uniqueness of human mobility traces is high, thereby emphasizing the importance of the idiosyncrasy of human movements for individual privacy. This uniqueness means that little outside information is needed to re-identify the trace of a targeted individual even in a sparse, large-scale, and coarse mobility dataset. These results should inform future thinking in the collection, use, and protection of mobility data. Going forward, the importance of location data will only increase and knowing the bounds of individual's privacy will be crucial in the design of both future policies and information technologies.

Session 11

Information propagation

The use of mobile phone call record data for malaria control and elimination strategic planning

Tatem A.J. [1,2*], Huang Z. [3,4], Kumar U. [3,5], Pindolia, D.[3,4], Kandula, D.[6] and Lourenco, C. [6,7]

1. Department of Geography and Environment, University of Southampton, UK 2. Fogarty International Center, National Institutes of Health, Bethesda, USA., 3. Department of Geography, University of Florida, Gainesville, USA., 4. Emerging Pathogens Institute, University of Florida, Gainesville, USA., 5. Department of Computer Science, University of Florida, Gainesville, USA., 6. Clinton Health Access Initiative, Boston, USA, 7. National Vector-borne Diseases Control Programme, Windhoek, Namibia.

E-mails: andy.tatem@gmail.com, seenhzj@gmail.com, udayan.kumar@gmail.com, dkandula@clintonhealthaccess.org, clourenco@clintonhealthaccess.org

Introduction

The control and elimination of an infectious disease from a local area or country is often complicated by human movement bringing infections in from endemic areas. In the case of malaria, parasite importation to transmission-receptive areas has resulted in the resurgence and reestablishment of the disease in previously free areas, and represents a significant challenge for the 36 countries and regions now planning for elimination. The design of strategic plans for controlling, eliminating and preventing malaria reestablishment should therefore ideally account for human and parasite movement patterns. Here we describe how anonimized mobile phone call data records can be analysed to provide information to support planning using the example of Namibia.

Data

Mobile phone call data records covering October 2010 to September 2011 were provided by the leading mobile phone service provider in Namibia, MTC, who reported a 90% market share. Daily locations were calculated for anonymous subscribers using the location of calls and texts at one of 402 regions across the country, following methods outlined in other similar studies [1-5]. Movements between locations were calculated by examining the temporal sequences of communications and assigning a movement to a new location and a time of this move when the region through which the call/text was routed changed.

To characterize spatial variations in malaria transmission intensity across the country, and the uncertainties that exist in these measurements, 25%, 50% and 75% quartile *P. falciparum* prevalence estimates in the 2-10 year age group for 2010 were obtained [6]. Previously defined mathematical models [7] were used to adjust the malaria risk maps to the 15-30 age group, to match that of the majority of subscribers and travellers [8, 9]. To adjust estimated malaria risks to capture the strong seasonality that occurs in Namibia, reported outputs from the national malaria control program surveillance system in 2011 were obtained to provide quantitative risk adjustments for low and high transmission season months.

Movements of infections were calculated for two types of traveller, following previous approaches [5, 10]: (i) 'Returning residents'; Residents of a location who visited an endemic area then returned to their home location, potentially bringing an infection with them, and (ii) 'Visitors'; Residents of an endemic area who visited a new location and potentially carried an infection with them.

Results

Human movements

Figure 1(a) shows that population movements in Namibia follow patterns seen elsewhere [1, 3, 11], but with a deviation at around 400km from a smooth power law distribution, corresponding to the distance apart of the two major regions of high population density.

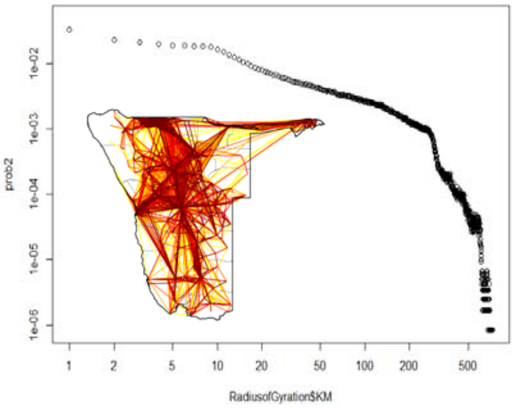


Figure 1. (a) Plot of radius of gyration for all movements in Namibia, with inset map showing movement totals between regions over the Oct 2010-Sept2011 period. Rates of movement are coloured from yellow (lowest) to red (highest).

Communities

Clusters of communities that are connected to one another through high levels of human or malaria parasite movement were mapped to aid in the regional planning of interventions, identifying how an intervention in one community may impact imported case numbers in surrounding communities. The identification of distinct communities within the weighted networks of human and malaria parasite movements was undertaken using a modularity optimization algorithm [12], with results for human travel shown in figure 2.

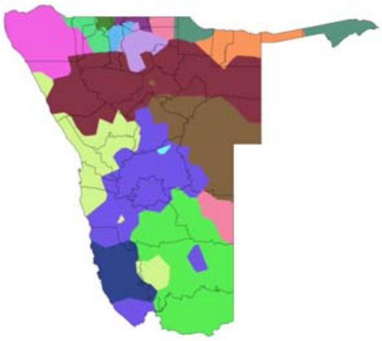


Figure 2. Mapped communities of human travel - regions mapped with the same colour are ones for which movements within them are stronger than movements between surrounding communities.

Sources and sinks

Targetting the largest exporter communities ('sources') of infections is likely to have an impact on the numbers of infections seen in surrounding areas that are net importers of infections ('sinks') and be a better use of limited resources. For each location, the estimated total number of infections exported or imported annually, and by month, for visitors and returning residents were summed, and the differences between these values calculated to assess whether locations were likely to be sources or sinks of infections (figure 3).

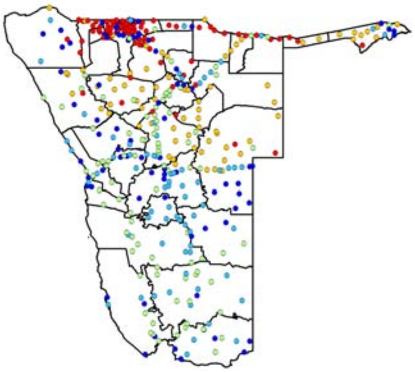
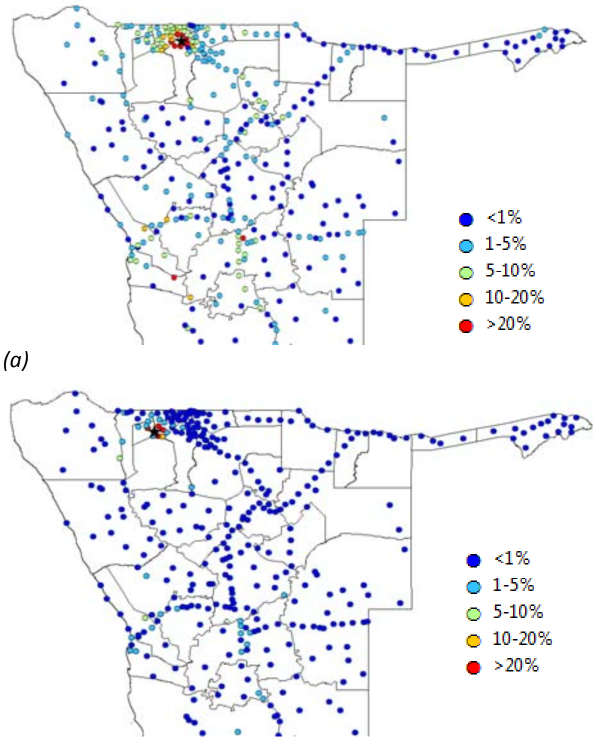


Figure 3. Map of 'sources' (net exporters) and 'sinks' (net importers) of malaria infections. Areas coloured red are infection sources and those coloured blue are sinks. Those coloured green are neither sinks nor sources.

The likely regional effects of targeting control efforts on the largest source and sink areas are shown in figure 4, with substantial differences evident, due to differing levels of connectivity through parasite movement.



(a)

(b)

Figure 4. The effects of human and parasite mobility on control targetting effectiveness. (a) The percentage reduction in estimated imported case numbers through reducing parasite exportation numbers to zero in the area marked with a star, which is one of the major source regionsin figure 3; (b) The percentage reduction in estimated imported case numbers through reducing parasite exportation numbers to zero in the area marked with a star, which is one of the major sink regions in figure 3.

Conclusions

As Namibia makes progress towards elimination, imported cases will make up an increasingly large proportion of total cases seen, both in terms of cases imported into Namibia from Angola, and cases seen in districts close to elimination that are imported from high transmission districts. Such imported cases will become increasingly important in terms of threatening success in achieving elimination. With mobile phone usage proliferation continuing and data continually being collected by network providers, a huge potential exists to make operational use of such valuable data in infectious disease control and elimination.

References

1. Gonzalez MC, Hidalgo CA, Barabasi AL. Understanding individual human mobility patterns. Nature. 2008 Jun 5;453(7196):779-82.

2. Tatem AJ, Qiu Y, Smith DL, Sabot O, Ali AS, Moonen B. The use of mobile phone data for the estimation of the travel patterns and imported *Plasmodium falciparum* rates among Zanzibar residents. Malar Journal. 2009;8:287.

3. Song C, Qu Z, Blumm N, Barabasi AL. Limits of predictability in human mobility. Science. 2010 Feb 19;327(5968):1018-21.

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

5. Wesolowski A, Eagle N, Tatem AJ, Smith DL, Noor AM, Snow RW, et al. Quantifying the impact of human mobility on malaria. Science. 2012;338:267-270.

6. Gething PW, Patil AP, Smith DL, Guerra CA, Elyazar IR, Johnston GL, et al. A new world malaria map: *Plasmodium falciparum* endemicity in 2010. Malar J. 2011;10:378.

7. Smith DL, Guerra CA, Snow RW, Hay SI. Standardizing estimates of the *Plasmodium falciparum* parasite rate. Malaria Journal. 2007 Sep 25;6.

8. Communications Regulatory Authority of Namibia. Universal Service Baseline Study. Windhoek: Communications Regulatory Authority of Namibia,; 2011.

9. MEASURE DHS. Namibia Demographic and Health Survey. Washington DC: USAID; 2006.

10. Le Menach A, Tatem AJ, Cohen JM, Hay SI, Randell H, Patil AP, et al. Travel risk, malaria importation and malaria transmission in Zanzibar. Sci Rep. 2011;1:93.

11. Simini F, Gonzalez MC, Maritan A, Barabasi AL. A universal model for mobility and migration patterns. Nature. 2012 Apr 5;484(7392):96-100.

12. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech-Theory E. 2008 Oct.

Time-varying networks and the weakness of strong ties

Márton Karsai *†1, **Nicola Perra** *2, **Alessandro Vespignani** *‡§3
*Laboratory for the Modeling of Biological and Socio-technical Systems, Northeastern University, Boston MA 02115 USA
†Department of Biomedical Engineering and Computational Science, Aalto University, P.O. Box 12200, Finland
‡Institute for Scientific Interchange Foundation, Turin 10133, Italy
§Institute for Quantitative Social Sciences at Harvard University, Cambridge, MA, 02138
[1] m.karsai@neu.edu, [2] n.perra@neu.edu, [3] a.vespignani@neu.edu

In this work we analyze a large scale mobile call dataset to investigate the temporal evolution of the egocentric network of active individuals. We empirically observe a simple quantitative statistical law characterizing the memory of agents and encode the observed dynamics in a reinforcement process defining a generative computational network model with time-varying connectivity patterns. This reinforced activity-driven network model spontaneously generates the basic dynamic process for the differentiation between strong and weak ties. The model is used to study the effect of time-varying heterogeneous interactions on the spreading of information on social networks. We observe that the presence of strong ties may severely inhibit the large scale spreading of information by confining the process among agents with recurrent communication patterns. Our results provide the counterintuitive evidence that strong ties may have a negative role in the spreading of information across communities.

In the last ten years the access to high resolution datasets from mobile devices, communication, and pervasive technologies has propelled a wealth of developments in the analysis of social networks [1, 2, 3]. Particular efforts have been devoted to characterize how their structure influences the critical behavior of dynamical processes evolving on top of them. However, the large majority of the approaches put forth to tackle this subject utilise a time-aggregated representation of the interactions and neglect their time-varying nature. Indeed, the concurrency, and time ordering of interactions, even if the social network contains stable relationships, are crucial and may have considerable effects [4, 5, 6].

The characterization, and modeling of time-varying networks are still open, and active areas of research [7]. A simplification of this framework has been recently proposed by the activity-driven generative algorithm for time-varying networks [5]. This approach defines the activity potential, a time invariant function characterizing the agents' interactions, and constructs an activity-driven model capable of encoding the instantaneous temporal description of the network dynamics. However, this framework is memoryless, and it misses important features of real world systems. In this work we propose to extend the activity-driven modeling framework in order to

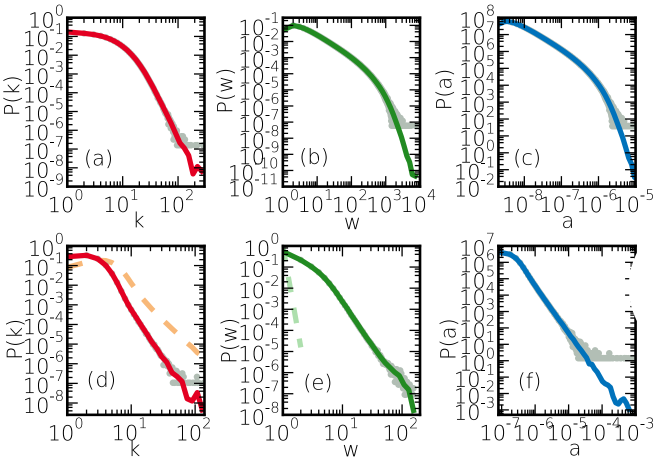


Figure 1: *Distributions of the characteristic measures of the aggregated mobile phone call network, and the reinforced activity-driven network. In panels (a), and (d) we plot the degree distributions. In panels (b), and (e) we plot the edge-weight distributions. Finally, in panels (c), and (f) we plot the node-activity distributions.*

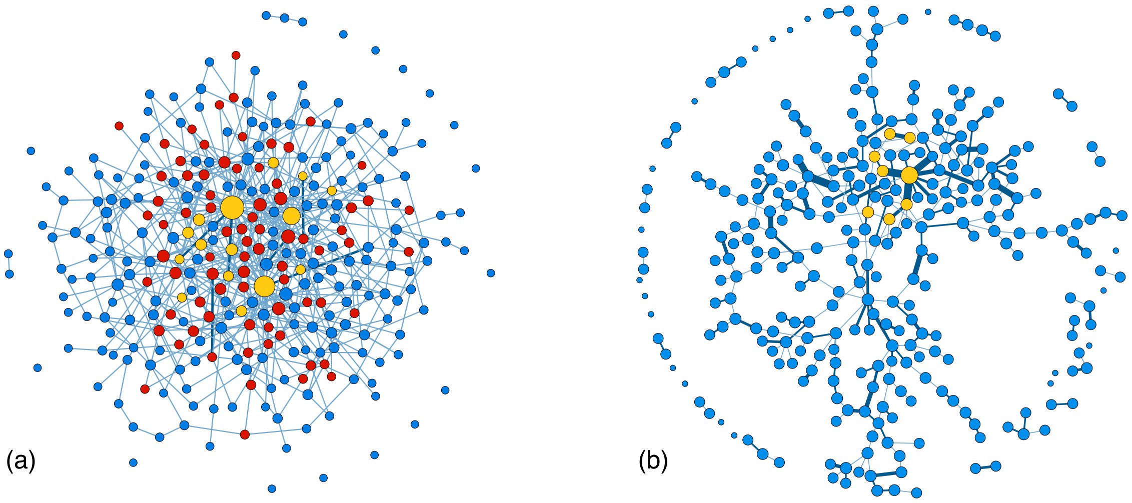


Figure 2: *Rumor spreading processes in (a) memoryless and (b) reinforced activity-driven networks of the same parameters. Nodes colors assign the actual node states as ignorant (blue), spreader (red) and stifler (green) states after the same number of evaluation steps. Node sizes, color, and width of edges represent the corresponding degrees and weights.*

explain the effect of the repetitive emergence of interactions within one's social circle. We perform a thorough analysis of a large-scale mobile phone-call dataset, which contains the records of time-stamped communication events of millions of individuals. We show that recurrence interaction of connected individuals can be explained by simple memory effects synthesized by non-Markovian reinforcement processes. The introduction of this mechanism in the activity-driven model allows us capturing the evolution of he egocentric network of each actor in the system, covering also its global dynamics. In particular, induces the proper weight heterogeneities reducing the degree diversity of the emerging structure in good agreement with the empirical observations as it is shown in Fig.1.

Using this model we study the effect of repetitive time-varying interactions on a family of rumor spreading models [8]. We assume that the time scales of the contact patterns evolution, and the spreading process are comparable. Interestingly, our findings clearly show that memory in the interaction dynamics hamper the rumor spreading process. Strong ties have an important role in the early cessation of the spreading dynamics. They favor interactions with agents already aware of the rumor and allow the rumor to diffuse only locally. This is illustrated in Fig.2 where the difference in the contagion level between the memoryless and reinforced process spreading is apparent. This evidence points out that strong ties may have an active role in weakening the spreading of information by constraining the dynamical process in clumps of strongly connected social groups. We validate the basic assumption, and modeling framework against results of data-driven simulations performed in the actual mobile call time-varying network.

References

[1] Albert, R & Barabási, A. L. (2002) *Rev. Mod. Phys.* **74**, 47–97.

[2] Newman, M. (2010) *Networks: An Introduction.* (Oxford University Press, USA), 1 edition.

[3] Barrat, A, Barthélemy, M, & Vespignani, A. (2008) *Dynamical Processes on Complex Networks.* (Cambridge University Press), 1 edition.

[4] Karsai, M., et.al. (2011) *Phys. Rev. E* **83**, 025102(R).

[5] Perra, N, et.al. (2012) *Sci. Rep.* **2**, 469.

[6] Perra, N, et.al. (2012) *Phys. Rev. Lett.* **109**, 238701.

[7] Holme, P & Saramäki, J. (2012) *Phys. Rep.* **519**, 97–125.

[8] Daley, D. J & Kendall, D. G. (1964) *Nature* **204**, 1118.

Limited communication capacity unveils strategies for human interaction

Giovanna Miritello,[1, 2] Rubén Lara,[2] Manuel Cebrian,[3, 4] and Esteban Moro[1, 5, *]

[1] *Departamento de Matemáticas & GISC, Universidad Carlos III de Madrid, 28911 Leganés, Spain*
[2] *Telefónica Research, 28050 Madrid, Spain*
[3] *NICTA, Melbourne, Victoria 3010, Australia*
[4] *Department of Computer Science & Engineering, University of California at San Diego, La Jolla, CA 92093, USA*
[5] *Instituto de Ingeniería del Conocimiento, Universidad Autónoma de Madrid, 28049 Madrid, Spain*

Social connectivity is the key process that characterizes the structural properties of social networks and in turn processes such as navigation, influence or information diffusion. An empirical limitation of studies to date is the definition of what constitutes an active tie that retains capacity to transfer information in the future. The limited observational length of existing human interaction datasets, together with the bursty nature of dyadic communications lie at the core of this problem, hindering the discrimination of inactive ties from large inter-event times in active ones. Here we develop a method for the detection of tie activation/deactivation, and apply it to a large longitudinal, cross-sectional communication dataset ($\approx$ 19 months, $\approx$ 20 million people). Contrary to the perception of ever-growing connectivity, we observe that individuals exhibit a finite communication capacity, which limits the number of ties they can maintain active. In particular we find that men have an overall higher communication capacity than women and that this capacity decrease gradually for both sexes over the lifespan of individuals (16-70 years). We are then able to separate communication capacity from communication activity, revealing a diverse range of tie activation patterns, from stable to exploratory. We find that, in simulation, individuals exhibiting exploratory strategies display longer time to receive information spreading in the network those individuals with stable strategies. Our principled method to determine the communication capacity of an individual allows us to quantify how strategies for human interaction shape the dynamical evolution of social networks.

Many different forces govern the evolution of social relationships making them far from random. In recent years, the understanding of what mechanisms control the dynamics of activating or deactivating social ties have uncovered forces ranging from geography to structural positions in the social network (*e.g.* preferential attachment, triadic closure), to homophily [2]. These finding are pervasive in empirical analyses across cultures, communication technologies and interaction environments [3–6, 10–15].

However, the incorrect assumption that time, attention and cognition are elastic resources has blurred the study of how individuals manage their social interactions over time [17–19]. Understanding such social strategies is not only of paramount importance to make progress in the characterization of human behavior, but also to improve our current description of social networks as evolutionary objects against the (aggregated) ever-growing or static pictures of the social structure.

Several reasons have hampered the observation of tie activation/deactivation dynamics in social networks at large scale: on the one hand, studies of diffusion based on datasets from pre-electronic eras have safely assumed that tie activation/deactivation is a much slower process than interactions within a tie, and thus their dynamics might be safely neglected [20–22]. However, the current ability to communicate faster and further than ever accelerates tie dynamics in an unprecedented manner to the point that tie activation/deactivation may rival in time with processes like information spreading. On the other hand, available data about how ties form or decay were restricted to egocentric, small social networks and/or short periods of time which made it difficult to assess the universality of the results obtained and their extension to other situations [6, 8, 9]. Finally, although in some online social networks there are explicit rules for the establishment of social ties, in most cases activity is the only way to assess the existence or not of the tie [23, 24]. Online social networks are plagued with this problem due to the cheap cost of maintaining "friends" which are ir fact already deactivated relationships [25]. However, u ing activity as proxy for tie presence is a problem in mo communication channels like mobile phone calls, emails, electronic social networks etc., since tie activity is very bursty [26] and so far there is no clear method to discriminate those social ties that are already inactive from large-inter even times within active relationships [27].

In this communication we study the formation and decay of communication ties using a large (both cross-sectional and longitudinal) database of of the anonymized voice calls of about 20 million users that form 700 million communication ties. We use a novel method to determine the presence of a tie based on activity data (see figure 1) and apply it to understand the dynamics of formation and destruction of ties. While previous work has analyzed datasets over insufficient time-spans (thus confounding tie dynamics with the bursty nature of ties) and/or small cohorts (thus unable to obtain statistically

[*] Corresponding author emoro@math.uc3m.es

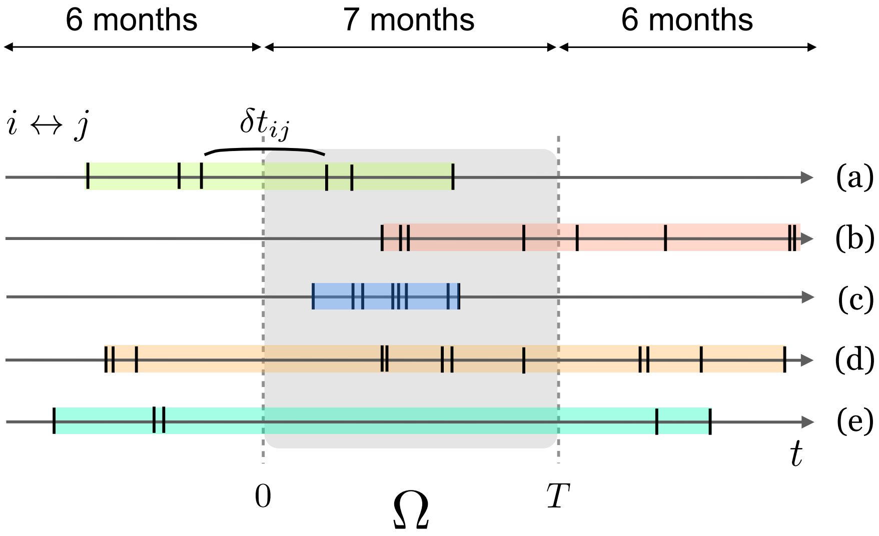


FIG. 1. **Detection of tie activation/deactivation** Schematic view of the time intervals considered in our database and the different situations of tie activation/deactivation and the interplay between the tie communication patterns and tie activation/deactivation for a given observation time window Ω of length $T = 7$ months (shadowed area). Each line refers to a different tie while each vertical segment indicates a communication event between $i \leftrightarrow j$ and δt_{ij} is the inter-event time in the $i \leftrightarrow j$ time series. Our method is based on the *observation* of tie activity in a time window before/after Ω: if tie activity is observed in the 6 months before Ω then it is considered an old tie [cases (a) and (d)]; on the other hand, if activity is observed in the 6 months after Ω we will assume that the tie persists [cases (b) and (d)]. In any other case, we will consider that the tie is activated and/or deactivated in Ω [cases (a), (b) and (c)].

significant results to discriminate different communication strategies), our method and extensive dataset allow us to determine, for the first time, the social strategies which result from the limited communication capacity of individuals. We have also made some analysis of how social strategies depend on sex and age of the individuals and about how the strategies are correlated with the (dynamical) topological structure of the network and on the dependence of the access to information with the strategy displayed.

Our contribution shows a number of important results which we summarize as follows:

- We develop a new methodology to investigate temporal networks, an issue which is convoluted with the way social networks are observed and modeled, and which has been recently been pointed out as a problem in the field of social networks [28].

- Contrary to conventional wisdom in which social connectivity is an ever-growing quantity, we have found that instantaneous connectivity is bounded: humans manage their interactions so that the number of open connections is kept constant through time.

- We find that men have an overall higher communication capacity than women and that this capacity decrease gradually for both sexes over the lifespan of individuals (16-70years).

- A principled methodology to determine the communication capacity of an individual enables us to characterize individual communication strategies: while some individuals are *social keepers* (they maintain a clusterized static network around them) other seem to undergo a *social exploring* strategy (in which many ties are formed and destroyed in time).

- Finally we use computer simulations to investigate the possibility that social explorers have a competitive advantage towards information advantage because of their fast and distant tie formation dynamics. Counterintuitively, we found that social keepers receive information before social explorers do. The answer to this paradox is that, despite having a large variability in the social structure around a social explorer, the amount of time allocated to their contacts is low enough to produce countervailing effects and hinder information diffusion.

This result is important as it provides conclusive evidence for the divergence between the static and dynamic characterizations human interaction. Fine-grained, longitudinal and cross-sectional data as the one presented in this study are then needed to fully understand processes such as navigation, influence and information diffusion as they happen concurrently and possibly entangled to the unfolding of social strategies in time.

[*] emoro@math.uc3m.es

[2] Rivera, M.T., Soderstrom, S.B. and Uzzi, B., (2010) Dynamics of Dyads in Social Networks: Assortative, Relational, and Proximity Mechanisms. Annual Review of Sociology, 36(1), pp.91-115.

[3] R.S. Burt, (2000) Deactivation functions, Social Networks **22**, 1-28.

[4] Martin, J. L. and Yeung, K.-T. (2006) Persistence of close personal ties over a 12-year period. Social Networks 28:331-362.

[5] Hidalgo, C. & Rodriguez-Sickert, C., (2008) The dynamics of a mobile phone network. Physica A, 387(12), pp. 3017-3024.

[6] Raeder, T., Lizardo, Chawla N.V., Hachen, D. (2011) Predictors of short-term deactivation of cell phone contacts in a large scale communication network. Social Networks Volume 33, Issue 4, Pages 245-257.

[7] J. Saramäki *et al.* (2012) The persistence of social signatures in human communication, arXiv:1204.5602.

[8] N. Eagle, A. Pentland, (2006) Reality mining: sensing complex social systems, Personal and Ubiquitous Computing 10 255?268.

[9] C. Cattuto, W. van den Broeck, A. Barrat, V. Colizza, J.-F. Pinton, A. Vespignani, (2010) Dynamics of person-to-person interactions from distributed RFID sensor networks, PLoS One 5 e11596.

[10] Kossinets, G., Watts, D. J. (2009) Origins of homophily in an evolving social network. American Journal of Sociology, 115(2):405-450.

[11] Crandall, D., Cosley, D., Huttenlocher,D., Kleinberg, J. and Suri, S., (2008) Feedback effects between similarity and social influence in online communities, in KDD'08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. New York, NY, USA: ACM, pp. 160-168.

[12] Romero, D.M., Meeder,B., Barash,V., Kleinberg, J., (2011) Maintaining Ties on Social Media Sites: The Competing Effect of Balance, Exchange and Betweenness, Proceedings of ICWSM 2011: the fifth AAAI conference on Weblogs and Social Media, Barcelona, Spain.

[13] Wang, J., Suri, S., Watts, D.J.(2012) Cooperation and assortativity with dynamic partner updating, Proc Natl Acad Sci U S A 109: 14363-14368.

[14] Apicella, C. L., Marlowe, F.W., Fowler, J.H., Christakis, N.A. (2012) Social networks and cooperation in hunter-gatherers, Nature 481: 497-501.

[15] Rand, D.G., Arbesman, S., Christakis, N.A. (2012) Dynamic social networks promote cooperation in experiments with humans, Proc Natl. Acad Sci U S A 108: 19193-19198.

[16] Krings, G. Karsai, M., Bernharsson, S., Blondel, V.D., Saramäki, J. Effects of time window size and placement on the structure of aggregated networks, EPJ Data Science, 2012, Volume 1, Number 1, 4. [3–6, 10–15].

[17] Wu, F., Huberman, B.A. (2007) Novelty and collective attention, Proc Natl Acad Sci U S A, 104,17599-17601.

[18] Hodas, N.O., Lerman, K. (2012) How visibility and divided attention constrain social contagion, arXiv preprint arXiv:1205.2736.

[19] Backstrom, L., Bakshy, E., Kleinberg, J., Lento, T.M., Rosenn, I. (2011) Center of attention: How facebook users allocate attention across friends, Proc. 5th International Conference on Weblogs and Social Media.

[20] Hagerstrand, T. (1968) Innovation diffusion as a spatial process, Chicago, USA: Univ. Chicago Press.

[21] Rogers, E.M. (1995) Diffusion of innovations, Simon and Schuster.

[22] Christakis, N.A. and Fowler, J.H. (2007) The spread of obesity in a large social network over 32 years, he New England journal of medicine, 357(4):370-379.

[23] Huberman B.A., Romero D.M. & Wu F. (2009) Social networks that matter: Twitter under microscope. First Monday 14: 1.

[24] Grabowicz, P.A., Ramasco, J.J., Moro, E., Pujol, J.M. & Eguíluz,V.M. (2011) Social features of online networks: the strength of weak ties in online social media, PLoS ONE, 7(11), e29358.

[25] Sibona, C., Walczak, S. (2011) Unfriending on Facebook: Friend request and online/offline behavior analysis, System Sciences (HICSS), 44th Hawaii International Conference on, pp.1-10.

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

[27] Song, C., Wang, D. and Barabasi, A.-L., (2012). Joint Scaling Theory of Human Dynamics and Network Science. arxiv preprint, arXiv:1209.1411v1.

[28] Krings, G. Karsai, M., Bernharsson, S., Blondel, V.D
Saramäki, J. Effects of time window size and placeme
on the structure of aggregated networks, EPJ Data Sc
ence, 2012, Volume 1, Number 1, 4.

Is Social Influence Always Positive? Evidence from a Large Mobile Network

Rodrigo Belo[*], Pedro Ferreira[†]

Carnegie Mellon University

1 Introduction

The design of a new product often involves incorporating *viral features* that increase the likelihood that adopters influence their peers to adopt the product as well (e.g., Aral and Walker, 2011). Network effects are a specific class of viral features that are frequently present in networked industries and more so in mobile networks, where pairwise communication still represents the major form of communication. We analyze a subset of products deployed by a large European mobile carrier and look at their characteristics in terms of viral features and network externalities. We develop a model to help identify the incentives in the adoption of these products.

We use randomization methods (e.g., Anagnostopoulos et al., 2008; La Fond and Neville, 2010; Belo and Ferreira, 2012) to identify social influence from observational data and find that network externalities contribute to an increase in the adoption rate of some products. For other products their contribution is negative.

These results have important management and policy implications. Social influence is not always positive and depends on the viral features of the products considered. Therefore, it is important to carefully design products that exhibit the correct characteristics for adoption to occur and spread as expected.

2 Identifying social influence

Social influence can be defined as the degree by which an action from an individual changes the behavior of someone else. For the purpose of this paper we look at how the adoption of a given promotion influences peers to adopt the same promotion. Social influence plays an important role in many diffusion models (e.g., Kermack and McKendrick, 1927; Bass, 1969; Granovetter, 1978; Watts and Dodds, 2007). Common to all these models is the characteristic that, under the right conditions, a small number of initial adopters may lead to a large number of adoptions.

However, these models are also consistent with other phenomena, such as heterogeneity in the propensity to adopt, homophily, correlated unobservables, or simultaneity. These phenomena pose significant challenges to the identification of social influence in observational data (Shalizi and Thomas, 2011). 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 of no influence (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.

3 Incentives to adopt with network externalities

We develop a model for how social influence can affect the adoption rate of products that exhibit network externalities. We focus on the specific case of free-calls within the network and look at the adoption incentives of two slightly different products.

By default with this carrier people pay for all the calls they make. Received calls are free. Consider now a product that offers free calls among members of a

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

network for a flat rate fee.

Consider a family of products k that allow calling for free. Type I promotions allow calling for free subscribers that also adopt the same product. Type II promotions allow calling for free any subscriber in the same network. Subscribers can choose either of these options. Both of them require subscribers to pay a fixed fee.

We assume each user has a fixed number of friends. Two subscribers are friends they exchange at least 3 calls in the same calendar month. Subscribers derive utility from calling each friend, u_{ij}, independently of how much she talks to other friends. We assume quadratic pairwise utility for the calls between user i and user j, c_{ij}:

$$u_i = \sum_{j \in F_i} u_{ij} = \sum_{j \in F_i} [a(c_{ij} + c_{ji}) - b(c_{ij} + c_{ji})^2 - pc_{ij}]$$

where p represents the price per call. If user i adopts type I promotion, she will pay a fixed fee, f_k, and not pay for calls:

$$u_i|A_k = \sum_{j \in F_i} [a(c_{ij} + c_{ji}) - b(c_{ij} + c_{ji})^2] - f_k$$

The user will choose to adopt product k if the utility derived from adopting is higher than the utility derived from not adopting, i.e., if $u_i|A_k \geq u_i$. Thus, i will adopt k iff:

$$d_{A_k i} \geq \frac{4b}{ap} f_k$$

where $d_{A_k i}$ represents the number of friends of i that will eventually adopt product k. User i will adopt if there is a minimum number of friends that will also adopt product k.

The incentives to adopt type II products are quite different. In this case, adoption occurs as long as there are at least a minimum number of friends that will not adopt:

$$d_{N_k i} \geq \frac{4b}{ap} f_k$$

where $d_{N_k i}$ represents the number of friends of i that will not adopt.

Thus, apparently somewhat similar products can generate different adoption incentives.

4 Data and Methods

We use an 11-month anonymized panel of data comprising of detailed information about all subscribers in a large mobile European network provider. The data include detailed call and SMS data records, pricing plans, and adoption of products and promotions.

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, location, 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, 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. We currently limit our analysis to a sample of 10,000 randomly selected subscribers and their direct neighbors.

We apply the shuffle test described by Anagnostopoulos et al. (2008) and use randomized versions of the data to infer an empirical distribution for the null scenario of no influence. We then compare the coefficient obtained from the original data with the null empirical distribution. This procedure has been shown to yield a lower bound for the magnitude of social influence and is described in detail in Belo and Ferreira (2012).

5 Results

We identify social influence on seven different products, one of them corresponding to a type I product — with free calls only to same-product adopters, while the other six correspond to type II products — free calls for all users in the network. Figure 1 depicts the results for the type I product and for one of the type II products (all other type II products are similar). Table 1 shows the total number of adopters for each of these products, the coefficient obtained from the original data, the average coefficient obtained from randomization and respective standard deviation, and our estimate of social influence.

All the empirical distributions have positive mean and are statistically different from zero, meaning that confounding factors, such as homophily are at play.

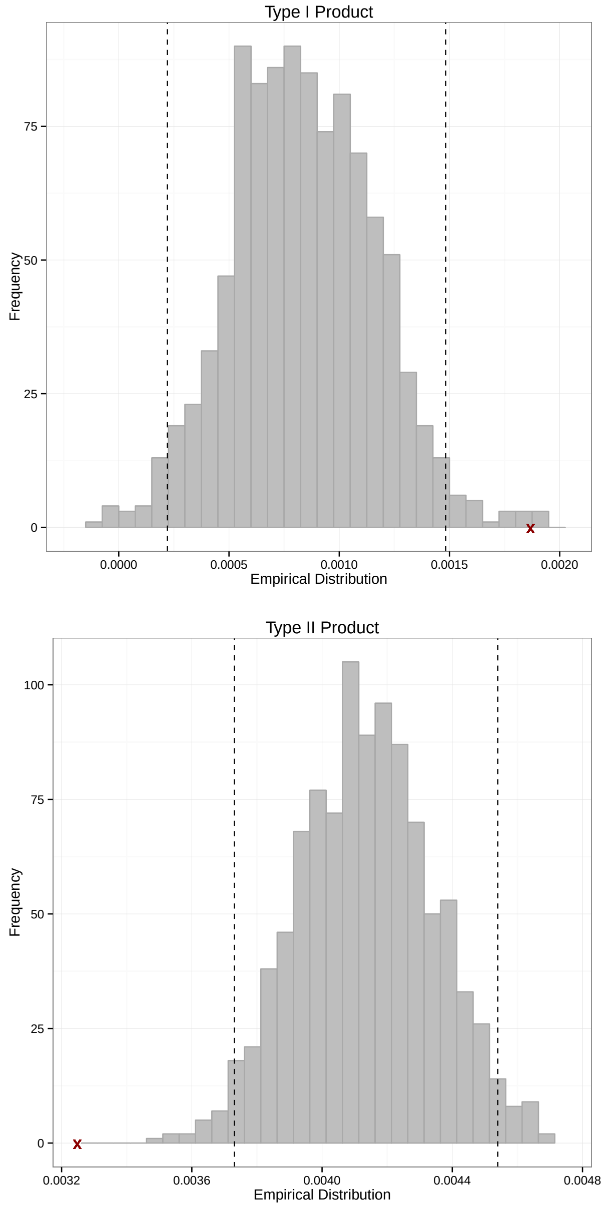


Figure 1: Type I and Type II Product coefficients over 1,000 adoption date shuffles. The '×' mark represents the coefficient obtained from the original data. Dashed lines represent 95% confidence intervals.

Prod. Type	Adp.	Orig. Coeff.	Emp. Dist	Marg. Eff.	Extra Adopt.
I	1126	.0019*** (4.5e-04)	8.4e-04 (2.1e-04)	1.0e-3***	148 (13%)
II (1)	534	.004*** (4.4e-04)	.0048 (1.7e-04)	-8.0e-4***	-33 (-6%)
II (2)	357	.0029*** (4.0e-04)	.0036 (1.6e-04)	-7.0e-4***	-22 (-6%)
II (3)	341	.0028*** (5.0e-04)	.0037 (2.0e-04)	-8.0e-4***	-29 (-9%)
II (4)	299	.0032*** (4.7e-04)	.0032 (1.6e-04)	-6.0e-5	
II (5)	205	.0015*** (4.0e-04)	.0018 (1.4e-04)	-2.5e-4**	-6 (-3%)
II (6)	110	.0015*** (4.1e-04)	.0011 (2.1e-04)	-3.6e-4	

Table 1: Influence estimates for type I and type II products.

Despite the existence of these effects, our estimates are aligned with the model outlined in section 3: social influence is positive for the type I product and negative for most type II products. For instance, in the case of the first type I product, the average coefficient obtained from randomization is 0.00084. Given that the coefficient obtained using the original data is .0019, 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 13% of the total observed adoption. In the case of the first type II product, peer influence reduces observed adoption in 6%.

6 Conclusion and Future Work

We show that peer influence can be either positive or negative in the adoption of products with network externalities. Previous literature on network externalities shows that the incentive to adopt increases with the number of friends that adopt. In this paper we show that network externalities can also trade-off with word-of-mouth, resulting in a decrease in the incentive to adopt as the number of friends that adopt increases. The direction of such incentives is partly determined by the design of the product.

This research is work in progress and we plan to improve on it by separating the network externalities incentives from the effect of word-of-mouth in the adoption of type II products. We intend to do this by looking at users with a relatively large number of non-adopter friends and test whether there is a positive total effect from social influence. In such a case the positive effect would have to come from word-of-mouth. We also plan to analyze under which conditions type I and type II products are profitable for the carrier, and to find the optimal fees for such products depending on the network topology.

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 and D. Walker. Creating social contagion through viral product design: A randomized trial of peer influence in networks. *Management Science*, 57(9):1623–1639, 2011.

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.

F. Bass. A New Product Growth for Model Consumer Durables. *Management Science*, 15(5):215, 1969.

R. Belo and P. Ferreira. Using Randomization Methods to Identify Social Influence in Mobile Networks. The Fourth IEEE International Conference on Social Computing, SocialCom 2012, 2012.

M. Granovetter. Threshold models of collective behavior. *American journal of sociology*, pages 1420–1443, 1978.

W. Kermack and A. McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A*, 115(772): 700–721, 1927.

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. *Sociological Methods & Research*, 40(2):211–239, 2011.

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

D. Watts and P. Dodds. Influentials, networks, and public opinion formation. *Journal of Consumer Research*, 34(4):441–458, 2007.

Poster

1

Mobility

Human Mobility and Predictability enriched by Social Phenomena Information

Nicolas B. Ponieman[1,2], Alejo Salles[2], and Carlos Sarraute[1]

[1]Grandata Labs, Argentina
{nico,charles}@grandata.com
[2]Physics Dept., Universidad de Buenos Aires, Argentina
alejo@df.uba.ar

1 Introduction

The information collected by mobile phone operators can be considered as the most detailed information on human mobility accross a large part of the population [1]. The study of the dynamics of human mobility using the collected geolocations of users, and applying it to predict future users' locations, has been an active field of research in recent years [2, 3].

In this work, we study the extent to which social phenomena are reflected in mobile phone data, focusing in particular in the cases of urban commute and major sports events. We illustrate how these events are reflected in the data, and show how information about the events can be used to improve predictability in a simple model for a mobile phone user's location.

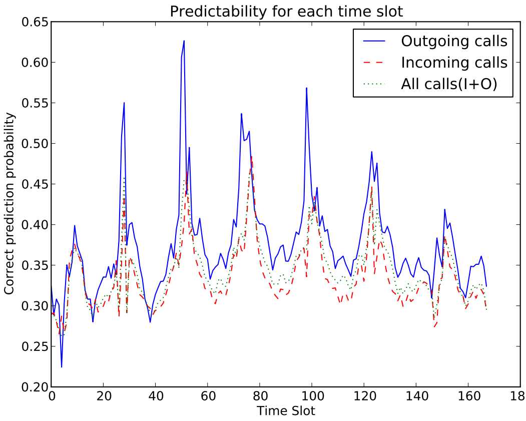


Figure 1: Users' location predictability by time slot. Blue: Outgoing calls. Red: Incoming calls. Green: All calls.

2 Mobile Data Source

Our data source is anonymized traffic information from a mobile operator in Argentina, focusing mostly in the Buenos Aires metropolitan area, over a period of 5 months. We use Call Detail Records (CDR) including time of the call, users involved, direction of the call (incoming/outgoing), the antenna used in the communication, and its position. The raw data logs contain around 50 million calls per day. CDRs are an attractive source of location information since they are collected for all active cellular users (about 40 million users in Argentina), and creating additional uses of CDR data incur little marginal cost.

3 Mobility Model

To predict a user's position, we use a simple model based on previous most frequent locations. In order to compute these locations, we split the week in time slots, one for each hour, totalizing $7 * 24 = 168$ slots per week. Since humans tend to have very predictable mobility patterns [1, 4, 5], this simple model turns out to give a good predictability baseline, achieving an average of around 35% correct predictions for a period of 2 weeks, training with 15 weeks of data, including peaks of above 50% predictability. This model was used as a baseline in [6], with which our results agree. In Figure 1 we show the average predictability for all time slots.

It is important to notice that it is possible to improve the accuracy of this simple model by clustering antennas, instead of defining each antenna as a location.

4 Urban Commute

The phenomenon of commuting is prevalent in large metropolitan areas (often provoking upsetting traffic jams and incidents), and naturally appears in mobile phone data. For instance, in [7] the authors study commute distances in Los Angeles and New York areas. Mobile data can lead to quantification

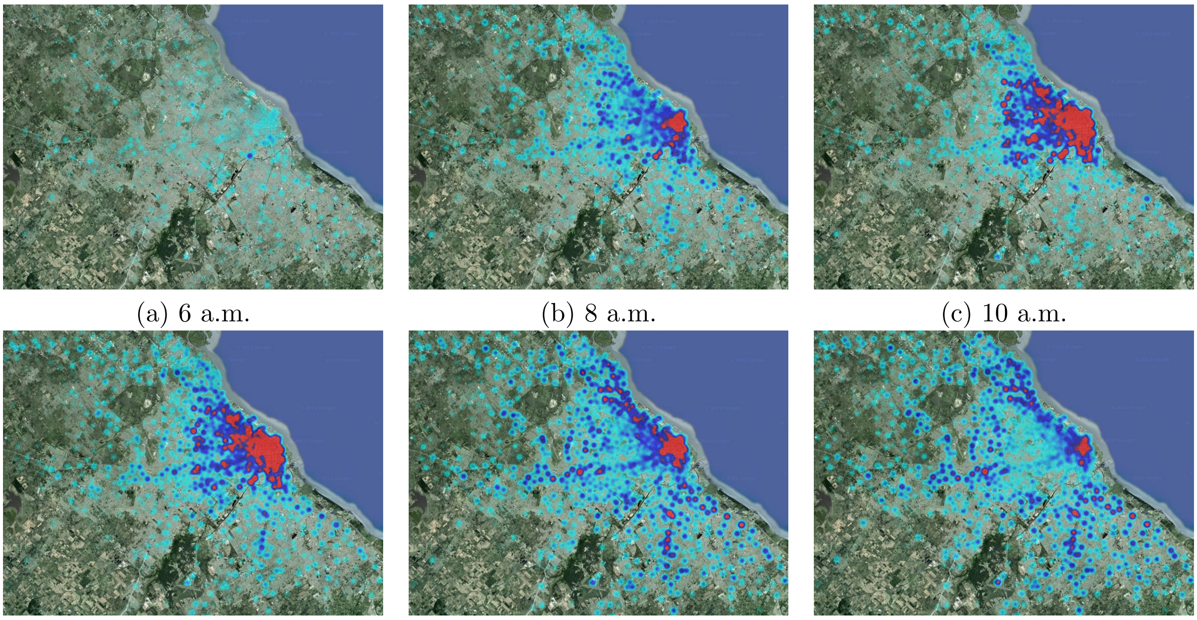

(a) 6 a.m.
(b) 8 a.m.
(c) 10 a.m.
(d) 5 p.m.
(e) 7 p.m.
(f) 8 p.m.

Figure 2: Commute to Buenos Aires city from the surrounding areas on a weekday, for different hours. Red color corresponds to a higher number of calls, whereas blue corresponds to an intermediate number of calls and light blue to a smaller one.

of this phenomenon in terms of useful quantities, which are much harder to measure directly. We include a series of call patterns illustrating the Buenos Aires commute in Figure 2[1].

From the data, we can estimate the radius of the commute (the average distance traveled by commuters). Considering the two most frequently used antennas as the important places for each user (home and work, see [8]), we get an average commute radius of 7.8 km (as a comparison, the diameter of the city is about 14 km, and the diameter of the considered metropolitan area is 90 km).

5 Sports Events

As in the urban commute case, we study human mobility in sports events as seen through mobile phone data. In Figure 3, we show how assistants to a Boca Juniors soccer match converge to the stadium in the hours prior to the game, and disperse outwards[1].

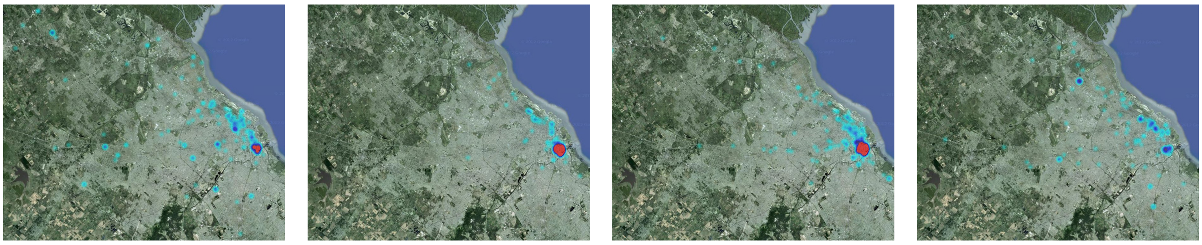
Note that postselecting the users attending the event necessarily produces the effect of having no calls outside the chosen area during the match, however, the convergence pattern observed is markedly different from the one seen for the same time slot of the week on a day with no match.

Improving Predictability with External Data

So far, our results allow us to understand (and quantify) social events through the analysis of mobile phone data. This understanding can be in turn used to improve the mobility model. Social relations among individuals have been used to improve predictability in mobility models before, as in [6], where social links learned from the mobile phone records are used to this end. Here, instead of peer to peer links learned from the mobile data, we show how an external data source can be used to improve the model.

We illustrate this effect using as proof of concept the case study of soccer matches. By taking the soccer fixture, we tag users as “Boca Juniors fans” if they make calls using antennas around the stadium and time slot where Boca plays for three consecutive matches (which include both home and away matches). Using this tagging, we can dramatically improve predictability for this group of Boca fans, even predicting positions that had never been visited by a user before. The predictability of the model for these users considering the fixture data rises for the matches to 38% – which doubles the 19% accuracy achieved by our previous model for the same set. Moreover, the initial model is only able to make predictions in 63% of events in the given set (as a consequence of a lack of information from the training set data), whereas the socially enriched model tries to predict 100% of the events during match days, which make the previous results

[1] Greater resolution versions of these maps, as well as additional figures, are available in the Labs section at www.grandata.com.



(a) 5 hours before (b) 1 hour before (c) 1 hour after (d) 3 hours after

Figure 3: Convergence to Boca Juniors stadium on hours prior to a soccer match, and dispersal after its end. Red color corresponds to a higher number of calls, whereas blue corresponds to an intermediate number of calls and light blue to a smaller one.

even more significant.

In order to understand these results, we illustrate with an example where the enriched model outperforms the simple model: the simple model would rarely predict a user's location on a different city, whereas the enriched model would do so if the user is a Boca fan, and Boca has an away match in that city.

6 Conclusion

We illustrated how social phenomena can be studied through the lens of mobile phone data, which can be used to quantify different aspects of these phenomena with great practicity. Furthermore, we showed how including external information about these phenomena can improve the predictability of human mobility models.

Although we showed this in a specific case as a proof of concept experiment, we note that this procedure can be extended to other settings, not restricted to sports but including cultural events, vacation patterns and so on (see [3] for a specially relevant application). The tagging obtained is useful on its own and is of great value for mobile phone operators. The big challenge in this line of work is to manage to include external data sources in a systematic way.

Lastly, the tag-based predictions can be taken to the community level. Defining, for instance, the "Boca Juniors fans" community, we can predict that if some users of this community make or receive calls in a certain location, other users in the community will do it as well.

References

[1] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

[2] Manlio De Domenico, Antonio Lima, and Mirco Musolesi. Interdependence and predictability of human mobility and social interactions. *Nokia Mobile Data Challenge Workshop*, 2012.

[3] Xin Lu, Linus Bengtsson, and Petter Holme. Predictability of population displacement after the 2010 haiti earthquake. *Proceedings of the National Academy of Sciences*, 109(29):11576–11581, 2012.

[4] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[5] Shan Jiang, Joseph Ferreira, and Marta C González. Clustering daily patterns of human activities in the city. *Data Mining and Knowledge Discovery*, pages 1–33, 2012.

[6] Eunjoon Cho, Seth A Myers, and Jure Leskovec. Friendship and mobility: user movement in location-based social networks. In *ACM SIGKDD*, pages 1082–1090. ACM, 2011.

[7] Sibren Isaacman, Richard Becker, Ramón Cáceres, Stephen Kobourov, Margaret Martonosi, James Rowland, and Alexander Varshavsky. Identifying important places in people's lives from cellular network data. *Pervasive Computing*, pages 133–151, 2011.

[8] Balázs Cs Csáji, Arnaud Browet, VA Traag, Jean-Charles Delvenne, Etienne Huens, Paul Van Dooren, Zbigniew Smoreda, and Vincent D Blondel. Exploring the mobility of mobile phone users. *Physica A: Statistical Mechanics and its Applications*, 2012.

Pisa Tourism Fluxes Observatory: deriving mobility indicators from GSM calls habits

Barbara Furletti, Lorenzo Gabrielli, Salvatore Rinzivillo, Chiara Renso
KDDLAB - ISTI CNR, Pisa, Italy
name.surname@isti.cnr.it

Abstract
The necessity to improve the management of the resources, urged many local governments to adhere to European initiatives in the context of competitiveness and sustainability, for creating the right balance between the welfare of tourists, the needs of the natural and cultural environment and the development and competitiveness of destinations and businesses. For many Italian Municipalities, this requirements become concrete with the establishment of a tourism monitoring systems that aims at survey these phenomenon through the analysis of heterogeneous data ranging from information of the territory, energy consumption, use of the land, and linked data (arrival and departure from the airport, bus, hotels etc). We describe the permanent observatory of touristic fluxes we realized in the town of Pisa where the standard indicators have been extended with an indicator of people presence extracted from mobile GSM call data and other exploratory analyses made by using the mobile phone data.we developed a method to partition the users into residents, commuters, in transit and visitors starting from a spatio-temporal profile inferred from people call habits.

Extending the TFO with GSM-based analysis
Art cities attract many tourists and visitors and these incoming flows concur to consume the cities resources both natural (energy, water, air, ...) and of services (parkings, public transportations, garbage collectors, ...). For these reasons some conflicts with residents and visitors can arise in the use of the limited resources. It is an hard task and a duty for the local administrations to guarantee general welfare for both citizen and visitors and the *Sustainable Tourism Planning* is more and more implemented in order to preserve local resources. In general, not only the tourism but also commuters flows cause criticism in the typical dynamics of a city (traffic increasing, public transportation congestion, pollution).
The Tourism Fluxes Observatory (TFO) carried out in cooperation with the Municipality of Pisa, aims at studying the fluxes of tourist visiting the town in order to evaluate the overall quality of the reception system on the territory. This is supposed to be a means of support for the Administrations and the private operators in assessing the overall quality of the reception system planning.
The data used in a TFO come from heterogeneous sources and represent observations of several socio-economic and environmental phenomena. Some of the indicators included in the TFO are directly or indirectly influenced by the presence of tourists on the territory like the number of tourist buses that stop at a parking, the consumption of water and energy, or the presence at the hotels.
The TFO we realized in Pisa enriches the standard studies of tourists flows integrating and using unconventional data like GSM and Social Media data (e.g. Flickr, Facebook, Twitter data). The use of these data supply a different view of the phenomena related to behaviors or cultural interests of the individuals (e.g., the call habits or the shared photos of the visited places).
It is known from the literature the ability of these new data to give quali-quantitative estimations of several phenomena as for example the traffic flows [1], the city dynamics, and an estimation of people presence at the points of interest [2]. Furthermore the use of GSM and Social Media data is particularly interesting because they overcome the problems of the massive data collection as for example the high costs for surveys and the limitation related to the availability of both huge amount and up-to-date data.
Besides the quantitative indicators extracted from the data provided by the local administrations and operators in the tourism sector, we extended the TFO with a set of analysis conduced on a GSM dataset. This dataset consists of Call Data Records (CDR) collected in the urban area of Pisa of about 232.200 users with a national mobile phone contract (no roaming users are included in the dataset). These users have been observed from January 9th to February 8th 2012, and generated around 7.8 million of tracks.
A CDR records is recorder for each call fo the user and has the following information:
<*Caller ID, ID Cell Start, Start Time, ID Cell End,Duration*>
where: *Caller ID* is the anonymous identifier of the caller, *ID Cell Start* and *ID Cell End* are the identifiers of the cell where the call starts and ends respectively, *Start Time* is the date and time when the call starts, and *Duration* is the call duration.

Taking inspiration from the literature, we used CDRs to estimate the presence of people in the city and to identify which areas are particularly popular and attractive. By studying the calls density, we are able to answer to some non-trivial questions about people habit: *How many...?* (the "volume" of people in the area of interest), *Where...?* (the origin and destination of people), and *When...?* (the temporal profile).
Figure 1, shows the density of people presence in the area of Pisa. The thick lines delimit the main districts, while the polygons are the Voronoi tessellation of the GSM coverage. The density is shown by using the colors in the map: the

higher density are red. The red cells labeled with (1) contain “Piazza dei Miracoli” (the square with the leaning Tower) and the Hospital; the red cell labeled with (2) contains the Galilei International Airport, and the cell labeled with (3) contains the National Research Council (CNR). These three areas actually attract many from tourist/visitors and workers (both resident and commuters). Starting from this analysis, giving just an idea on how people are distributed on the territory, it is possible to focus on a particular area (GSM cell) in order to analyse the origin and destination of the individuals. Considering the cell numbered as (4), that contains “Ponte di mezzo” (the Middle Bridge), we identified where people go after having traversing that area, and where people come from. Figure 2 shows the origin and destination cells involved in the movements to and from the cell (4): the thickness of the arrows identify the flow volumes, and the direction of the movements. The presence in each area can be further investigated by time. In particular, grouping the calls per hours we can plot the chart that represents how the presence change over the day. Figure 3 shows the variation of the presences in the cell number (4). Each separate group is a different day of the week starting from Monday. The presence, during each day, have the typical form with the two peaks early in the morning and in the middle of the afternoon i.e., the hours in which people move to reach the work/school places and when they leave. As introduced above, an important aspect when we try to monitor the territory is to understand what kind of people are moving on. Very often the local Administrations do not have exact estimations of tourist flows, but only partial data collected from the tourist offices or census statistics. In the TFO of Pisa we implemented a method for people profiling using the CDR [3]. We identify, with a certain degree of approximation, which calls may correspond to predefined users categories among residents, commuters, in transit and tourists. Starting from these profiles we extract and indicator of presence that enrich the set of variables of the TFO. Identifying tourists/visitors is essential to study how the city is receiving people from outside and how their movements are affecting the city. Again, being able to combine the mobility of resident population with the temporary population (like commuters, visitors or people in transit) may give a measure of the sustainability of the incoming population with respect to resident one. The population on a territory consumes resources like water, air and produces negative effects on the surroundings, like garbage, pollution, noise. In the cases where these resources are limited, the incoming tourist population may break the sustainable equilibrium of such resources. Thus, the ratio between residents and incoming people should be monitored in order to prevent critical situations.

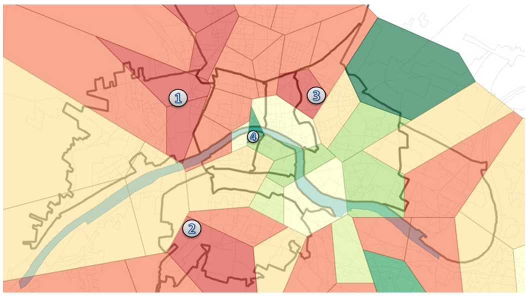


Figure 1: **Density map in the city center of Pisa computed with the CDRs.**

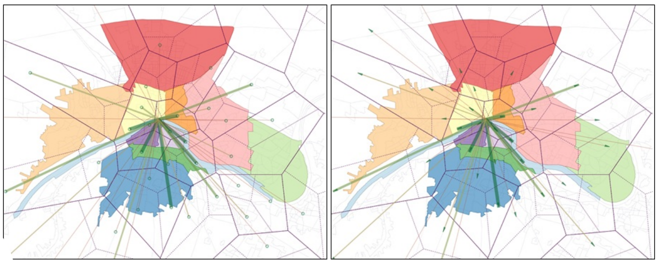


Figure 2: **Origin and destination map of people seen in the cell that contains Ponte di Mezzo. (Left) Cells of Origin; (Right) Cells of the Destinations.**

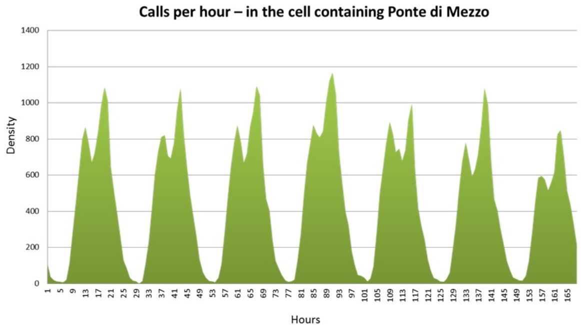


Figure 3: **Call distribution over the time, in the cell that contains Ponte di Mezzo.**

The profiling methodology uses an inductive machine learning step based on the SOM [4] starting from a spatio-temporal user profiles extracted from people call habits.

Starting from the set of calls made by each user in the time window, a Space Constrained Temporal Profile of each user is reconstructed. According to the definition given in [3], a *Temporal Profile* is a vector of call statistics according to a given temporal discretization, and the *Space constrained Temporal Profile* is a Temporal Profile where only the calls performed in the cells contained within the a certain area, are considered.

Since our aim is to study the mobility of residents and visitors in the area of Pisa, from the whole network we first selected the cells overlapping the urban area of the city (Figure 4-Left). The urban center of the city and its corresponding cells are highlighted in pink, while the larger gray area corresponds to the administrative territory of the city. The time projection is built by performing two temporal operations (Figure 4-Right):

1) the aggregation of the days in weekday and weekend slots;

2) the splitting of each slot in time bands representing 3 interesting time windows during the day:

t1 = [00:00:00 - 07:59:59], Early in the morning when people are usually still at home;

t2 = [08:00:00 - 18:59:59], Middle day when people are out for work/school or other activities;

t3 = [19:00:00 - 23:59:59], Late in the evening and night when people are back to home.

Starting from the calls along the days, the presences over t1, t2 ,t3 are computed and then aggregated over the weekdays and the weekend summing up all of them. The result is a sort of compact representation of the user's behaviors measured by his calls.

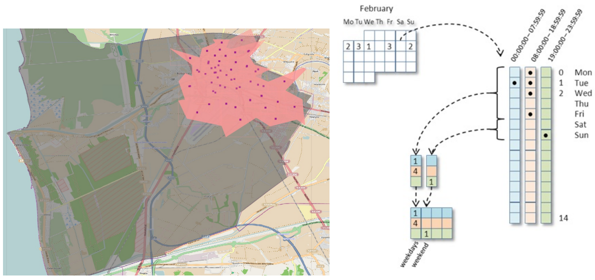


Figure 4: (Left) GSM Cell coverage in the area of Pisa; (Right) Reconstruction of the Temporal Profile for the users in Pisa.

The dataset is then processed by using the SOM algorithm in order to extract the typical global profiles. A SOM is a type of neural network based on unsupervised learning that produces a one/two-dimensional representation of the input space using a neighbourhood function to preserve the topological properties of the input space [4]. In our case, the SOM output is a set of nodes representing groups of users with similar temporal profiles. The SOM tends to highlight similar and compatible presence profile of longer stay people, allowing to separate commuters and residents, by exploiting the calling habits of these users in particular during the weekends. In particular, as shown in Figure 5, on the bottom left corner (identified with blue points) there are the temporal profiles corresponding commuter-like pattern with high frequency during the workdays and a smaller activity during weekends. On the upper right, there are instead the profiles describing residents with high presence during the whole time windows. In the central part (identified with yellow points) there are the profiles corresponding to short visits of the city.

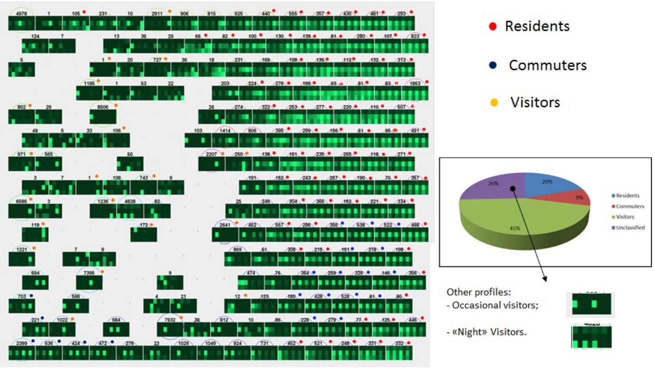


Figure 5: SOM Result: The user profiles.

Counting the instances in each group, we estimates the percentage of residents, commuters and visitors as 20%, 9%, and 45%, respectively. The 26% of the individuals are unclussified at first, because they do not match with any predefined profiles. Carrying out a more accurate analysis of the results, we are able to extend the cathegories with new profiles such as for example the "Occasional visitors" and "The night visitors". While the former are essentially people that come to Pisa only few times along a month, the latter visit Pisa almost regularly but only the night (maybe for the nightlife in the pubs).

The population rates have been integrated in the TFO as a particular socio-demographic indicator.

The profiling methods through the SOM, permits also to further discovery particular situation or events that can enrich the global description of a city dynamics. Analyzing the different profiles we can discover special cases that can hide interesting information, as the one shown in Figure 6 (Left). This profile groups a huge amount of people that appear in Pisa only in a unique slot. This slot corresponds to the weekdays from Monday 23rd to Friday 27th 2012. When plotting the daily time distributions of the calls, we discovered a peak of calls during Friday 27th around 4 p.m. as shown in Figure 6 (Right). Actually this day some minutes before 4 p.m. a strong earthquake affected the Tuscany and Pisa. In this case the GSM data are a good proxy to capture extraordinary events.

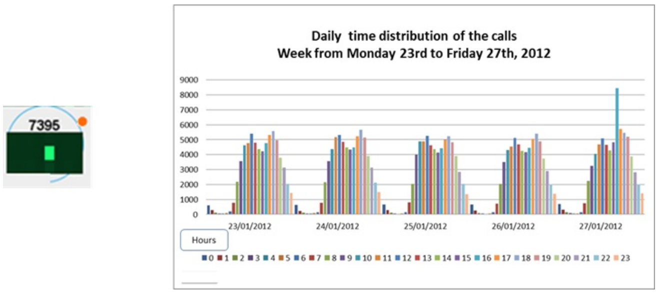


Figure 6: (Left) Particular profile and the corresponding daily time calls distribution.

Conclusions

In this abstract we sketched some features of the Tourist Fluxes Observatory we have set up in collaboration with the Municipality of Pisa. Apart from integrating the basic standard data like hotel and parking presence, airports arrivals and car rentals, the main innovative point of this Observatory is the ability to analyse GSM data. We have developed mobility analysis including density based presence of people during time, to more sophisticated analysis to infer the user profile among residents, commuter and tourists.

References

[1] H. Bar-Gera. Evaluation of a cellular phone-based system for measurements of traffic speeds and travel times: A case study from israel. *Transportation Research Part C*, pages 380–391, 2007.

[2] F. Calabrese, M. Colonna, P. Lovisolo, D. Parata, and C. Ratti. Real-time urban monitoring using cell phones: A case study in rome. *IEEE Transactions on Intelligent Transportation Systems*, 12:141–151, 2011.

[3] B. Furletti, L. Gabrielli, C. Renso, and S. Rinzivillo. Identifying users profiles from mobile calls habits. In *In the Prooc. of UrbComp'12*, 2012.

[4] T. Kohonen. *Self-Organizing Maps*. Springer Series in Information Sciences. Springer, 2001.

Hierarchical Exploration of Human Mobility Regularities

Zhenhui Jessie Li

Penn State University, University Park, PA
jessieli@ist.psu.edu

ABSTRACT

The spatial and temporal regularities are the most fundamental patterns in human movements. In this work, we address the problem of analyzing human regularity in *a hierarchical way*. Our intuition is that, a person may show different temporal patterns with respect to locations at different spatial granularities. When looking at the raw trajectory of a person, there could be several cities/towns that he/she usually goes to. And the person could have yearly regularity visiting some of these cities. And we could further zoom into those cities and find certain specific locations that this person frequently visits, such as home, office or local grocery stores. Taking these locations as references, we may further observe the daily and weekly patterns.

We propose to use *reference spots*, which are the frequent locations, to observe human movements. The challenge lies in to how to effectively detect reference spots at different spatial granularities. To solve this, we first calculate the density map over all locations and *iteratively* breaks the region into reference spots. The reference spots form a hierarchy and the temporal patterns in a 24-hour window and 7-day window will reveal the mobility regularities respect to a reference spot. The experiments are carried on Nokia dataset, mainly using the GPS data. We will show some case studies and all the results are put online.

. INTRODUCTION

With the advanced position technology, massive amounts of object movement data have been collected. Human movement data collected from mobile phones is a particular interesting moving object data. The mobility patterns mined from human movement are useful for urban planning, traffic forecasting, and the spread of biological and mobile viruses.

An important analysis on human movement data is to find the locations that he/she frequently visits and show the temporal regularity w.r.t. these locations. Intuitive examples include daily behaviors in-and-out of home and weekly visits to local grocery store.

There are several interesting studies showing the potential of using mobile or positioning technologies to study the human mobility regularity [2, 1, 6]. In our recent work [4], we study the problem of detecting periods in movement data. *Our idea is to find dense regions, namely reference spots, and examine the in-and-out pattern w.r.t. reference spots. Periods can be detected using Fourier transform and autocorrelation on the binary in-and-out sequences.*

In this work, we will study human movement regularity in *a hierarchical structure of spatial locations*. For example, if we take a town as one big reference spot, we could find that the person has yearly regularity visiting the town. If we take a person's office as one small reference spot, the weekly regularity visiting the office could be revealed. Reference spots could be a state, a city, a downtown region, or a building. By looking at reference spots at different spatial granularity, we gain better insights into human mobility from various aspects.

The reference spots in hierarchy are *data-dependent and should be detected automatically*. Our previous work [4, 5] proposes to use kernel-based method to detect reference spots. The method is limited to detect only one-level reference spots. Now the challenge lies in how to detect the *hierarchical structure of the reference spots*. We propose to a top-down iterative way to detect reference spots. In each level, we first calculate the density map over all the locations. If there are more than two peaks on the density map, we will find the lowest density threshold to break them into separate reference spots. Then, we further look into each reference spot and iteratively do the same for the locations in every reference spot. Since we know human usually follow daily or weekly calendar behavior, for each reference spot, we will examine the regularity of the times visit this reference spot in a 24-hour window, 7-day window and all-time window. The skewed time distribution will reveal the human movement regularities.

2. METHOD

Let $D = \{(loc_1, time_1), (loc_2, time_2), \dots (loc_n, time_n)\}$ be the original movement database for a moving object, where loc_i is a spatial point represented as a pair $(loc_i.x, loc_i.y)$. A *reference spot* is a dense area that is frequently visited in the movement. It is important to find reference spots at different spatial granularities. In this section, we will first describe how to calculate the density map and then discuss our method to detect reference spots in a hierarchical way.

2.1 Kernel-based Density Map Calculation

Intuitively, reference spots are those dense regions containing more points than the other regions. While computing the density for each location in a continuous space is computationally expensive, we discretize the space into a regular $w \times h$ grid and compute the density for each cell. The grid size is determined by the desired resolution to view the spatial data.

To estimate the density of each cell, we adapt a popular kernel method [7], which is designed for the purpose of finding home ranges of animals. If an animal has frequent activities at one place, this place will have higher probability to be its home. This actually aligns very well with our definition of reference spots.

For each grid cell c, the density is estimated using the bivariate normal density kernel,

$$f(c) = \frac{1}{n\gamma^2}\sum_{i=1}^{n}\frac{1}{2\pi}\exp(-\frac{|c-loc_i|^2}{2\gamma^2}),$$

where $|c - loc_i|$ is the distance between cell c and location loc_i. In addition, γ is a smoothing parameter which is determined by the following heuristic method [7],

$$\gamma = \frac{1}{2}(\sigma_x^2+\sigma_y^2)^{\frac{1}{2}}n^{-\frac{1}{6}},$$

where σ_x and σ_y are the standard deviations of the whole sequence LOC in its x and y-coordinates, respectively. The time complexity for this method is $O(w \cdot h \cdot n)$.

2.2 Hierarchical Detection of Reference Spots

After obtaining the density values, a reference spot can be defined by a contour line on the map. A contour line joins the cells of equal density. We use contour line to define the boundary of a reference spot. Any point within the reference spot has higher density value than that of the boundary. So the reference spot is essentially an area with high density.

When we set density threshold p equal to 0, all the grids will be in one single big contour. If we gradually increase p, the size of the contour will shrink. If there is only one peak on the density grids, the contour with density p will eventually shrinks to empty as p decreases. If there is more than one peak on the density grids, there will be multiple contours with density p when p decreases to certain value. Therefore, we can increase p from 0 to the maximal density value over all the grids, the first time when we get more than one contours with density p, we take these contours as the reference spots at the current level and further look into each contour separately.

2.3 Summarization of Periodic Behaviors

Given a reference spot, we have a set of locations fall in this spot: $\{(loc_{i_1}, time_{i_1}), (loc_{i_2}, time_{i_2}), \ldots (loc_{i_m}, time_{i_m})\}$. Since human usually follow either daily periodicity or weekly periodicity, we examine the time distributions of the reported locations in this reference spot in a 24-hour window and 7-day window. For example, in the 24-hour window, if the frequency of some particular hours are significantly higher than other hours, it means this person has high daily regularity visiting this reference spot.

To examine the time distributions, we could look at the frequencies per hour or per day. However, it is worth noting that the original time distribution in the data collection is not even. As we found in Nokia dataset, there are very few locations reported from late night to early morning. For example, if there are 100 locations reported at 1 a.m. in the whole movement history, 80 out 100 are found to fall into at one specific reference spot. Then the probability that the person is at this reference spot is 0.8. Probabilities normalize the the biased collection of the raw data.

3. A CASE STUDY

We analyze a user trajectory in the Nokia dataset [3]. The raw trajectory of this user is plotted on Google Earth as

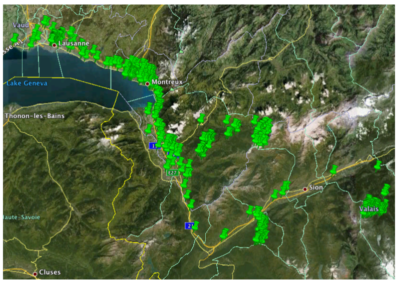


Figure 1: Raw trajectory of a person.

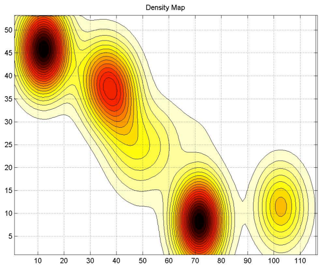


Figure 2: Density map of the trajectory.

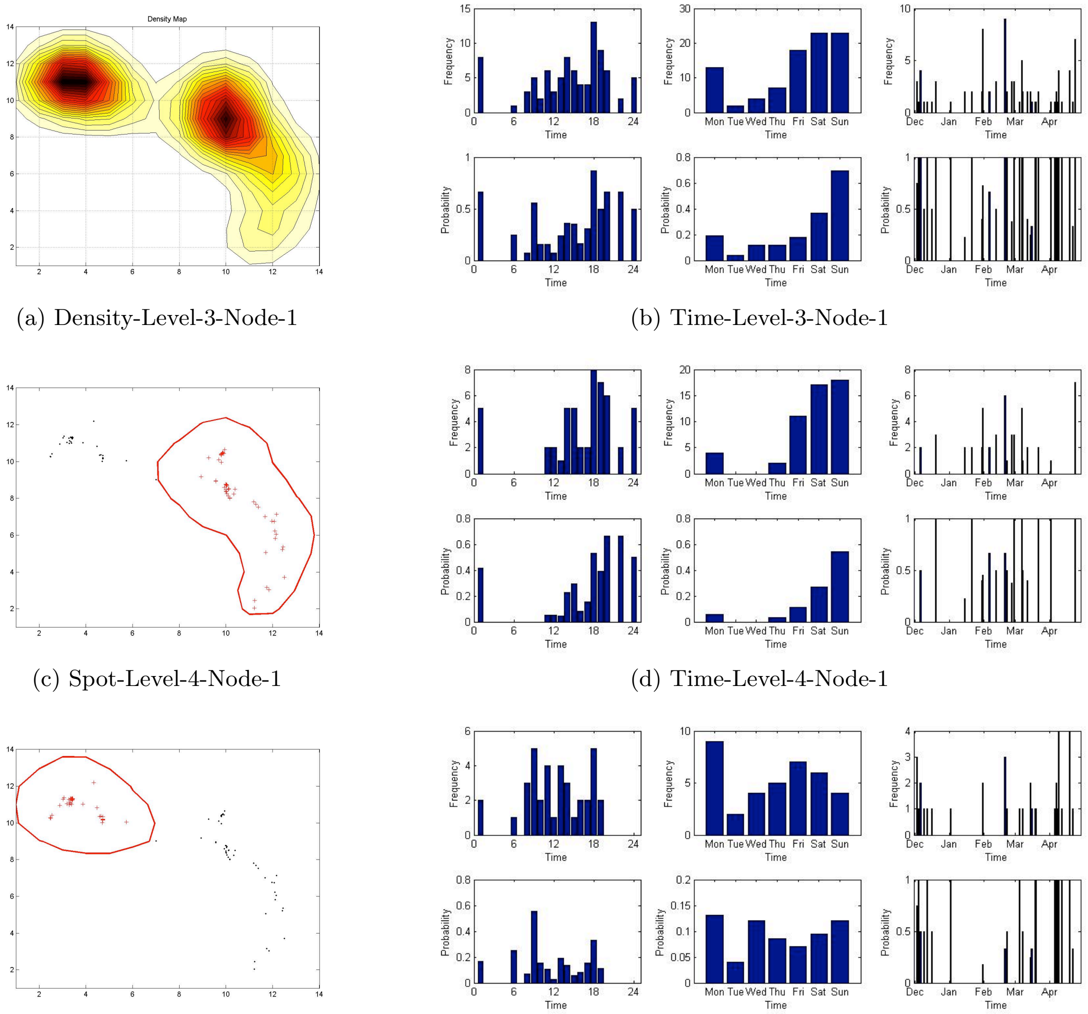
shown in Figure 1. Each green pin is one recorded location of this user. The density map over all the locations is shown as Figure 2.

When we reach to node 1 at level 3, the density map is shown as Figure 3(a). We see from Figure 3(b) that the person is *more likely to visit the region on weekends with skewed time distribution in 7-day time window.* When we go to level 4, it breaks into two reference spots. And the time distributions of these two reference spots are showing very different characteristics. The reference spot as shown in Figure 3(c) showing strong weekly regularity. *The person has high probability to be in this reference spot on weekends and sometimes showing up in this spot at nights.* However, for the other reference spot as shown in Figure 3(e), the *time distribution in a 7-day window is more even and the person is observed in that spot mostly during daytime as seen in the 24-hour window.* Therefore, we gain insights into the mobility pattern of this person in level 4 than level 3. Only when we treat these two reference spots separately, we will observe this person is likely to be in Spot-Level-4-Node-1 on weekends, which indicates this could be the region that this person is usually spending his spare time. And Spot-Level-4-Node-2 is likely to be a working place that this persor usually shows up at daytime.

4. CONCLUSION AND DISCUSSIONS

In this work, we present the method of analyzing human movement in a hierarchical way. We propose using reference spots at different spatial granularities to observe human movement. The discovery of human movements can be revealed from the temporal regularities in reference spots.

It will also be interesting to study the regularity in the trajectory paths. This would be more challenging since the frequent trajectory paths are more difficult to be detected than reference spots.



(a) Density-Level-3-Node-1

(b) Time-Level-3-Node-1

(c) Spot-Level-4-Node-1

(d) Time-Level-4-Node-1

(e) Spot-Level-4-Node-2

(f) Time-Level-4-Node-2

Figure 3: A Case Study.

In our next step, we want to integrate this method into our online demo system, MoveMine[1]. Currently, MoveMine is providing data mining functions to analyze animal movements. We will implement this method to enable our system analyze human movements regularity in a hierarchical way.

5. REFERENCES

[1] Nathan Eagle and Alex Pentland. Eigenbehaviors: identifying structure in routine. In *Behavioral Ecology and Sociobiology*, pages 1057–1066, 2009.

[2] Marta C. González, Cesar A. Hidalgo R., and Albert-László Barabási. Understanding individual human mobility patterns. *Nature*, 453:779–782, 2008.

[3] Juha K. Laurila, Daniel Gatica-Perez, Imad Aad, Jan Blom, Olivier Bornet, Trinh-Minh-Tri Do, Olivier Dousse, Julien Eberle, and Markus Miettinen. The mobile data challenge: Big data for mobile computing research. In *Mobile Data Challenge by Nokia Workshop in conjunction with Int. Conf.. on Pervasive Computing*, June 2012.

[4] Zhenhui Li, Bolin Ding, Jiawei Han, Roland Kays, and Peter Nye. Mining periodic behaviors for moving objects. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'10)*, pages 1099–1108, 2010.

[5] Zhenhui Li, Jingjing Wang, and Jiawei Han. Mining periodicity for sparse and incomplete event data. In *Proc. of 2012 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'12)*, Beijing, China, Aug. 2012.

[6] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of predictability in human mobility. *Science*, 327:1018–1021, 2010.

[7] B. J. Worton. Kernel methods for estimating the utilization distribution in home-range studies. In *Ecology*, volume 70, 1989.

[1] dm.cs.uiuc.edu/movemine

Mining User Mobility Features for Next Place Prediction in Location-based Services

Anastasios Noulas, Salvatore Scellato, Neal Lathia, Cecilia Mascolo
Computer Laboratory, University of Cambridge
email: firstname.lastname@cl.cam.ac.uk

1 Introduction

Understanding human mobility has been a long-standing subject in academic research due to the multitude of potential applications. Those range from the better grasp of human behavior and migration patterns, to the evolution of epidemics and spread of disease, or the understanding of the mechanisms that shape social networks. Besides, studying human movement and geographic activity is increasingly a focal point of research in computer sciences. The rise of the Mobile Web and the provision of Internet scale applications and services to millions of smartphone users is bringing geography and location to the spotlight; knowing how people move and choose to visit specific places in a city can benefit a plethora of applications, including mobile web browsing, local search and content discovery. Up until recently most attempts towards the analysis and modeling of human mobility were relying on mining spatio-temporal datasets sourced from GPS sensors [3], WiFi logs or Cellular Data [2]. These data sources describe with fine grained geographic and temporal granularity user movements, where each data point is pair of longitudinal coordinates coupled with a timestamp that informs us *where* a user is at a particular time. The existence of these datasets has led to the development of numerous statistical frameworks that aim to predict the whereabouts of users, the characterization of spaces based on local human activity or the detection of significant locations in urban environments [1].

Nonetheless, with the introduction and increasing popularity of location-based services, the opportunity to study human movement in a qualitatively different setting is provided. Mobile applications such as Foursquare, where users *check in* broadcasting their visits to places, allow us not only to know the geographic coordinates of a user at a given time, but also the *exact* places they may go. As with cellular data, the position of the user is becoming known when they explicitly use the service, yet the multiple layers of data offered open new avenues for addressing previously unanswered research questions. A library, a cinema or an airport terminal are only a few examples amongst the millions of places which are accessible through these services. The knowledge about the specific places users visit, which goes beyond plain geographic coordinates, can be exploited as an additional dimension to describe human mobility. Insights about the type and time of users' visits can greatly improve the development of recommendation systems. For instance, advertisers who want to push offers to users would greatly benefit from *knowing the next location a user is going to visit*, so to offer the right coupon or the right recommendation in a timely manner. However there are many challenges involved in the prediction of the next visited location, such as user preferences, place properties as well as spatio-temporal conditions. In this work we address this research question by mining user data generated on a popular location-based service and studying the predictive power that different dimensions of the data offer. We formalize the *Next Check-in Problem*, where we aim to predict the exact place a user will visit next given historical data and the current location.

The challenge posed in this context is to rank all the potential target places in the prediction

Poster 1 4

scenario, which could easily contain thousands of candidates, so that the actual place visited *next* by the user is ranked as high as possible. This represents a highly imbalanced prediction scenario, where a single correct instance has to be found (the place a user is going to) amongst thousands of candidate instances. We study what factors may drive user behavior by analyzing a large dataset sourced from the most popular location-based social network, Foursquare. We have collected approximately 35 million user check-ins over a period of 5 months in 2010, taking place over a set of five million geo-tagged venues. We focus our prediction problem over 33 cities, the most active in our dataset in terms of check-in number, treating each city as a different prediction scenario. Our contributions can be summarized as follows:

- We define a set of prediction features that exploit different information dimensions about users' movements: those include information tailored specifically to an *individual user*, such as historical visits or social ties, and features extracted by mining *global knowledge* about the system such the popularity of places, their geographic distance and user transitions between them. Moreover, we employ a set of features that leverage explicitly *temporal information* about users' movements. We assess the predictability of individual features and we discover that the most effective features are those which leverage the popularity of target venues and user preferences.
- We combine the predictive power of individual features in a supervised learning framework. By training two supervised regressors, a simple linear model and M5 model trees, on past user movements, we demonstrate how a supervised approach can significantly outperform single features in the prediction of future user movements, indicating that user behavior in location-based services is driven by multiple factors who may act synchronously. Notably, M5 Model Trees rank constantly one in two user *check-ins* in the top 50 predicted venues.
- We study the performance of features and classifiers over time, finding that prediction performance is higher over lunchtime and weekdays. In all cases, a strong temporal periodicity is apparent in the prediction task, but features based on the geographic distance amongst places are achieving higher scores at nighttime, unlike other features. This shows how the factors driving human mobility can vary over time and highlights the importance adding spatio-temporal context to the prediction task.

We envision a number of applications for which our work may prove beneficial, including mobile recommendations and content delivery for mobile Web users. In particular, the decomposition of users' movements into a set of distinct features is central to our approach and evaluation strategy, as one of the principal goals of this work is to understand how mobile users are driven by different factors in their choice of places. Insights on this process can be offered both for research scientists in human mobility and urban planning and for mobile application developers. In the following sections, we begin by analyzing the Foursquare check-in dataset. Subsequently we define twelve mobility prediction features and evaluate them individually and in a supervised learning framework. We close with remarks on related work and conclusions.

References

[1] D. Ashbrook and T. Starner. Using GPS to Learn Significant Locations and Predict Movement Across Multiple Users. *Journal of Personal and Ubiquitous Computing*, 7(5):275–286, October 2003.

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

[3] V. W. Zheng, Y. Zheng, X. Xie, and Q. Yang. Collaborative location and activity recommendations with GPS history data. In *Proceedings of WWW '10*, 2010.

Extracting People's Stays from Cellular Network Data

Mori Kurokawa,[1] Takafumi Watanabe,[1] Shigeki Muramatsu,[1]
Hiroshi Kanasugi,[2] Yoshihide Sekimoto,[2] and Ryosuke Shibasaki[2]
[1]KDDI R&D Laboratories Inc., Saitama, Japan
[2]The University of Tokyo, Chiba, Japan
mo-kurokawa@kddilabs.jp, tk-watanabe@kddilabs.jp, mura@kddilabs.jp,
yok@csis.u-tokyo.ac.jp, sekimoto@csis.u-tokyo.ac.jp, shiba@csis.u-tokyo.ac.jp

Abstract— Behavior analysis of mobile phone users has increased in importance for mobile phone carriers in accordance with mobile phone traffic increases. Call detail records (CDRs) known as cellular network data are an important data source in the inspection of human behavior. We propose a novel method of segmenting CDR time sequence into staying and moving via Mean-shift to find stay time intervals from CDRs. We also show accuracy through experiments using CDRs and web diaries obtained from the experimental survey conducted for 25 days with 162 examinees.

Keywords-Call detail records, segmentation, extracting stays

I. Introduction

Mobile phone traffic has been increasing due to the widespread use of smartphones and smartphone applications. Mobile phone carriers have found it more important to continuously survey actual human behavior of using mobile phones in order to plan resource deployment and capacity.

Call detail records (CDRs) known as cellular network data are an important data source in the inspection of human behavior. CDRs are recorded on cellular network equipment, mainly to detect and resolve problems with the equipment. Each record is generated through a call, sending of a text message, or browsing the Internet via a mobile phone, and the record contains a timestamp and location related to connected base stations.

Global positioning systems (GPS) are an alternative way of acquiring locations. Although embedding GPS functionality into mobile phones has become common, issues of power consumption and indoor positioning remain. Furthermore, some applications using GPS functionality send additional traffic to cellular networks.

We utilize CDRs to survey actual human behavior. The difficulty of analyzing CDRs lies in its spatiotemporal resolution. The temporal resolution of CDRs differs for each person according to the mobile phone communication pattern. Additionally, the spatial resolution of CDRs is lower than that of existing positioning devices, such as GPS, because such devices depend on the coverage area of base stations.

In this paper, we focus especially on segmenting CDR time sequence into staying and moving because we believe there is a significant difference in mobile phone usage between the behaviors.

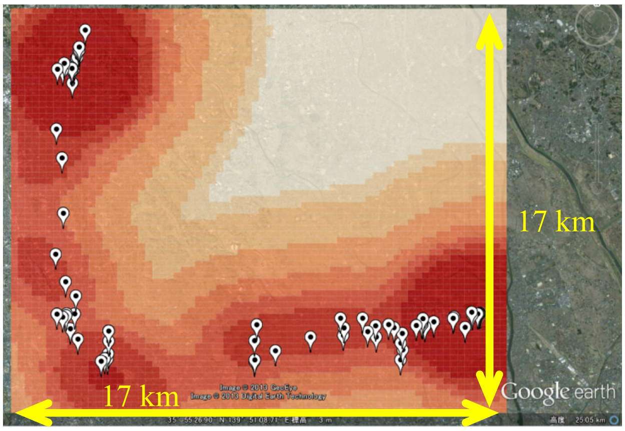


Figure 1. An example of candidate stay locations

We propose a novel method of segmenting CDR time sequence and extracting people's stays, which is robust with regard to the spatiotemporal sparseness of CDRs. We apply Mean-shift to location data on time segments to find candidate stay time intervals and locations, and then apply Mean-shift again to cluster candidate stay time intervals and locations. The proposed method is assessed by CDRs and web diaries obtained from an experimental survey conducted for 25 days with 162 examinees.

The remainder of this paper is as follows. In section II, we explain the proposed method. In section IV, we present the results of the experiment. In section IV, we describe related studies based on CDRs. Finally, in section V, we conclude this paper and suggest further directions for research.

II. Proposed Method

In this section, we describe our method of extracting stays. Our method consists of two steps: extracting candidate stays and clustering. Figure 1 shows an example of candidate stay locations for one examinee (represented by white markers) and the density estimated using Kernel density estimation via Gaussian Kernel with a bandwidth equal to 1 km.

A. Extracting candidate stays

Here, we describe how to extract candidate stay time intervals and locations from a personal location history $\{p(t)$:

$t = t_0,\dots,t_n\}$, where $p(t)$ represents a two-dimensional spatial point at time t in CDRs.

We assume that the connected base station locations are distributed around the stay locations while staying. In other words, the mode of spatial distribution of base station locations is supposed to be close to the stay location. Thus, we try to seek the mode of spatial distribution of connected base stations in each time interval.

We apply a sliding time window to extract stay time intervals because we do not know the stay time intervals a priori. We employed a sliding time window with width T and shift S, where the first time segment is $[t_0, t_0+T)$, the second time segment is $[t_0 + S, t_0 + T + S)$, and so on.

We apply Mean-shift [1] to each time segment which includes no less than N points. Mean-shift is a popular mode seeking method, and each iteration calculates the mean $m(p(t))$ of nearby points of $p(t)$ within the window determined by the window function $K(.)$, shifts $p(t)$ to $m(p(t))$, and then proceeds to the next iteration.

$$m(p(t)) = \frac{\sum_{p(t')} K(p(t') - p(t))p(t)}{\sum_{p(t')} K(p(t') - p(t))}$$

As for the window function $K(.)$, we use a rectangular window defined as follows: $K(x) = 1$ if $\|x\| < th_1$, and 0 otherwise. The threshold parameter th_1 corresponds to a bandwidth of the density estimation. Convergence is checked by evaluating the difference in the mean points $\|m(p(t)) - p(t)\| < th_2$.

Then, we determine candidate stay time intervals as a set of time segments in which the resulting mean points are concentrated within the range of a circle with a radius of th_2. We then determine candidate stay locations as the resulting mean points within each candidate stay time interval.

B. Clustering

Here, we describe how to cluster extracted candidate stay 'ocations and time intervals.

1) Clustering stay locations

Extracted candidate stay locations are noisy due to few observed points within each time segment, and thus we apply Mean-shift with the same parameters as used in extracting candidates again to cluster stay locations. We determine stay locations by the resulting mean points.

2) Clustering stay time intervals

We determine stay time intervals as a series of time segments that meet the following conditions.

a) The stay location in the first time segment and the one in the last time segment are equal.

b) The stay location in the intermediate time segment does not belong to the stay locations other than the stay location determined in a) above.

III. EXPERIMENTS

In this section, we explain the experimental methods and results with practical CDRs and web diaries from an experimental survey.

A. Experimental Data

From November 28, 2011, to December 22, 2011, we conducted an experimental survey with 184 examinees to obtain activity data that included CDRs, GPS logs, actual activity data from a web diary and personal attributes via a questionnaire. Examinees consented to the privacy policy and terms of the experiment. The average number of accumulated CDRs in a day was 454.9 for each examinee. That is, one telecommunication event occurred every 3 minutes. The Android application for GPS logging acquired positions every 5 minutes and sent logs to the web server every 15 minutes.

We prepared a web diary system for examinees to register actual activities and to collect ground truth data. In order to facilitate the entry of activities, the web diary system automatically suggested some candidate stay locations that the system preliminarily estimated based on GPS logs. The average number of registered stay locations was 4.6 per day.

Among 184 examinees, there were 162 examinees whose CDRs and activity states were both available. For the experiments described below, we used the available results of these 162 examinees.

B. Experimental Method and Evaluation Metrics

We compared parameter settings of our method. We compared settings of the sliding time window when $th_1 = 4$ km, $th_2 = 1$ km, and $N = 4$. Additionally, we compared the threshold parameter th_1 when $(T, S) = (40, 10)$, $th_2 = 1$ km, and $N = 4$.

We also compared our method with the following baseline method, which was based on [2], where stay locations were determined by timestamps and locations under the following conditions.

a) There are two or more CDR records within 24 h.

b) Sequentially recorded positions of the base stations are within 4 km.

c) Duration of sequential positions under condition b) is less than 20 minutes.

We employed F-measure of stay time intervals for an evaluation metric. We calculated the F-measure for each examinee.

Let T_w be a set of estimated stay time intervals of each examinee and let T_c be a set of correct stay time intervals that each examinee registered in the web diary; precision and recall are defined as follows and F-measure is the harmonic mean. We eliminated time intervals about which the examinee did not register activities from time intervals for evaluation.

$$\text{Pecision} = \frac{|T_w \cap T_c|}{|T_w|}$$

$$\text{Recall} = \frac{|T_w \cap T_c|}{|T_c|}$$

$$\text{F-measure} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{(\text{Precision} + \text{Recall})}$$

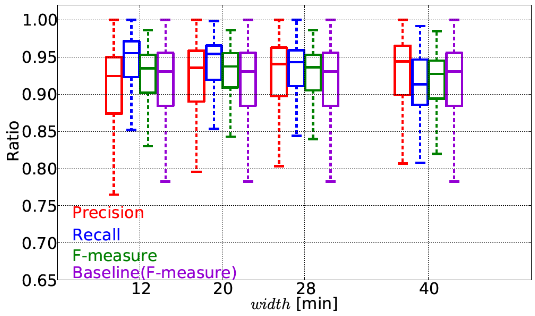


Figure 2. Comparison of width of sliding time window

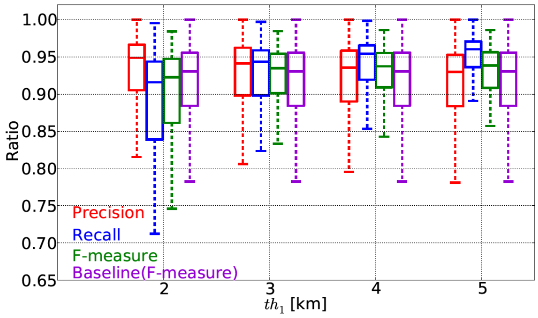


Figure 3. Comparison of threshold parameter of Mean-shift

C. *Experimental Results*

Figure 2 and Figure 3 show box-plots of precision, recall and F-measure of stay time intervals. The results shows that our methods exceeded the baseline except the case of a width equal to 40 minutes in Figure 2 and the case of a threshold parameter th_1 equal to 2 km in Figure 3.

With regard to settings of the sliding time window, Figure 2 showed that a width equal to 20 minutes was the best. In that case, the F-measure of stay time intervals was 93.7% on median. The precision increased and the recall declined with a longer width. The width of the sliding time has the effect of increasing the sensitivity to switching between staying and moving and increasing false positive stays, when short. The obtained data contained a relatively high number of points per unit time and thus the sliding time window with a width equal to 20 minutes showed sufficient accuracy.

The threshold parameter of Mean-shift th_1 has the effect of decreasing false negative stays and increasing false positive stays, when large. Figure 3 showed that th_1 equal to 5 km was the best. In that case, the F-measure of stay time intervals was 93.8% on median, which is nearly the same as in the case of th_1 equal to 4 km.

IV. Related Work

While there are numerous studies related to mobility estimation using GPS logs, some studies are applicable to CDR-based estimation [2][3][4][5]. On the other hand, by regarding CDRs as footprints of personal mobility, some studies attempted to extract significant locations [6][7][8]. Among them, [6] uses Leader algorithm which is a location-based clustering algorithm. The others apply time-based clustering: [7] clusters points based on distance between temporally adjacent points and filters small clusters where little time was spent, and [8] clusters points by setting a threshold for switching counts of cell towers.

V. Conclusion and Further Study

In this study, we attempted to estimate stay time intervals using CDRs. The proposed methods are assessed by CDRs and web diaries obtained by an experimental survey. According to the results, the best case was the case with a width of the sliding time window equal to 20 minutes and with a threshold parameter of Mean-shift equal to 5 km, and in that case, the F-measure of stay time intervals was 93.8% on median.

In the future, we would like to conduct detailed experiments to tune our method. We would also like to try to estimate more detailed human behavior, such as stay objectives and transportation modes.

References

[1] Y. Cheng, “Mean Shift, Mode Seeking, and Clustering”, IEEE Trans. Pattern Anal. Mach. Intell, Vol. 17, No. 8, pp. 790-799, 1995

[2] J. Liu, O. Wolfson, H. Yin, “Extracting Semantic Location from Outdoor Positioning Systems”, Proc. of the 7th International Conference on Mobile Data Management (MDM 2006), pp. 73, 2006

[3] D. Ashbrook and T. Starner, “Using GPS to learn significant locations and predict movement across multiple users”, Personal and Ubiquitous Computing, Vol. 7, pp. 275-286, 2003

[4] C. Zhou, D. Frankowski, P. Ludford, S. Shekhar, L. Terveen, “Discovering personally meaningful places: An interactive clustering approach”, ACM Trans. Inf. Syst, Vol. 25, No. 3, 2007

[5] P. Nurmi, and S. Bhattacharya, “Identifying meaningful places: The non-parametric way”, Proc. of the 6th International Conference on Pervasive Computing, pp.111-127, 2008.

[6] S. Isaacman, R. Becker, R. Caceres, S. G. Kobourov, M. Martonosi, J. Rowland, and A. Varshavsky, “Identifying Important Places in People's Lives from Cellular Network Data”, Proc. of the 9th International Conference on Pervasive Computing, pp. 133-151, 2011

[7] J. H. Kang, W. Welbourne, B. Stewart, G. Borriello, “Extractin Places from Traces of Locations”, Mobile Computing an Communications Review, Vol. 9, No. 3, pp. 58-68, 2005

[8] M. A. Bayir, M. Demirbas, and N. Eagle, “Mobility profiler: A framework for discovering mobility profiles of cell phone users”, Proc. of the International Conference on Pervasive and Mobile Computing, Vol. 6, No. 4, pp. 435-454, 2010

Building Energy-Efficient Spatio-Temporal Mobility Profiles for Mobile Users

Kuldeep Yadav, Vinayak Naik, Amarjeet Singh
Indraprastha Institute of Information Technology, New Delhi, India
Email : {kuldeep,naik,amarjeet}@iiitd.ac.in

***Abstract*—Fine grained mobility information of mobile phone users are required for many context-aware application. Most of research in this space uses location interfaces such as GPS and WiFi which results in high power consumption and also have limited availability. GPS and WiFi is not available on many low-cost phones (popularly called as feature phones). Also, GPS does not work indoors and WiFi infrastructure is not widespread especially in developing countries.**

In this paper, we propose a framework to find different places visited by a person solely using GSM (Cell ID) data and then use them to build spatio-temporal mobility profiles for the users. Also, proposed framework incorporates data from other location data sources (i.e. WiFi) to improve the accuracy of only Cell ID-based clustering. We have done comprehensive evaluation of proposed algorithms on two real-world datasets i.e. self collected dataset (16 users, 4 weeks) and Nokia MDC dataset (45 users, 50 weeks) and found it very accurate.

I. Introduction and Motivation

A mobility profile for a user consists of all the places visited by her with accurate arrival and departure time information for the respective places. An accurate mobility profile can enable many context aware applications such as location-based notifications, targeted advertisements, content-sharing decisions [4], pollution impact report and many others. Mobile Phones have various kind of location interfaces such as Global Positioning System (GPS), WiFi [3], Cell ID-based (GSM information) [2] etc which can be used for high resolution location tracking of users. Each of these interfaces differ in terms of accuracy, availability, and power consumption. Most phones in developing countries are feature phones which have limited capabilities, e.g. they lack sensors such as Global Positioning System (GPS) and WiFi. Due to limited capabilities, feature phones are not able to use context aware applications which uses mobility profiles. For smart phones too, building mobility profile using GPS and WiFi requires continuous tracking of location which drains the phone battery very quickly. Friedman et al [6] found that scanning modes of WiFi and Bluetooth consume significant power and quickly drain the battery in continuous location sensing applications.

Researchers have worked on using GPS and WiFi data to find places visited by the user automatically and then building accurate spatio-temporal mobility profile. Bayir et al [2] proposed a framework which discover places using Cell ID data in reality mining dataset but it take help of manually tagged Cell IDs for clustering. There is a lack of an algorithm/framework that can discover places visited by a user using Cell ID data without human intervention/tagging.

In this paper, we propose a framework to build mobility profiles of users using energy-efficient and widely available location sensing frameworks i.e. GSM (Cell ID) and WiFi (if available). Our clustering algorithm automatically learn places visited by a user solely using Cell ID data. Evaluations on our self collected dataset [1] (Location : India, 16 users, 4 weeks of duration) showed that, framework correctly learn places with nearly 80% accuracy when compared to places learnt using WiFi data. Further, we develop an algorithm which uses an initial training of WiFi data to learn places using Cell ID data for some days and later use Cell ID data only. With the help of WiFi training, accuracy improves further i.e. 87% if 8 days of training is provided. Further, we have applied same framework on a publicly available dataset which has data of 38 users for 40 week collected in Switzerland. In Nokia MDC datase, proposed framework were able to find places with an accuracy of 86.06% without any training.

II. User Mobility Profiling Framework

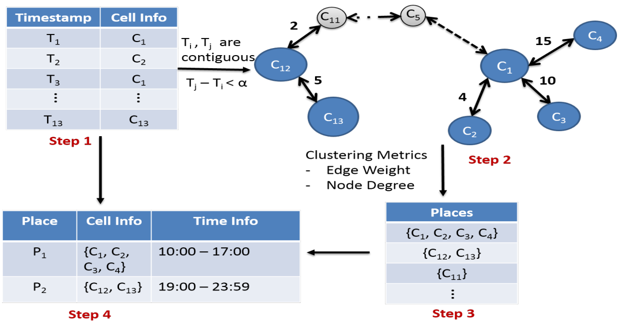
A place is defined as a location, where the user stays for a significant amount of time , e.g., “Home” and “Workplace”. For building mobility profile, one of first step and challenging task is to find different places that a person visited using raw location data and then use this place information to find arrival and departure time specific to those places. In this section, we will present algorithms to find places using GSM information and combination of GSM and WiFi. Also, we will use “clusters” and “places” interchangeably from here on.

A. Cell ID (GSM) based Mobility Profiling :

Finding distinct places using only GSM information has several challenges. Previous work [2] has shown that even if a user stays at the same place, the Cell ID may change due to various reasons such as network load, small time signal fading, and inter-network ($2G$ to $3G$ or vice versa) handoff. This change in Cell ID at the same place is called as an “oscillating effect”.

Assuming that $\{C_1,C_2,C_3,....C_k\}$ are the distinct time-ordered Cell IDs observed in a day (Step 1 of Figure 1), we build an undirected graph, called as *movement graph*,

[1] We will be releasing our dataset publicly very soon.



Timestamp	Cell Info
T_1	C_1
T_2	C_2
T_3	C_1
⋮	⋮
T_{13}	C_{13}

Places
$\{C_1, C_2, C_3, C_4\}$
$\{C_{12}, C_{13}\}$
$\{C_{11}\}$
⋮

Place	Cell Info	Time Info
P_1	$\{C_1, C_2, C_3, C_4\}$	10:00 – 17:00
P_2	$\{C_{12}, C_{13}\}$	19:00 – 23:59

Figure 1: A snapshot of different steps in mobility profiling framework

$G(V,E)$ where $\forall_{i\in\{1,k\}} C_i \in V$ and there exist an edge $e(C_i, C_j)$ between C_i and C_j, if both of the following conditions are satisfied:

1) C_i and C_j are contiguous in time ordered cell records
2) Time difference between start time of C_j and end time of C_i is less than α.

As an example in Figure 1, C_1 and C_2 occurred contiguously and t_2-$t_2 \leq \alpha$, so there will be an edge between C_1 and C_2 in the corresponding movement graph of the user. Multiple edges between C_i and C_j are merged into a single edge with weight equal to the number of edges between C_i and C_j. α ensures that an edge occurs only across neighboring (in time) Cell IDs and other cell records, that may be neighboring but with a high time difference between them due to reasons such as switching off of the phone, unavailability of the network, and loss of location updates, are pruned. An example of movement graph created from user X's data is shown in step 2 of Figure 1.

As seen in step 3 of Figure 1, even same place Cell IDs (i.e.$\{C_1$,C_2,C_3,$C_4\}$) have many fluctuations among themselves (i.e. oscillating effect) which is modeled as edge weight in the movement graph. To cluster Cell IDs, accounting for the oscillating effect, into different places visited by the user, we propose a three phase algorithm as described in Algorithm 1. *Graph Clustering Algorithm* takes movement graph as an input and produces Cell ID clusters as an output, where each cluster will represent a different place and can be used to build mobility profile (step 4 in Figure 1). The detailed description of algorithm can be found in [5]. *Graph Clustering Algorithm* also takes two parameters into account i.e. one of them is *oscillation parameter* η, which measures the number of fluctuations between a pair of Cell IDs in a day and η', which measures the number of transitions from a Cell ID to any other Cell IDs. We will empirically derive the good value of η and η' in evaluation section.

B. WiFi Trained Cell ID Clustering (WTCA)

GCA inherently assumes that different places in a user's profile will have non-overlapping sets of Cell IDs. As a result, for a person visiting distinct places that are in close proximity, e.g. a student staying in a dorm that is close to the academic building, GCA will merge the two different places if a common Cell ID is observed at each of the two places. This merging effect of GCA is also observed in our collected data as some of our users also live in campus residence. For instance, if an user saw Cell IDs {a,b,c,d} at P_1 and Cell IDs {d,e,f,g} at place P_2. Though, Cell ID d remains overlapping across places P_1 and P_2.

Algorithm 1: Pseudocode of *Graph Clustering Algorithm*

```
Algorithm: Graph-based Cell Clustering Algorithm
Input: Movement Graph $G(V,E)$ where $V$ is set of vertices and $E$
       is the set of edges
Output: Set of Cell ID Clusters $CG$
begin
    Rank all the edges in $E$ into decreasing order of their weight;
    $CG = \phi$ ;
    while ($\forall e_k \in E$) AND $w(e_k) \geq \eta$ do
        if $v_i \in CG_j$ where $v_i \in e_k$, $i \in (1,2)$, $CG_j \in CG$ then
            $CG_j = CG_j \cup v_{k1} \cup v_{k2}$;
        else
            Create new cluster $CG_n = v_{k1} \cup v_{k2}$ and add it to
            $CG$ ;
    while ($\forall v_j \in CG_k$) where $CG_k \in CG$ do
        if ($degree(v_j) \geq \eta'$) then
            $CG_k = CG_k \cup neighbors(v_j)$ ;
    $CG' = \phi$ ;
    while ($\forall CG_i \in CG$) do
        $isExist = false$ ;
        while ($\forall CG_j \in CG'$) do
            if ($CG_i \cap CG_j$) $\neq \phi$ then
                $CG_i = CG_i \cup CG_j$; $isExist = true$; break ;
        if $\neg(isExist)$ then
            Add $CG_i$ to $CG'$ ;
    while ($\forall v_i \in V$) do
        if ($v_i \notin \exists CG_k$) where $CG_k \in CG'$ then
            Create new cluster $c_n = v_i$ and add it to $CG'$ ;
    return $CG'$ ;
```

While such geographically close places may have overlapping Cell IDs, it is unlikely that they will have overlapping WiFi APs. Further, not all Cell IDs will overlap across the two distinct places. In the above example, we can take into account other Cell IDs such as a and e to distinguish between different places. We use these two insights to exten GCA by training it with WiFi based Cell ID clustering For training purpose, we take help of WiFi mobility profile to determine corresponding Cell ID clusters, accounting for arrival and departure time at a place. As user is likely to visit same places again, a few days of training is sufficient to learn conflicting Cell IDs and GCA use that input while clustering Cell IDs. The detailed description of this algorithm is in [5].

C. Evaluation

Cellular network assigns a group of cell base station in a location area with the same identifier, known as Location

Area Code (LAC). One of the basic way to cluster Cell IDs are by considering LAC information which is called as *LAC-based Clustering Algorithm* [5], which will be used here for comparison purpose. Using WiFi mobility profiles, we associate Cell IDs with actual physical places which is used as ground truth for evaluation of Cell ID based clustering approaches, GCA and LCA. For instance, by using WiFi mobility profiles, we discover that user stayed at a place P_i from $9:00$ AM to $5:00$ PM, the Cell IDs which are seen in this time duration will form a cluster.

Using WiFi mobility profile, we find equivalent Cell ID clusters (say CW) from one day's Cell ID data. For GCA, we empirically found η and η' to be equal to 3 and used it for performing all experiments related to GCA. For every day, we define Cell ID clusters made using WiFi (ground truth), GCA and LCA as CW, CG and CL respectively. We further define a pair-wise comparison metrics, called *Correct Pair* for our evaluation. A Cell ID pair (i.e. C_i and C_j) is counted as *Correct Pair*, if their occurrence within the same or across different clusters in CW is reflected accordingly in the Cell ID based approach evaluated.

Correspondingly, to evaluate GCA, we first calculate the $\%$ of Correct Pairs, out of total pairs, of Cell IDs across CW and CG, and then take an average across different days to compute the final accuracy of GCA. Similar process is followed w.r.t CW and CL and accuracy of LCA is computed. As shown in Figure 2, GCA produces 80.29% correct pairs (on an average across all users and all days) while LCA produces 70.82% correct pairs (on an average). Errors in GCA occurred since it mistakenly merged places that were geographically close (as discussed earlier). Our evaluations on Nokia MDC dataset found that GCA produces 86.06% correct pairs as compared to ground truth which was generated using GPS stay points.

With 8 days of traning, WTCA either equals or improves $\%$ of correct pairs across all users as compared to GCA or LCA. On an average across all users and days, WTCA produces 87.30% correct pairs as compared to GCA which produces 80.29% correct pairs. WTCA improves upon the overall accuracy of clustering, when compared to GCA, since it can split merged places and put them into different clusters, using the training data.

III. Discussion and Future Work

Due to large heterogeneity among mobile devices and available location interfaces on them, there is lack of a generic method to build mobility profiles. Also, current location interfaces are power hungry. We proposed a framework which uses location interfaces such as GSM to build mobility profile. There are two main advantages of using GSM based interface for mobility profiling, (1) It consume very less energy as compared to current alternatives (WiFi, GSM), (2) It is available on all programmable mobile devices and can work in smartphones as well as feature phones.

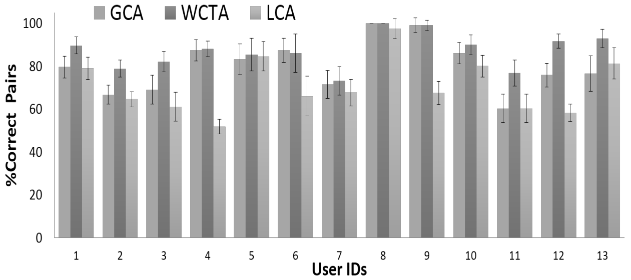


Figure 2: Comparison of GCA,WTCA, and LCA w.r.t. ground truth in our self collected dataset. On an average, WTCA produce more correct pairs (87.30%) than GCA (80.29%) and LCA (70.82%)

Our evaluations on two big diverse datasets confirmed that proposed framework can work in real-world while giving good accuracy.

We are working to create a cloud service of proposed framework so that mobile application developers can utilize these algorithms to build user's mobility profiles, some of them are following:

1) LifeMap: Logging and visualization of all the places that a user visits with low energy. It can be further extended with location-based reminders, meeting scheduling, logging encounters with other users.
2) MobiShare : It is a system designed to support opportunistic content search and sharing with limited Internet (2G) connection. It uses proposed framework to predict encounters between a pair of users [5].
3) Unity : A social collaborative content downloading application which predicts the time at which a group of friends will be co-located.

Finally, we believe that the work described in this paper is an interesting direction and in future, we envision mobile applications/systems using our energy-efficient mobility profiling algorithms to deliver context based information to mobile users.

References

[1] Barabi A., Understanding individual human mobility patterns," Nature 453, 779-782.
[2] Bayir M.A. et al, Discovering spatiotemporal mobility profiles of cellphone users, WOWMOM 2009.
[3] Vu et al, Jyotish: Constructive approach for context predictions of people movement from joint Wifi/Bluetooth trace, PerCom 2011.
[4] Talipov et al, Content Sharing Over Smartphone-based Delay-tolerant Networks, IEEE TMC,2012.
[5] Yadav et al. MobiShare: Cloud-enabled Opprtunistic Content Sharing among Mobile Peers, IIITD-TR-2012-009.
[6] Friedman et al, On Power and Throughput Tradeoffs of WiFi and Bluetooth in Smartphones, INFOCOM 2011.

Properties of the Positioning Error of Cell Phone Trajectories

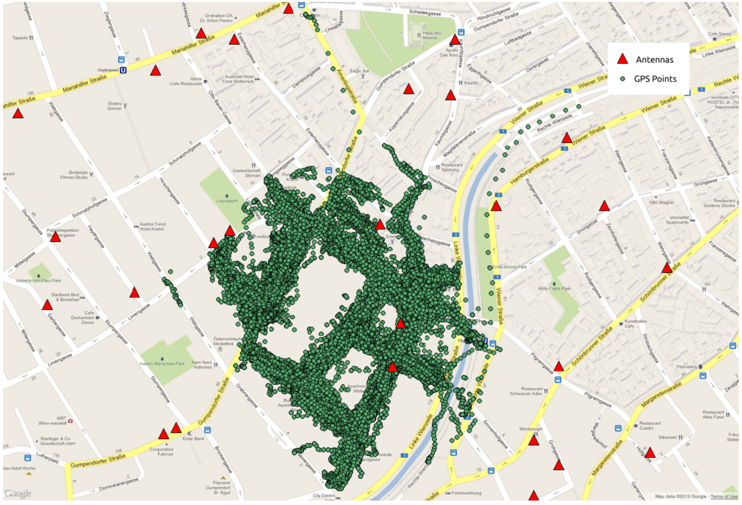
Michael Ulm, Peter Widhalm
Austrian Institute of Technology, Mobility Department
michael.ulm@ait.ac.at, peter.widhalm@ait.ac.at

Abstract

Cell phone data has become a major source for the scientific community to analyze human mobility behaviour. A typical task in this context is to estimate motion trajectories based on sequences of antenna locations. In this paper we examine the question: given that a mobile device is logged in to an antenna with known position and characteristics, what can be said about the probability distribution of the position of the device?

We examine three datasets of trajectory data, where the location was determined using a GPS logger. Based on this data, we give estimates on the distribution of the locations of the users given that they are logged in to one antenna. Finally, we evaluate different strategies for estimating user location from the antenna position.

Figure 1. GPS Positions of mobile devices, logged in to Antenna 52024



Privacy in Computational Social Science: An overview

Riccardo Pietri
Technical University of Denmark
s110913@student.dtu.dk

Arkadiusz Stopczynski
Technical University of Denmark
arks@dtu.dk

Sune Lehmann
Technical University of Denmark
sljo@dtu.dk

ABSTRACT

In recent years the amount of information collected about human beings has increased dramatically. Users store their data in online social networks or collect it for self-tracking purposes; our environments sense and record us with embedded RFIDs, WiFi access points, traffic monitoring. Concurrently, we are seeing an increase in dedicated Computational Social Science (CSS) studies, where researchers collect data on human behavior with unprecedented resolution and scale, providing important insights into human nature. The more powerful the data collection (bitrate, number of users, duration, etc) and analysis, the more important privacy becomes.

One serious threat to CSS as a field is a 'privacy catastrophe', where the participants' rights or data are dramatically compromised as a result of malevolent parties or gross misunderstandings between researchers and experiment participants. Such a 'catastrophe' carries the potential lead to a loss of public confidence and hence a decreased ability to carry out future experiments and research. In order to avoid such a negative scenario, we urge a renewed focus on privacy, users' rights, and data security.

Here, we argue that the current state-of-the-art on these privacy related issues can be improved significantly. For example, study purposes are often not made explicit, 'informed consent' is difficult to define in many cases, security and sharing protocols are only partially disclosed, etc. Below we provide a survey of the work related to privacy issues in CSS tudies. In particular, focus on topics of *informed consent, nonymization*, and *data security*. We also include our reflections on the key problems and provide some recommendations for future work.

POLICY AND INFORMED CONSENT

For CSS studies *informed consent* consists of an agreement between researchers and the data producer (user, participant) by which the latter agrees to understand the procedures applied to his data (collection, transmission, storing, sharing, and analysis). Users need to comprehend through *informed consent* which information will be collected, who will have access to them, what is the incentive, and for which purposes the data is used [1].

We begin by noting a scarcity of available examples and best practices for *informed consent* in the literature; the majority of the reviewed studies do not mention any consent procedures [2]–[7]. While this does not necessarily imply that experiment participants did not consent to the data collection procedures, it is difficult to produce comparisons and create useful models applicable for future studies. In cases when the procedure for achieving *informed consent* was reported, the agreement was carried out using forms, similar to `http://green-way.cs.illinois.edu/GreenGPS_files/ConsentForm.pdf`, containing users' rights [8]–[10]. However, simply stating all the information does not guarantee that *informed consent* is implemented sufficiently: years of EULAs and other lengthy legal agreements show that most individuals tend to blindly accept form that appear before them and to unconditionally trust the validity of the default settings which are perceived as authoritative [11]. Such an *all-or-nothing* approach does not allow the user to select subsets of the permissions, making it only possible to either participate in the study fully or not at all [12].

One improvement would be to allow users to gradually grant permission over time. The efficacy of this approach is not clear: some studies have shown that users understand the issues about security and privacy more clearly, when individual requests are presented gradually [13]; others argue that too many warnings distract users [12], [14]. The literature contains only few analyses of whether the consenting participants are, in fact, *informed*. Evaluating how people understand their privacy conditions can be done by conducting feedback sessions throughout the duration of the experiment [8]. If we wish to increase the focus on participants' rights, approaches such as this should be the norm, not the exception. One option is for participants to be more closely involved in shaping the privacy policies. This view gains support from studies showing that people do not realize smartphone sensing capabilities nor the consequences of privacy decisions [15], [16]. Additionally, we suggest to carefully consider special cases where the participants may not have the competence or authority to fully understand the privacy aspects [17], [18].

Since so little is understood about the precise nature of conclusions that may be drawn from highly detailed data collection, we need to constantly work to improve *informed consent* as our understanding continues to grow. We recommend that the paradigm should move from a one-time static agreement to dynamic consent management [19]. Furthermore, the concerns related to privacy are context-specific and vary across different cultures [20]. The need for a way to let the users easily understand and specify which kinds of data they would like to share and under what conditions was foreseen in 2002 for the Internet purposes by the W3C group, where the aim was to define a *Platform for Privacy Preferences (P3P)* (suspended in 2006), in 2003 by *Kagal et al.* [21], and also in

2005 by *Friedman et al.* [1], all shaping dynamic models for *informed consent*. Recent studies such as [22] have worked to design machine learning algorithms that automatically infer policies based on user similarities. These frameworks can be seen as a mixture of recommendation systems and collaborative policy tools where default privacy settings are suggested to the user and then modified over time.

ANONYMIZATION

The datasets created in CSS studies often contain highly sensitive information about the users. Their privacy needs to be protected either for the purpose of disclosing the data to public scrutiny [23]–[25] or to guarantee that users can not abuse the the provided services [6], [26], [27]. This can be achieved by various anonymization techniques, where the *Personal Identifiable Information* (PII) is removed from the data. Making data anonymous (or de-identified) decreases the data utility by reducing resolution or introducing noise [28].

The most common practice in the data anonymization field is to one-way hash all the PII such as MAC addresses, network identifiers, logs, names, etc. This breaks the direct link between a user in given dataset to other, possibly public datasets (e.g. *Facebook* profile). There are two main methods to achieve this. The first - used in the *LDCC* study (a generic data collector framework developed for the *Lausanne Data Collection Campaign* [8])—is to upload raw data from the smartphone to an intermediate proxy server where algorithms hash the collected information. Once anonymized, the data can be transferred to a second server which researcher have access to. We argue that a less vulnerable option is to hash the data directly on the smartphones and then upload the result the final server for being analyzed. This alternative design has been selected for many MIT studies [3]–[5] and for the *SensibleDTU* project (`http://www.sensible.dtu.dk/`). In principle, hashing does not reduce the quality of the data (provided that it is consistent within the dataset), but it makes easier to control what data is collected about the user and where it comes from. However, it does not guarantee that users cannot be identified in the dataset. Finally, some types of raw data - like audio samples - can be *obfuscated* directly on the phone without losing the usability before uploading [8].

Another frequent method employed for anonymization is ensuring k-anonymity [29] for a published database. This technique ensures that is not possible to distinguish a particular user from at least $k-1$ people in the same dataset. *AnonySense* - a general framework for opportunistic task reports [30] - and the *LDCC* platform both create k-anonymous different-sized tiles to preserve users' location privacy, outputting a geographic region containing at least $k-1$ people instead of single user's location. Nevertheless, later studies have shown how this property is not well suited as a privacy metric [31]: first, *Machanavajjhala et al.* tryed to solve k-anonymity weaknesses with a different privacy notion called l-diversity [32]; then, *Li et al.* proposed a third metric, t-closeness, arguing the necessity and the efficacy of l-diversity [33]. Although these two techniques seem to overcome most of the previous limitations, they have not been deployed in any practical framework to date.

A more complex option is to employ homomorphic encryption, an advanced cryptographic technique that allows an entity to perform some operations on a ciphertext which corresponds to others on the plaintext. It allows users to send anonymous information (the ciphertext) to a central server which can compute meaningful operations. The results are sent back to the users who can finally decrypt and obtain the plaintext. In *VPriv* (a privacy-aware toll system) a central server first collects anonymous tickets produced when cars exit the highways; then by homomorphic transformations it computes the total amount that each driver has to pay at the end of the month [26]. Another similar example is *HICCUPS* which is a health system that keeps patient records encrypted, but at the same time gives doctors access to aggregated information through homomorphic encryptions [34].

DATA SECURITY

The security of the collected data, although necessary for ensuring privacy goals, is rarely discussed in the studies, with only the most obvious issues addressed [2], [9], [10], [35].

Because it is easier to deploy for small-scale experiments, the centralized architecture has been the preferred solution in the surveyed frameworks [2]–[5], [8], [10], [26]. If the main server is subject of *denial-of-service* attacks, it can not guarantee the availability of the service [30]. This might result in the smartphones having to retain a more information locally with consequential privacy risks. More importantly, however, single server can compromise all user data in an instant.

A complementary approach is to deploy a decentralized architecture where algorithms can run with inputs coming from different nodes. One possible way to achieve this is to upload data from the mobile device, not to a single server, but onto personal datasets [36], like a personal home computer, or cloud-based virtual machines, lowering users' concern about systems that centralize data. On one hand, users would feel—and possibly be—more in control of their personal data, having this electronic aliases. On the other hand, part of the security of the data would inevitably rely on the user.

Given the amount of sensitive information present on mobile devices, it is our recommendation that social science researchers should team up with engineers to develop robust portable applications in order to avoid possible privacy violations [37] due to viruses and malware. Some of the studied frameworks reduce the time that the sensed raw information is kept on the phone. For example, in [2] the data records are discarded once the classification task has been performe Since most of the sensing applications use an opportunist w of uploading the data to the servers they might still store qu a lot of data temporarily on external memory [4]. This introduces a security threat if the device does not procure an encrypted file-system by default. A possible way to tackle this problem is employing frameworks like *Funf*, an open-source sensing platform for Android devices developed in [3] and also used in the *SensibleDTU* project. *Funf* provides the developers with a reliable storing system that encrypts the files before moving them to special archives on the SD card. Then an automatic process uploads what archived keeping a temporary (encrypted) backup. This layer of defense also contrasts unintended disclosures of information if the smartphone gets stolen/is lost. In this case the last resort is to provide a remote access to delete the data off the phone.

REFERENCES

1. B. Friedman, P. Lin, and J.K. Miller. Informed consent by design. *Security and Usability*, pages 495–521, 2005.
2. E. Miluzzo, N.D. Lane, K. Fodor, R. Peterson, H. Lu, M. Musolesi, S.B. Eisenman, X. Zheng, and A.T. Campbell. Sensing meets mobile social networks: the design, implementation and evaluation of the cenceme application. In *Proceedings of the 6th ACM conference on Embedded network sensor systems*, pages 337–350. ACM, 2008.
3. N. Aharony, W. Pan, C. Ip, I. Khayal, and A. Pentland. Social fmri: Investigating and shaping social mechanisms in the real world. *Pervasive and Mobile Computing*, 2011.
4. A. Madan, K. Farrahi, D. Gatica-Perez, and A. Pentland. Pervasive sensing to model political opinions in face-to-face networks. *Pervasive Computing*, pages 214–231, 2011.
5. A. Madan, S.T. Moturu, D. Lazer, and A.S. Pentland. Social sensing: obesity, unhealthy eating and exercise in face-to-face networks. In *Wireless Health 2010*, pages 104–110. ACM, 2010.
6. P. Ypodimatopoulos and A. Lippman. 'follow me': a web-based, location-sharing architecture for large, indoor environments. In *Proceedings of the 19th international conference on World wide web*, pages 1375–1378. ACM, 2010.
7. N. Eagle, A.S. Pentland, and D. Lazer. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, 106(36):15274–15278, 2009.
8. N. Kiukkonen, J. Blom, O. Dousse, D. Gatica-Perez, and J. Laurila. Towards rich mobile phone datasets: Lausanne data collection campaign. *Proc. ICPS, Berlin*, 2010.
9. R.K. Ganti, N. Pham, H. Ahmadi, S. Nangia, and T.F. Abdelzaher. Greengps: A participatory sensing fuel-efficient maps application. In *Proceedings of the 8th international conference on Mobile systems, applications, and services*, pages 151–164. ACM, 2010.
10. D. Kotz, S. Avancha, and A. Baxi. A privacy framework for mobile health and home-care systems. In *Proceedings of the first ACM workshop on Security and privacy in medical and home-care systems*, pages 1–12. ACM, 2009.
11. R. Böhme and S. Köpsell. Trained to accept?: A field experiment on consent dialogs. In *Proceedings of the 28th international conference on Human factors in computing systems*, pages 2403–2406. ACM, 2010.
12. A.P. Felt, K. Greenwood, and D. Wagner. The effectiveness of application permissions. In *Proc. of the USENIX Conference on Web Application Development*, 2011.
13. S. Egelman, A.P. Felt, and D. Wagner. Choice architecture and smartphone privacy: Theres a price for that. In *Workshop on the Economics of Information Security (WEIS)*, 2012.
14. A. Kapadia, T. Henderson, J. Fielding, and D. Kotz. Virtual walls: Protecting digital privacy in pervasive environments. *Pervasive Computing*, pages 162–179, 2007.
15. P. Klasnja, S. Consolvo, T. Choudhury, R. Beckwith, and J. Hightower. Exploring privacy concerns about personal sensing. *Pervasive Computing*, pages 176–183, 2009.
16. A.P. Felt, E. Ha, S. Egelman, A. Haney, E. Chin, and D. Wagner. Android permissions: User attention, comprehension, and behavior. In *Proceedings of the Eighth Symposium on Usable Privacy and Security*, page 3. ACM, 2012.
17. M.K. Underwood, L.H. Rosen, D. More, S.E. Ehrenreich, and J.K. Gentsch. The blackberry project: Capturing the content of adolescents' text messaging. *Developmental psychology*, 48(2):295, 2012.
18. S. Vosoughi, M.S. Goodwin, B. Washabaugh, and D. Roy. A portable audio/video recorder for longitudinal study of child development. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 193–200. ACM, 2012.
19. P. Shabajee. Informed consent on the semantic web-issues for interaction and interface designers. In *3rd International Semantic Web User Interaction Workshop.¡ http://swui. semanticweb. org/swui06/papers/Shabajee/Shabajee. pdf¿(retrieved 15.11. 10)*, 2006.
20. M. Li, K. Sampigethaya, L. Huang, and R. Poovendran. Swing & swap: user-centric approaches towards maximizing location privacy. In *Proceedings of the 5th ACM workshop on Privacy in electronic society*, pages 19–28. ACM, 2006.
21. L. Kagal, T. Finin, and A. Joshi. A policy based approach to security for the semantic web. *The Semantic Web-ISWC 2003*, pages 402–418, 2003.
22. E. Toch, N.M. Sadeh, and J. Hong. Generating default privacy policies for online social networks. In *Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems*, pages 4243–4248. ACM, 2010.
23. A. Narayanan and V. Shmatikov. Robust de-anonymization of large sparse datasets. In *Security and Privacy, 2008. SP 2008. IEEE Symposium on*, pages 111–125. IEEE, 2008.
24. M. Barbaro, T. Zeller, and S. Hansell. A face is exposed for aol searcher no. 4417749. *New York Times*, 9(2008):8For, 2006.
25. L. Sweeney. Simple demographics often identify people uniquely. *Health (San Francisco)*, pages 1–34, 2000.
26. R.A. Popa, H. Balakrishnan, and A. Blumberg. Vpriv: protecting privacy in location-based vehicular services. In *Proceedings of the 18th conference on USENIX security symposium*, pages 335–350. USENIX Association, 2009.
27. B. Hull, V. Bychkovsky, Y. Zhang, K. Chen, M. Goraczko, A. Miu, E. Shih, H. Balakrishnan, and S. Madden. Cartel: a distributed mobile sensor computing system. In *Proceedings of the 4th international conference on Embedded networked sensor systems*, pages 125–138. ACM, 2006.
28. T. Li and N. Li. On the tradeoff between privacy and utility in data publishing. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 517–526. ACM, 2009.
29. L. Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.
30. C. Cornelius, A. Kapadia, D. Kotz, D. Peebles, M. Shin, and N. Triandopoulos. Anonysense: privacy-aware people-centric sensing. In *Proceedings of the 6th international conference on Mobile systems, applications, and services*, pages 211–224. ACM, 2008.
31. R. Shokri, G. Theodorakopoulos, J. Le Boudec, and J. Hubaux. Quantifying location privacy. In *Security and Privacy (SP), 2011 IEEE Symposium on*, pages 247–262. IEEE, 2011.
32. A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):3, 2007.
33. N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*, pages 106–115. IEEE, 2007.
34. A.D. Molina, M. Salajegheh, and K. Fu. Hiccups: health information collaborative collection using privacy and security. In *Proceedings of the first ACM workshop on Security and privacy in medical and home-care systems*, pages 21–30. ACM, 2009.
35. E. Miluzzo, C.T. Cornelius, A. Ramaswamy, T. Choudhury, Z. Liu, and A.T. Campbell. Darwin phones: the evolution of sensing and inference on mobile phones. In *Proceedings of the 8th international conference on Mobile systems, applications, and services*, pages 5–20. ACM, 2010.
36. J.I. Hong and J.A. Landay. An architecture for privacy-sensitive ubiquitous computing. In *Proceedings of the 2nd international conference on Mobile systems, applications, and services*, pages 177–189. ACM, 2004.
37. Y. Altshuler, N. Aharony, Y. Elovici, A. Pentland, and M. Cebrian. Stealing reality: when criminals become data scientists (or vice versa). *Security and Privacy in Social Networks*, pages 133–151, 2011.

Participant Behavior in PHONELAB

Anandatirtha Nandugudi, Anudipa Maiti, Fatih Bulut, Sonali Batra, Taeyeon Ki
Geoffrey Challen, Murat Demirbas, Steven Y. Ko, Tevfik Kosar, and Chunming Qiao

Department of Computer Science and Engineering
University at Buffalo, The State University of New York
team@phone-lab.org

1 Introduction

This abstract examines the behavior of the participants in PHONELAB, a public smartphone testbed being developed at SUNY Buffalo. Currently consisting of 191 participants using Nexus S 4G smartphones, PHONELAB aims to provide a combination of unique features desirable for smartphone experimentation. This abstract briefly introduces PHONELAB and presents some of the early results of a usage measurement study conducted with 115 participants.

1.1 PHONELAB Overview

PHONELAB is designed to provide the following features necessary for smartphone research—open access, scale, power, realism, locality, and relevance:

- **Open Access:** After the initial approval process, PHONELAB allows any researcher to deploy their research prototype on the participants' smartphones.
- **Scale:** By 2014, PHONELAB will grow to 700 participants already incentivized and recruited to participate in experiments; participants of PHONELAB receive discounted voice, data, and messaging.
- **Power:** By utilizing the Android open-source smartphone platform, PHONELAB allows application-level experiments as well as platform-level, i.e., the OS kernel, middleware, and libraries.
- **Realism:** Participants use the phones as their primary device.
- **Locality:** Most participants live in Buffalo near SUNY campuses, enabling research requiring device-to-device interaction.
- **Relevance:** PHONELAB allows researchers to stop relying on out-of-date datasets. Instead, new data can be collected in the most appropriate way for the experiment.

PHONELAB application-level experiments are distributed through the Play Store; participants are notified of new experiments and install the experimental applications directly from the Play Store. On the other hand, PHONELAB platform-level experiments are distributed through the PHONELAB control software that runs on each participant's phone; this control software is capable of updating platform components, e.g., libraries and kernel modules. To the best of our knowledge, PHONELAB is the only testbed that provides all the above features together.

Affiliation			
Freshman	64	Masters	5
Sophomore	33	PhD	53
Junior	1	Faculty/Staff	29
Senior	1	None	5
Gender			
Female	51	Male	140
Age			
Under 18	12	30–34	15
18–19	74	35–39	6
20–21	12	40–49	13
22–24	22	50–59	7
25–29	29	60+	1

Table 1: Demographic breakdown of 191 PHONELAB participants. Date ranges are inclusive.

1.2 PHONELAB Demographics

Currently, PHONELAB consists of 191 participants. Roughly half of our participants are first- and second-year undergraduates, a quarter PhD students, and a fifth faculty, staff and other professionals. However, males greatly outnumber females, and the young outnumber the middle-aged and older, both unrepresentative features we will try and rectify in the future years. For management reasons we limited participation to people with a SUNY Buffalo affiliation except for several exceptions: a local reporter, a technology writer, and an international rock star. Table 1 summarizes our demographics.

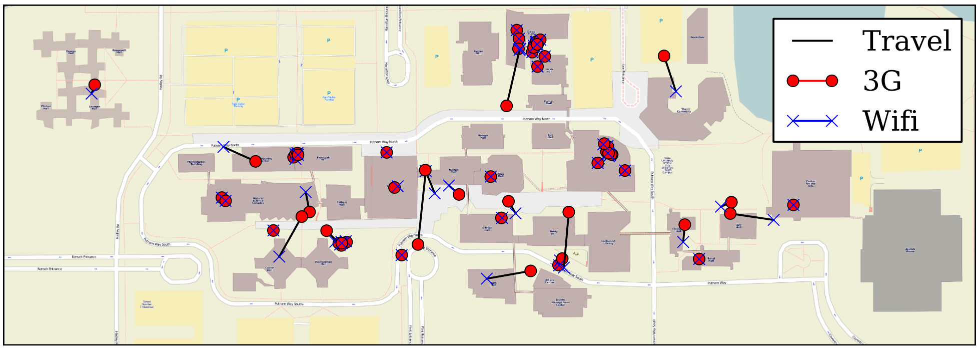


Fig. 1: 3G to Wifi transition locations. The map indicates that there are several common areas where network hand-offs occur.

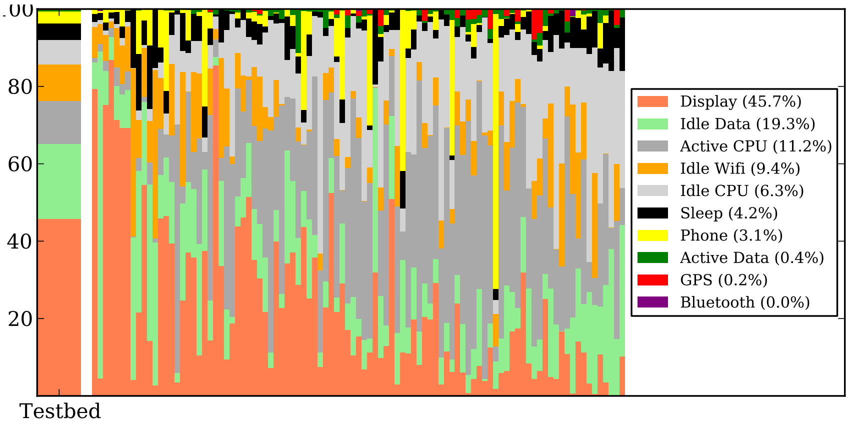


Fig. 2: Power usage by component. The large bar at left shows an aggregated breakdown for all participants. The participant bars are scaled against the participant with the most energy usage.

2 Participant Behavior

We have conducted a usage measurement study with 115 participants over 21 days. For this purpose, we have developed a measurement application that collects all salient features of smartphone usage: networking, mobility, power consumption, and application usage.

This section presents some of the early results of this study. We show the network transition behavior between 3G and WiFi first and the battery and charging behavior next.

2.1 Mobile Network Transitions

Mobile devices like smartphones move through a complex network environment. Providing the illusion of seamless connectivity requires negotiating hand-offs both between Wifi access points and between Wifi and 3G radios. We were interested in observing hand-offs between 3G (provided by Sprint, PHONELAB's operational partner) and Wifi and found many in the dataset collected by our usage experiment. Since the Android `ConnectivityService` frequently switches network interfaces for exploration purposes, we have defined a transition as two one-minute or longer sessions on different interfaces separated by less than one minute. We further limit ourselves to cases where we received a location update during the transition.

Figure 1 plots the location of transitions that occurred on or near SUNY North Campus. We notice that many cluster in expected locations: near the entrance and exits of buildings where participants are likely to be moving from campus Wifi to 3G.

2.2 Energy Breakdown

A single-day component-by-component breakdown is shown in Figure 2. Our results are similar to those reported by a previous smaller-scale study [4], and indicate that mobile data (labeled as “Idle data” and “Active data” depending on the state), the screen, and CPU usage are the main sources of smartphone power consumption. The per-participant bars also show a great deal of variation, with differences in both the amount and the breakdown of energy consumed by each participant.

One supposedly power-hungry component that has less of an impact than we had expected is the GPS. This is particularly surprising given the large amount of location-monitoring work motivated by GPS power consumption. One of several factors may be at work. First, the Android platform estimates the GPS chipset current consumption at 50 mA. This number is used by the standard "Fuel Gauge" battery monitor and by our calculations. However, it is lower than the data sheet for the Broadcom 4751 GPS receiver [1] and may represent a best-case average. Still, even if the GPS current consumption is off by as much as a factor of five, it does not represent a significant contribution. Other hypotheses are that Android network location is providing location with sufficient accuracy for many applications, eliminating the need for GPS, or participants and applications may simply be conscious of GPS power consumption and taking steps to control it.

2.3 Opportunistic Charging

One way that users work around the battery limitations of their smartphone devices is by finding new times and places to charge their phones: plugging in at their desk at work, in the car during their commute, or at home before a long night out. We refer to these charging sessions as *opportunistic* to distinguish them from *habitual* nightly charging. Assuming that many smartphone users encounter plug points throughout the day, engaging in opportunistic charging becomes an additional sign of energy awareness, and understanding opportunistic charging becomes necessary to improving energy management on mobile devices. Others have analyzed this behavior before [2, 3] and our goal is to examine the battery charging behavior of PHONELAB partipants.

Figure 3 shows that many users engage in opportunistic charging. We define a charging session as opportunistic if is long enough to not be spurious (over 10 minutes) but does not bring the battery to a fully-charged state, indicating that the user disconnected the device before charging could finish. For a representative day during our experiment, of the 245 charging sessions we observed that day, 96 (39%) were opportunistic by this definition. 50 of 95 active participants engaged in opportunistic charging at some point during our experiment an average of once per day.

Opportunistic charging may be a response to an anticipated need for more smartphone battery power: the student who plugs her smartphone in for a brief charge before a night out. Our data also allowed us to examine how many of these opportunistic charging sessions were necessary to bridge the gap to the next full charge. We found that 24 of the 96 (25%) of the opportunistic charges we observed were necessary. We believe that this indicates that participants have responded to their smartphones' battery limitations by engaging in conservative charging behavior, grabbing power whenever possible even if they do not anticipate needing it later.

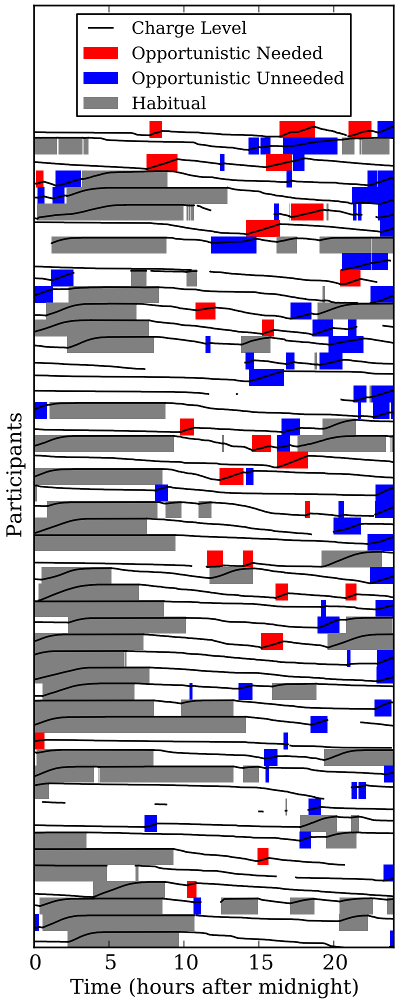


Fig. 3: Patterns of opportunistic charging. Many users perform opportunistic charging multiple times during the day.

3 Conclusions

This abstract introduced PHONELAB, a new large-scale programmable smartphone testbed operated by SUNY Buffalo and presented the participant behavior in terms of network transitions, energy, and charging.

References

[1] Broadcom BCM4751 Integrated Monolithic GPS Receiver. http://www.broadcom.com/products/GPS/GPS-Silicon-Solutions/BCM4751.

[2] N. Banerjee, A. Rahmati, M. D. Corner, S. Rollins, and L. Zhong. Users and Batteries: Interactions and Adaptive Energy Management in Mobile Systems. In *UbiComp*, 2007.

[3] A. Rahmati, A. Qian, and L. Zhong. Understanding Human-Battery Interaction on Mobile Phones. In *MobileHCI*, 2007.

[4] A. Shye, B. Scholbrock, and G. Memik. Into the Wild: Studying Real User Activity Patterns to Guide Power Optimizations for Mobile Architectures. In *MICRO*, 2009.

Exploring the Use of Urban Greenspace through Cellular Network Activity

Ramón Cáceres[†], James Rowland[†], Christopher Small[‡], Simon Urbanek[†]
[†]AT&T Labs [‡]Columbia University
{ramon, jrr, urbanek}@research.att.com csmall@columbia.edu

Knowing the extent to which people make use of urban greenspace is central to our understanding of urban ecology. While the type and location of greenspace in urban areas is well documented, we lack accurate, quantitative measures of when and where people occupy it.

Cellular telephone networks can provide a wealth of information about the use of urban greenspace. Mobile phones are carried by a large portion of the population and are used throughout the day. A measure of how many phones are active in which geographic areas can thus serve as a proxy for human density in those areas.

In this work, we use anonymous records of cellular network activity to quantify the spatiotemporal patterns of human density within a major US metropolitan area. More specifically, we use counts of voice calls and text messages handled by cellular antennas as a measure of how many people are in the geographic areas covered by those antennas. Because of the close-knit spacing of antennas in urban areas, variations in these counts can shed light on the use of individual green spaces.

We aim to characterize how the density of network activity changes over time, and how these density patterns relate to greenspace and microclimate. By aggregating activity into density maps at different times of day, week, and season, we hope to enhance our understanding of when people occupy different types of greenspace. This paper presents the vision and some early results of this effort.

1. DATASET

We have gathered from a major US communications service provider a dataset of anonymous cellular network activity in the New York metro area. We identified the set of ZIP codes within 50 miles of downtown Manhattan, then obtained a list of cellular antennas that were active in those ZIP codes during our study. We grouped into a *sector* the set of antennas that reside on the same cellular tower and that point in the same compass direction, or *azimuth*. For each of those sectors and for each minute of each day, we gathered counts of how many new voice calls and how many text messages were handled by the antennas in that sector.

Finally, we sum the voice calls and text messages to arrive at a single measure of cellular network activity that we term *call volume*. Similarly, we use *call density* to denote call volume per geographic area, and treat call density as a proxy for human density.

Our current dataset spans the six months between February 1 and July 31, 2011. It contains one record per minute for more than 12,000 sectors, yielding more than 3 billion call-volume samples. We are currently gathering data for a full year.

We have been careful to preserve privacy throughout this work. In particular, this study uses only anonymous and aggregate data. There is no personally identifying information in the data records described above.

2. TESSELLATION

Our cellular network data gives us estimates of human activity levels, but we need a way to assign that activity to geographic areas. Voronoi tessellation has been used to associate spatial regions with cellular towers [4, 12]. However, basing the tessellation only on tower locations results in coarse regions, and therefore coarse assignments of activity to geographic areas.

We have developed an algorithm that performs a finer-grained tessellation by making use of antenna directions in addition to tower locations. Figure 1 illustrates the result of our refined tessellation based on sector azimuths, with added edges drawn as dashed lines. Making use of the azimuth information clearly improves the granularity. For the New York metropolitan area, the median area of a Voronoi region resulting from the extended tessellation is roughly one quarter of the median area resulting from the regular tessellation.

3. EOF ANALYSIS

In this study, we seek to quantify the spatiotemporal patterns of call volumes in order to infer the spatiotemporal distribution of people in the New York metro area. We can then analyze these patterns in the context of the spatial distribution of greenspace and temporal variations in weather conditions. We determine the greenspace distribution using vegetation maps derived

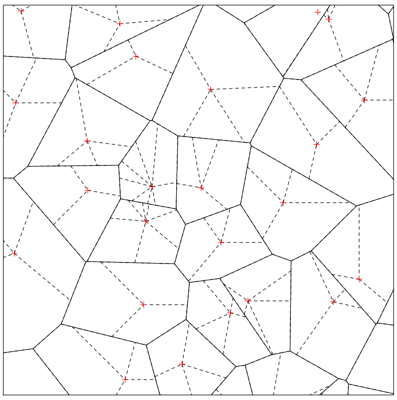


Figure 1: Voronoi tessellation for a sample 60-km^2 area. Red crosses denote cellular tower locations. Taking into account antennas and their directions produces finer-grained estimates of coverage areas.

from visible and infrared satellite imagery [9, 10]. We capture weather variations using data from a regional network of weather stations.

One objective of this coanalysis is to quantify the relationship(s) between outdoor ambient environmental conditions and the spatiotemporal distribution of people within the urban area. A primary challenge in this analysis is to distinguish between indoor and outdoor activity. A related challenge is to distinguish between regular patterns of activity (e.g. the dominant daily and weekly cycles) and the variations in these patterns that may be related to environmental conditions (e.g. indoors on cold days, outdoors on temperate days).

We will approach both of these challenges by mapping deviations from regular patterns as anomalies in time and space. We will accomplish this mapping using Empirical Orthogonal Function (EOF) analysis, a tool commonly used to quantify spatiotemporal patterns in meteorology and oceanography [13]. EOF analysis is a form of Principal Component (PC) analysis.

We will treat the call volume data as instantaneous spatial snapshots of call volumes, then analyse the spatiotemporal patterns in these time series of call volume maps. Our approach is similar to how PC analysis is used to reduce the dimensionality of multispectral imagery in remote sensing applications (e.g., [3, 5, 8]). Because variables in high-dimensional data are often correlated, PC transforms provide an efficient low-dimensional projection of the uncorrelated components of the data. The same property applies to temporal dimensions.

Generally, EOFs are spatial patterns intended to represent spatially continuous modes of variability of physical processes, while the PCs are the weights representing the temporal contribution of the corresponding spatial patterns [6, 13]. In this study, we reverse the convention so that EOFs represent temporal patterns and PCs represent spatial weights. We consider daily, weekly, and seasonal trends that result from deterministic processes such as commuting, as well as higher-frequency day-to-day variability presumably related to ambient environmental conditions and isolated transient events. Additional details of the approach are given by [11].

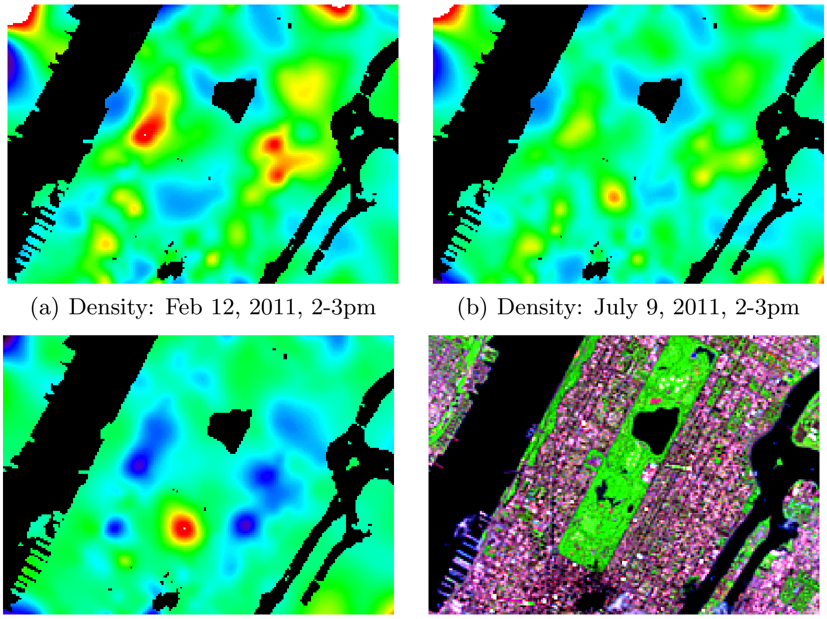
Our EOF analysis is mathematically related to the methods used in [2] and [7] to analyze cellular network data. However, we use the EOFs to identify and remove the dominant temporal periodicities in the data—thereby revealing any non-periodic spatial patterns related to greenspace and temporal patterns related to weather. In addition, we are experimenting with the combined use of EOF analyses and linear mixture models as described by [11].

4. EARLY RESULTS

We are continuing to refine our analysis approach and apply it to our data. However, we can already see relevant patterns of human behavior emerge from preliminary analysis of selected subsets of the data.

As an example of a seasonal change in human activity around greenspace, we compared Saturday-afternoon call density in central Manhattan between February 12 and July 9, 2011. We summed the per-minute call volumes between 2pm and 3pm for each sector within this area, then normalized by the sector area to produce density maps for each date. Figure 2 shows these two maps and their difference, along with a satellite image that highlights greenspace in the same area. From winter to summer, we find that density increases in the greenspace of southern Central Park, and decreases in the residential areas of the Upper East and West Sides.

Our analysis strategy for the complete New York metro area is based on the identification of spatiotemporal regularities and anomalies. We will use EOF analysis to quantify the spatial form of the dominant daily and weekly cycles associated with commuter migration, as well as any seasonal components that emerge in the low-order dimensions. Once identified, we will remove these components by inverse transformation of the remaining dimensions to produce a spatiotemporal representatior of any anomalies that are distinct from the dominant periodicities. We can then directly compare the spatial components of these anomalies to maps of greenspace and thermal microclimate. We can likewise compare the temporal components of the anomalies to time series of air temperature, precipitation, and humidity, in order to quantify whatever relationships may exist between the residual call volumes and the spatiotemporal variations in microclimate and ambient environmental conditions. Please see our longer paper [1] for additional details.



(a) Density: Feb 12, 2011, 2-3pm

(b) Density: July 9, 2011, 2-3pm

(c) Density change: July minus Feb

(d) Landsat 5: July 7, 2011, 11am

Figure 2: Spatiotemporal change in Saturday afternoon call density for central Manhattan. From winter to summer, call density increases in the greenspace of Central Park, but decreases in residential areas on the Upper East and West Sides. The visible-infrared satellite image shows parks and other greenspace as shades of green.

5. REFERENCES

[1] R. Cáceres, J. Rowland, C. Small, and S. Urbanek. Exploring the use of urban greenspace through cellular network activity. In *2nd Workshop on Pervasive Urban Applications (PURBA)*, 2012.

[2] F. Calabrese, F. Pereira, G. DiLorenzo, L. Liu, and C. Ratti. The geography of taste: analyzing cell-phone mobility and social events. In *International Conference on Pervasive Computing*, 2010.

[3] A. Green, M. Berman, P. Switzer, and M. Craig. A transformation for ordering mutispectral data in terms of image quality with implications for noise removal. *IEEE Transactions on Geoscience and Remote Sensing*, 26:65–74, 1988.

[4] T. Horanont and R. Shibasaki. An implementation of mobile sensing for large-scale urban monitoring. In *Workshop on Urban, Community, and Social Applications of Networked Sensing Systems*, 2008.

[5] J. Lee, A. Woodyatt, and M. Berman. Enhancement of high spectral resolution remote sensing data by a noise-adjusted principal components tranform. *IEEE Transactions on Geoscience and Remote Sensing*, 28:295–304, 1990.

[6] R. Preisendorffer. *Principal component analysis in meteorology and oceanography*. Elsevier, Amsterdam, 1988.

[7] J. Reades, F. Calabrese, and C. Ratti. Eigenplaces: Analysing cities using the space-time structure of the mobile phone network. *Environment and Planning B: Planning and Design*, 36:824–836, 2009.

[8] A. Singh and A. Harrison. Standardized principal components. *International Journal of Remote Sensing*, 6:883–896, 1985.

[9] C. Small. Estimation of urban vegetation abundance by spectral mixture analysis. *International Journal of Remote Sensing*, 22:1305–1334, 2001.

[10] C. Small. High spatial resolution spectral mixture analysis of urban reflectance. *Remote Sensing of Environment*, 88:170–186, 2003.

[11] C. Small. Spatio-temporal dimensionality and characterization of multitemporal imagery. *To appear in Remote Sensing of Environment*, 2012.

[12] V. Soto and E. Frias-Martinez. Robust land use characterization of urban landscapes using cell phone data. In *Workshop on Pervasive Urban Applications*, 2011.

[13] H. von Storch and F. Zwiers. *Statistical Analysis in Climate Research*. Cambridge University Press, Cambridge, UK, 1999.

Citizen as a Sensor for the SmartCity: The Barcelona Urban Mobility Use-case

Daniel Villatoro, Marc Pous, Jetzabel Serna, Francesco Ronzano and Marc Torrent-Moreno
Fundació Barcelona Digital Centre Tecnològic (BDigital)
Barcelona, Spain
{dvillatoro,mpous,jserna,fronzano,mtorrent}@bdigital.org

ABSTRACT

Recently, and with the penetration of the smart city paradigm, urban mobility is becoming an active research area. Moreover, the explosion of mobile technology enables the transformation of citizens into active and passive sensors increasing the amount of information of our scope of study: the city. The usage of citizens as sensors results in a low-cost infrastructureless framework that not only profits from the mobile facet of humans but also from their common sense labelling contextual information such as anomalous situations or venue categories. These aspects are treated in two different scenarios specific for the city of Barcelona: one that allows us to obtain information about the state of the public transportation network and the other that detects clusters of activity within the urban environment.

1. INTRODUCTION

Researchers from different scientific areas have worried about the dynamics of urban environments, specifically focusing of how humans transit the city [2, 3, 1].

Dedicated sensors have been installed in cities to capture some data and try to understand the behaviour of such complex system, determined by the aggregation of the individual decisions.

Moreover, in recent years with the explosion of mobile technologies citizens have been given the capability to continuously share information anytime anywhere. This ubiquitous and continuous sensing capability combined with the humans' "common sense" transforms any device-holder in a potential proactive intelligent sensor, which extends the classical continuous-sensing sensors installed in a certain location with an specific scope.

In this extended abstract we propose two complementary ways to exploit the human sensing capabilities in order to understand their location in the urban environments, and their mobility patterns within it (where do citizens go and how they go there). Namely, the macro vision obtained through the analysis of proactive citizen-sensed information retrieved from social media sites (in this paper Twitter) allows us to understand a wide range of aspects of the city (because of the free text accompanied with GPS metainformation) although generated by a specific segment of the population. Consequently this macro vision provides us with intuitions of the mobility patterns of the city (captured with consecutive geopositioned tweets). In order to obtain deeper insights and representative information arises the necessity of a dedicated sensor able to acquire more accurate and representative information on this matter. This link with the micro vision led us to develop Commutio, a participatory sensing mobile app targetted for commuters. Commutio data will provide us with real-time access to the load of the public transportation network. This information is essential (1) to have a complete overview of the public transpotration system dynamics, (2) recommend alternative and more efficient trips trips for individuals, (3) and provide real-time information of unexpected incidents and alternatives for afected commuters.

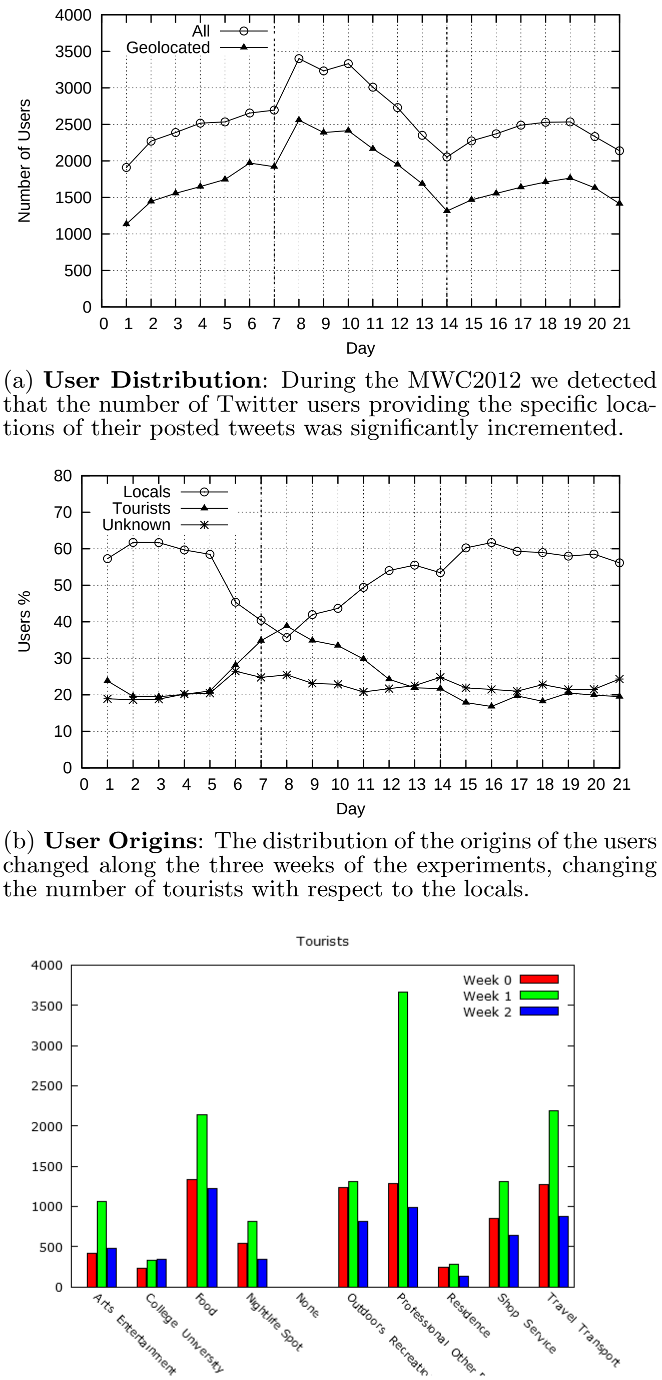
In this contribution we describe the barriers encountered when deploying our solution with the goal of understanding urban processes such as the identification of transitory areas of activity and mobility patterns in the public transportation network (PTN).

2. URBAN MACRO VISION THROUGH SOCIAL NETWORKS

Because of its popularity, geo-positioning capabilities and penetration of its mobile app[1], we opted to use Twitter as the first source of information. This social sensor provides us with information (in the form of geolocalized tweets shared through users mobile devices) to observe and detect alterations of several parameters such as areas of activity, temporal patterns or mobility routes, which define the Urban Chronotype of the city. In order to obtain another dimension of information, we enrich the information obtained from Twitter with information from Foursquare. This combination results in a fully crowdsourced sensor where the meta-information provided through "tweets" is augmented w that of "check-ins".

The processes of tweet-collection, foursquare cross-refe and analysis presented certain technical barriers that h been solved with the development of our Urban Sensing Platform. This platform ensures the acquisition in near real-time of all the tweets, without the requirement of any private partnership with Twitter as commonly done by other researchers, allowing an agile deployment in any system. The obtained geopositioned tweets are enriched with Foursquare

[1] According to an study released by Semiocast in January 2013, 74% of the tweets are generated using a mobile device.



(a) **User Distribution**: During the MWC2012 we detected that the number of Twitter users providing the specific locations of their posted tweets was significantly incremented.

(b) **User Origins**: The distribution of the origins of the users changed along the three weeks of the experiments, changing the number of tourists with respect to the locals.

(c) **Tourist Behaviour with Foursquare Augmented Info**: We can quantify that during the MWC2012 (Week 1) certain venue categories activity are significantly altered, specifically those that relate directly to such type of event.

Figure 1: Participatory Sensing during MWC2012 through Twitter and Foursquare.

data, obtaining a distribution of venue categories. Moreover, it is important to understand the relationship of our citizen-based sensors with the city to better understand the samples they generate. For such task, our platform has been enriched with a semantic-intelligence module that contextualize the geographical origin of users, obtained from their Twitter profile. Categorizing users as locals or foreigners with respect to the city of study allows for a detailed behavioural pattern analysis differentiating their urban habits.

We performed an experiment during the expected although exceptional situation of the celebration of the Mobile World Congress (MWC) in Barcelona (whose main results are presented in Fig. 1). Our knowledge about this specific event allowed us to compare the average urban chronotype of the city (in the normal state without external influences), with that obtained during the event, allowing us to observe different behavioural patterns within the city. The empirical results obtained have given us the possibility to mine and compare the behavioural patterns of the city in its normal state and during the event: the number of sensed users increases during the event as it can be seen in Fig. 1(a); on the other han, we can infer that these users attending the event are tourists as Fig. 1(b) shows; and finally, Fig. 1(c) shows how the types of venues change accordingly to the type of event in the city, affecting mainly proffesional, travel and food venues. Substantial differences have been observed in the areas of activity of the city after applying geospatial data mining techniques (geo-clustering), thus providing a first proof of validity of our urban social sensing approach. For a more descriptive explanation of this experiment and further experimental results we refer the reader to [4].

3. URBAN MICRO VISION USING CITIZENS AS SENSORS

However, profiting from the information proactively shared by the citizens in social media applications only gives us access to a restricted subset of the potential information that represents the Urban Chronotype of the city: a partial understanding of the location of the users can be made remaining still unclear how people get to those places. This is where the macro vision of the social sensed connects with the micro necessity of undertanding how people transit the city, specifically, what the transit dynamics of the public transportation network are? Unfortunately, the transportation company's existing infrastructure cannot provide specific real-time information about the number and location of users within the network: in cities such Barcelona it is only controlled the access stations of travellers, but not the exits neither the load of the vehicles or the transportation lines. Our vision assumes that citizens in the network (through the usage of their mobile devices) can provide with this information in a crowdsourced way. However social media applications (as the previously used) do not provide the incentives for users to continuously share their position within the network.

For such reasons, and to obtain more accurate real-time information about the near real-time state of the PTN, we have created an ad-hoc sensor in the form of a app (shown in Fig.2) that profits from the sensing capabilities of travellers and, while open, continuously share the citizen geoposition (latitude, longitude and timestamp) with our centralized server. However, we faced two problems: why will users open an app that sacrifices part of their battery life? and

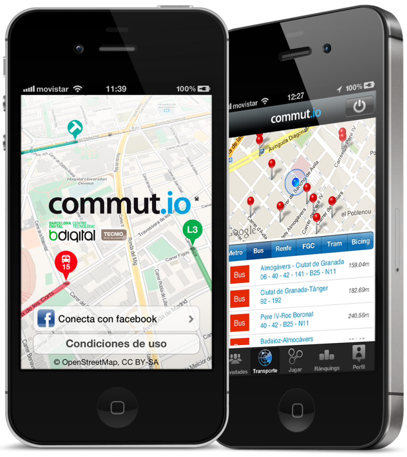


Figure 2: Commutio: Through this mobile application we are able to track Barcelona's public transportation network travellers in real-time.

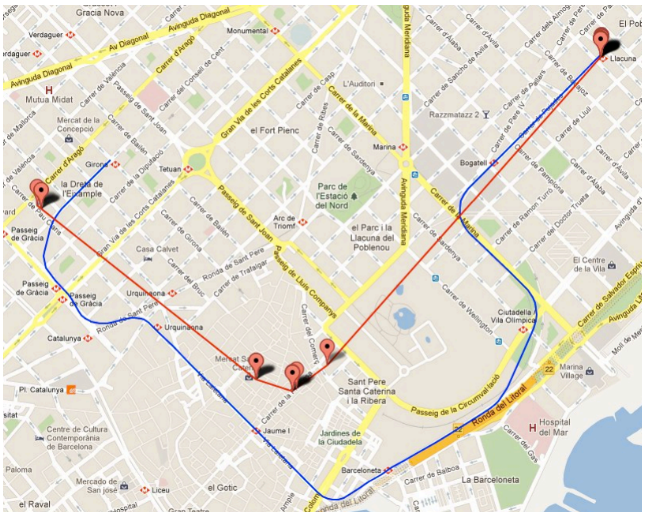


Figure 3: One Metro-commuter Trace: The red icons in the map are the position provided by the client application to the server and the red lines are the connections between two consecutive geolocations. These positions belong to a user that was travelling in the metro line specified with the blue line. We can easily notice how divergent are the acquired data from that of the real user location. With the usage of the adequately trained Case-Based Reasoning algorithm, our framework will be able to associate the captured GPS positions to the stations or edges associated to each of them.

how can we obtain enough critical mass of users to make this crowdsourced sensor work? To solve both questions we have created an upper layer using gamification techniques. Thanks to a competition-based real-time trivia-like game, we incentivate our users to play with our app: the game constitutes an individual incentive to open the app, and the real-time competition with other users and friends ensures the virality and the long-term usage of the app.

With the real-time data gathered from the commuters of Barcelona we are able to understand the load of the PTN, obtaining also a commuter behavioural profile in the system in various situations (e.g. demonstrations, or unexpected incidents). Our modelization using a dynamic network data structure allows us to apply graph theory algorithms to optimize the behaviour of the system in several aspects and eventually build a route recommender system. Amongst the challenges we have faced in this project, it is interesting to emphasize how the lack of GPS-signal in underground transportations have been solved using a Case-Based Reasoning algorithm as shown in Fig. 3.

4. CONCLUSIONS

These two specific applications have given us different but complementing visions of the urban flows: (1) a macro vision, captured from the proactively shared social media contents, allowing us to understand the location of a specific segment of the population; and a micro vision, with dedicated apps sensing a specific subject such as the PTN, allowing us to obtain complete information of how citizens transit the city.

Nowadays, the concept of smart city appears with a top-down vision where citizens receive the services implemented by the city and the government. As we proposed, the proper usage of the collected information would provide a more complete picture (micro and macro) with near real-time information about the target city allowing for a significant improvement of the urban performance, with no specific investment on hardware infrastructure.

Acknowledgements

This work has been completed with the support of ACC1Ó, the Catalan Agency to promote applied research and innovation; and by the Spanish Centre for Development of Industrial Technology under the INNPRONTA program, project IPT-20111006, "CIUDAD2020".

5. REFERENCES

[1] C. F. Daganzo. *Fundamentals of Transportation and Traffic Operations*. Elsevier Science, 1997.

[2] K. Lynch. *The Image of the City*. The MIT Press, J 1960.

[3] C. E. Mandl. Evaluation and optimization of urban public transportation networks. *European Journal oj Operational Research*, 5(6):396–404, December 1980.

[4] D. Villatoro, J. Serna, V. Rodríguez, and M. Torrent-Moreno. The tweetbeat of the city: Microblogging used for discovering behavioural patterns during the mwc2012. In J. Nin and D. Villatoro, editors, *Citizen in Sensor Networks*, volume 7685 of *Lecture Notes in Computer Science*, pages 43–56. Springer Berlin Heidelberg, 2013.

Trace-based Analysis to Identify Popular Locations Visited in Mobile Ad hoc Networks

Aarti Munjal
Department of Biostatistics and Informatics,
Colorado School of Public Health
Aurora, CO 80045
Aarti.munjal@ucdenver.edu

Thyago Mota
Department of Electrical Engineering and
Computer Science, Colorado School of Mines
Golden, CO 80401
tmota@mines.edu

Tracy Camp
Department of Electrical Engineering and
Computer Science, Colorado School of Mines
Golden, CO 80401
tcamp@mines.edu

Research shows that human movement patterns are predictable to some extent [CITE2]. Finding patterns in human mobility is of interest for several reasons, e.g., knowledge of contacts made by humans carrying mobile devices can be used in making efficient routing decisions. The movement trace collected from a real scenario reveals several interesting statistical features present in human mobility. We note, however, that the extracted statistical patterns are based on the dataset used for analysis. As a result, observations made from one dataset may not be applicable to another scenario. Our recent work towards understanding human mobility can be found in [CITE1, CITE3]. Our submission titled *"Exploring Social Interactions via Multi-Modal Learning"* presents results obtained from the analysis of the Nokia Mobile Data Challenge (MDC) dataset. As part of Nokia MDC'12, we were released a dataset for 38 participants. (See [CITE1] for a detailed description of our analysis of the Nokia MDC'12 dataset.)

In this work, we propose to extend our analysis done in [CITE1] and explore the impact of different input values used in the analysis methods. Specifically, we plan to analyze the Nokia MDC dataset and investigate the following research questions:

1. Song et al. in [CITE2] analyzed a large dataset of human movement traces and concluded that while moving, humans revisit a location based on the probability given by Equation 1 in [CITE1]. We note that the values of parameters a and b (Equation 1 in [CITE1]) obtained from the Nokia MDC'12 dataset are different than the values of a and b obtained in [CITE2]. In particular, a and b values extracted from Nokia dataset are 0.98 and 0.0008, respectively, while a and b obtained from [CITE2] are 0.6 and 0.21, respectively. We note that these values are based on the dataset as well as the definition used for a "*previously visited*" location. In [CITE2], the authors track the cell phone towers a user is connected to during his movement. Thus, if a user reconnects to a previously visited cell phone tower, it is recorded as a visit to a *"previously visited"* location. In [CITE1], however, we define X as a "*previously visited*" location if the user is within *5* meters of a location Y previously visited by him. In other words, the scale in our work in

[CITE1] is much more precise than in [CITE2]. In this work, we plan to extend our analysis by defining a "*previously visited*" location similar to [CITE2]. We then propose to compare our results with the results obtained in [CITE2] to validate the results in [CITE2] and our methodology in [CITE1].

2. In [CITE1], we analyzed the MDC'12 dataset for 38 users to explore social interactions among the users. (See [CITE1] for details.) During analysis, we found that users that belong to the same social group have a strong correlation between their visited locations. In other words, we found a significant overlap between the locations visited by users that showed strong social ties. In this work, we are interested in investigating the *type* of social interaction between a pair of nodes. In the MDC'12 dataset, for example, user 94 visits 1.75% of the locations visited by user 23; however, user 23 visits 67.44% of the locations visited by user 94. With this observation, we may conclude that user 94 does not move as often as user 23. For example, perhaps user 94 is a bank teller and user 23 is visiting the bank as a customer; in other words, perhaps user 94 only has a few locations that are visited (e.g., bank, store, home) and user 23 visits two of them (e.g., bank and store). Similarly, user 2 contacts user 75 more frequently than contacts user 51 and user 68. Thus, based on the contact information (e.g., frequency of contacts, duration of contacts, etc.) and/or the amount of overlap between the locations visited by the users, we are interested in exploring the *type* of social interactions that may exist between a pair of users.
3. We plan to simulate the Nokia MDC scenario on our mobility model called SMOOTH [CITE3] and then compare the movement traces generated by the scenario simulated on SMOOTH with the real (dataset) traces for their several statistical features listed in [CITE3].

[CITE1] A. Munjal, T. Mota, and T. Camp. Exploring Social Interactions via Multi- Modal Learning. *Proceedings of Mobile Data Challenge by Nokia Workshop*, 2012.

[CITE2] C. Song, T. Koren, P. Wang, and A.-L. Baraba´si. Modelling the scaling properties of human mobility. *Nature Physics*, pages 818–823, 2010.

[CITE3] A. Munjal, T. Camp, and W. C. Navidi, "SMOOTH: A simple way to model human walks," *Proceedings of MSWiM*, pp.351-360, 2011.

2

Poster

Network

Simulation of Epidemic Spread using Cell-Phone Call Data: H1N1 Case Study

Enrique Frías-Martínez [‡], Graham Williamson [#], Vanessa Frías-Martínez [‡],

[‡] *Telefónica Research, Madrid – Spain*

[#] *School of Computer Science , University College Dublin – Ireland*
{efm,graham,vanessa}@tid.es

***Abstract*—Ubiquitous computing technologies enable capturing large amounts of human behavioral data. The digital footprints computed from these datasets provide information for the study of social dynamics, including social networks and mobility patterns, key elements for the effective modeling of virus spreading. Traditional epidemiologic models do not consider individual information and hence have limited ability to capture the inherent complexity of the disease spreading process. In this paper we propose an agent-based system that uses social interactions and individual mobility patterns extracted from call detail records to accurately model virus spreading. The proposed approach is applied to study the 2009 H1N1 outbreak in Mexico.**

I. Introduction

Traditional epidemiological approaches base their solutions on using differential equations that divide the population into subgroups based on socio-economic and demographic characteristics. Although these models fail to capture the complexity and individuality of human behavior, they have been extremely successful in guiding and designing public health policies. The recent adoption of agent-based modeling (ABM) approaches has allowed to capture individual human behavior and its inherent fuzziness by representing every person as a software agent.

The adoption of ubiquitous computing technologies by very large portions of the population (*e.g.* GPS devices, ubiquitous cellular networks or geolocated services) has enabled capturing large scale human behavioral data. These datasets contain information that is critical to accurately model the spread of a virus, such as human mobility patterns or the social network characteristics of each individual

In this paper, we propose an ABM system designed to simulate rus spreading using agents that are characterized by their in-vidual mobility patterns and social networks as extracted from ll phone records. We carry out simulations with data collected uring the 2009 Mexican H1N1 outbreak and measure the impact that government calls had on the mobility of individuals and the subsequent effect on the spread of the H1N1 virus. An extended description of our system and its evaluation using the 2009 H1N1 outbreak can be found in [1].

We have used call detail records(CDR) to compute: (1) a *mobility user model* and (2) a *social user model* that identifies each agent's social network. This approach of capturing and modeling agent behavior from CDRs sets our work apart from others because: (1) we model agents from real individual data and not from census or surveys; and (2) we capture behavioral adaptations to the spread of the disease.

II. ABM of Virus Spreading using CDRs

We propose an ABM architecture with two main components: (1) a set of agents that are modeled using the information contained in call detail records; and (2) a discrete event simulator (DES) that simulates the virus propagation over time based on the agents' models.

Agent Generation

We define the behavior of each agent with three models: (1) a mobility model extracted from CDR data; (2) a social network model computed from CDR data; and (3) a disease model that characterizes the progression of the disease through its various states in each agent.

The mobility model provides the position (at the BTS level) where the agent is at each moment in time. This model is used by the event simulation process to predict the location of each agent at each simulation step. We propose a mobility model that divides each day into a set S of i non-overlapping equal-length time slots. The mobility model of agent n, M_n, is defined as:

$$\begin{array}{l} M_n = \{M_n^{wday}, M_n^{wend}\} = \\ \{\{M_n^{wday,0}, .., M_n^{wday,i}\}, \{M_n^{wend,0}, .., M_n^{wend,i}\}\} \quad \forall i \in S \\ M_n^{wday,i} = \{p_n^{wday,i,0}, \ldots, p_n^{wday,i,j}\} \quad \forall j \in B \\ M_n^{wend,i} = \{p_n^{wend,i,0}, \ldots, p_n^{wend,i,j}\} \quad \forall j \in B \end{array} \tag{1}$$

where B is the number of BTS towers that give coverage to a geographic area; and $p_n^{wday,i,j}$ and $p_n^{wend,i,j}$ denote the probability that agent n may be found at BTS j in timeslot i during a weekday or weekend, respectively. Given a CDR dataset, the mobility model is built by associating with each time slot i the set of BTSs where each person has been *observed* during weekdays or weekends during the period of time under study.

Note that each individual might be assigned to more than one BTS in a specific time slot i. In this case, the event simulator assigns the position of the tower with the highest probability, *i.e.*, the BTS that the individual has used the most over the training period. Since people tend to show monotonic behaviors, an average person typically has very few BTS towers in his/her mobility model.

We compute the social network of an agent as the set of individuals with whom there was at least one reciprocal contact during the time period under study:

$$\begin{array}{l} S_n = \{S_n^{wday}, S_n^{wend}\} = \\ S_n^{wday} = \{list\ of\ reciprocal\ contacts\ in\ wdays\} \\ S_n^{wend} = \{list\ of\ reciprocal\ contacts\ in\ wends\} \end{array}$$

where S_n^{wday} is the social network during the weekdays and S_n^{wends} the social network during the weekends. Given the social networks of an agent, we assume that the probability of being physically close to another agent will be higher if that other agent is part of its social network. To model physical proximity within a BTS coverage area we define two probabilities: (1) p_1 is the probability that two agents that are in the same BTS at the same time of the simulation and are part of the same social network are physically close enough for the virus to be possibly transmitted; and (2) p_2 the probability that two agents that are in the same BTS and are *not* in the same social network at the same moment in time are physically close for the virus to be possibly transmitted.

The disease model captures the progression of the disease in each agent. We follow a similar approach to that of Barret *et al.* [2] and define a disease model that is composed of two parts: the *between hosts* transmission model and the *within host* progression model. In Figure 1 we observe that the *between hosts* transmission model happens at a probability p_i and represents the probability that an agent goes from Susceptible to Exposed. The *within host* model represents the evolution from Exposed to Infective in a given period of time ϵ, and from Infected to Removed in period of time β.

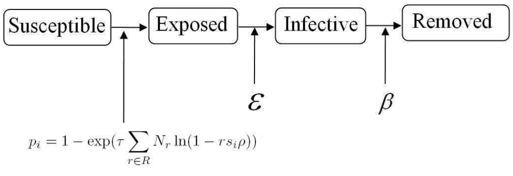

Susceptible
Exposed
Infective
Removed
$\mathcal{E}$
β
$p_i = 1 - \exp(\tau \sum_{r \in R} N_r \ln(1 - r s_i \rho))$

Fig. 1. Disease Model composed of Between hosts and Within hosts models.

Discrete Event Simulator

The Discrete Event Simulator (DES) simulates the evolution of the epidemic spreading for a set of agents over a specific period of time. To bootstrap the epidemic spreading, we assume that an initial agent is Infected and starts the transmission. Specifically, the DES does the following consecutive tasks: (1) It identifies the geographical area (BTS) where each agent is located using the mobility model; (2) it identifies the geographical areas where there is, at least, one Infective agent; (3) for each Infective agent, it takes all the Susceptible agents of his social network that are located in the same geographical area (BTS coverage) and applies probability p_1 that they will be physically close for the virus to be transmitted; (4) for each Infective agent and the rest of Susceptible agents included in its geographical area (not part of its social network), it applies the probability p_2 that they will be physically close for the virus to be transmitted; (5) for the set of agents physically close obtained from steps (3) and (4), it applies the *between hosts transmission probability* to go from Susceptible to Exposed; (6) for the agents that are already in the Exposed or Infective state of the disease model, it applies the corresponding progression; and at last (7) it removes from the simulation all agents that have reached the Removed state.

III. Experiments: The case of H1N1 in Mexico

In case of a pandemic, the World Health Organization (WHO) recommends authoritative bodies to consider the suspension of activities in educational, government and business units as a measure to reduce the transmission of the disease. The actions implemented by the Mexican government to control the H1N1 flu outbreak of April 2009 constitute an illustrative example. The actions consisted in three stages: (a) a medical alert issued on Thursday, April 16th, which was triggered by the diagnosis of the first H1N1 flu cases; followed by (b) the closing of schools and universities, enacted from Monday April 27th through Thursday, April 30th; and (c) the suspension of all non essential activities, implemented from Friday, May 1st to Tuesday, May 5th.

Period	Date Range	Description
preflu	1/1 – 16/4	Period before any H1N1 case has been discovered. Agents will move largely unaffected and showing their usual mobility patterns.
alert	17/4 – 26/4	April 16th - Diagnosis of H1N1 cases and medical alert triggered the following day. People may be reacting to the news and modify their usual mobility patterns.
closed	27/4 – 31/4	Schools and Universities closed. Normal behavior disrupted as people change their usual mobility patterns.
shutdown	1/5 – 5/5	Closure of all non-essential activities.
reopened	6/5 – 31/5	Restrictions lifted.

TABLE I. Time periods of study.

Experimental Setting

In order to examine the impact of government restrictions we evaluate changes in the mobility and disease models in five chronological periods. Table I presents the timeline under study. We generate agents (with corresponding mobility and social models) for each of these time periods. In order to measure behavioral changes, we define two scenarios: a *baseline* scenario and an *intervention* scenario. The *baseline* scenario is built using the mobility and social models obtained during the pre-flu period, when individuals show normal – not affected by medical alerts – mobility behavior. The *intervention* scenario considers the models that are built with data from the alert, closed, shutdown and reopened periods. In this case, depending on the moment of the simulation, the DES will jump from one set of models to the next. The evaluation is done by comparing the results obtained by both scenarios. Due to the inherent randomness of the spreading process we run each scenario 10 times and average the results.

Generation of Agents

We collected CDRs from January 1^{st} to May 31^{st} of 2009 of one of the most affected Mexican cities. The dataset contains 1 billion CDRs and 2.4 million unique cell phone numbers. Each cell phone number is associated with one agent and we compute the mobility, social and disease models for both the *baseline* and the *intervention* scenarios. The mobility models are computed with a granularity of one hour. Following Song *et al.* [3], we only consider the agents that (1) are assigned to at least two BTSs; (2) have a minimum average calling rate of 0.25 calls/hour; and (3) have at least 20% of the hourly time slots filled. These requirements narrow down the final number of agents to $25,000$.

We also build the social network models for the *baseline* and the *intervention* scenarios. As part of these models, we needed t define values for the contact probabilities p_1 and p_2. In orde compute their values, we make use of the work by Cruz-Pac *et al.* [4], where the authors examined the effect of the govern intervention measures on the epidemic spread using SIR. De can be found in [1]. Our search determined that the best value were $p_1 = 0.9$ and $p_2 = 0.1$.

To build each agent's disease model, we use the parameters reported in the literature related to the H1N1 outbreak: $R_0 = 1.75$ (Estimated Reproduction number), $\epsilon = 26.4^{-1}$ hours (Expected duration latent period), $\beta = 60^{-1}$ hours, (Expected duration infectious period) and $\rho = 34^{-1}$ hours (Expected time before infecting another agent). We initialize our simulations with one infected agent on April 17th (the first day a case was detected) [4] and run the simulation for 30 days.

Analysis of the Results

In this Section, we compare the results of the *intervention* scenario with the *baseline* scenario from a mobility perspective and from a disease model perspective.

Agent Mobility: In order to measure the changes in mobility due to government mandates, we computed for each scenario the percentage of agents that moved from one BTS coverage area to another one at each step of the simulation (1 step = 1 hour). Figure 2 shows the results.

Both the *baseline* and the *intervention* plots show similar cyclical changes. However, there are a number of important differences. There is a significant decrease in mobility on April 27^{th}, precisely when the *alert* period finishes and the *close* period starts. This decrease in mobility continues until the beginning of the *shutdown* period. On May 1^{st} and throughout the *shutdown* period, there is an even larger decrease in mobility ($< 30\%$) that lasts until all restrictions are lifted on May 6^{th}. To sum up, during the *intervention* scenario there is a reduction in the mobility of the agents of 10% during the alert period and of up to 30% during the closing and shutdown periods, when compared to the baseline. These differences in the agents' mobility disappear once the *reopen* period starts (from May 6^{th} onwards).

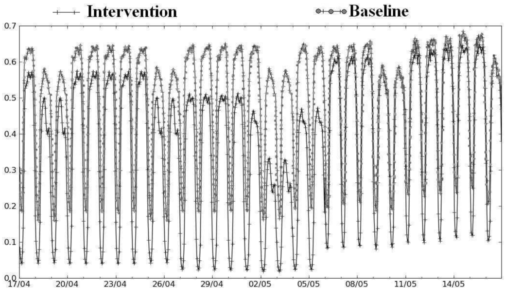


Fig. 2. Percentage of agents that move between BTSs for the *intervention* and *baseline* scenarios. The temporal granularity is 1 hour.

Disease Transmission: In this section we study the evolution the disease focusing on the number of susceptible and infected agents in the *intervention* and *baseline* simulations.

Figure 3 displays the percentage of the population that is in the susceptible stage of the disease model for a specific date and time. Results are shown for both the *intervention* and the *baseline* scenarios. We observe that at the beginning of the simulation all agents are susceptible of being infected. As time passes, the evolution of susceptible agents is described by a sigmoid function. The number of susceptible agents decreases faster in the *baseline* scenario, *i.e.* the number of infected agents grows faster than in the *intervention* scenario. This result supports the hypothesis that the government measures taken during the *intervention* scenario had an impact on the agents' mobility patterns and hence managed to reduce the number of infected agents when compared to the *baseline* scenario. The largest difference between both sigmoid functions takes place during the peak of the epidemic, with approximately a 10% less of susceptible agents in the *intervention* scenario.

Figure 4 shows the percentage of infected agents during the simulation for both scenarios. We observe that the peak of the epidemic in the *intervention* scenario happens *later* in time than in the *baseline*, and has a *smaller* absolute value. The reduction in mobility and the closure of public buildings delayed the peak of the epidemic by 40 hours. Also, in our simulations, the total number of infected agents was reduced by 10% in the peak of the epidemic in the *intervention* scenario when compared to the *baseline* scenario. These results are in agreement with the ones reported in [4].

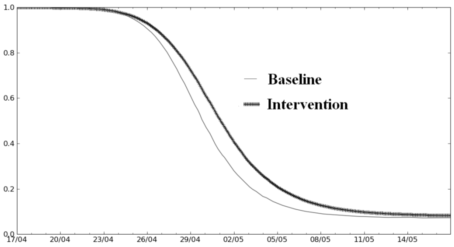


Fig. 3. Fraction of susceptible agents in the population over time. These curves are an average of all simulation runs.

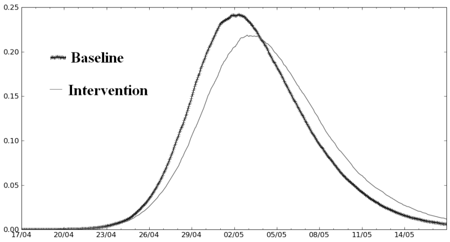


Fig. 4. Fraction of infected agents over time. These curves are an average of all simulation runs.

References

[1] E. Frias-Martinez, G. Williamson, and V. Frias-Martinez. An agent-based model of epidemic spread using human mobility and social network information. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pages 57–64. IEEE, 2011.

[2] Christopher L. Barrett, Keith R. Bisset, Stephen G. Eubank, Xizhou Feng, and Madhav V. Marathe. EpiSimdemics: an efficient algorithm for simulating the spread of infectious disease over large realistic social networks. In *SC'08: Proceedings of the 2008 ACM/IEEE conference on Supercomputing*, 2008.

[3] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–21, 2010.

[4] G Cruz-Pacheco, L Duran, L Esteva, A A Minzoni, M López-Cervantes, P Panayotaros, A Ahued Ortega, and I Villaseñor Ruíz. Modelling of the influenza A(H1N1)V outbreak in Mexico City, April-May 2009, with control sanitary measures. *Eurosurveillance*, 14(26), 2009.

Network-behavior Dynamics in a Medium Size Mobile Phone Network

Cheng Wang	David S. Hachen
University of California, Irvine	Unviersity of Notre Dame
Department of Public Health	Department of Sociology
chengw5@uci.edu	dhachen@nd.edu

Abstract

The stochastic actor-based model (SABM) has been widely used to explore the co-evolution of friendship networks and behavior dynamics simultaneously in recent years. This study analyzes data from the Netsense project, a longitudinal survey and smartphone data collection of 196 college students over 4 semesters since 2011. And we find that selection effects play an important and consistent role in creating peer clusters with similar political tastes in a dynamic context, but friends were not found to influence political tastes, net of other sociodemographic, network, or family factors. In case of obesity contagion both selection and influence effect are detected: friendships are more likely to be found among those with similar body mass index (BMI) and friendships also drive those involving in them to have similar BMIs. We also study the network-behavior dynamics of academic achievement, smoking behavior, drinking behavior, depressive feeling, and religious preferences.

On Information Propagation in Mobile Call Networks

Kirill Dyagilev KIRILLD@TX.TECHNION.AC.IL
Shie Mannor SHIE@EE.TECHNION.AC.IL
EE department, Technion, Haifa, Israel

Elad Yom-Tov ELADYT@MICROSOFT.COM
Microsoft Research, Herzliya, Israel

In the last decade the accessibility of large-scale social interaction data has led to an explosion of research in the field of social networks (Jackson, 2008).

In particular, a substantial body of work has focused on telecommunication data, e.g., (Eagle et al., 2009; Hill et al., 2006; Nanavati et al., 2006; Richter et al., 2010). A common approach in this research is to rely on a static **call graph** constructed from aggregates of calls given in **Call Data Records (CDRs)**, which record, among other details, the calling and called numbers of the subscribers. This graph contains nodes that represent different subscribers and edges that connect subscribers that have social relation. The existence of a meaningful social connection between users A and B is assumed if the total interaction of calls between these two subscribers over some period of time is significant. The volume of calls can be measured, e.g., by the total number of calls over a designated period, the total duration of calls, or a fraction of calls made by user A to user B out of calls made by user A. The edges in the call graph can be directed and undirected depending on the context. It was shown in Eagle et al. (2009) that this graph accurately approximates the actual social network.

However, this approach disregards the information encapsulated in the dynamics of the interactions. In contrast, we focus on the actual sequences of information-passing events.

We propose a method to identify sequences of calls that are likely to be related to the same topic or propagate the same information. It is evident that without knowledge of the actual content of these calls, this task is impossible. Therefore we heuristically confine our search to a specific mode of information diffusion in which once the information is received, it is either transferred to somebody else during a relatively short period of time or not transferred to anyone. We refer to this mode of information propagation as **Rapid Propagation of Information (RPI)**. As an illustration of RPI, one can consider the following two sequences of information diffusion calls:
Scenario 1. The sequence begins with Alice calling Bob. Once their conversation is over, Bob immediately calls Clare. Shortly after the end of the second call, Clare dials Daniel and so on. Assuming that this chain of calls is long enough and that time intervals between consecutive calls are short, this sequence represents a rare event. In this scenario, calls seem to trigger other calls, hence could be related to the same topic.
Scenario 2. In this scenario, Alice makes calls to all of her friends within a relatively short period of time. Our measurements show that such a "burst" of calls is a very unusual subscriber behavior that might suggest that Alice transfers the same piece of information to all of her friends, e.g., an invitation to a party.

We design an algorithm for identification of RPIs in the call data records and apply this method to data sets **DS-1** and **DS-2** from two providers from different parts of the world. Data set DS-1 contained all calls logged by an operator in a country with population of over 7 million people in a period of 35 days. The data contains more than 600 million calls involving approximately 3 million distinct subscribers from the analyzed operator and over 5 million subscribers belonging to other service providers. Data set DS-2 contained calls in a city with a population of over 2 million people in 24 days out of period. The data contains more than 50 million calls involving 5.4 million distinct subscribers, out which approximately 2.1 million belong to the analyzed operator.

We study the properties of RPIs found in these data sets and compare them. In particular, we consider the properties of **information flow trees**, namely, the paths in which information propagates from one subscriber to another within RPI. We show that the typical topologies of these trees are the same in both data sets (see Figure 1), indicating that these topologies represent universal modes of information propagation. However, the fractions of RPIs belonging to each

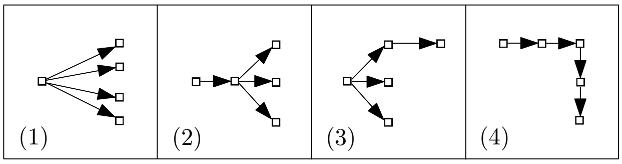


Figure 1. Typical topologies of information flow trees. Arrows point from parent node to child node.

of these topologies differs between two data sets. We conjecture that this difference can be attributed to the difference in the type of clients (pre-paid vs. post-paid customers, private vs. business clients). We further show that a significant fraction of RPIs have a single subscriber propagating the information to majority of other users. We refer to these users as "dissemination-leaders" and conjecture that they play a significant role in diffusion of information in the mobile phone network.

We develop two generative models of RPIs that have a relatively small number of parameters. The first model generates sequences of calls that yield an RPI. The second model describes the emergence of different topologies of information flow trees. We show that these models accurately predict features of RPIs found in the real-world data. The simplicity and locality of these models allows them to be a computationally simple way to generate synthetic call data records that capture the RPI phenomenon.

We justify our heuristic definition of RPI by providing evidence that some of the RPIs propagate geo-spatial information. In particular, we show that appearance in the same RPI increases the chances that two subscribers will visit the same geographical location during the same day.

Finally, we consider the problem of **churn prediction**, namely, identification of subscribers that are considering transferring their business to a competing mobile phone service provider. Since it is usually significantly cheaper to retain an existing subscriber than to acquire a new one, churn prediction has become a central business intelligence application for telecommunication operators (Fildes & Kumar, 2002).

The mainstream approach to churn prediction (e.g., Coussement & Van den Poel (2008); Datta et al. (2000)) considers each customer separately. Essentially, each subscriber is characterized by a numerous features based on their socio-economic characteristics and call behavior. These features are then used by some learning regression algorithm to calculate subscriber's likelihood to churn.

One drawback of this approach is fairly obvious: it relies on the assumption that the decision to churn is made by each user individually and is not affected by a subscriber's social circle. However, it is well-known that there are social aspects to churn. E.g., Nitzan & Libai (2010) considered a social network of mobile phone subscribers and showed that having a neighbor that churned increases the chances of churn by 80%. Thus, it seems that leveraging social relations may lead to churn prediction systems with better performance. While commercial solutions have started exploring this direction, to the best of our knowledge the only published work directly in this context is the diffusion-based algorithm introduced in Dasgupta et al. (2008) and machine-learning based algorithm in Richter et al. (2010).

Dasgupta et al. (2008) introduced the **Spreading Activation algorithm (SPA)**. Its underlying assumption is that recent churners are known and they are likely to affect the churning decisions of their social neighborhood. The network of subscribers is then modeled as a weighted directed call graph. Next, a *diffusion process* is used to model the flow of churn propensity from recent churners to their social environment. Specifically, each node in the call graph that corresponds to a recent churner is assigned an initial weight which propagates to the call graph network according to a decaying diffusion process. Once diffusion process converges, each subscriber in the call graph has some associated weight corresponding to the amount of churn propensity that has reached him. The individual **churn scores**, namely, the likelihood of a user to churn, are then derived directly from these weights.

The method proposed by Richter et al. (2010) is called Group-First Churn Prediction (GFCP) and proceeds in the following manner. It employs an information-theory based measure to quantify the strength of the social connection between pairs of subscribers. By keeping only the strongest connection, this algorithm identifies closely-knit groups of subscribers and the leader of each group. GFCP then establishes a churn score for each group using a novel group Key Performance Indicators. It finally assigns individual churn score to each subscriber based on the corresponding group churn score and subscriber's personal characteristics.

We propose a new churn prediction method that relies on subscribers' behaviors in RPIs. We thus introduce *the first algorithm for churn prediction that is based on subscribers' dynamic, rather than static, social behavior.* Our algorithm proceeds through the following steps. It receives call data records (CDRs) of all calls made over an **input period** of several days, along with identities of all users that churned during this period. Our goal is to accurately identify users that will churn in the **prediction period** that follows the input period.

As the first step, we identify RPIs in the input period data. We then characterize all subscribers that belong to the analyzed carrier with attributes of their behavior in RPIs. E.g., we count the number of RPIs the user participated in, whether the user had a central location in one or more information flow tree, the number of churners that participated in the same RPIs etc.

As the second step, we use a pre-trained regression-classification tree (Duda et al., 2001) to map between the values of these features of the user and the churn score. This churn score indicates the likelihood of a user to churn during the prediction period. The regression-classification tree mentioned above is trained using historical data. Namely, we designate input and prediction periods in data from the past. We use CDRs from the input period to extract subscribers' features and use the $0-1$ churn signal observed during the prediction period as the desired output of the trained tree. The regression-classification tree is trained once and used in all future predictions.

We compared the performance of our algorithm (RPI-CP) to the performance of the SPA algorithm that assumes knowledge of similar information. We show that the RPI-CP algorithm outperforms the SPA algorithm on both data sets. In particular, we take the top 1% of subscribers, as ranked most likely to churn by each algorithm, and we check the actual number of churners in each set. In DS-1, the RPI-CP algorithm finds *five times* more actual churners than the SPA algorithm. In DS-2, the RPI-CP algorithm finds *1.5 times* more actual churners than the SPA algorithm.

Further work of immediate interest includes leveraging geographical information to get additional insights on the social structure of mobile call networks. We also note that the actual content of calls was unavailable to us, therefore our research was solely based on temporal features of calls. However, there exist other social media, e.g. Twitter, in which both temporal and content data are available. The extension of our approach to these media can provide us with additional insights to dynamics of RPI.

Acknowledgments

The full version of this paper can be found in Dyagilev et al. (to appear). Kirill Dyagilev and Elad Yom-Tov were also with IBM Haifa Research Lab, Haifa, Israel during part of this research.

References

Coussement, K. and Van den Poel, D. Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. *Expert systems with applications*, 34(1):313–327, 2008. ISSN 0957-4174.

Dasgupta, K., Singh, R., Viswanathan, B., Chakraborty, D., Mukherjea, S., Nanavati, A.A., and Joshi, A. Social ties and their relevance to churn in mobile telecom networks. In *EDBT'08*, 2008.

Datta, P., Masand, B., Mani, DR, and Li, B. Automated cellular modeling and prediction on a large scale. *Artificial Intelligence Review*, 14(6):485–502, 2000. ISSN 0269-2821.

Duda, R.O., Hart, P.E., and Stork, D.G. *Pattern classification.* Wiley New York, 2001.

Dyagilev, K., Mannor, S., and Yom-Tov, E. On Information Propagation in Mobile Call Networks. *Social Network Analysis and Mining*, to appear.

Eagle, N., Pentland, A., and Lazer, D. Inferring social network structure using mobile phone data. *PNAS*, 106(36), 2009.

Fildes, R. and Kumar, V. Telecommunications demand forecasting–a review. *International Journal of Forecasting*, 18(4):489–522, 2002. ISSN 0169-2070.

Hill, S., Provost, F., and Volinsky, C. Network-based marketing: Identifying likely adopters via consumer networks. *Statistical Science*, pp. 256–276, 2006.

Jackson, M.O. *Social and Economic Networks.* Princeton University Press, Princeton and Oxford, 2008.

Nanavati, A.A., Gurumurthy, S., Das, G., Chakraborty, D., Dasgupta, K., Mukherjea, S., and Joshi, A. On the structural properties of massive telecom call graphs: findings and implications. In *ICIKM'06*, 2006.

Nitzan, I. and Libai, B. Social effects on customer retention. Marketing Science Institute, working paper 10-107, 2010.

Richter, Y., Yom-Tov, E., and Slonim, N. Predicting customer churn in mobile networks through analysis of social groups. In *ICDM'10*, 2010.

Using telephone network data to study the effects of socio demographics on ego network structure

by

Binh Phan[1], Kenth Engø-Monsen[2], and Øystein D. Fjeldstad[1]

Abstract

The effects of socio demographics on ego network structure have captured much research attention in the 1980s and 1990s (such as, Beggs, Haines and Hurlbert (1996), Marsden (1987)). The network data in these studies was collected using interviews (survey). Extant research has shown issues associated with such a network data. Specifically, people face memory constraints which make them report much less partners than they actually have (Marsden, 1993, 2003), and are biased toward recent events, physically proximate and close partners (Eagle, Pentland, & Lazer, 2009).

With the support of technology, recent social network studies use automatically recorded communication networks, such as an email correspondence network (Ebel, Mielsch, & Bornholdt, 2002; Guimera, Danon, Díaz-Guilera, Giralt, & Arenas, 2003), and a telephone network (Eagle et al., 2009; Lambiotte et al., 2008; Onnela et al., 2007), as reliable proxies for real social networks. The use of these networks partly overcomes the issues associated with the survey network data. Among recorded communication networks, the telephone network, including fixed-line call, mobile phone call and SMS networks, is considered to be a better proxy for the real social network (Hidalgo & Rodriguez-Sickert, 2008; Onnela et al., 2007; Pool & Kochen, 1978) because the mobile phone call service has been used as one of the main communication mediums in societies over decades (Kwan, 2007), and has the high penetration rate, almost 100 percent in developed countries and more than 90 percent in fast-growing countries.[3]

In this study, we examine the effects of socio demographics with a sample of 509 anonymous mobile users which was randomly collected in Norway in 2010. We use the telephone network as a proxy for the social network, and Ordinary Least Squared (OLS) estimation. This network data consists of 12529 nodes and 17737 ties. We found that (i) age has a U-shaped effect on ego network clustering and an inverse U-shaped effect on ego network size, (ii) geographic centrality has a negative effect on ego network clustering but no effect on ego network size.

Our findings, although generally consistent with prior research using survey data, nuances prior findings. The effects of age and geographic location have been examined in the studies of Marsden

[1] Department of Strategy and Logistics, BI Norwegian Business School, 0484 Oslo, Norway. Emails: tbinhphan@yahoo.com and oystein.fjeldstad@bi.no
[2] Telenor Group, Research and Future Studies, Snarøyveien 30, N-1331 Fornebu, Norway. Email: Kenth.Engo-Monsen@telenor.com
[3] According to International Telecommunication Union (ITU):http://www.itu.int/ITU-D/ict/facts/2011/index.html

(1987) and Beggs et al. (1996). These studies used the 1985 General Social Survey (GSS)'s network data in the US. In this data, interviewees were asked to give the names of five or more people whom they discusses "important matters" with and provide the information about the connections among partners. In this data, the average size and clustering of ego networks are 3.01 and 0.61 respectively (Marsden, 1987). In our data, the average size and clustering of ego networks are 23.48 and 0.37 respectively. Marsden (1987) found that age has a positive effect on ego network clustering and has a inverse U-shaped effect on ego network size. Marsden (1987)'s finding suggests that at early ages, people have small and sparse ego networks. This is probably because in the data, nearly a quarter of respondents report to have ego networks of size 0 or 1 (Marsden, 1987). Such a small size is not adequate to construct ego network clustering. In our data, all ego networks have a size of at least 5 alters, which is adequate to capture the true ego network clustering (Marsden, 1993). With a different type of data, our findings extend the findings of Marsden (1987) by showing that at early ages, people have small and clustered ego networks.

Further, our findings and Beggs et al. (1996)'s findings are consistent on a negative effect of geographic centrality on ego network clustering. Beggs et al. (1996) found a significantly positive effect of geographic centrality on ego network size while we find no support. This difference is conceivably due to the different time points of data collection and the different types of network data used. Networking is costly (Burt, 1992). People face a physical constraint on networking, especially with geographically distant partners. The cost of networking and the physical constraints limit the number of ties that people can maintain (Burt, 1992: 17). Communication technologies reduce the cost of networking and the physical constraint (Wellman, Haase, Witte, & Hampton, 2001). Beggs et al. (1996) used the GSS data collected in 1985, when the gap of the use of communication technologies was significant between rural and urban areas. The use of communication services, especially mobile services, in rural areas has been recently dramatically growing. This gap in developed countries remains, but has been significantly mitigated over time (LaRose, Gregg, Strover, Straubhaar, & Carpenter, 2007). The significant mitigation in the gap would reduce the difference in ego network size between urban and rural people. The use of the telephone network in this study is able to capture the effect of this gap mitigation. Moreover, the individual participation to social and non-economic organizations and associations where people can establish and maintain non-work ties has been shown to decline over the past 40 years (Putnam, 2010). This decline is especially high in urban areas where people face a large pressure of money and work (Gellis, Kim, & Hwang, 2004). This high decline in urban areas significantly reduces the number of non-work ties in the ego networks of people in these areas. People in rural areas with a less pressure of money and work can remain a high participation to these non-economic organizations and associations, and therefore maintain a significant number of non-work ties. The combination of these two reasons may explain why a positive effect was found to be significant with the data collected in 1985, but is non-significant with the data collected in 2010.

In summary, our findings imply that the use of the telephone network data as a proxy for the social network in the research on the antecedents of social network structure may extend the findings of prior studies using the survey network data, and is also able to capture the effects of communication technologies on people's networking and network structure. We call for future research to use the telephone network data to reexamine the origins and effects of social network structure.

References:

Beggs, J. J., Haines, V. A., & Hurlbert, J. S. 1996. Revisiting the Rural-Urban Contrast: Personal Networks in Nonmetropolitan and Metropolitan Settings1. ***Rural Sociology***, 61(2): 306-325.

Burt, R. S. 1992. ***Structural Holes***: Cambridge: MA: Harvard University Press.

Eagle, N., Pentland, A., & Lazer, D. 2009. Inferring Friendship Network Structure by Using Mobile Phone Data. ***Proceedings of the National Academy of Sciences of the United States of America***, 106(36): 15274-15278.

Ebel, H., Mielsch, L.-I., & Bornholdt, S. 2002. Scale-Free Topology of E-Mail Networks. ***Physical Review E***, 66(3): 035103.

Gellis, Z., Kim, J., & Hwang, S. 2004. New York State Case Manager Survey: Urban and Rural Differences in Job Activities, Job Stress, and Job Satisfaction. ***The Journal of Behavioral Health Services & Research***, 31(4): 430-440.

Guimera, R., Danon, L., Díaz-Guilera, A., Giralt, F., & Arenas, A. 2003. Self-Similar Community Structure in a Network of Human Interactions. ***Physical Review E***, 68(6): 065103.

Hidalgo, C. A. & Rodriguez-Sickert, C. 2008. The Dynamics of a Mobile Phone Network. ***Physica A: Statistical Mechanics and its Applications***, 387(12): 3017-3024.

Kwan, M.-P. 2007. Mobile Communications, Social Networks, and Urban Travel: Hypertext as a New Metaphor for Conceptualizing Spatial Interaction*. ***The Professional Geographer***, 59(4): 434-446.

Lambiotte, R., Blondel, V. D., de Kerchove, C., Huens, E., Prieur, C., Smoreda, Z., & Van Dooren, P. 2008. Geographical Dispersal of Mobile Communication Networks. ***Physica A: Statistical Mechanics and its Applications***, 387(21): 5317-5325.

LaRose, R., Gregg, J. L., Strover, S., Straubhaar, J., & Carpenter, S. 2007. Closing the Rural Broadband Gap: Promoting Adoption of the Internet in Rural America. ***Telecommunications Policy***, 31(6–7): 359-373.

Marsden, P. V. 1987. Core Discussion Networks of Americans. ***American Sociological Review***, 52(1): 122-131.

Marsden, P. V. 1993. The Reliability of Network Density and Composition Measures. ***Social Networks***, 15(4): 399-421.

Marsden, P. V. 2003. Interviewer Effects in Measuring Network Size Using a Single Name Generator. ***Social Networks***, 25(1): 1-16.

Onnela, J.-P., Saramaki, J., Hyvonen, J., Szabo, G., de Menezes, M. A., Kaski, K., Barabasi, A.-L., & Kertesz, J. 2007. Analysis of a Large-Scale Weighted Network of One-to-One Human Communication. ***New Journal of Physics***, 9: 27.

Pool, I. d. S. & Kochen, M. 1978. Contacts and Influence. ***Social Networks***, 1(1): 5-51.

Putnam, R. D. 2010. ***Bowling Alone: The Collapse and Revival of American Community***. New York: Simon & Schuster Inc.

Wellman, B., Haase, A. Q., Witte, J., & Hampton, K. 2001. Does the Internet Increase, Decrease, or Supplement Social Capital?: Social Networks, Participation, and Community Commitment. ***American Behavioral Scientist***, 45(3): 436-455.

Understanding Social Influence Using Combined Network Analysis and Machine Learning Models

Dhaval Adjodah, Alex ‘Sandy’ Pentland

dhaval@mit.edu, sandy@media.mit.edu

MIT Media Lab, Cambridge, MA

Abstract:

In this paper, the drivers of social influence in a social network during a phone-tracked intervention will be investigated. The funf dataset[1], which comprises detailed high-frequency data gathered from 25 mobile phone-based signals from 130 people over a period of 15 months, will be used to test the hypothesis that people who are closer to each other have a greater ability to influence each other. Various metrics of closeness will be investigated such as self-reported friendships (collected through surveys), call logs and Bluetooth co-location signals. A graph of all participants is then created using such metrics as edge weights, and the network constraint[2] of each pair of participants is calculated as a measure of not only the direct friendship (or number of calls or Bluetooth co-locations) between two participants but also the indirect friendships through intermediate connections that form closed triads with both the participants being assessed. To measure influence, the results of the live funf intervention will be used where behavior change of each participant to be more physically active was rewarded - the reward being calculated live. There were three variants of the reward structure: one where each participant was rewarded for her own behavior change without seeing that of anybody else (the control), one where each participant was paired up with two 'buddies' whose behavior change she could see live but she was still rewarded based on her own behavior, and one where each participant who was paired with two others was paid based on their behavior change that she could see live. As a metric for social influence, it will be considered how the change in slope and average physical activity levels of one person follows the change in slope and average physical activity levels of the buddy who saw her data and/or was rewarded based on her performance. Finally, a linear regression model that uses the various types the network constraints (self-reported friendship, call logs, Bluetooth) will be created to predict the behavior change of one participant based on her closeness with her buddy. In addition to explaining and demonstrating the causes of social influence with unprecedented detail, this paper will also briefly discuss the policy implications of this technology such as privacy, moral hazard, misuse and effectiveness in the long term.

[1] Aharony, N., Pan, W., Ip, C., Khayal, I., Pentland, A., & Nadav Aharony, Wei Pan, Cory Ip, Inas Khayal, A. P. (2011). Social fMRI: Investigating and shaping social mechanisms in the real world. *Pervasive and Mobile Computing*, *7*(6), 643–659. doi:10.1016/j.pmcj.2011.09.004

[2] Burt, Ronald S. 1992. Structural Holes: The Social Structure of Competition. Cambridge, MA: Harvard University Press.

Accelerating internet growth in Asia using viral spreading

Pål Sundsøy[1a], Johannes Bjelland[1b], Geoffrey Canright[1c], Kenth Engø-Monsen[1d], Asif M.Iqbal[1e], David Lazer[23f]

[1] Telenor ASA, Research & Future Studies, 1360 Fornebu, Norway
[2] Harvard University, John F. Kennedy School of Government, Cambridge, MA 02138
[3] NorthEastern University, Lazer Lab, Boston MA 02115

E-mails: [a]pal-roe.sundsoy@telenor.com, [b]johannes.bjelland@telenor.com, [c]geoffrey.canright@telenor.com , [d]kenth.engo-monsen@telenor.com, [e] asifmiqbal@outlook.com, [f]david_lazer@harvard.edu

For many in Asia, the mobile phone is their only gateway to the Web. The penetration of internet in these countries is nevertheless very small, causing a large digital discrepancy between more and less developed countries - the so called digital divide. Increased mobile internet adoption may help to bridge this gap.

In this talk we examine how viral campaigns can boost the mobile internet adoption in one developing Asian country. By sending a unique offer code to 70 000 people, forwardable to friends, we are able to observe how the offer spreads from customer to customer. The spreading can be observed by coupling adoption data with call data records, which can provide a good proxy for the social network. By counting both direct and indirect hits we observe a strong adoption rate of 53%. On average we find that each person recruits 8.3 other customers, while the most extreme people recruit over 350 others.

We also introduce molecular targeting - where we aim to reach socially connected pairs of people. Our results indicate that individuals in these ‘molecules’ adopt more often together than expected from the single-individual hit rate - leading us to believe that the neighbours boost each other’s awareness of mobile internet, due to the so called “up-in-air effect”. Early results also indicate that this approach increases the overall spreading as measured by the indirect hitrates.

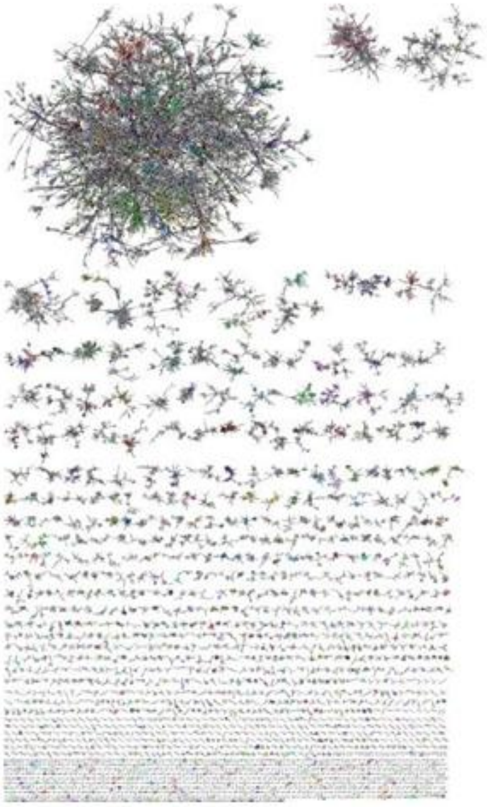


Fig 1 The social network among the mobile internet adopters 5 days after campaign launch. 45% of the connected adopters can be found in the largest connected component. The connections are based on voice + SMS communication.

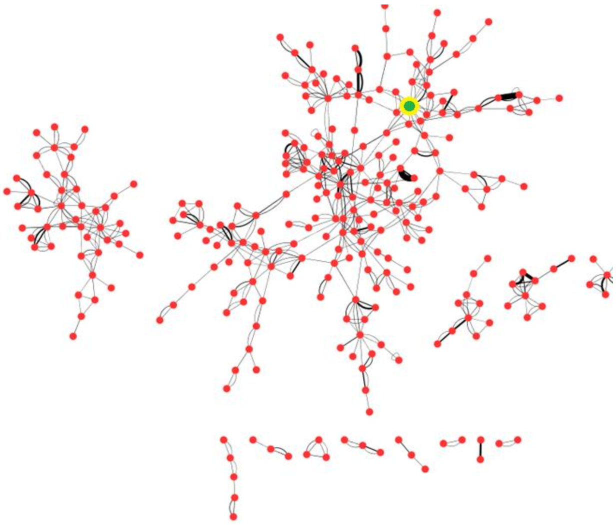


Fig 2 The extreme case where one single customer (in green) recruits 360 other customers. The missing links indicate that the offer is also spread via other channels (which our call data does not pick up)

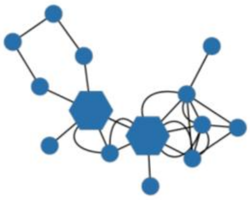


Fig 3 Molecular targeting: A component where two customers (hexagons) are targeted with an offer. They both adopt, and also recruit 12 other customers

Determinants of Subscriber Churn in Wireless Networks: The Role of Peer Influence*

[Extended Abstract]

Qiwei Han*†
qiweih@cmu.edu

Pedro Ferreira*‡
pedrof@cmu.edu

*Department of Engineering and Public Policy
Carnegie Mellon University
Pittsburgh, PA 15213
United States

‡Heinz College
Carnegie Mellon University
Pittsburgh, PA 15213
United States

†Instituto Superior Técnico
Universidade Técnica de Lisboa
Lisbon, 1049-001
Portugal

ABSTRACT

Subscriber churn is a top challenge for wireless carriers. Understanding the determinants of churn is key for carriers to identify potential churners and apply effective retention strategies to reduce subscriber loss. In this paper, we apply generalized propensity score matching to separate peer influence from other confounding factors that might affect churn. Our empirical analysis, developed over a large scale wireless network, confirms that peer influence plays a role in churn. The estimated marginal influence of having a first friend churn is roughly 3%. While the marginal effect of friends' churn decreases significantly as more friends do so, contagious churn is still a significant part of the story beyond high churn rates in the mobile industry.

1. INTRODUCTION

In today's competitive wireless industry, subscriber churn is considered to be the "biggest issue for all wireless carriers" [7]. Preserving the existing subscriber base is of crucial importance for carriers to ensure their profitability. Understanding determinants of churn becomes fundamental for carriers so that they can identify potential churners and apply appropriate retention strategies to reduce subscriber loss. However, the complex nature of churn poses significant challenges to carriers that pursue effective churn management solutions to deal with all kinds of churn problems. As a consequence, most carriers focus only on retaining their most valuable subscribers.

Advances in studying the effect of social influence on subscriber churn in wireless networks have received much attention in recent times. [5] found that the likelihood of churn increases with the number of friends who have already churned. [6] also confirmed that the "word-of-mouth" effect has a positive impact on subscriber's churn. However, work that identifies contagious churn on a causal sense and separates it from confounding effects such as homophily still lacks. Correlation in the behavior among people who share social ties can be explained by both peer influence and their inherent similarities [10]. Therefore, misattribution of homophily to contagion, or-and vice versa, needs to be carefully thought from an empirical point of view.

Numerous studies on identification of peer influence in other networked context have been proposed (e.g. [1, 4, 13]). [2] used dynamic propensity score matching (PSM) to estimate the effect of contagion in the adoption of an online service by analyzing a community of instant messenging users. Their findings suggest that homophily accounted for much of the adoption previously perceived as peer influence. However, [2] dichotomize the treatment due to the binary nature of treatment regime. The applicability of PSM therefore is confined, as effects of different numbers of adopter friends are overlapped. To overcome this problem, in this paper, we apply a generalized propensity score matching (GPSM) method to separate peer influence from homophily. We perform our empirical analysis on a massive dataset from a major European wireless carrier (hereafter refer to as EURMO). We have call detail records (CDR) and tariff plan information from EURMO. The GPSM method allows us to estimate the magnitude of the contagion effect given different numbers of friends who churn. This can provide us with more information on the marginal effect of peer influence and thus help us better understand the role of peer influence on churn.

2. DATA

The EURMO dataset includes CDRs for roughly 4 million prepaid subscribers between August 2008 and June 2009. For each call we know the caller and the callee, the duration and time of the call. For each SMS we know the sender and receiver and the time of the SMS. Subscribers are identified by their anonymized phone number. For each subscriber, we know their tariff plan at all times. Understanding subscriber churn for prepaid consumers is quite different from postpaid subscribers. First, we have little demographic information on prepaid subscribers. Second, the usage pattern of prepaid subscribers is more irregular than that of postpaid subscribers. Third, prepaid subscribers churn by ceasing usage whereas postpaid subscribers explicitly inform the carrier when they want to do so. After consulting with the carrier, we use its definition of churn and thus assume that a prepaid subscriber churns if she does not place a call or sends a SMS for three consecutive months.

*Support for this research was provided by the Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) through the Carnegie Mellon Portugal Program under Grant No. *SFRH/BD/51153/2010.*

Variable	Description	MD_C	SD_C	MD_NC	SD_NC
Time Invariant Individual					
PLAN_ID	The ID for the tariff plan	1.09	1.15	1.35	1.95
Time Variant Individual					
#CALL	Number of calls made or received per day	0.24	2.29	1.41	3.16
AIRTIME	Duration of calls made or received (in min)	0.22	4.68	1.59	7.32
#NEIGHBOR	Number of friends	10	46.37	64	104.40
#SMS	Number of SMS sent or received	0.012	4.51	0.30	20.14
LIFETIME	Duration since subscription to carrier (in month)	3.67	12.08	15.37	18.58
RCO	Ratio of calls to other networks	0.2	0.31	0.15	0.23

Table 1: List of covariates extracted from EURMO, MD is median and SD is standard deviation, C stands for churner and NC stands for non-churner

We use a random sample of 10,000 subscribers together with their 690,000 friends (430,000 in the same network). Two subscribers are called friends if at least they exchange one call in the same calendar month. We observe network dynamics: every month new subscribers join EURMO, existing subscribers leave EURMO and subscribers call and/or text different friends. Therefore, we aggregate time-varying individual subscriber usage and time-invariant characteristics at the monthly level (Table 1). Over the eleven months in our period of analysis, the 10,000 subscribers in our sample placed 6.5 million calls. 2,282 of them left EURMO, which amounts to an average monthly churn rate of 2.07%.

We find that the subscribers that churn have much less usage and fewer friends than the subscribers who do not both in terms of number of calls and airtime. Moreover, we also observe that both subscribers who churn and do not have much more usage within the network. This is sensible because calls across carriers cost more as carriers pass on to subscribers part of the interconnection charges. We also find that the conditional churn probability decreases with the subscriber's lifetime with the carrier. One possible explanation can be that subscribers become loyalty to carriers over time. We also note that subscribers exhibit a significantly higher churning rate during the first three months they sign up with the carrier. This implies that carriers should pay particularly attention to these subscribers who just join the network and thus build up good customer relationships with them to keep them in the firm, as they become more valuable with time.

3. METHODOLOGY

Propensity Score Matching (PSM) is a widely used method to evaluate the causal treatment effect from observational data in various empirical research fields [11], in particular when the assignment of a binary treatment is not random and counterfactual outcomes are unavailable. With PSM units from a treated group are matched to those in a control group using a propensity score. Differences in the behavior of these pairs of units measure the effect of the treatment. [8] extended and proposed this framework to allow for continuous levels of treatment. Formally, consider a set of N subscribers and let i denote a single subscriber. Let $P = \{1, \dots p\}$ represent a set of time periods. We observe a vector of pre-treatment covariates X_{ip} as shown in Table 1 at each time period. We define the treatment at each period for each subscriber as her exposure to a certain number of friends who churned in the last time period τ_{ip-1}. Very few subscribers in our sample have more than 3 friends who churn. Therefore, we decided to use four levels of treatment: $T \in \{0, 1, 2, 3\}$, to indicate whether 0, 1, 2 or 3 or more friends churn, respectively. The outcome of interest is whether subscriber i churns in time period p: $Y_{ip} \in \{0, 1\}$.

GPSM requires weak unconfoundedness: $Y(t) \perp\!\!\!\perp T|X$. In our case, though the number of friends who churn at time $p-1$ is not randomly assigned, we observed all variables that can affect both the subscriber's churn at time p and the likelihood of receiving treatment (justification of this assumption is discussed in the next section). We estimate the conditional distribution of the number of friends who churn given these covariates to estimate the generalized propensity score (GPS) for each subscriber, R_i (we assume that the logarithm of the number of friends who churn is normally distributed). We also investigate the balancing property for our covariates adjusted by GPS by testing whether the mean of one of the four treatment levels was different from the mean of the other three treatment levels. We generally observe moderate evidence against the balancing properties according to a two-sided t-test.

We denote the dose response function as a set of potential outcomes given the treatment level t: $\{Y_{ip}(t)_{t \in T}\}$ where T is the set of potential treatment values. Then the conditional expectation of churn is a function of number of friends who churn T and of the GPS R:

$$\beta(t, r) = E[Y_{ip}(t)|r(t, X_{ip}) = r] = E[Y_i|T_{ip-1} = t, R_i = r]$$

We use a polynomial approximation of order two to regress the subscribers' churn Y_i on the number of churned friends T_i, and the GPS R_i.

$$Y_i = \alpha_0 + \alpha_1 T_i + \alpha_2 T_i^2 + \alpha_3 R_i + \alpha_4 R_i^2 + \alpha_5 T_i R_i$$

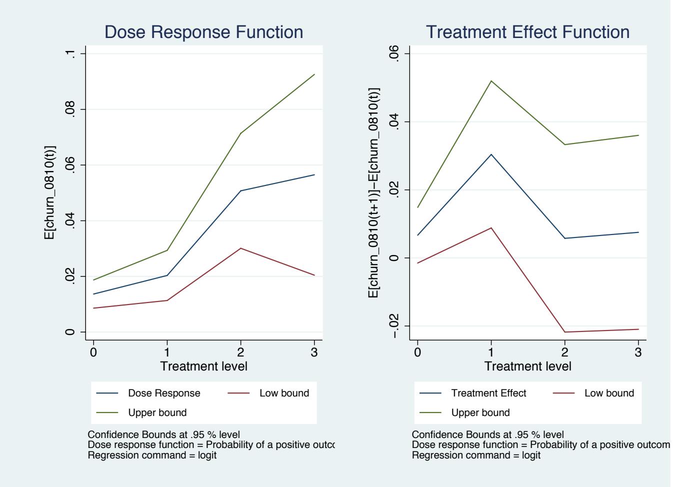
Therefore, the effect of peer influence on churn is the average conditional expectation over GPS at a particular number of friends who churn:

$$\mu(t) = E[\beta(t, r(t, X_i))]$$

Taking derivatives, we can easily obtain the marginal ef associated with one more friend churn on the subscril churn for different levels of treatment.

4. RESULTS AND DISCUSSIONS

For each month we evaluate the dose response function separately (we use the Stata package provided by [3]). We use



Confidence Bounds at .95 % level
Dose response function = Probability of a positive outc
Regression command = logit

Confidence Bounds at .95 % level
Dose response function = Probability of a positive outcom
Regression command = logit

Figure 1: Dose response function and treatment effect function in October 2008

bootstrapping to calculate the asymptotic standard errors and confidence intervals. Figure 1 shows the average dose response and treatment effect for October 2008 (estimates for other months exhibit similar characteristics). The figure on the left shows that the effect on subscribers churn increases with the number of friends who churn. For example, the probability of churn for subscribers who have two friends who churned in September 2008 is 4% higher than that for subscribers who had no friends who churned in September 2008. The figure on the right shows that the marginal influence (the effect of having one more friend who churns) decreases as more friends churn. For example, having one friend who churns compared to having no friends who do so will lead to an increase of 3% in the probability of churn but having two or more friends who churn compared to having one friend who churns will increase the probably of churn by only 1%. Our results confirm the positive effect of peer influence on churn. When we remove the selection bias due to the homophily, we still observe contagious churn.

We notice the argument made by [12] that the plausibility of the unconfoundedness assumption remains unidentifiable from observational data. As long as there are systematic differences in unobserved covariates, we cannot safely conclude that the unconfoundedness assumption holds. Therefore, we acknolwedge that our results may still be biased. As future work, we will perform sensitivity analysis to check the robustness of our results. One possible way to do so is to relax the unconfoundedness assumption, introduce an artificial unobserved variable and reestimate the dose response function to check whether the presence of unobserved heterogeneity may significantly change our results [9].

5. REFERENCES

[1] Anagnostopoulos, A., Kumar, R., and Mahdian, M. Influence and correlation in social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (2008), KDD '08, ACM, pp. 7–15.

[2] Aral, S., Muchnik, L., and Sundarajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of National Academy of Science 106*, 51 (2009), 21544–21549.

[3] Bia, M., and Mattei, A. A stata package for the estimation of the dose-response function through adjustment for the generalized propensity score. *The Stata Journal 8*, 3 (2008), 354–373.

[4] Bramoullé, Y., Djebbari, H., and Fortin, B. Identification of peer effects through social networks. *Journal of Econometrics 150*, 1 (2009), 41–55.

[5] Dasgupta, K., Singh, R., Viswanathan, B., Chakraborty, D., Mukherjea, S., Nanavati, A. A., and Joshi, A. 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* (2008), EDBT '08, ACM, pp. 668–677.

[6] Dierkes, T., Bichler, M., and Krishnan, R. Estimating the effect of word of mouth on churn and cross-buying in the mobile phone market with markov logic networks. *Decision Support Systems 51*, 3 (2011), 361–371.

[7] FCC. Annual report and analysis of competitive market conditions with respect to commercial mobile services. WT Docket 08-27, Federal Communication Commission, 2009.

[8] Hirano, K., and Imbens, G. W. *The Propensity Score with Continuous Treatments*. John Wiley & Sons, Ltd, 2005, pp. 73–84.

[9] Ichino, A., Mealli, F., and Nannicini, T. From temporary help jobs to permanent employment; what can we learn from matching estimators and their sensitivity? *Journal of Applied Econometrics 23* (2008), 305–327.

[10] McPherson, M., Smith-Lovin, L., and Cook, J. Birds of a feather: Homophily in social networks. *Annual Review of Sociology 27* (2001), 415–444.

[11] Rosenbaum, P., and Rubin, R. The central role of the propensity score in observational studies for causal effects. *Biometrika 70*, 1 (1983), 41–55.

[12] Shalizi, C., and Thomas, A. Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods and Research 40*, 2 (2011), 211–239.

[13] Steglich, C., Snijders, T., and Pearson, M. Dynamic networks and behavior: Separating selection from influence. *Sociological Methodology 40*, 1 (2010), 329–393.

Understanding Quota Dynamics in Wireless Networks

Matthew Andrews Glenn Bruns Mustafa K. Doğru Hyoseop Lee
Bell Labs, Alcatel-Lucent
Email: {matthew.andrews,glenn.bruns,mustafa.dogru,hs.lee}@alcatel-lucent.com

I. Introduction

Wireless service providers would like to learn more about how users respond to prices and the utility users derive from various network transactions. This kind of information is valuable in developing models of user behavior and in designing pricing schemes, especially dynamic pricing plans [2],[3].

However, obtaining such information is difficult, because in today's wireless networks users rarely make a separate payment for each network transaction (e.g., voice call, SMS, or data download), nor are transactions offered at multiple price points. Instead, users pay for a quota, which can take two forms. With a *prepaid* plan, the user pays for a certain quantity of voice or data service, which can then be consumed over a flexible time frame. Alternatively, with a *capped* plan, the user pays a flat fee for a fixed voice or data quota, and any unused quota at the end of the month perishes.

The idea we explore here is to see how much of what we would like to know about users can be obtained via a study of *quota dynamics*: the change in user behavior as a function of quota balance and/or time to the end of the quota period. For example, in a prepaid voice plan, how does a user's call frequency change with decreasing balance? Or, in a capped data plan, how much of a user's quota is typically wasted in each period?

Understanding the dynamics of such quota use can (i) provide insights into user price sensitivity, (ii) enable us to develop models of user behavior that can be fed into a variety of pricing models (like the two-sided pricing model in [1]), and (iii) shed light on the underlying utility that end users apply to different network transactions.

Our work on quota dynamics has two parts. First (Section II), we have performed data analysis on call detail records (CDRs) from an operator that services prepaid voice and SMS users. This analysis shows that user behavior changes markedly as the remaining quota gets low. Second, we have developed user models to describe the observed behavior. In the first model (Section III-A), there is an underlying set of "potential" network transactions. The probability that the user actually performs a transaction depends on the current balance as well as the time to the end of the quota period. In the second model (Section III-B), we aim to explain this type of behavior via a utility maximization framework where the user has different utilities for different types of transactions and her goal is to maximize the total utility received over the quota period.

We believe that, although our existing CDR data set is for voice and SMS users only, an equivalent analysis could be carried out for wireless data users. Moreover, that setting is likely to give richer behavior due to the greater heterogeneity of content. We are in discussions with an operator to obtain CDRs for mobile data users in order to perform such a study.

II. Data Analysis

We consider CDRs from a population of prepaid wireless users that make both voice calls and SMSs. The data set spans 6 weeks and corresponds to about 1.6 million subscribers. However, some of the subscribers have complex plans with various types of "free" calls that are bundled in with the monthly subscription. For simplicity, we focus on users whose plan has a simple monetary quota. For these users, the data set contains about 60 million outgoing calls and 150 million SMSs. Each voice call or SMS decrements the balance by a predetermined amount. In the remainder of this section we discuss how call duration, inter-call time and inter-SMS time vary as a function of remaining balance.

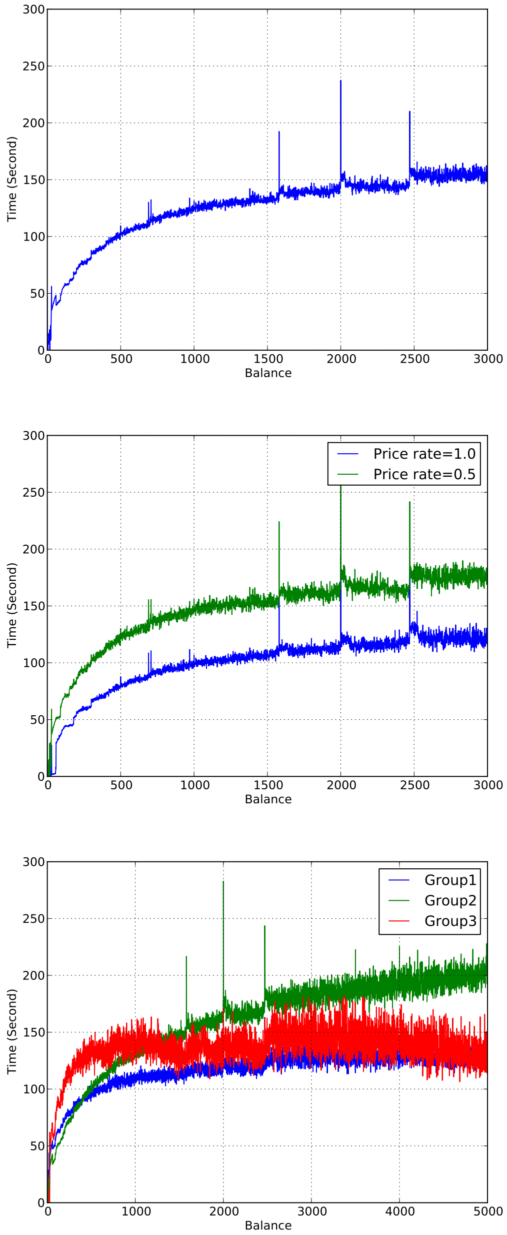
Fig. 1 (top) shows how call duration varies with current user balance. Clearly, call duration decreases as the balance gets low. The spikes around the balances 1,000, 1,700, and 2,000 are closely related to the most frequent top-up amounts. One could conclude that the first call after top-up typically lasts longer than any other calls.

There are different price rates for a single call depending on subscribers' plans. Among them, the price rates 1.0 and 0.5 monetary units per second appear most frequently in the data. (We use the term "monetary unit" so as not to hint at the source of data.) Fig. 1 (middle) depicts the behavior of subscribers under different price rates. The longer call durations of subscribers with the lower price rate is obvious. Note however that the half-price rate does not give a two-fold increase in duration, but extends call duration by about 50%. Nevertheless, this behavior suggests some degree of price sensitivity among the users.

Users' different responses to remaining balances is observed in Fig. 1 (bottom). In order to differentiate responses, each user's call duration is modeled as a power function of remaining balance. The exponent is estimated from each individual's call history, and then used as a classifier. The figure only includes the individuals whose history of call durations can successfully yield parameters of the model. Group number 1, 2, and 3 stands for the group having approximately zero, positive, and negative exponent, respectively. Each point for a group in the figure is the average of all call durations for that group. Note that all three groups behave similarly when the balance is low, particularly below 1,000. However, the usage patterns show different trends as the balance increases. particular, Group 2 continues responding to increasing bala by having longer calls. This demonstrates there are diffe responses to remaining balance within the user population this suggests different price sensitivities for different users.

Fig. 2 shows the time between consecutive outgoing-calls (blue) and outgoing-SMSs (red) as a function of current user balance. Sharply increasing inter-call time indicates the users' reluctance for topping up. For the inter-SMS time, since the raw data is noisy, we applied a smoothing filter based on

H. Lee's work was supported by the Seoul Metropolitan Government R&BD Program WR080951. The authors would like to thank J. Borger, A. Grover and K. Nithi for valuable conversations.



g. 1: Call durations (top), call durations at different rates niddle), and call durations of different user groups (bottom)

the Bartlett window function with window size 51. The inter-SMS time decreases as current user balance decreases. This is exactly the opposite of the trend for the inter-call time and indicates that users consider an SMS as a cheap substitute to a call when their balance is low. (For the plans we consider, the cost of an SMS is about the same as a one or two-second call.) Overall, the behavior observed in Fig. 2 provides further empirical guidance in developing models of user behavior responding to the current balance.

III. User Modeling

In this section we present models that could be used to explain user behavior under prepaid and capped pricing plans. A main difference between these two kinds of plans is that, in a capped plan, wireless usage is a perishable good: what is not used at the end of the billing period is lost. Thus, one way to try to learn about users' price sensitivities is to examine how much of their quota is typically "wasted". On the other hand, the two kinds of plans both have a notion of balance. In the prepaid case, balance is the monetary balance in a user account; in the capped case, balance is the amount of service remaining for the billing period. In both cases we might expect usage to decline as the balance decreases.

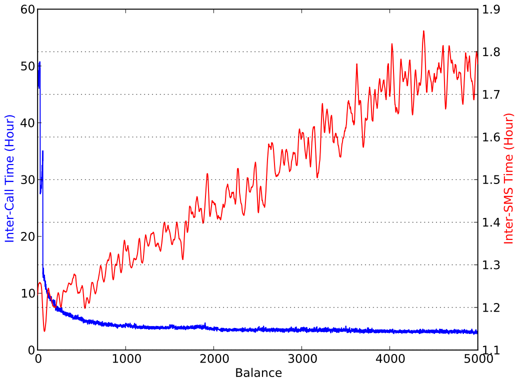


Fig. 2: Inter-call time (blue) and inter-SMS time (red)

A. *Modeling Usage Directly*

We begin describing the model for the prepaid pricing case, and then discuss how it can be extended to capped plans for voice and data. Let us consider a user who feels the need/urge to make calls that arrive following a Poisson process with rate $\lambda \geq 0$ per day. Each need is either satisfied with probability $p(q)$, which is a function of the remaining balance q, or not satisfied with probability 1-$p(q)$. We assume $p(q)$ is nondecreasing in q and the probability becomes 1 beyond a threshold: $p(q) = 1$ for $q \geq q_0$. In other words, the user does not take the balance into account and satisfies all calling needs provided that the balance is high enough. It can be shown that the expected inter-call time for a given q, denoted by $\bar{\tau}(q)$, is the inverse of $\lambda p(q)$. If the urge/need is satisfied, the call duration is $X(q)$, which is a non-negative random variable with a known distribution.

Depending on the service (voice/data, prepaid/capped), $p(q)$ may take different forms for $q < q_0$. Next, we analyze the data from Section II to estimate the parameters of the model described above. Fig. 2 gives the inter-call times as a function of q, which we use to estimate $p(q)$. More explicitly,

$$\bar{\tau}(q) = \frac{1}{\lambda}\frac{1}{p(q)}$$

is utilized to estimate the probabilities and the arrival rate. A log-log plot suggests a power law between the probability of acceptance and the remaining balance. Thus, the particular form of $p(q)$ for the prepaid call data is

$$\mathsf{Pr}\{\text{accept}|q\} = p(q) = \begin{cases} aq^b & \text{if } q < q_0 \\ 1 & \text{if } q \geq q_0 \end{cases}$$

where a, b, and q_0 are parameters to be estimated. We fit the data with nonlinear least squares method that yields $a = 0.097$, $b = 0.296$, $q_0 = 2656$, and $\lambda = 7.142$. Also, we can use the data for Fig. 1 (top) to fit a model for call duration $X(q)$, which we omit due to space restrictions.

The model described here can be extended to model usage in capped voice or data services. Now, the probability of satisfying a need/urge to use the service is a function of the remaining quota q and the remaining time in the billing cycle t. We expect $p(q,t)$ to increase as q increases and decrease as t increases. For capped data services, we can enrich the model by incorporating how the usage pattern of different applications (web surfing, video, e-mail, etc.) change in the course of a billing cycle. One way to accomplish this with our model is to assume two arrival streams for the need/urge to access the network: λ_1 for low bandwidth applications like e-mail and λ_2 for high bandwidth applications like video. The probability to satisfy the need/urge can be considerably different for each stream within the billing cycle. Moreover, if satisfied, each need would consume highly different amounts from the remaining quota, $X_1(r,t)$ and $X_2(r,t)$ for low and high bandwidth applications, respectively.

B. Modeling Usage Via Utility

So far we have assumed that the need to access the network is an exogenous process. We now discuss how this need can be explained and modeled by the actions of a utility-maximizing user. Our second model focuses on usage under capped plans. The intuition behind the model is that users know the utility they will derive from a service at the current time, but uncertain about this utility in the future, since they do not know with certainty which content they might want to access. A user's pattern of consumption over the billing cycle follows from the user attempting to maximize the total expected utility for the billing cycle in the face of this uncertainty.

We formalize the model as a Markov Decision Process (MDP). Recall that an MDP consists of states, actions, a state transition function, and a reward function. Here, a state consists of quota balance q, a number i of points remaining in the quota period, and a parameter c of utility function $u_c(x) = 1 - e^{-cx}$. An action a (with $a \leq q$) is the amount of voice or data consumed by a user at a state. If $i > 0$ (so that the quota period is not finished), the probability of of a transition from a state (q, i, c) to a state $(q-a, i-1, c')$ on action a is simply the probability $\mathsf{Pr}(c')$ that the utility function parameter takes value c', based on a discrete prob. mass function. In other words, an amount a is consumed of the balance, the next time period is reached, and the utility function of the next time period is established. The reward with an action a is the utility $u_c(a)$, regardless of source or destination state.

In this framework, the problem we solve is to compute a user strategy that will maximize the expected utility over the quota period. We do this by solving a set of Bellman equations, working backward from the end of the quota period.

We have run simulations to see how patterns of user consumption are affected by the probability distribution of utility function parameter c. We observed that the average consumption pattern does not depend on the distribution for c. In every distribution we have tested, the average consumption at each point in the quota period is approximately equal. Fig. 3 (top) shows simulated consumption over 30 quota periods, for a uniform distribution over five possible values of c.

Thus, by varying the distribution over c, we were unable to generate user behaviors similar to Fig. 1. However, such patterns of consumption can be generated by our model by allowing users to be mistaken about the actual distribution over parameter c. For example, Fig. 3 (bottom) shows the pattern of user consumption when an approximately normal distribution is used to guide decision making, but a distribution skewed towards higher values of c is used to determine the actual c values encountered by the user. In other words, in this scenario the user underestimates the utility she will obtain from data at a future point within the quota period and therefor uses up too much of her quota earlier in the period.

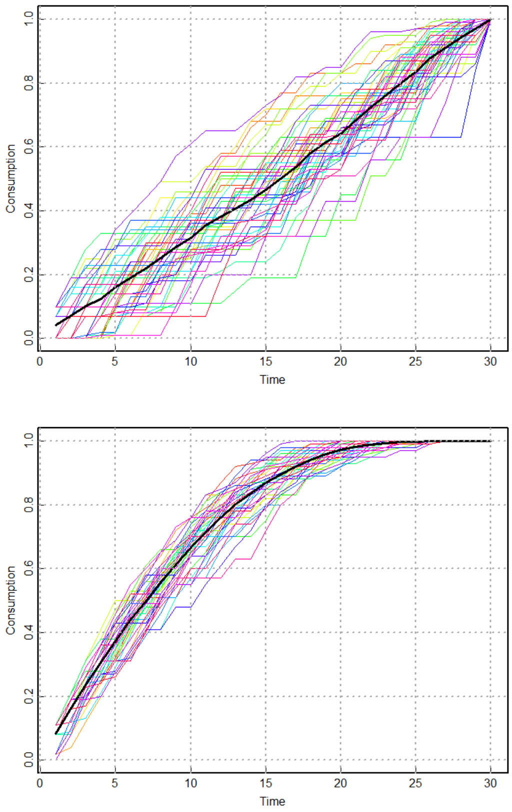


Fig. 3: Data consumption under uncertain utility (top), and by an "underestimating" user (bottom)

In our utility-maximizing model, a rational user will consume all available service in every billing period. This is contrary to everyday experience. We are investigating extensions to our model to account for wasted quota, including negative utility (sometimes the work to use a service outweighs the derived benefit) and partially-observed state (users don't always know their balance or the date at which the quota period ends).

References

[1] Matthew Andrews, Ulas Ozen, Martin I. Reiman, and Qiong Wang. Economic models of sponsored content in wireless networks with certain demand. In *Proc. of 2nd IEEE Workshop on Smart Data Pr (SDP2013)*, 2013.

[2] C.A. Gizelis and D.D. Vergados. A survey of pricing scheme wireless networks. *IEEE Communications Surveys Tutorials*, 13(1):1 145, 2011.

[3] S. Sen, C. Joe-Wong, S. Ha, and M. Chiang. A survey of broadband data pricing: Past proposals, current plans, and future trends. to appear in ACM Computing Surveys, 2013.

Quantitative Analysis of Community Detection Methods for Longitudinal Mobile Data

Syed Agha Muhammad and Kristof Van Laerhoven
Embedded Sensing Systems
Technische Universität Darmstadt
Germany
{muhammad,kristof}@ess.tu-darmstadt.de

***Abstract*—Mobile phones are now equipped with an increasingly large number of built-in sensors, that can be utilized to collect long-term socio-temporal data of social interactions. Moreover, the data from different built-in sensors can be combined to predict social interactions. In this paper, we perform quantitative analysis of 6 famous community detection algorithms to uncover the community structure from the mobile data. We use Bluetooth, WLAN, GPS, and contact data for analysis, where each modality is modelled as an undirected weighted graph. We evaluate community detection algorithms across 6 inter-modality pairs, and use well-know partition evaluation features to measure clustering similarity between the pairs. We compare the performance of different methods based on the delivered partitions.**

I. Introduction

The problem of finding community structures within a complex network is familiar in social sciences, and different community detection methods have been introduced to accomplish this task. However, the methods developed till now are not acknowledged to be fully reliable, due to the lack of shared definition of the term community and partition. Mostly community detection algorithms are analysed on benchmark datasets [2]. To our knowledge, there has been no work done on the evaluation of community detection methods on real networks data. Community structures in a real world environment can change abruptly, and benchmark datasets lack the skewness found in real networks. Generally, real networks lack reference community structures to evaluate algorithms. To collect the complete ground truth from a large number of participants is an inefficient and impractical solution. A feasible alternative is to utilize different built-in sensors within mobile phones. Combining the data from different sensors can provide points to the possible structures, and the possibility of more accurate prediction.

In this paper, we are interested in quantitative analysis of community detection methods on the Nokia Mobile Dataset [1]. The provided data does not have complete reference structure of the community. We utilize Bluetooth, WLAN, GPS, and contact data of the participants. We have anonymized Bluetooth MAC address and mobile number of the participants. Our goal is to model graphs for different modalities, apply community detection methods and determine the partition similarity of different modality combinations. We utilize certain attributes from each modality to build their graphs. Later on, we test them with community detection methods. We evaluate the partition results across 6 different pairs (Bluetooth and GPS, Bluetooth and contact, Bluetooth and WLAN, GPS and contact, GPS and WLAN, contact and WLAN). We compare the partition results from every pair to measure clustering similarity. Figure 1 shows the overview of our approach. For the detailed discussion of creating graphs for each modality, we refer the reader to our previous work [3].

II. Results and Discussions

The plots of Figure 2 and 3 illustrate the experimental results for the pair of modalities and the community detection methods. In the figures, each column represent the community detection method and each row represent the combination of modalities. BT and CT in the figures represent Bluetooth and contact. To measure the similarity between modalities, we created different pairs of modalities and later tested clustering evaluation features against them. The blue bars represents computed Rand indices, green bars represents the Jaccard indices, red bars represents the distance measures, and small blank spaces between the measures are given for the convenience of the readers. We performed tests on the weighted and unweighted graphs to evaluate the performance of the community detection methods under different conditions. We found that the weighted graphs have better community structures and indices values are higher than the unweighted graphs. The communities detected by dynamic algorithms have followed the power law distribution, one or two large communities along with many small communities were detected. Modularity optimization techniques have performed well for some modalities. Similarly, edge betweenness and infomap have detected either a single big outlier or a big group with many isolated nodes for all modalities. We will discuss every method and the effect of adding and removing weights upon the communities detected.

Edge betweenness has a poor overall performance for both weighted and unweighted graphs. For some modalities a single super community, and for some single big outlier along with many isolated nodes were detected. The plots for the weighted and unweighted graphs shows Rand and Jaccard indices have fluctuated values. The higher values come from the result of comparing two super communities with many common nodes, and lower values represent cases of comparing super communities with not too many common nodes. Similarly, it shows individual big bars along many small ones for distance measure. Normally, edge betweenness produces fair results for a sparse graph or graphs with an average degree of 16, but in our case the graphs are more dense with the minimum average degree of 15 for single modality, and up to 29 for contact

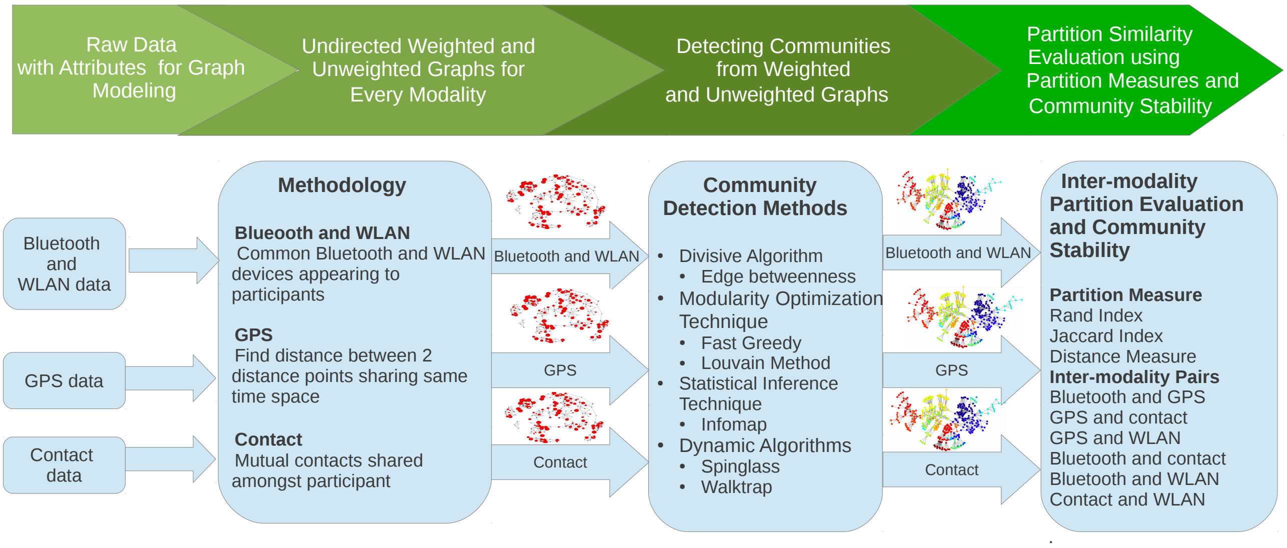


Fig. 1: Overview of our method. After preprocessing the raw data, a graph is generated for each modality that is used as an input for community detection methods to detect the community structures. The community structures results of inter-modality pairs are evaluated by partitioning measures to find the clustering similarity.

modality. Additionally, to study the effect of the number of clusters on the grouping formed and indices, we varied the number of clusters. The results showed no significant improvements. For the unweighted graphs, the results are no more different than for the weighted graph. The detected groups were of bigger size along with many individual isolated communities except for the contacts, where only a single big group was detected.

The Fast greedy method has produced heterogeneous groups for some modalities, with some small and a single or two large communities for modalities. For the contacts modality; however, a single super community was detected. The most probable reason could be due to a well known resolution limit, that is biasness of the modularity towards the bigger communities, which often yield a poor values of the modularity maxima. The rand indices for the weighted graph are between 0.47 and 0.57. For the combinations of BT/WLAN, the Rand index is slightly higher, and distance measure shows many similar clusters.

The Louvain method has performed better on weighted graphs. The detected community structures followed the power law, except for contact modality. For contact modality, a single super group was detected. It has higher Rand and Jaccard index of 0.68, and 0.66 for the combination of ’GPS and WLAN’, similarly distance measure shows many groups with similarity around 0.40. The Rand indices for the remaining combinations ranges between 0.50 and 0.68. For distance measure, detected groups have the similarity around 0.30. For the unweighted graph, the detected groups were slightly bigger in size. The average values for Rand and Jaccard indices have come down for the combinations. The only exception can be found for the case of ‘GPS and BT‘, where the Rand index value is 0.65.

Infomap has produced weak results for weighted and unweighted graphs. For the weighted graph, lots of small groups were detected, Rand and Jaccard indices values were very low. The detected groups have very small similarity. For the unweighted graph, for every combination a single big community was detected and this fact be easily observed in fig 3 .

Walktrap has fair results for weighted and unweighted graphs. The Rand indices for combinations, such as GPS/BT, GPS/WLAN, and WLAN/BT, are 0.61, 0.62, 0.75 respectively. The distance measure for those combinations were around 44%. The detected groups for WLAN, BT, GPS were almost the same; however, there was a single big community for contacts, and that resulted in comparatively lower indices values for those combinations. For the unweighted graph, for contacts and WLAN super communities with many isolated nodes were identified. The overall performance for weighted graphs were better as compare to unweighted graphs. The length of random walks to perform can be a decisive factor for the performance of this method. The variation of the length showed that for larger steps single big communities with m isolated nodes were detected and similarly for smaller v of steps a single big community was detected.

Spinglass has performed efficiently for the unweighte and weighted graph. The Rand, Jaccard indices, and distance measure for both weighted and unweighted graphs are the almost same. We change some parameters of the algorithm to find its effects on the results. We found for the lower spin states, bigger communities were detected. By increasing the

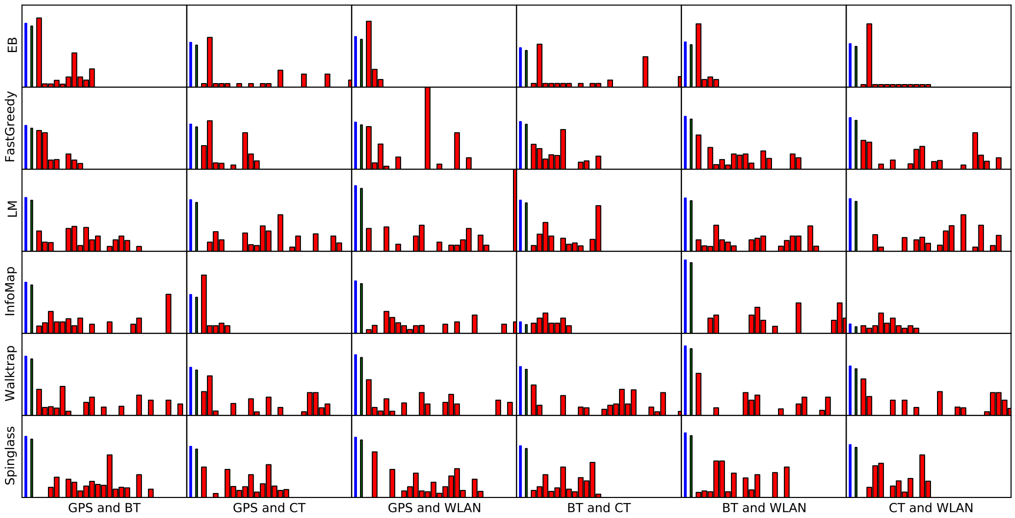


Fig. 2: Community detection methods and the applied modality combinations for weighted graphs.

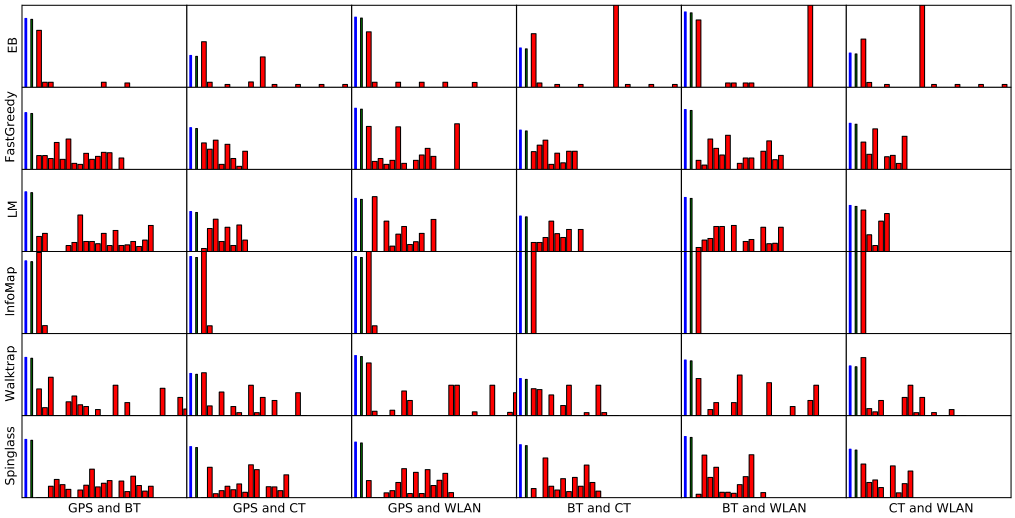


Fig. 3: Community detection methods and the applied modality combinations for unweighted graphs.

spin states, the size of the groups tends to be smaller. Until a certain point there was a changing behavior in the formation of the clusters, but then the detected groups were the same and the only difference was in orders. The maximum value of the rand, and Jaccard indices were found for BT/WLAN modality with 0.68 and 0.65, and similarly the distance measure also represents matches amongst the groups. We tested graphs against the negative implementation, which assigns negative values to the edges. The computation time was slightly higher. For our particular case, negative edge implementation had a poor performance and for most of the modalities single bigger communities were produced. We also verified the effect of updating the spins of the vertices in parallel on the results, and we found that it effected the results by predicting single super communities. For the unweighted graph, we found similar results for the Rand, Jaccard indices, and distance measure. The communities detected for the unweighted graphs were slightly bigger than the weighted graphs.

In conclusion, our evaluation shows that spinglass has produced better results for mobile data. It has performed equally better for weighted and unweighted graphs. This fact is further cemented by the higher values of Rand, Jaccard indices and a consistent similarity found by distance measure. Walktrap has also performed good; however, the number of random walks are a crucial factor for the clustering. Then at the third and fourth place comes Louvain and fast greedy methods. These methods have performed for some modalities, but they have also produced outliers for some modalities.

Infomap and edge betweenness performed poor as compared to the other methods. Pairwise combinations of GPS, Bluetooth, and WLAN data provide points about reference structures. For the combination of modalities, we found that GPS/BT, WLAN/GPS, BT/ WLAN have the best results. Mostly features have given a higher value for the pairs of these modalities. Table I features some key results from our observations. It represents results only from weighted graphs.

	GPS/BT		WLAN/GPS		BT/WLAN	
	RI	JI	RI	JI	RI	JI
Spinglass	0.63	0.608	0.64	0.62	0.68	0.65
Walktrap	0.61	0.59	0.62	0.605	0.75	0.72
Louvain Method	0.55	0.535	0.68	0.66	0.59	0.56
Fast Greedy	0.47	0.42	0.48	0.46	0.51	0.49

TABLE I: Key results derived from the results (JI, and RI stands for Jaccard and Rand Index).

References

[1] N. Kiukkonen, Blom J., O. Dousse, Daniel Gatica-Perez, and Laurila J. Towards rich mobile phone datasets: Lausanne data collection campaign. In *Proc. ACM Int. Conf. on Pervasive Services (ICPS,',',')*, *Berlin.*, 7.

[2] Andrea Lancichinetti and Santo Fortunato. Community detection algorithms: a comparative analysis. September 2010.

[3] Syed Agha Muhammad and Kristof Van Laerhoven. Discovery of user groups within mobile data. In *Nokia Mobile data challenge workshop*, june 2012.

Calling Patterns in Human Communication Dynamics

Zhi-Qiang Jiang,[1,*] Wen-Jie Xie,[1] Ming-Xia Li,[1] Boris Podobnik,[2,3] Wei-Xing Zhou,[1,†] and H. Eugene Stanley[2,‡]

[1] *School of Business, School of Science,*
and Research Center for Econophysics,
East China University of Science and Technology, Shanghai 200237, China

[2] *Center for Polymer Studies and Department of Physics,*
Boston University, Boston, Massachusetts 02215, USA

[3] *Zagreb School of Economics and Management, 10000 Zagreb, Croatia*

Abstract

Modern technologies not only provide a variety of communication modes, e.g., texting, cellphone conversation, and online instant messaging, but they also provide detailed electronic traces of these communications between individuals. These electronic traces indicate that the interactions occur in temporal bursts. Here, we study the inter-call durations of the 100,000 most-active cellphone users of a Chinese mobile phone operator. We confirm that the inter-call durations follow a power-law distribution with an exponential cutoff at the population level but find differences when focusing on individual users. We apply statistical tests at the individual level and find that the inter-call durations follow a power-law distribution for only 3460 individuals (3.46%). The inter-call durations for the majority (73.34%) follow a Weibull distribution. We quantify individual users using three measures: out-degree, percentage of outgoing calls, and communication diversity. We find that the cellphone users with a power-law duration distribution fall into three anomalous clusters: robot-based callers, telecom frauds, and telephone sales. This information is of interest to both academics and practitioners, mobile telecom operator in particular. In contrast, the individual users with a Weibull duration distribution form the fourth cluster of ordinary cellphone users. We also discover more information about the calling patterns of these four clusters, e.g., the probability that a user will call the c_r-th most contact and the probability distribution of burst sizes. Our findings may enable a more detailed analysis of the huge body of data contained in the logs of massive users.

[*] zqjiang@ecust.edu.cn

[†] wxzhou@ecust.edu.cn

[‡] hes@bu.edu

Evolution of Communities with Focus on Stability

Carlos Sarraute
Grandata Labs
Buenos Aires, Argentina
charles@grandata.com

Gervasio Calderon
Grandata Labs
Buenos Aires, Argentina
gerva@grandata.com

1 Introduction

The detection of communities is an important tool used to analyze the social graph of mobile phone users. Within each community, customers are susceptible of attracting new ones, retaining old ones and/or accepting new products or services through the leverage of mutual influences [1]. The communities of users are smaller units, easier to grasp, and allow for example the computation of role analysis – based on the centrality of an actor within his community.

The problem of finding communities in static graphs has been widely studied (see [2] for a survey). However, from the point of view of a telecom analyst, to be really useful, the detected communities must evolve as the social graph of communications changes over time – for example, in order to perform marketing actions on communities and track the results of those actions over time. Additionally the behaviors of communities of users over time can be used to predict future activity that interests the telecom operators, such as subscriber churn or handset adoption [3]. Similary group evolution can provide insights for designing strategies, such as the early warning of group churn.

Stability is a crucial issue: the analysis performed on a given community will be lost, if the analyst cannot keep track of this community in the following time steps. This is the particular use case that we tackle in this paper: tracking the evolution of communities in dynamic scenarios with focus on stability.

We propose two modifications to a widely used static community detection algorithm. We then describe experiments to study the stability and quality of the resulting partitions on real-world social networks, represented by monthly call graphs for millions of subscribers.

2 Data Sources

Our raw data source is anonymized traffic information from a mobile operator. The analyzed information ranges from January 2012 to January 2013, and contains for each communication the origin, target, date and time of the call or sms, and duration in the case of calls.

For each month T, we construct a social graph $\mathcal{G}_T =< \mathcal{N}_T, \mathcal{E}_T >$. This graph is based on the aggregation of the traffic of several months, more concretely $\mathcal{G}_T$ depends on the traffic of three months: T, $T-1$ and $T-2$. The raw aggregation of the calls and messages gives a first graph with around 92 M (million) nodes and 565 M edges (on a typical month). The voice communications contribute 413 M edges and the messages contribute 296 M edges to this graph.

We then perform a symmetrization of the graph, keeping only the edges (A, B) whenever there are communications from A to B and from B to A. This new graph has around 56 M nodes and 133 M (undirected) edges, and represents stronger social interactions between nodes. Additionally we filter nodes with high degree (i.e. degree greater than 200) since we are interested in the communications between people (and not call centers or platform numbers).

3 Dynamic Louvain Method

Our first experiment to detect evolving communities was to run the original Louvain algorithm [4] on the graphs at time T and $T+1$, and compare the two partitions, method that resulted very unstable. Our second experiment was to run the Louvain algorithm modified by Aynaud and Guillaume [5] to obtain a more stable evolution. As we show in Section 4 the results were still unsatisfying in terms of stability.

In our use case (e.g. telecom analysts performing actions on the communities), the stability of the partition is our main concern. With this goal in mind, we propose two modifications to the Louvain method, that give the partition at the previous time step a sort of “momentum”, and make it more suitable to track communities in dynamic graphs.

Before describing them, we introduce some notations. As stated in the previous section, we consider snapshots of the social graph constructed at discrete time steps (in our case every month). Let $\mathcal{G}_T =< \mathcal{N}_T, \mathcal{E}_T >$ be a graph that has already been analyzed and partitioned in communities. Let $\Gamma =< C_1, \dots, C_R >$ be such partition in R communities. Given a new graph $\mathcal{G}_{T+1} =< \mathcal{N}_{T+1}, \mathcal{E}_{T+1} >$ our objective is to find a partition of $\mathcal{G}_{T+1}$ which is stable respect to Γ.

The first idea is to have a set of *fixed nodes* $\mathcal{F}$. Let $\mathcal{R} = \mathcal{N}_T \cap \mathcal{N}_{T+1}$ be the set of nodes that remain from time T to $T+1$. The set $\mathcal{F}$ is a subset of $\mathcal{R}$, whose nodes are assigned to the community they had at time T. In other words, noting γ the function that assigns a community to each node, we require: $\gamma_{T+1}(x) = \gamma_T(x)\ \forall x \in \mathcal{F}$.

We experimented with different distributions of the fixed nodes, ranging from no fixed nodes ($\mathcal{F} = \emptyset$) to all the remaining nodes ($\mathcal{F} = \mathcal{R}$). For the experimental results, we used a parameter p that represents the probability that a node belongs to $\mathcal{F}$ (i.e. $|\mathcal{F}| = p \cdot |\mathcal{R}|$).

The second idea is to add a probability q of "preferential attachment" to pre-existing communities. With probability q, the new nodes will prefer to attach to a community existing at time T instead of attaching to a community formed at time $T+1$. We give the details below.

The Louvain Method [4] is a hierarchical greedy algorithm, composed of two phases. During phase 1, nodes are considered one by one, and each one is placed in the neighboring community (including its own community) that maximizes the modularity gain. This phase is repeated until no node is moved (that is when the decomposition reaches a local maximum). Then phase 2 consists in building the graph between the communities obtained during phase 1. Then the algorithm starts phase 1 again with the new graph, in the next hierarchical level of execution, and continues until the modularity does not improve anymore.

We construct a set $\mathcal{P} \subseteq \mathcal{N}_{T+1}$ such that $|\mathcal{P}| = q \cdot |\mathcal{N}_{T+1}|$. For every node x, we consider its neighbors that belong to a community existing at time T, that is the set $A(x) = \{z \in \mathcal{N}_{T+1} \,|\, (x,z) \in \mathcal{E}_{T+1} \wedge \gamma_{T+1}(z) \in \Gamma_T\}$. During phase 1 of the first iteration of the algorithm (i.e. during the first hierarchical level of execution), the inner loop is modified. For all node $x \in \mathcal{N}_{T+1}$, if $x \in \mathcal{P}$ and $A(x) \neq \emptyset$ then place x in the community of $A(x)$ which maximizes the modularity gain (whereas if $A(x) = \emptyset$ proceed as usual).

4 Experimental Results

In our experiments, we computed the social graph (constructed as described in Section 2). Since we are interested in the real-world application of our method, we preferred to evaluate it on real data.

Given two months T and $T+1$, we calculated a partition in communities of $\mathcal{G}_T$ using the Louvain Method (with the modification of [5]) that we note $\Gamma =< C_1, \dots, C_R >$; and a partition of $\mathcal{G}_{T+1}$ using our dynamic version of the Louvain Method, with different values of the parameters p and q. Let $\Gamma' =< C'_1, \dots, C'_S >$ be the partition of $\mathcal{G}_{T+1}$. We are interested in comparing Γ and Γ' in terms of stability and quality of the partition. To this end, we measure: (i) the mutual information between the two partitions; (ii) the number of matching communities (i.e. such that the proportion of nodes in common is greater than a parameter r); (iii) the final modularity of Γ' (as defined in [6]).

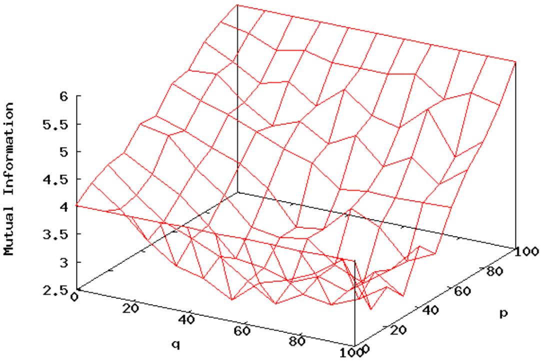


Figure 1: Mutual Information as a function of p and q (expressed as percentages).

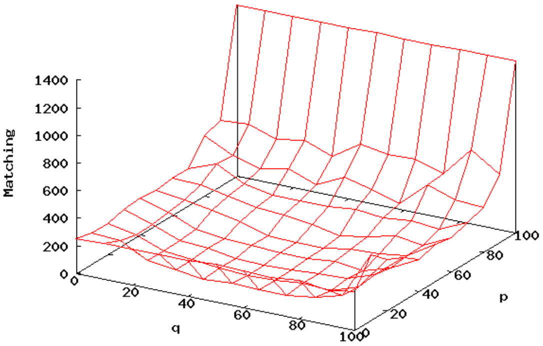


Figure 2: Matching communities.

The number of matching communities is computed as follows: for each community $C'_j \in \Gamma'$, we evaluate whether there is a community $C_i \in \Gamma$ such that $|C_i \cap C'_j| > r \cdot |C_i|$ and $|C_i \cap C'_j| > r \cdot |C'_j|$, where r is a fixed parameter verifying $r > 0.50$ (for instance we used $r = 0.51$). In that case, we say that C'_j matches C_i. The matching communities are of particular interest, because C'_j can be considered as the evolution of C_i (although the community may have grown or shrank) and can be individually followed by a human analyst.

The mutual information for two partitions of communities (see [7, 3] for definitions[1]) is computed

[1] Since nodes can change between time T and $T+1$, we

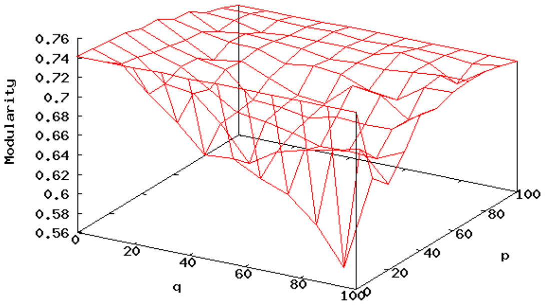


Figure 3: Modularity.

as:

$$MI(\Gamma,\Gamma') = \sum_{i=1}^{R}\sum_{j=1}^{S} P(C_i,C'_j)\log\frac{P(C_i,C'_j)}{P(C_i)\cdot P(C'_j)}.$$

To analyze the effect of p and q, we made those parameters vary from 0 to 1. The baseline, for $p = 0$ and $q = 0$, corresponds to the Louvain Method with the modifications of [5].

Fig. 1 shows the effect on the mutual information between the two partitions. We can clearly observe that the mutual information increases as p increases, and reaches its maximal values at $p = 100\%$. The effect of varying q is not so clear, since it produces fluctuations of the mutual information without a marked tendency.

Fig. 2 shows the number of matching communities (according to our criterion). In this graph we see that the number of matching communities increases dramatically when p approaches 100%. The effect of varying q is again not clearly marked, although the increase of q produces higher matching communities for smaller values of p.

Fig. 3 shows the effect on the modularity of the new partition. We can observe that the modularity decreases slightly as p increases for small values of q. For greater values of q (closer to 100%), varying p produces fluctuations with a decreasing tendency.

As a conclusion, we can see that increasing the probability p of fixed nodes has a clear effect on increasing the mutual information between the two partitions, and the number of matching communities. The trade-off with quality is good, since the decrease in modularity is relatively low.

On the other side, increasing the probability q of preferential attachment to pre-existing communities has not a clear effect on mutual information or matching communities. It does not seem advisable to use this second modification for generating evolving communities.

5 Conclusion and Future Steps

The detection of evolving communities is a subject that still requires further study from the scientific community. We propose here a practical approach for a particular version of this problem where the focus is on stability. The introduction of fixed nodes (with probability p) increases significantly the stability of successive partitions, at the cost of a slight decrease in the final modularity of each partition.

As future steps of this research, we plan to: (i) study the evolution of communities with finer grain, using smaller time steps; (ii) evaluate the proposed method on publicly available datasets, to facilitate the comparison of our results; (iii) refine the matching criteria, and consider additional events in the evolution of dynamic communities (such as birth, death, merging, splitting, expansion and contraction [3]).

References

[1] Bin Wu, Qi Ye, Shengqi Yang, and Bai Wang. Group CRM: a new telecom CRM framework from social network perspective. In *CNIKM'09*, pages 3–10. ACM, 2009.

[2] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3):75–174, 2010.

[3] Derek Greene, Dónal Doyle, and Pádraig Cunningham. Tracking the evolution of communities in dynamic social networks. In *ASONAM'10*, pages 176–183. IEEE, 2010.

[4] Vincent Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[5] Thomas Aynaud and Jean-Loup Guillaume. Static community detection algorithms for evolving networks. In *WiOpt'10*, pages 513–519. IEEE, 2010.

[6] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

[7] Daniel J Fenn, Mason A Porter, Peter J Mucha, Mark McDonald, Stacy Williams, Neil F Johnson, and Nick S Jones. Dynamical clustering of exchange rates. *Exchange Organizational Behavior Teaching Journal*, 2012.

only consider the intersection $\mathcal{N}_T \cap \mathcal{N}_{T+1}$ for the mutual information computation.

Clustering of smartphones ownership in developing countries

Rich Ling
IT University of Copenhagen
rili@itu.dk

Johannes Bjelland
Telenor
Johannes.Bjelland@telenor.com

Geoff Canright
Telenor
geoffrey.canright@telenor.com

Kenth Engø-Monsen
Telenor
kenth.engo-monsen@telenor.com

Pål Roe Sundsøy
Telenor
Pal-Roe.Sundsoy@telenor.com

We are seeing the development of a large number of applications for smart phones that focus on the needs of people in developing countries (Alam, Khanam, & Khan, 2010). These developments promise to provide users with a wide spectrum of services and functionality ranging from m-health, to agricultural information, the ability to report problems and the ability to participate in political life (Kulkarni & Agrawal, 2008). A critical issue in this situation is, however, limited access to smart-phones that can use these applications. It is not clear as to whether people in developing countries have access to these devices.

In this paper we examine the degree to which smart phones are used in a developing country (Bangladesh) among strong-tie clusters. The experience of adoption in developed countries shows that there are strong network effects when considering the diffusion of smart phones (Sundsøy et al., 2011). Analysis shows that in developing countries, there are few smart phones. In the Scandinavian countries the adoption of smart phones (here defined as having an open OS and using GPRS), has reached approximately 60% of users. By contrast, in the poorer countries of southern Asia there are only about 3% of the users who have a smart phone (Telenor, 2012). This has implications in relation to the functionality of apps that are, in some cases, the threshold for use of m-health services, m-agriculture and m-inclusion.

In this paper we examine the degree to which the existing users of smartphones in Bangladesh are clustered. To do this we will examine the adoption of smart phones among 100 000 users in Bangladesh to determine the current adoption rates. Further we will examine the adoption of smart phones for the 10 top links for each of these users. Our hypothesis is that smart phone users cluster with other smart phone users As we move down the scale from smart phones to feature phones and to basic phones that there will be decreased clustering. We will examine this clustering in the light of subscription type, urban/rural location and intensity of use.

This analysis will provide a baseline from which to examine the diffusion of this technology. Also it will help us to understand the potential for app adoption and use.

Alam, M., Khanam, T., & Khan, R. (2010). Assessing the scope for use of mobile based solution to improve maternal and child health in Bangladesh: A case study. In *Proceedings of the 4th ACM/IEEE International Conference on Information and Communication Technologies and Development* (p. 3). Retrieved from http://dl.acm.org/citation.cfm?id=2370755

Kulkarni, S., & Agrawal, P. (2008). Smartphone driven healthcare system for rural communities in developing countries. In *Proceedings of the 2nd International Workshop on Systems and Networking Support for Health Care and Assisted Living Environments* (pp. 8:1–8:3). New York, NY, USA: ACM. doi:10.1145/1515747.1515758

Sundsøy, P., Bjelland, J., Canright, G., Engø-Monsen, K., & Ling, R. (2011). Comparing and visualizing the social spreading of products on a large social network. In T. Ozyer (Ed.), *The Influence of Technology on Social Network Analysis and Mining*. Springer.

Telenor. (2012). *Handset adoption* (Internal material). Fornebu: Telenor.

Mobile phones and digital generations in EU5 countries: a comparison of the 1996 and 2009 survey data.

Leopoldina Fortunati[1] and Sakari Taipale[2]

[1] Department of Human Sciences, University of Udine, Italy (fortunati.deluca@tin.it)

[2] Department of Social Sciences and Philosophy, University of Jyväskylä, Finland (sakari.taipale@jyu.fi)

Abstract

Mobile phone traffic data sets provide much-needed information on actual usage of mobile phones, such as the type, volume, place and time of usage. However, compared with traditional survey studies, traffic data involves only a small amount of information on users' socio-economic background (e.g. they do not say anything about users' education, marital status, occupation, income, family type or place of residence) and in some cases information on users, for instance their age, may be unreliable as Ling, Bertel and Sundsøy (2012) have shown. Furthermore, traffic data typically contains no motivational and attitudinal material. It is against this backcloth, that we will build our study on the analysis of two consecutive telephone surveys funded by Telecom Italia. These surveys, based on the same questionnaire (the questionnaire was slightly updated for the second survey) were carried out in Italy, France, the United Kingdom, Germany and Spain (EU5 countries) in 1996 (N=6,609) and 2009 (N=7,255). What makes these data sets rich is that they contain much cross-national information and allow a comparison between 1996 and 2009 in EU5 countries.

In this paper we will investigate, with a special focus on mobile phones, if it makes sense to talk about digital generations in EU5 countries. Several studies have already questioned whether the difference between digital native and digital immigrant generations, originally proposed by Prensky (2001), is justified (e.g. Herold 2012). It has been shown that both groups are internally incoherent, and other factors (e.g. breadth of use, experience, gender, education) have expressed in some cases more predictive power than age/generation (Selwyn 2004; Hargittai 2010; Helsper 2010; Helsper & Eynon 2010). In addition, it has been proposed that a so-called second generation of digital natives (born after 1990) could be separated from the first generation of natives (born in the 1980s) owing to their greater immersion in the social media (Helsper & Eynon 2010; Fortunati 2011). The above-mentioned studies have typically dealt with single countries and have been premised on cross-sectional data sets.

This study aims to see, firstly, whether the first generation of digital natives was the most technologically equipped generation in 1996 and did its relative position sustain until 2009. Secondly, we investigate if the first and second generation of digital natives differ from each other as regards to the use of digital technologies, especially mobile phones. We will use both bivariate statistics and Multiple Regression Analysis to analyse the data sets.

With regard to the first aim, our preliminary results show that the youngest respondents were the most equipped with mobile phones (and personal computers) in 1996, but no longer in 2009. Instead, it seems that the youth and young adults, who belong to the first generation of digital natives – and who also were the first group to adopt these devices in 1996 – maintained their position in 2009. With regard to the second aim, our data shows that there were actually no substantial differences between digital natives and immigrant adults in

relation to the made/received mobile (and fixed) phone calls in 1996. However, it appears that in 2009 the first generation of natives made and received more mobile phone calls and send more SMS than the second generation of natives.

References

Fortunati L. (2011) General Native Generations and the New Media. In F. Colombo and L. Fortunati (Eds.) *Broadband Society and Generational Changes*. Berlin: Peter Lang, 201–220.

Hargittai, E. (2010) Digital Na(t)ives? Variation in Internet Skills and Uses among Members of the “Net Generation”, *Sociological Inquiry*, 80(1), 92–113.

Helsper, E. J. & Eynon, R. (2010) Digital Natives: Where is the Evidence?, *British Educational Research Journal*, 36(3), 503–520

Helsper, E. J. (2010) Gendered Internet Use across Generations and Life Stages, *Communication Research*, 37(3), 352–374

Herold D. (2012) Digital Natives: Discourses of Exclusion in an Inclusive Society. In E. Loos, L. Haddon and E. Mante-Meijer (Eds.) *Generational Use of New Media*. London: Ashgate, 71–88.

Ling, R. Bertel, T. F. & Sundsøy, P. R. (2012) The socio-demographics of texting: An analysis of traffic data. *New Media and Society*, 14(2), 281–298.

Prensky, M., 2001. Digital Natives, Digital Immigrants. *On the Horizon* MCB University Press, 95. Available from: http://pirate.shu.edu/~deyrupma/digital%20immigrants,%20part%20I.pdf [Accessed 20 December 2009]

Selwyn, N., 2004. The Information Aged: A Qualitative Study of Older Adults’ Use of Information and Communications Technology. *Journal of Aging Studies*, 18(4), 369-384.

Protecting Cellular Networks from Attacks Launched Via Mobile Applications

E. Suissa, Y. Elovici, B. Chizi
Department of Information Systems Engineering Ben-Gurion University of the Negev P.O.B. 653, Beer-Sheva, 84105. Israel
Deutsche Telekom Laboratories at Ben-Gurion University Beer-Sheva, Israel
{einatsui, elovici, chiziba}@post.bgu.ac.il

Malicious mobile applications pose a serious risk to cellular networks because compromised mobile devices operating as a collaborated group can lead to denial of service attacks against parts of the cellular network infrastructure.

In this study we propose mitigating attacks against the cellular network by load balancing the signaling traffic between cells. The proposed method was evaluated using two datasets: User's movement data set (containing 196,591 users and 6,442,890 movement records), and an Antenna's location and orientation data set (containing 24,462 antenna locations from the Verizon wireless provider). The evaluation results demonstrate the feasibility of such attacks even after load balancing the signaling load between the cells. The evaluation also demonstrates how changing dynamically the cell's location can prevent overloading the network when there is high signaling traffic in parts of the network. This mitigation method is used to create a mechanism, which allows for the natural immunity of the network, prevents the denial of service on parts of the cellular network and therefore increases the network's resistance to signaling attacks launched via malicious mobile applications.

Mobile devices permit the downloading of many apps from various repositories, and therefore make the task of spreading new and possibly malicious applications extremely easy [3]. While a single compromised device threatens a single user, a large number of infected devices become a threat to the cellular network operator infrastructure.

Malicious and ill programmed mobile applications can cause serious damage to the cellular networks by performing many kinds of attacks as described in [2]. In fact, any application that runs on many mobile devices may cause DOS (denial of service) to users in the cellular network.

The GSM networks contain many cell sites, which sometimes called a "cell towers", Base Transceiver Stations (BTS), or "base stations". A base station is a physical structure that holds radio antennas, and a sector refers to a direction from a given cell tower.

When a mobile device attempt to establish a connection with the network (i.e., initiate a service), the base station assigns Standalone Dedicated Control Channels (SDCCH), over which the signaling with the HLR (Home Location Register) itself is performed. This resource is shared between all devices in an area [6].

In each cell in the cellular network, many devices are able to communicate simultaneously with the base station. Each cell has a fixed maximum service range of 35 kilometers. As the user moves with his/her mobile device, different cells connect him/her to the cellular network; hence the same device can overload different base stations while the user is moving to another location.

Researchers have already suggested different protections mechanisms against denial of service attacks on the cellular network at the device level, where each device "decides" on its own, the level of the signaling allowed without taking into consideration the threat coming from other nearby mobile devices (which seriously impacts the overall risk level on the close cellular infrastructure) and the movement of other users. We argue that a collaborative approach aimed at load balancing the signaling traffic between the cells will reduce overload and prevent blocking legitimate users from getting service much more effectively than limiting the signaling at the device level.

We took several assumptions in our work. One of them is that each mobile device equipped with a malicious application causes the same damage to the cellular network. Moreover, we assume that the accumulated damage of all the infected mobile devices is the sum of each one of them in a given time slot (i.e. the diminishing marginal proceeds law is not taken into consideration in our model). We also assume no dependency between users, which means that the only way users can get a malicious application is by downloading it from the market independently; therefore users do not infect each other by Bluetooth or MMS messages, etc. Our last assumption is that users can get service from each base station in the radius of 35km without taking into consideration the height of the base stations and other factors, which may affect the working range of the cell site – the range within which mobile devices can connect to it reliably.

In our study, according to [7], we first balanced the signaling traffic between the cells, i.e. we found the optimum base station for each mobile device at a specific time in a specific location. Then, we assessed the feasibility and potential damage of the attacks caused by mobile applications on the cellular network infrastructure. At last, we suggested how to mitigate the risks caused by mobile applications on the cellular network infrastructure by changing part of the cell's locations, i.e. attack oppression is performed based on data on users' movement and the spread of the malicious mobile application. The mitigation method provided in this study is very useful and prevents DOS attacks on the cellular network infrastructure and assures that users will receive service in any place at any time.

Most of the signaling services provided by the cellular network can be delivered over either the CCH (Common Control Channel) or the DCH (Data Channel). One of the most critical wireless bottlenecks is the Standalone Dedicated Control Channels (SDCCH). If available, the base station assigns an SDCCH, over which the signaling with the HLR itself is performed. By repeatedly triggering radio channel allocations and revocations to complete the data transfer a DOS attack may occur.

Sectors in a cellular network typically allocate 8 or 12 SDCCHs. The hold time of this channel is 2.7 seconds (as modeled by previous researches), so 0.37 users hold this channel every second. Each base station has 3 sectors, for each sector 12 SDCCH channels and each one of these channels is occupied by 0.37 users every second. Therefore the number of signaling messages per second over which an attack may be distributed in each base station is calculated as follows:

$$\mathrm{msg/sec} = \frac{\mathrm{sectors} * \mathrm{SDCCH}}{\rho_{\mathrm{SDCCH}}} = \frac{3 * 12}{2.7}$$

In order to assess the feasibility and potential damage of the attacks caused by mobile applications on the cellular network infrastructure, we used a mathematical optimization technique. This gives us precise mathematical formulations for the practical performance optimization task of load balancing, and the ability to create sensitivity analysis automatically. We introduce the following method, which is formalized as an integer linear program (d - device, s – BTS):

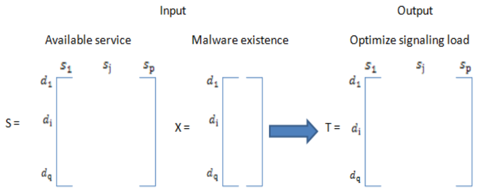


Figure 1- Input and output of the method

Each value in the available service matrix is a function of the distance between the antenna si and the device di (i.e. the value will be higher as di is closer si). The malware existence matrix describes which of the devices are infected by malware (1 for infected). The proposed target matrix indicates which base station si gives service to each device di.

Based on optimization techniques we found for each device the closest base station possible, under the constraints regarding limitations of the cellular network, without overloading base stations. The optimization allows us to balance the load automatically, using mathematical tools. It also provides us the ability to make a sensitivity analysis and to understand the relationships between input and output variables in the model (i.e. how many cells required to deal with the given load, how each cell loads the network etc). We calculated the percentages of overloaded base stations for each percentage of users running the malicious applications, and the percentages of legitimate users affected by the DOS attack for each percentage of users running the malicious applications. The evaluation results based on mobile phone dataset are summarized in Figure 2 and Figure 3 respectively:

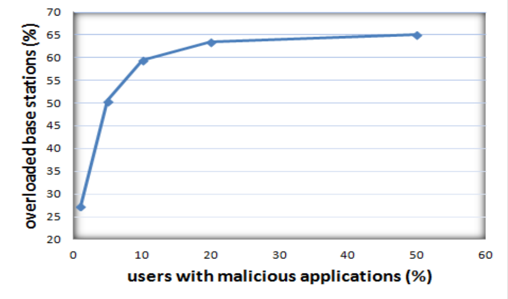


Figure 2- Percentage of overload base stations for a given percentage of users running the malicious applications.

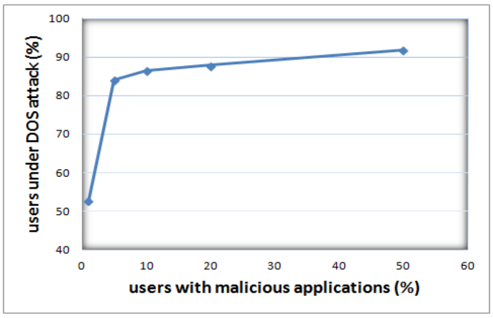


Figure 3 - Percentages of legitimate users under DOS attack for a given percentages of users running the malicious applications.

After evaluating the potential damage of such attacks, we would like to ensure that each user will get service at any place and any time, and in order to do so we need to add antennas in specific places and in accordance to the timing of attacks.

If we add antennas to each base station under attack, we will use an unnecessary amount of resources instead of simply moving antennas from one base station to another base station at a different time, according to the users' movement. To evaluate the exact number of cellular antennas required to mitigate such attacks, we used an algorithm, which takes as input the time of the attack, base station's locations and number of antennas to add to each overloaded base station. The results are described in Figure 4 (no time to move antenna from S2 to S3).

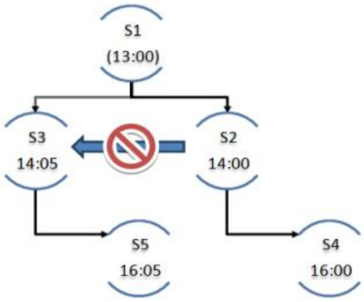


Figure 4 – Base stations under attack Graph, the connectivity of this graph determine whether it is possible to transfer antennas to other base stations according to the time and distance.

Our method was applied on our mobile phone dataset and returns the minimum number of cellular antennas that we need to add in order to immunize the network and prevent it from being overloaded, and their paths (movement of antennas from one location to another). Figure 5 present a small example of the algorithm result, where each base station needs amount of added antennas (described in parentheses) to deal with the overload in the specific time mentioned. The number on each arrow presents the number of antennas moving in that path.

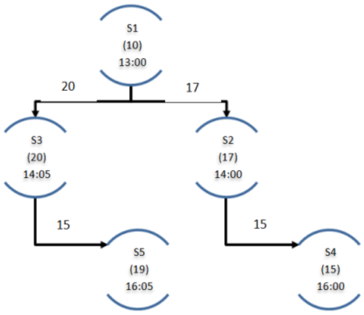


Figure 5 – Output of the mobility algorithm. It determines the amount of cellular antennas to add and their paths.

The outcome of this research will provide advanced tools and methodology, which will allow us to evaluate the loads caused by mobile applications on the cellular network. These results will include a demonstration of possible denial of service attacks against the cellular network infrastructure for a given movement data and antennas' locations and their potential damage. This can be represented by heat maps of signaling traffic loads on the mobile infrastructure.

The findings of this work can be used in cases of abnormal signaling traffic to block attacks and improve the quality of user data communications. Moreover, the outcome of this research can be formulated as a set of actionable insights, which can be taken into account on to the everyday activity of any cellular provider.

REFERENCES

[1] Wang, P., González, M. C., Hidalgo, C. A., & Barabási, A. L. (2009). Understanding the spreading patterns of mobile phone viruses. *Science*, *324*(5930), 1071-1076.

[2] Mulliner, C., Liebergeld, S., Lange, M., & Seifert, J. P. (2012, June). Taming Mr Hayes: Mitigating signaling based attacks on smartphones. In *Dependable Systems and Networks (DSN), 2012 42nd Annual IEEE/IFIP International Conference on* (pp. 1-12). IEEE.

[3] Hypponen, M. (2006). Malware goes mobile. *Scientific American*, *295*(5), 70-77.

[4] Pollini, G. P., Meier-Hellstern, K. S., & Goodman, D. J. (1995). Signaling traffic volume generated by mobile and personal communications. *Communications Magazine, IEEE*, *33*(6), 60-65.

[5] Traynor, P., Lin, M., Ongtang, M., Rao, V., Jaeger, T., McDaniel, P., & La Porta, T. (2009, November). On cellular botnets: Measuring the impact of malicious devices on a cellular network core. In *Proceedings of the 16th ACM conference on Computer and communications security* (pp. 223-234). ACM.

[6] Rappaport, T. S. (1996). *Wireless communications: principles and practice*. IEEE press.

[7] Mathar, R., & Niessen, T. (2000). Optimum positioning of base stations for cellular radio networks. *Wireless Networks*, *6*(6), 421-428.

Customer churn prediction: Integration of sociometric theory of cliques into a diffusion model

Uroš Droftina[A], Andrej Košir[B]

[A]Telekom Slovenije, d.d., Slovenia,
[B]Faculty of Electrical Engineering, University of Ljubljana, Slovenia
[A]uros.droftina@telekom.si, [B]andrej.kosir@ldos.fe.uni-lj.si

Abstract

This paper presents current state-of-the-art diffusion model used in churn prediction modelling and faces with some of its drawbacks. A more sophisticated approach to determining initial energy of users is proposed, based on direct connections between users with focus on maximal sociometric cliques. The proposed approach is experimentally evaluated on real world data from Slovenian mobile service provider. Results indicate that substantially improved performance is achieved using proposed approach for initial energy determination in SPA model compared to the basic model.

I. Introduction

The area of churn prediction modelling has received much attention in the last two decades, especially in the telco market. The first known approach in this area uses machine learning through data mining techniques [5]. A prediction model is constructed based on previous user behaviour and used to predict events, such as churn. An exhaustive overview of these techniques is presented in [7]. The problem of this approach is that it only considers individual user attributes. It is believed that user decisions such as churn are commonly influenced by other connected users.

With the evolving of social networks in the recent years considering social ties has proven to be very promising in churn prediction. Some researchers in this area recommended and evaluated the usage of diffusion models for the purpose of modelling influence spread and consequently spread of churn. Dasgupta et al. [3] proposed a spreading activation-based technique (further addressed as "SPA" algorithm) that predicts potential churners based on their corresponding social network, including information on users that already churned. The idea is that users, who recently churned, influenced with their decisions the other users in their social neighbourhood. Underlying algorithm assigns all recent known churner the same initial positive non-zero energy (e.g. 1) while all other users start with zero energy. An energy spreading technique is then initialized where in each iteration active nodes (with non-zero energy) transfer a portion of their energy to their neighbours, relative to the strength of connections between nodes. Overall amount of energy stays the same throughout the whole process. Iterating process continues until a stable state is achieved. Afterwards a simple threshold-based technique is used where a threshold T is fixed and nodes with energy greater than T are labelled as potential churners while other are labelled as non-churners.

II. Methods

Presented SPA algorithm is an effective approach to determining potential churners influenced by recent churners that are strongly connected to them. However certain issues arise after a detailed study of SPA algorithm. One of the issues we address and try to solve in this paper is that all seed churners are assigned the same energy on the beginning. It is known that users have very different influence in their corresponding social neighbourhoods [4] which means that starting energy should be different among users. An old saying says that it is better to have a few true friends than lots of fake friends. In that aspect we hypothesise that users with fewer connections are more influential than users with many connections. This is already partially implemented in original SPA diffusion model at transfers of seed-node energy to their connections, where smaller number of connected nodes means more transferred energy per user, but our results show that influence transfer ratio between less and more connected users should be greater.

Users churn because of different reasons. Reasons can be basically divided into two categories, local reasons (e.g. non-optimal price, poor coverage) and social reasons (influence from other users in social network). We believe that if user's reasons to churn are social, these users will also have greater influence on other users in their neighbourhood as opposed to users that churned due to local reasons. This is our second hypothesis.

Our third hypothesis says that even if users have many connections, they can also have a few very good friends that have considerably greater influence on each another than on other users outside their small groups. These groups can be determined by searching maximal cliques [1] on a corresponding social network graph. In the area of graph theory, cliques are such subsets of undirected graphs that every two nodes in the subset are connected by an edge. Furter maximal clique is a clique that cannot be extended by including another adjacent node. In a social network of people where connections represent acquaintances, a clique is a group of people who all know each other. Social science theory states that people in the same clique have greater influence on each other than on or from connections outside cliques [2]. In this paper we propose a more sophisticated approach to determining initial energy of users, based on their direct connections with focus on maximal cliques. The goal of this paper is to introduce and evaluate the improvement of SPA model using the application of social science clique theory for initial energy determination.

Our proposed model assigns all users three different types of contribution to initial energy where each contribution is given as a pair $(energy, weight)$:

1. self contribution
2. clique contribution
3. out-of-clique contribution

Self contribution is the same energy contribution as in original SPA model, i.e. if a user recently churned, we assign him a self-contribution energy $E_{self} = 1$, else $E_{self} = 0$. Weight of a self contribution is determined as 1.

Clique contribution is an energy contribution from all maximal cliques a user is a part of. Due to simplicity reasons we further address maximal cliques only as cliques. Energy contribution of each user in a clique is determined by equation (1)

$$E_{cl\,n} = 2\frac{n_c}{n_u} - 1 \tag{1}$$

where n_c is the number of churners in a clique and n_u in the number of all users in a clique. Each clique assigns a separate contribution to the energy of a user. Weight of a clique contribution is usually greater than 1 (e.g. proportional to number of users in a clique).

Out-of-clique contribution is an energy contribution of all users that are not in any clique. Out-of-clique energy contribution is also calculated by equation (1), where a user with its directly connected nodes is considered instead of a clique. We determine the value of weight of an out-of-clique contribution as 1. Users that are members of at least one clique do not have an out-of-clique contribution.

$E_{cl\,n}$ in equation (1) can take values on interval $[-1, 1]$ where value -1 describes a clique without churners and value 1 describes a clique with all churners. As a consequence a negative influence is also introduced, which symbolically spreads influence against churn decision.

When all the contributions are assigned, an initial energy value is calculated as a scalar product of energy- and weight-vector (2)

$$E_0 = \mathbf{e} \cdot \mathbf{w} \tag{2}$$

where $\mathbf{e}$ is an energy vector $\mathbf{e} = [e_{self}, e_1, e_2, ..., e_n]$ and $\mathbf{w}$ is a weight vector $\mathbf{w} = [w_{self}, w_1, w_2, ..., w_n]$. The sign of initial energy symbolically determines positive and negative churn influence.

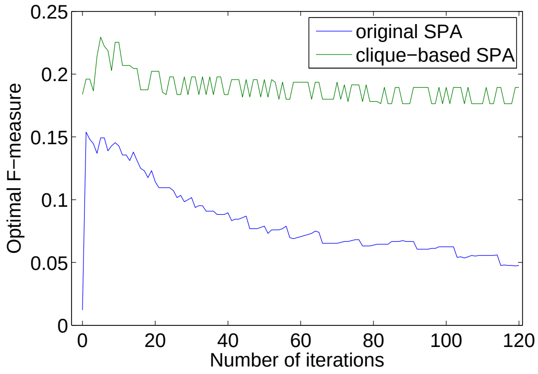


Figure 1: *F-measure values in each iteration of SPA*

III. Results

In this study, we compare the classification performance of SPA model using original initial energy determination as in [3], and in this paper proposed algorithm. We evaluate the results using standard evaluation method, F-measure (or F-score).

For our experiment we used the data from the largest Slovenian mobile service provider with more than one million users. Due to the computational complexity of the problem only a subset of over 5000 connected users was used in our experiment. A social graph was constructed from a one month period call detail records (CDR) where nodes presented users and connections presented communication between users. Connections were weighted by sum of successful calls and sent short text messages where only connections with both-way communication were included (i.e. user A at least once called or sent an SMS to user B, and vice versa).

Evaluation of both approaches is presented in figure 1. Figure presents how optimal F-measure value changes with the number of SPA iterations applied.

IV. Discussion

Our preliminary results clearly show a significant improved in churn prediction compared to basic SPA diffusion model. An inspection of original SPA model trend line in figure 1 shows that highest prediction value is already achieved after first iteration. This indicates that users are considerably more influenced by their directly connected users rather than other users that are not in their neighbourhood. Iterating original SPA diffusion model until convergence is therefore not optimal.

Prediction evaluation of our approach is already relatively high when using initial values as prediction scores. After running first few iterations of SPA algorithm using our initial values prediction result also slightly improves. We believe that improvement of our algorithm can be made where the best prediction accuracy will be achieved before applying a process of energy spreading.

In computer science, finding cliques in a graph is known to be NP-hard problem. Therefore even by avoiding an iterative process of SPA algorithm completely, clique discovery still remains computationally most expensive part of algorithm. Using advanced fast algorithms for finding cliques in social graphs addresses this issue [6].

References

[1] Alba, R.D.: A graph-theoretic definition of a sociometric clique. The Journal of Mathematical Sociology **3**(1), 113–126 (1973)

[2] van den Berg, Y.H., Cillessen, A.H.: Computerized sociometric and peer assessment an empirical and practical evaluation. International Journal of Behavioral Development **37**(1), 68–76 (2013)

[3] Dasgupta, K., Singh, R., Viswanathan, B., Chakraborty, D., Mukherjea, S., Nanavati, A.A., Joshi, A.: 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 (2008)

[4] Kempe, D., Kleinberg, J.: Influential nodes in a diffusion model for social networks. Automata, Languages and (2005)

[5] Mozer, M.C., Wolniewicz, R., Grimes, D.B., Johnson, E., Kaushansky, H.: Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry. IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council **11**(3), 690–6 (2000)

[6] Östergård, P.R.: A fast algorithm for the maximum clique problem. Discrete Applied Mathematics **120**(1-3), 197–207 (2002)

[7] Verbeke, W., Dejaeger, K., Martens, D., Hur, J., Baesens, B.: New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. European Journal of Operational Research **218**(1), 211–229 (2012)

Characterizing social network spatio-temporal structure through Cell Phone Records

Luis G. Moyano, Enrique Frias-Martinez,
Telefónica Research, Madrid, Spain
moyano,efm@tid.es

Oscar R. Moll Thomae
MIT, Boston, MA, USA
rimoll@mit.edu

Abstract

The connection between human mobility and social networks has gained much attention lately, but research in this area is limited due to lack of adequate data that describes both mobility and social interactions. In this work, we characterize spatio-temporal features of social networks taken from a comprehensive, national-wide dataset of cell phone records. Our results show a non-trivial dependence between social network structure and the spatial distribution of its elements (Fig 1., left panel). Moreover, we describe and quantify precisely the probability of parties in a call to be at a given distance. Our results show that this probability is well described by the framework of gravity models, but with different decaying rates at urban and interurban scales (Fig. 1, right panel). Finally, we discuss how the structural information gained may be used to estimate a class of unknown features of the network.

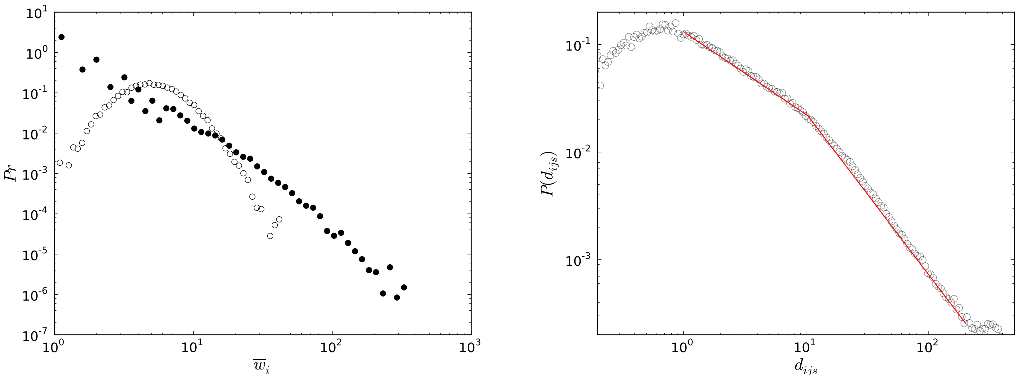


Figure 1: Left panel: Probability distribution of the mean number of calls $\bar{n}_i$ given that $k_i \geq 45$ (open dots) and $k_i \leq 3$ (solid dots). Right panel: Probability distribution of the distance d_{ijs} (in km) associated to a call. The region approximately between 1 km and 10 km may be described by a power-law decay with exponent $\alpha = -0.77$, while the region from 10 km onward shows a faster decay consistent with an exponent $\alpha = -1.5$.

Ethnic segregation in daily spatial mobility: Call Detail Record based study in Estonia

Rein Ahas[1,2], Siiri Silm[1], Erki Saluveer[3]

[1]Department of Geography, University of Tartu, 46 Vanemuise St., Tartu 51014, Estonia, rein.ahas@ut.ee; siiri.silm@ut.ee

[2]Department of Geography, Ghent University, Krijgslaan 281, S8, B9000 Gent - Belgium

[3]Positium LBS, Õpetaja 9, Tartu, Estonia, erki.saluveer@positium.ee

Abstract

Ethnic segregation remains a topical issue, with globalisation leading to increasing migration flows and ethnic tensions escalating in a number of regions. The need to examine the entire scope of activities of individuals in segregation studies is often emphasised by scientists and policy makers. We used mobile phone Call Detail Records (CDR) for one year to compare the spatial mobility of native Estonians and the Russian-speaking minority in Estonia and international travel. The results show that ethnicity has a significant influence on the spatial mobility of individuals in Estonia. The biggest differences between the two population groups occur in Estonia outside the respondents' home city of Tallinn where the Russian minority were found to visit 45% fewer districts than Estonians. For international travel, the Russian-speaking minority visit fewer countries and have a 3.6 times higher probability of visiting former Soviet Union countries than Estonians. Our results show that ethnic segregation has less effect on everyday spatial mobility and a greater influence on the choices made regarding long-distance travel.

Reality Commons: Rich Datasets for Mobile Network Science

Human Dynamics Lab, MIT
http://realitycommons.media.mit.edu

Cell phones have become an important platform for the understanding of social dynamics and influence, because of their pervasiveness, sensing capabilities, and computational power. Many applications have emerged in recent years in mobile health, mobile banking, location based services, media democracy, and social movements. With these new capabilities, we can potentially be able to identify exact points and times of infection for diseases, determine who most influences us to gain weight or become healthier, know exactly how information flows among employees and productivity emerges in our work spaces, and understand how rumors spread.

There remain, however, significant challenges to making mobile phones the essential tool for conducting social science research and also supporting mobile commerce with a solid social science foundation. Perhaps the greatest challenge is the lack of data in the public domain, data large and extensive enough to capture the disparate facets of human behavior and interactions. Another major challenge lies in the interdisciplinary nature of conducting social science research with mobile phones. Software engineers need to work collaboratively alongside social scientists and data miners in various fields.

In an attempt to address these challenges, we release several mobile data sets here in "Reality Commons" that contain the dynamics of several communities of about 100 people each. We invite researchers to propose and submit their own applications of the data to demonstrate the scientific and business values of these data sets, suggest how to meaningfully extend these experiments to larger populations, and develop the math that fits agent-based models or systems dynamics models to larger populations. Data, code, and documentation can be found at http://realitycommons.media.mit.edu

These data sets were collected with tools developed in the MIT Human Dynamics Lab and are now available as open source projects (see the funf open-source sensing platform for Android phones,http://funf.media.mit.edu) or at cost (e.g., the sociometric badges for sensing organizational behavior, see http://sociometricsolutions.com)