NetMob 2015

7-10 April 2015
MIT MediaLab
http://netmob.org

Book of Abstracts :: Oral

Editors: Esteban Moro, Yves-Alexandre de Montjoye, Vincent Blondel, Alex 'Sandy' Pentland

March 26th version

Organized by

Universidad Carlos III de Madrid

UCL Université catholique de Louvain

Sponsored by

orange

RealImpact ANALYTICS

Contents

Session 1 :: Mobility

Analyzing the Influence of Phone Context Data on the Performance of Human Mobility Predictors

Paul Baumann, Silvia Santini
Embedded Systems Lab, TU Dresden, Germany
{paul.baumann, silvia.santini}@tu-dresden.de

Christian Koehler, Anind Dey
Carnegie Mellon University, USA
{ckoehler, anind}@cs.cmu.edu

1. INTRODUCTION

Understanding and predicting human mobility has been already for a long time in the focus and interest of researchers and practitioners [4, 2, 5]. Over the years, the potential sources of information about human mobility, e.g., temporal data, calendar information, or social ties, have been grown dramatically.

In the context of this extended abstract, we focus on the phone context data, e.g., number of recently used applications or time since the last received or made phone call. By considering this new source of data, we analyze its influence on the performance of 4 state-of-the-art and 3 baseline predictors for 3 human mobility prediction tasks. We capture the results of the potential influence by considering 3 well-known performance metrics. Our results highlight that our use of the phone context data *does not* lead to signification performance improvements. Statistical information from the data set shows that users used 7 applications, and made/received 5 phone calls, on average per day. This highlights the fact that the phones have been used intensively by their owners.

Our contributions are two-fold: (1) we derive an extended list of 28 phone context features, and (2) we analyze their potential influence on all combinations of the considered 3 prediction tasks, 7 predictors, and 3 performance metrics.

2. METHODOLOGY AND BACKGROUND

In order to investigate the influence of phone context data on the performance of several predictors for human mobility, we analyze a rich data set – Nokia Lausanne Data Collection Campaign (LDCC) [3] – that contains information from 141 users collected over 18 months. We first derive a list of meaningful phone context features based on the available information in the data set. After that, we run feature selection to reduce the amount of derived features for all combinations of predictors, prediction tasks, and metrics. We evaluate the performance of the resulting combinations of a predictor and a set of features on well-known metrics – accuracy, F1 score, and Matthews Correlation Coefficient (MCC).

2.1 Features

The human mobility predictors that are used throughout this work need adequate input data to be able to compute the required prediction. To this end, we derive a list of 51 features that contains 28 phone context features – indicated as set $\mathcal{F}_{pc}$. Table 1 shows them along with their corresponding description.

2.2 Prediction Tasks

In the context of this work, we consider 3 prediction tasks. The *Next-place (NP)* prediction task only considers the next place visited and treats as irrelevant when the user moves to the next location and how long she stays in each place. Timing information can easily be included in a prediction task by considering equally spaced time slots of length s and computing a new next place prediction for each time slot. We refer to this task as the *Next-slot place (NSP)* prediction task. Finally, the *Next-slot transition (NST)* prediction task consists of estimating, at time slot k, whether or not there will be a *transition* at time slot $k+1$. A transition occurs when the user moves between two places.

[1] We consider calls and messages for both incoming and outgoing directions.

[2] We consider date of creation, status, title, location, type, and confidence class.

Table 1: List of phone-context features considered in this study.

Label	Description
`c_time_call`[1]	Time since last call/sms made/received
`c_calllog_type`	Last calllog type
`c_calllog_direction`	Last calllog direction
`c_sms_status`	Last sms status
`c_last_call_duration`	Last call duration
`c_last_cal`[2]	Information about last calendar entry
`c_next_cal`[2]	Information about next calendar entry
`c_time_last_app`	Time since last application used
`c_last_app`	Last used application
`c_phone_charging`	Current phone charging status
`c_last_charge`	Time since last charge
`c_battery`	Current phone battery status
`c_ring`	Current ring profile
`c_profile`	Current user profile
`c_last_action`	Time since last phone interaction

2.3 Predictors

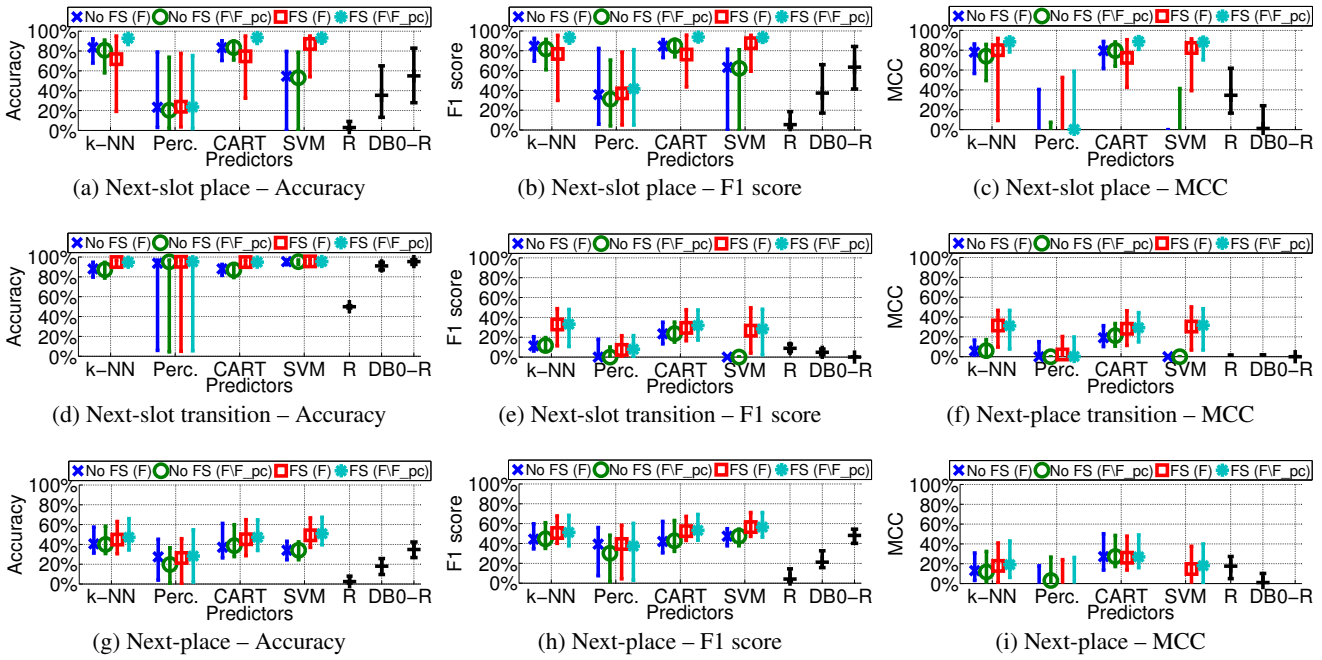
In this study, we consider 7 predictors that rely on different basic techniques, have different weaknesses and strengths, and require different amounts of computational and memory resources. The 7 predictors include 4 well-known predictors – Support Vector Machine (SVM) [6, 1], k-Nearest Neighbor (k-NN) [6, 1], Classification and Regression Trees (CART) [6], and Perceptron [1] – as well as 3 baseline predictors – Random predictor (R), Distribution-based predictor (DB), and 0-R predictor (0-R).

3. INFLUENCE OF PHONE CONTEXT DATA ON PREDICTORS' PERFORMANCE

For our analysis part, we adopt the Sequential Forward Floating Selection (SFFS) algorithm to identify the best performing features for all combinations of prediction tasks, metrics, and predictors. In all these cases, we differentiate between using the entire feature set $\mathcal{F}$ and using only a reduced feature set by excluding phone context features ($\mathcal{F} \backslash \mathcal{F}_{pc}$). We next present our findings.

3.1 Performance Results

The feature selection process returns an optimal feature subset $\mathcal{F}'$ for each combination of user, predictor, prediction task, and metric. Figure 1 shows the performance in terms of accuracy, F1



(a) Next-slot place – Accuracy (b) Next-slot place – F1 score (c) Next-slot place – MCC

(d) Next-slot transition – Accuracy (e) Next-slot transition – F1 score (f) Next-place transition – MCC

(g) Next-place – Accuracy (h) Next-place – F1 score (i) Next-place – MCC

Figure 1: Performance results for all 3 considered prediction tasks, 3 performance metrics, and 7 predictors.

score, and MCC achieved by the predictors with or without the feature selection ("FS" vs "No FS"). The markers indicate median values and the whiskers indicate the 5th and 95th percentiles.

Figure 1a, Figure 1b, and Figure 1c show the median performance for the NSP prediction task. The plots for the metrics accuracy and F1 score reveal significant performance differences for the extracted feature sets with phone context features ($\mathcal{F}_{pc}$) and those without. It holds for the predictors k-NN and CART. In these cases, the performance after applying feature selection is higher for the feature set $\mathcal{F}\backslash\mathcal{F}_{pc}$ than for $\mathcal{F}$. It is also the case for MCC, but with a lower performance difference in the case k-NN is used. Furthermore, SVM shows similar performance for both feature sets ($\mathcal{F}$ and $\mathcal{F}\backslash\mathcal{F}_{pc}$) after applying feature selection. We observe that on the one hand SVM is able to leverage SFFS to extract meaningful subsets of features. On the other hand, SVM reveals dramatic performance drops if no feature selection is applied. At the same time, Perceptron does not outperform the distribution-based (DB) and the 0-R predictors, which are both just baseline predictors.

For the NST prediction task, we observer in Figure 1d, Figure 1e, and Figure 1f that all predictors achieve a high accuracy. It is not surprising since the class of "No Transitions" dominates with almost 95%. Only Perceptron fails for a number of users to achieve a high accuracy. In the case of the F1 score and MCC the predictors k-NN, CART, and SVM achieve similar performance after applying feature selection to both sets – with and without the phone context data. It is worth to note that for both metrics – F1 score and MCC – the predictor CART tends to be much more robust in the cases with no feature selection by achieving at least twice as high performance than the next best performing predictor.

Last but not least, Figure 1g, Figure 1h, and Figure 1i demonstrate results for the NP prediction task. For k-NN, CART, and SVM we make two observations. First, the feature selection leads to performance improvements in terms of all 3 metrics. Second, the consideration of the phone context data does not show any signification improvements in terms of the considered metrics.

4. CONCLUSIONS AND FUTURE WORK

We summarize our results and conclude that our use of the phone context features *does not* lead to signification performance improvements. However, phone context data is in general a rich information source. Statistical information from the data set highlights the presence of the potentially meaningful context data and the fact that the phones have been used intensively by their owners. Thus, we believe that instead of capturing temporal phone information, e.g., time since last action X, analysis on the correlation between the appearance of phone data events, e.g., received a phone call, and a corresponding mobility behavior may uncover additional potential for prediction improvements. In the context of this work, we leave the proof of this hypothesis as future work.

5. ACKNOWLEDGEMENTS

This work has been partially supported by the Collaborative Research Center 1053 funded by the German Research Foundation.

6. REFERENCES

[1] C. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[2] J. Krumm and D. Rouhana. Placer: Semantic Place Labels from Diary Data. In *UbiComp'13*, 2013.

[3] J. Laurila *et al.* The Mobile Data Challenge: Big Data for Mobile Computing Research. In *Pervasive'12*, 2012.

[4] P. Baumann *et al.* The Influence of Temporal and Spatial Features on the Performance of Next-place Prediction Algorithms. In *UbiComp'13*, 2013.

[5] V. Srinivasan *et al.* Mobileminer: Mining Your Frequent Patterns on Your Phone. In *UbiComp'14*, 2014.

[6] X. Wu *et al.* Top 10 Algorithms in Data Mining. *Knowledge and Information Systems*, 14(1):1–37, 2008.

Evaluating urban sensing applications using actively and passively-generated mobile phone location data

Fabio Pinelli, Giusy Di Lorenzo, Francesco Calabrese
IBM Research – Ireland
Email: {fabiopin, giusydil, fcalabre}@ie.ibm.com

***Abstract*—Mobile phone location data from telecom operators in the form of Call Detail Record (CDR) has been widely studied, especially to extract insights into urban dynamics [5]. Such massive data can be useful to extract patterns of human mobility at an incredible scale. This data sample the location of a mobile device every time is actively interacting with the network, e.g. at call time, while sending and receiving a SMS, or while connecting to the Internet with Smartphones. The disadvantage of such data collection method is that the spatio-temporal sampling of each individual user trajectory over time might be very uneven, and perhaps biased to specific locations (e.g. home locations) or times (e.g. during the evenings). Moreover, some users might interact more or less with the network, resulting in more or less mobility information extracted from them. This could result in under-sampling the population, or more problematically, biasing the extracted insights. In the past decade, there has been a rising interest in using mobile phone location data to infer user trajectories [1], [8], and so study human mobility and their patterns [7]. Different types of data have been used in these studies. CDR data were used in [7],[2],[6]. CDR information, enriched with records from Internet access was exploited in [4]. Data from idle phones were also used in [9] to estimate the road traffic. However, to the best of our knowledge, no work so far has specifically compared the different types of datasets that a telecom operator can collect. Moreover, no work so far has analysed the limitations of using a specific dataset for a given urban sensing application. We then ask the question whether insights extracted from actively collected-mobile phone location data are a good proxy for human mobility. To answer this question, we compare such results, with results extracted by both actively and passively sampled user location, which constitute a richer set of location information.**

We used a real dataset collected from a telecom operator in Belgium, which had a system which allowed to collect both CDR, records of Internet connections (which we call IPDR), and passively generated data (which we call Signaling), generated because of location updates, radio access network or data sessions. We were able to decompose the dataset in three different ones: only CDR, CDR + IPDR, and all data, since each record event was tagged with the specific type of event generating it.

The dataset contains anonymised mobile phone location data from Mobistar, for users in the area around the city of Mons, Belgium. The data include users connected to 150 distinct cell towers in the city area. For each cell tower, we were given the coordinate and the azimuth of each cell sector. Thus we were able to derive a voronoi tessellation of the space, following the approach presented in [3]. Some of the cells cover the same area (i.e. 2G and 3G antennas installed on the same tower), therefore we were able to identify 58 distinct locations in the city.

The available data covers one week in October 2014. We use the available data to simulate 3 different scenarios:

- availability of only CDR information, in which we only use the CALL and SMS data items, and are representing cases in which only CDR information is provided.
- availability of CDR and IPDR, in which we use the above data, together with IPDR, to represent cases where all Event-driven signaling information is provided.
- availability of all signaling information (CRD+IPDR+Signaling), which will be our reference for comparison.

Figure 1 depicts an example of temporal sequence of events for a user in the dataset. We also depict the trajectory that we are able to detect, given the tree different scenarios. The example clearly shows that for this user, the availability of all information allows detecting 3 different visited locations, and an estimated stop time for locations 2 and 3. If no Signaling information is available, only two visited places could be detected, and the estimated stop time would also be reduced, with lowest accuracy in the case of only CDR information available.

At the general level, the advantages of using Network-driven location data (in addition to event-driven) include:

- sampling more users (people who are not making calls/SMS/Internet connections);
- having more samples of user locations, particularly at times where users are not too active, e.g. at night);

Motivated by this example, in the following sections we quantitatively and qualitative compare properties of the three datasets. This is firstly done by extracting application-independent characteristics. Then, we selected frequently used urban sensing applications designed for Telco data, and compared the accuracy of the extracted insights among the different datasets, highlighting in which cases one dataset is preferable compared to the others.

I. Application-independent comparison

We compare the three datasets along different dimensions. A first one is in terms of the set of users for which we have data. We counted, for each user, the number of the three different types of events (CDR, IPDR and Signaling). Figure 2 shows a Venn diagram of the unique users by data

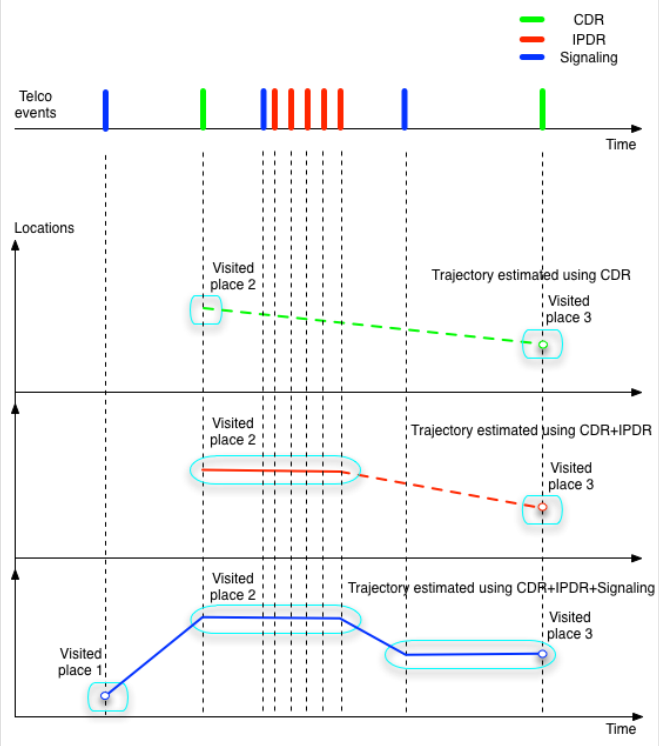


Fig. 1. Example of temporal sequence of events (top), and estimated trajectories using three different datasets (bottom). To simplify the reading, the locations have been drawn in one dimension (as opposed to the two dimensions (latitude and longitude).

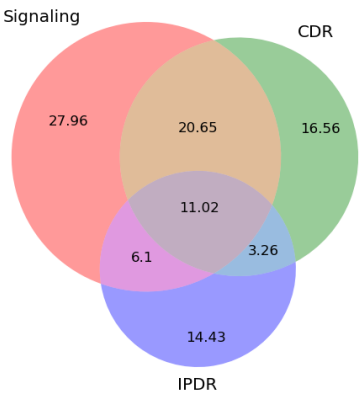
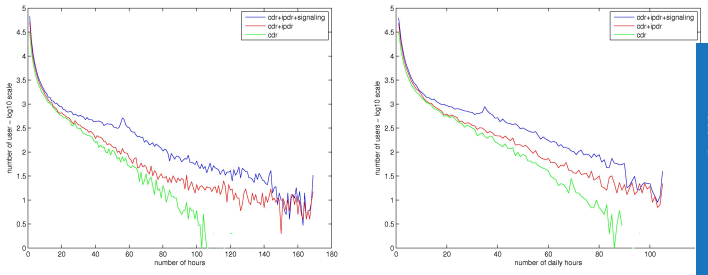


Fig. 2. Percentage of users per data type and relative intersections.

type. Only for about 11% of the users we can see all three different types of events. This is due to that fact that not all users have smartphones for which IPDR can be generated. Moreover, some users are only observed very temporary in the dataset (users only traversing the city), and so only Signaling information is available. Not all users generate the same number of events. The number of events by user follows a long tail distribution. Clearly considering only CDR or CDR+IPDR events, the average number of events per user is smaller. We



(a) all hours (b) daylight hours

Fig. 3. Distribution of user by number of hours for which at least one event

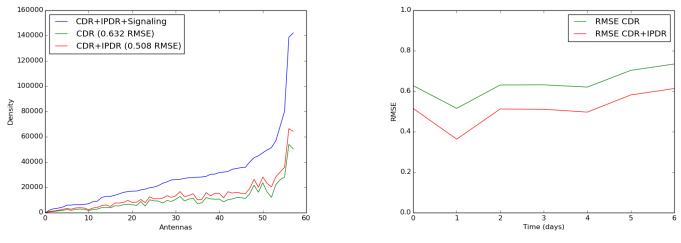
then ask the question whether this decrease in the number of events is concentrated in particular hours of the day, or is equally spread over time. Figure 3(a) show the distribution of users by number of distinct hours for which there is at least one event. Curves CDR and CDR+IPDR look very close, to show that the large amount of IPDR events are on average concentrated in the same hours as the CDR events. Signaling events are able to sample the user location over many more hours. However, if we only consider daily hours (from 6 to 22), as reported in Figure 3(b) we notice that difference decreases. This let is hypothesise that actively generated location data are a good sample of user location during daylight hours. We will verify this intuition in the next section.

II. Application-dependent comparison

We have taken a frequently used example of urban sensing application using mobile phone location data: the count estimation over time, such as the number of people being in a certain cell in a given time interval. This information is highly relevant for many sectors, such as Retail, Property, Leisure and Media, since it allows to compare locations in terms of expected crowd. Clearly, an accurate estimation of the time series of number of users by location is crucial to provide trustable insights. We computed user count time series for each location in the city, starting from the three different datasets.

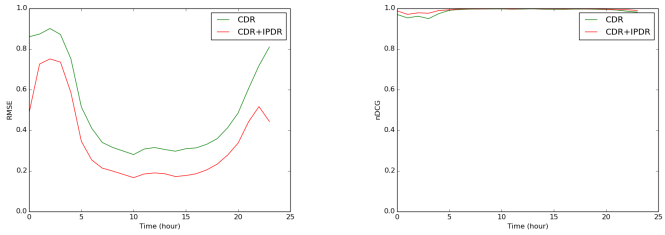
Cumulative count estimation by location is show in Figure 4(a), ranked by increasing value of count (based on the reference dataset). We computed two measures of error. The first measure is the root mean square error (RMSE) computed on the hourly estimates compared to the counts estimated using all data (CDR+IPDR+Signaling). The error ranges from 0 to 1, and low values correspond to low error. Moreover, we computed a measure of Normalised discounted cumulative gain (nDCG) [10] used in recommender systems to measure rank quality. We have chosen this measure to evaluate whether the estimated ranking of crowded locations is kept the same by using CDR or CDR+IPDR only information. The error ranges from 0 to 1, and high values correspond to low error. This measure is different from the RMSE, since it does not take into account the absolute estimate count for each location or time, but just the relative order of such counts by location. This measure is directly useful in application scenarios such as choosing the most crowded place between a set of locations.

As Figure 4(a) shows, the error is higher if we only consider CDR information. Moreover error measured in terms of nDCG is much lower, and there is no much difference in using



(a) by antenna (b) error by day of week

Fig. 4. Density estimation



(a) RMSE (b) nDCG

Fig. 5. Density estimation error by hour of day

CDR or CDR+IPDR data. This shows that, for the purpose of comparing user counts by locations, CDR information in on average a good source of information. Figure 4(b) shows the average RMSE computed on the 7 distinct days. As it can be seen, the error is higher over the weekend (last 2 days). Moreover, Figure 5 shows the errors as function of the hour of day (averaged over all days). It is interesting to see that error is large over the night hours, and quite low during the daylight hours. In conclusions, we can observer that using CDR or CDR+IPDR as proxy for user count per location works relatively well in application scenarios such as choosing the most crowded place between a set of locations, especially if the focus is on daylight hours over weekdays.

While some of the reported results might depend on the specific configuration of the monitoring system that the telecom operator used, the methodology we presented to evaluate the limitation of each dataset is general. Moreover, it can be applied by other telecom operators to measure the effectiveness of using each individual dataset for urban sensing applications. In addition, we plan to extend the analysis to others, and more complex, urban sensing applications, such as event detection, trajectory pattern extraction and trajectory clustering.

Acknowledgment

The authors would like to thank Mobistar for providing access to the anonymized data.

References

[1] T. Bao, H. Cao, Q. Yang, E. Chen, and J. Tian. Mining significant places from cell id trajectories: A geo-grid based approach. In *Mobile Data Management (MDM), 2012 IEEE 13th International Conference on*, pages 288–293, July 2012.

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

[3] R. Caceres, J. Rowland, C. Small, and S. Urbanek. Exploring the use of urban greenspace through cellular network activity. In *Proc. of 2nd Workshop on Pervasive Urban Applications (PURBA)*, 2012.

[4] F. Calabrese, G. Di Lorenzo, L. Liu, and C. Ratti. Estimating origin-destination flows using mobile phone location data. *Pervasive Computing, IEEE*, 10(4):36 –44, april 2011.

[5] F. Calabrese, L. Ferrari, and V. D. Blondel. Urban sensing using mobile phone network data: A survey of research. *ACM Comput. Surv.*, 47(2):25:1–25:20, Nov. 2014.

[6] G. di lorenzo, M. L. Sbodio, F. Calabrese, M. Berlingerio, R. Nair, and F. Pinelli. Allaboard: Visual exploration of cellphone mobility data to optimise public transport. In *Proceedings of the 19th International Conference on Intelligent User Interfaces*, IUI '14, pages 335–340, New York, NY, USA, 2014. ACM.

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

[8] S. Hoteit, S. Secci, S. Sobolevsky, G. Pujolle, and C. Ratti. Estimating real human trajectories through mobile phone data. In *Mobile Data Management (MDM), 2013 IEEE 14th International Conference on*, volume 2, pages 148–153, June 2013.

[9] 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, New York, NY, USA, 2012. ACM.

[10] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, Oct. 2002.

OpenStreetCab: Exploiting Taxi Mobility Patterns in New York City to Reduce Commuter Costs

Vsevolod Salnikov,[1] Renaud Lambiotte,[1] Anastasios Noulas,[2] and Cecilia Mascolo[2]
[1] *naXys, University of Namur, Belgium*
[2] *ComputerLab, University of Cambridge, UK*

The rise of Uber as the global alternative taxi operator has attracted a lot of interest recently. Aside from the media headlines which discuss the new phenomenon, e.g. on how it has disrupted the traditional transportation industry, policy makers, economists, citizens and scientists have engaged in a discussion that is centred around the means to integrate the new generation of the sharing economy services in urban ecosystems. In this work, we aim to shed new light on the discussion, by taking advantage of a publicly available longitudinal dataset that describes the mobility of yellow taxis in New York City. In addition to movement, this data contains information on the fares paid by the taxi customers for each trip. As a result we are given the opportunity to provide a first head to head comparison between the iconic yellow taxi and its modern competitor, Uber, in one of the world's largest metropolitan centres. We identify situations when Uber X, the cheapest version of the Uber taxi service, tends to be more expensive than yellow taxis for the same journey. We also demonstrate how Uber's economic model effectively takes advantage of well known patterns in human movement. Finally, we take our analysis a step further by proposing a new mobile application that compares taxi prices in the city to facilitate traveller's taxi choices, hoping to ultimately to lead to a reduction of commuter costs. Our study provides a case on how big datasets that become public can improve urban services for consumers by offering the opportunity for transparency in economic sectors that lack up to date regulations.

I. TAXI PRICE COMPARISON EXPERIMENT

The New York City Taxi Dataset. The Freedom of Information Law in United States encourages public authorities to release their data where appropriate to the benefit of the citizens. In 2014 the law was exploited by Chris Whong to acquire and post on the web one of the most comprehensive taxi mobility datasets available today. The dataset describes taxi journeys in New York City during the full course of 2013, and informs us not only on the origin and destination points of taxi trips, noted in the related jargon as pick up and drop off points respectively, but also on the financial costs incurred to the customer (trip fair) with unprecedented detail. This rather dense mobility dataset, containing hundreds of millions of trips is of gigabytes in size and can be downloaded here `http://chriswhong.com/open-data/foil_nyc_taxi/`. A sample of the traces generated by the data can be seen is in Figure 1, where we have drawn a black point for every pick up and drop off point of a taxi journey.

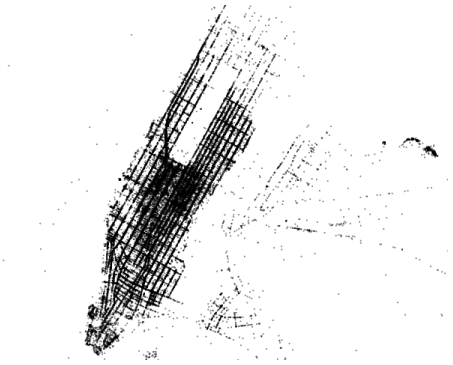


FIG. 1: Marking the traces of new york city yellow taxis. For every pick up and drop off point in a uniform sample of the data we draw a black point.

Comparing Taxi Prices In August 2014, Uber opened up an API with access to valuable information about its services. The occasion allowed us to perform a first head to head comparative analysis of prices between Uber and Yellow taxis in New York City. To achieve this we run the following experiment :

- 1. For every trip in the New York City Yellow Taxi dataset, record the geographic coordinates (latitude and longitude) of the pick up and drop off points.
- 2. Retrieve the total fare paid by the customer for the trip (including the tip).
- 3. Query Uber's API and ask how much they would charge for the same trip (same pick up and drop off points), considering the cheapest version of the service, Uber X.
- 4. Uber's API returns a value range indicating the minimum and maximum price estimate. We take the mean of the two values.
- 5. We then compare the prices from the two services.

As can be observed in Figure 2 where the distribution of prices for the two services is shown, despite the qualitative similarity of the two distribution, yellow taxi appear on average (median) 1.4 US dollars cheaper than Uber X. In Figure 3, we compare Uber and yellow taxis from another perspective: for every observed yellow taxi

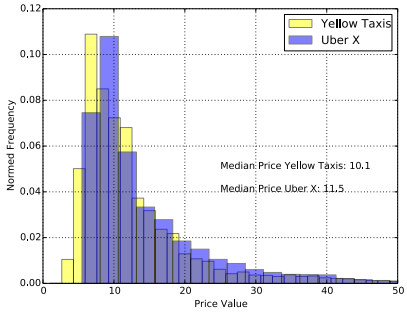


FIG. 2: Distribution of prices per journey for Uber X and Yellow Taxis in New York City.

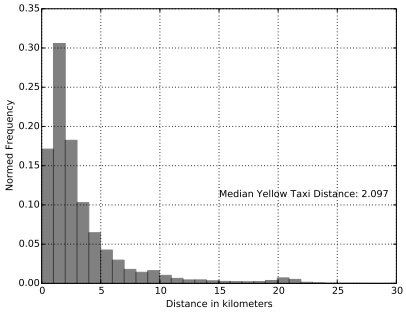


FIG. 4: Distribution of geographic distances between drop off and pick up points for Yellow Taxi journeys.

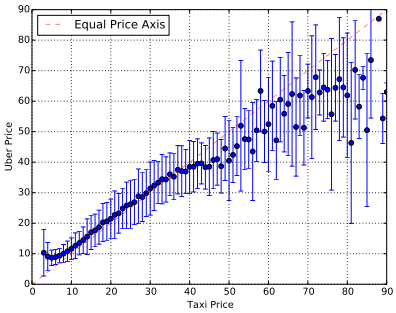


FIG. 3: Median Uber price for a given Yellow Taxi price.

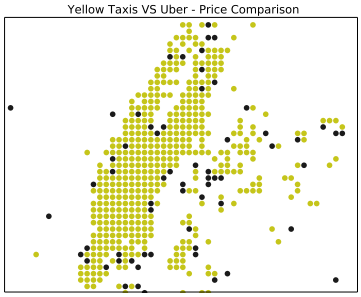


FIG. 5: Geographic comparison between Uber and Yellow Taxi prices. We paint an area black if Uber is cheaper by trip majority and yellow otherwise.

price, we show the median Uber X price. Uber appears more expensive for prices below 35 dollars and begins to become cheaper only after that threshold. As one would expect, the cheaper journeys are those that are in principle of shorter range. As observed in a variety of empirical data, human mobility tends to be characterised by a vast majority of short trips [1, 2]. This observation therefore suggests that Uber's economical model exploits this trend of human mobility in order to maximise revenue. We also confirm the skewed frequency distribution of movement distances in the present context by visualising it in Figure 4, where we note a mean distance for a yellow taxi trip in New York equal to 2.09.

The above experiment may involve a number of biases which we refer to here. The NYC Yellow taxi data corresponded to year 2013 whereas Uber to 2014. Although note that the prices for yellow taxis in the city had last changed in 2012 after 8 years [3]. So it should offer a good approximation of todays prices. Further, there was no control for time of the day/week for the API query, an additional dimension which should be incorporated when available. However, we argue that the process of comparing two different companies that provide the same service in the same geographic area is of value to commuters. Just as consumer have open access to airfares for a long time now allowing for transparency in a free, competitive, market we believe that similar approaches could benefit commuters in modern cities.

II. HELPING COMMUTERS

Our observations show that it might be financially advantageous on average for travellers to chose either Yellow Cabs or Uber depending on the duration of their journey. However the specific journey they are willing to take matters. In order to help users to take the right decision, we have developed a smartphone app, called OpenStreetCab, designed as follows.

One limitation for the design of our service is that only prices for trips with origins and destinations in the New York City Taxi Dataset can in principle be retrieved. In order to evaluate the price of any trip, as needed for a usable App, we have divided the NY region into a mesh with cells of size around 100m by 100m in order to index trips in the database efficiently. For each user query, we find a set of trips in our dataset with the origin in neighbouring cells of desired origin and, among them, we find the trip whose destination is closest to the desired one. This strategy has the advantage of being sufficiently fast to perform online queries and expected to provide reliable price estimates. For the same trip, Uber price is obtained through their API.

A real-time prototype has been designed and is currently launched on popular mobile platforms. Future improvements include the possibility to change predic-

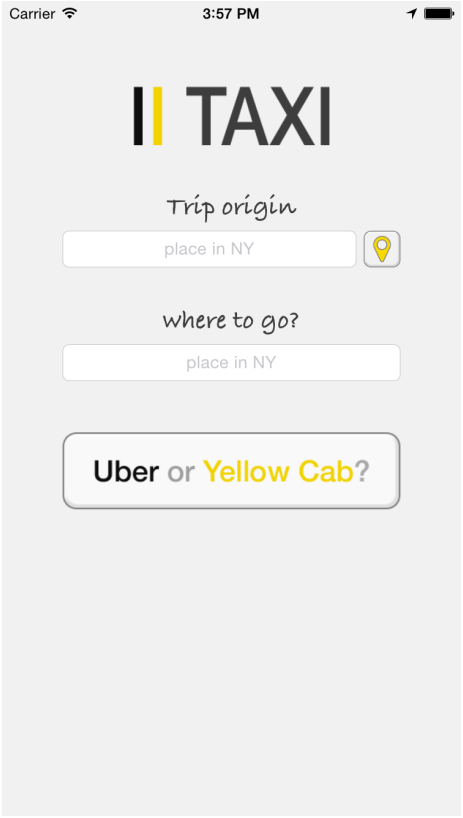


FIG. 6: The proof of the concept

tions depending on the time of the day, or on the expected traffic on the way, but also to suggest other types of transportations, such as walking when the distance is sufficiently short, or only part of the way, in situations when a small change in the origin point can lead to a significant change in the price quote. In the meanwhile the current version (Fig. 6) already provides a fully working solution, including geolocation services and address retrieval. We are planning to launch the application on the related stores very soon.

[1] Dirk Brockmann, Lars Hufnagel, and Theo Geisel. The scaling laws of human travel. *Nature*, 439(7075):462–465, 2006.
[2] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.
[3] Taxi Fares To Rise. New York Times, 2012. `http://cityroom.blogs.nytimes.com/2012/07/12/taxi-fares-in-new-york-to-rise-by-17/?_r=0`.

Assessing the impact of ride sharing on congestion using mobile phone data

Lauren Alexander[1,*], Jameson Toole[2], Marta González[1]

[1]Department of Civil Engineering and Environmental Engineering, MIT, 77 Massachusetts Ave., Cambridge, MA 02139, USA

[2] Engineering Systems Division, MIT, 77 Massachusetts Ave., Cambridge, MA 02139, USA

January 30, 2015

In recent years, smart-phone based technology has enabled ride-sharing services like UberPool and Lyft Line to match customers making similar trips in real-time [1, 2]. In this manner, ride-sharing services offer customers the option to share a ride with other users for a reduced rate and minimal increased travel time. But what are the impacts of such services on city-wide congestion? With policy makers faced with decisions about regulating the growing ride-share market, external impacts of such services are important considerations. This research explores the impact of ride-sharing on Boston area congestion using mobile phone data. We first extract average daily origin-destination (OD) trips from the mobile phone records and estimate the proportions of these trips made in vehicles. Next, we match spatially and temporally similar auto trips, and assume a range of adoption rates to distill hourly vehicles trips according to users' willingness to share rides. Finally, for each level of adoption, we evaluate roadway level of service characteristics, including volume over capacity (VOC) and congested travel time, to assess the network-wide impacts of ride-sharing.

1 Significance

Researchers have found significant overlap in taxi trips made in New York City, demonstrating the potential for real-time vehicle pooling with minimal increase in travel time [9]. Using mobile phone and social network data, researchers evaluated demand for ride-sharing between strangers, friends, and friends-of-friends in four cities in Spain and the US [4]. Unlike both of these studies, this work focuses on the network-wide congestion impacts of ride-sharing rather than determining the proportion of trips eligible for sharing given defined spatio-temporal constraints. Furthermore, we use aggregate, average daily trips in this study to estimate end-to-end[†] ride-sharing potential. Due to spatial and temporal limitations of our data set, we are not able to explicitly estimate en-route ride-sharing[‡].

2 Data

For this analysis, we use Call Detail Records (CDRs) containing more than eight billion mobile phone records for roughly two milion anonymized users in the Boston metropolitan area over two months in the Spring of 2010. Each record contains an anonymous user ID, longitude, latitude, and timestamp at the instance of a phone call or other types of phone communication (such as sending SMS, etc.). The coordinates of the records are estimated by service providers based on a standard triangulation algorithm, with an accuracy of about 200 to 300 meters.

Because CDRs contain traces of a user at approximated locations only when his/her phone communicates with a cell phone tower, they provide an inexact and incomplete picture of daily trip-making. Accordingly, much research has focused on developing methods to extract meaningful information about human mobility from mobile phone traces as well as understanding its limitations. It has been demonstrated that CDR data can be used to infer origin-destination (OD) trips using microsimulation and limited traffic count data [7]. At the level of the individual, daily trip chains/trajectories constructed from mobile phone data are consistent with household surveys [8, 10]. Further, road usage inferred from the CDR data has been validated against GPS speed data [12] and highway assignment results from a travel demand model [6]. Methodology to estimate aggregate average daily OD trips by time of day and purpose has been validated in Boston and Rio de Janeiro against national and local travel surveys [3, 5].

*corresponding author: lpalex@mit.edu

[†]ride-sharing between users with similar origins and destinations

[‡]ride-sharing between users sharing portions of their paths between dissimilar origins and/or destinations, such that additional passengers can be picked up en-route

3 Methods

We estimate average daily OD trips from triangulated mobile phone records by applying the methods developed by our team in [3, 5] to the Boston CDR dataset. CDRs are first converted into clustered locations at which users engage in activities for an observed duration. These locations are inferred to be *home*, *work*, or *other* depending on observation frequency, day of week, and time of day, and represent a user's origins and destinations. Since the arrival time and duration at these locations reflect the *observed* (based on phone usage) rather than *true* arrival time and duration of a user, we probabilistically infer departure time using survey data on trips in major US cities. Trips are then constructed for each user between two consecutive observations in a day. These trips are multiplied by expansion factors based on the population of a user's *home* Census Tract and divided by the number of days on which we observed the user, distilling average daily trips. Aggregating individuals' daily trips by Census Tract pair and hour of the day results in the trip matrices used in this study.

Expanding on these methods, we now propose a methodology to estimate vehicle trips and road usage under varying ride-share adoption rates. First, we estimate the proportion of incoming and outgoing person trips made using automobiles from Equations (1) and (2) below. To infer travel mode, we use the 2006-2010 Census Transportation Planning Products (CTPP) parts 1 and 2, which provide residence- and workplace-based characteristics of Census Tracts [11]. Auto mode share of outgoing (incoming) trips is approximated by the ratio of auto a_i (a_j) to total t_i (t_j) trips made by residents commuting *from* Tract i (workers commuting *to* Tract j). By multiplying these auto shares by the average daily Tract-pair person trips T_{ij} inferred from the CDR data and summing over distination Tracts j (origin Tracts i), we compute the number of outgoing V_i (incoming V_j) person trips made by automobile.

$$(1)\ V_i = \sum_{j=1}^{n} \frac{a_i}{t_i} * T_{ij} \qquad (2)\ V_j = \sum_{i=1}^{n} \frac{a_j}{t_j} * T_{ij}$$

The output from Equations (1) and (2) serve as marginal totals and the average daily Tract-pair person trips T_{ij} serve as the seed matrix used to estimate OD person auto trips V_{ij} with Iterative Proportional Fitting (IPF). Finally, we estimate OD-specific mode shares v_{ij}, which we use to compute OD person auto trip matrices V_{ijh} for each hour h using hourly OD person trips T_{ijh}, as shown in Equations (3) and (4), respectively.

$$(3)\ v_{ij} = \frac{V_{ij}}{T_{ij}} \qquad (4)\ V_{ijh} = v_{ij} * T_{ijh}$$

Next, we determine which trips have the potential to be shared using spatial and temporal resolution of Tract and hour, respectively. In effect, we assume that trips are uniformly distributed within each hour and Tract, and therefore trips occuring within the same Tract-pair and hour can be matched with minimal additional travel time incurred. Lastly, using Equation (5) we compute the number of hourly OD vehicles C_{ijha} for given adoption rate a and maximum carpool size s.

$$(5)\ C_{ijhas} = V_{ijh} * (1 - a + \tfrac{1}{s})$$

Finally, we can assess road usage and level of service characteristics by assigning vehicles to a directed network with edges representing road segments and nodes representing intersections. Each edge has a free flow travel time t_{FF} and hourly capacity c, based on the road segment's number of lanes and facility type. We use the Bureau of Public Roads (BPR) volume-delay function to capture the relationship between vehicle flow and travel time under congested conditions and Iterative Traffic Assignment (ITA) to assign vehicles to the network minimizing total travel time.

4 Results

We performed traffic assignment for trips computed using Equation (5) for the 5 pm hour ($h = 17$), a maximum carpool size of 4 ($s = 4$), and adoption rates ranging from 5 to 25% ($a = \{0.05, 0.10, 0.15, 0.20, 0.25\}$). Accordingly, we are able to assess the network-wide impacts of traffic conditions in the peak weekday hour depending on users' willingness to share rides with 3 other passengers. Table 1 summarizes the percent change in vehicles, congested travel time, and vehicle miles traveled (VMT). We find that, regardless of the adoption rate, the change in vehicles and VMT are inversely proportional to roughly 75% of the adoption rate. The reduction in time spent in traffic, however, is much larger in magnitude than the corresponding adoption rate (258% to 224%). This decrease in returns reflects the inherent relationship between road segment volume and travel time, such that a reduction in the number of vehicles has the most impact on travel time in highly-congested conditions.

Figure 1 (a) and (b) compares the distributions of road segment volume and volume over capacity (VOC) for adoption rates of 0% and 25%. We see that ride-sharing has the largest impact on the road segments that are the most used and most over capacity.

Adoption rate (%)	Vehicles (%)	Congested TT (%)	VMT (%)
5	-3.7	-12.9	-3.8
10	-7.5	-25.6	-7.5
15	-11.2	-36.9	-11.2
20	-15.0	-47.0	-15.0
25	-18.8	-56.0	-18.7

Table 1: Percent change in vehicles, congested travel time (TT), and vehicle miles traveled (VMT) compared with an adoption rate of 0% (i.e. 100% single occupancy vehicles). Results are for 5 pm hour ($h = 17$) a maximum carpool size of 4 ($s = 4$).

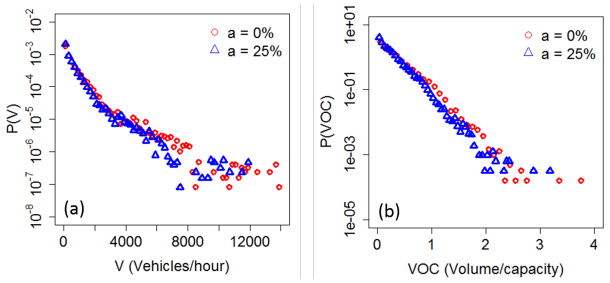


Figure 1: Distribution of (a) road segment volumes and (b) volume over capacity (VOC) ratio.

5 Conclusions

This research explores the extent to which ride-sharing services impact network-wide congestion using mobile phone records. To-date other research efforts have used rich data such as GPS taxi traces to explore potential ride-sharing demand under explicitly-defined spatio-temporal constraints. In contrast, we estimate aggregate, average daily ride-share demand using comparitively lower resolution CDR data. Despite the limitations, we are able to use CDR data to recreate reasonable travel demand for all vehicles, enabling us to evaluate the impact of ride-sharing on the entire road network. In Boston, we demonstrate that ride-sharing has the potential to significantly decrease time spent in traffic by all drivers, even for moderate levels of ride-sharing adoption.

References

[1] Lyft Line. https://www.lyft.com/line.

[2] UberPool. http://blog.uber.com/uberpool.

[3] Alexander, L., Jiang, S., Murga, M., and Gonzalez, M. C. Validation of origin-destination trips by purpose and time of day inferred from mobile phone data. *Transportation Research C* (2015).

[4] Cici, B., Markopoulou, A., Frias-Martinez, E., and Laoutaris, N. Assessing the potential of ride-sharing using mobile and social data: a tale of four cities. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ACM Press (2014), 201–211.

[5] Colak, S., Alexander, L., Alvim, B. G., Mehndiretta, S. R., and Gonzalez, M. Analyzing cell phone location data for urban travel: Current methods, limitations and opportunities. *Transportation Research Records* (2015).

[6] Huntsinger, L. F., and Donnelly, R. Reconciliation of regional travel model and passive device tracking data. In *Proceedings of the 93rd Annual Meeting of the Transportation Research Board* (2014).

[7] Iqbal, M. S., Choudhury, C. F., Wang, P., and González, M. C. Development of origin-destination matrices using mobile phone call data. *Transportation Research C 40* (2014), 63–74.

[8] Jiang, S., Yang, Y., Fiore, G., Jr., J. F., Frazzoli, E., and González, M. A review of urban computing for mobile phone traces: Current methods, challenges and opportunities. In *Proceedings of the ACM SIGKDD International Workshop on Urban Computing* (2013).

[9] Santi, P., Resta, G., Szell, M., Sobolevsky, S., Strogatz, S. H., and Ratti, C. Quantifying the benefits of vehicle pooling with shareability networks. *Proceedings of the National Academy of Sciences of the United States of America 111*, 37 (2014), 13290–13294.

[10] Schneider, C., Belik, V., Couronné, T., Smoreda, Z., and González, M. C. Unraveling daily human mobility motifs. *Journal of The Royal Society Interface 10* (2013).

[11] U.S. Department of Transportation Federal Highway Administration. CTPP 2006-2010 Census Tract Flows. http://www.fhwa.dot.gov/planning/census_issues/ctpp/data_products/2006-2010_tract_flows/index.cfm, 2013.

[12] Wang, P., Hunter, T., Bayan, A. M., Schectner, K., and González, M. C. Understanding road usage patterns in urban areas. *Scientific Reports 2* (2012).

Session 2 :: Cities (I)

Energy Consumption Prediction using People Dynamics Derived from Cellular Network Data

Andrey Bogomolov[1], Bruno Lepri[2], Roberto Larcher[3], Fabrizio Antonelli[3], Fabio Pianesi[2] and Alex Pentland[1]

[1]MIT Media Lab, USA [2]FBK, Italy [3]SKIL Telecom Italia, Italy

1 Introduction

Energy efficiency is a key challenge for building sustainable societies. Due to growing populations, increasing incomes and the industrialization of developing countries, the world primary energy consumption is expected to increase annually by 1.6%. Hence, this scenario raises issues related to the increasing scarcity of natural resources, the accelerating pollution of the environment, and the looming threat of global climate change.

Recently, several computational works have started in focusing on energy and sustainability problems. In particular, they have targeted energy issues with two main objectives: (i) modeling and predicting energy consumption behaviors [3] and (ii) inducing behavioral changes in energy consumption [1].

Our paper targets the energy consumption prediction task but adopting a novel approach. We use aggregated and anonymized human behavioral data, derived from mobile network activity, in order to predict the energy consumption of a given geographical area. More specifically, we deal with two different tasks: (i) *average daily energy consumption* and (ii) *peak daily energy consumption*. The method we propose in our paper has several advantages: (i) it is cheap in the sense that it uses existing data generated by telecom infrastructure. No modification to mobile phones or telecom equipment is required, and (ii) it has good scalability and so it is suitable for large populations, which is desiderable for city planning and energy management.

2 Methodology

We built our approach on a dataset spanning over a period of 2 months and a territory of 6000 square kilometers in the Northern Italy. The telecommunication and the energy consumption datasets have the same spatio-temporal aggregation. The temporal aggregation is ten minute intervals while the spatial one is obtained by partitioning the territory using a regular square grid, called *partitioning grid*. Each square of the grid measures approximately 1 square kilometer.

The *telecommunication dataset* is obtained from the Call Detail Records (CDRs) generated by the cellular network of a telecommunication operator offering its services on the territory under analysis. For the generation of this dataset CDRs recording sent SMSs, received SMSs, issued calls, received calls, and events related with Internet connections have been considered. The *energy consumption dataset* provides information about the structure of the electrical grid and the electrical current flowing through the 180 primary distribution lines serving the majority of the users living in the territory under analysis. Primary lines (medium voltage) are managed by a local company and bring energy from the national grid (high voltage) in order to distribute it among all the users. The dataset is composed by two sub-datasets: (i) the *structure of the electrical grid*, each line is described by providing the number of customer sites it serves in each square of the *partitioning grid* (customer sites often provide energy to more than one customer and they can also serve structures like condominiums, businesses and government organizations), and (ii) the *line measurement grid* that provides the instantaneous current flowing through each primary line every ten minutes.

Looking at the amount of electric current passing a point in an electric circuit per unit of time for each power line, we found that it has a number of cyclic characteristics and trends. We found predictable changes that repeat over daily and weekly periods. Based on these regularities we separated all power lines into 3 clustered areas: residential, touristic and city center/industrial area (see Fig. 1).

Then, we hypothesised that cellular communication patterns, which represent human dynamics in space and time, could be a good proxy for energy consumption prediction. To this end, we computed, from the anonymized and aggregated mobile network activity, a number of features characterizing diversity, regularity and general mobile phone usage in each part of the territory spatially aggregated by square grid. The discovered regularities, described above, were explicitly coded into the feature space by extracting of the number of hour in a day and the number of a weekday for each data source being processed.

The prediction tasks were solved for the next 7 days interval for each electric line ID and were designed as non-linear multiple regressions. We used the Random Forest algorithm proposed by Breiman [2]. The *consistency* of this algorithm has been proven: the algorithm adapts to sparsity in the sense that the rate of *convergence* depends only on the number of strong features and not on the number of noisy or less relevant ones.

To solve the problem of computational complexity of the huge amount of data samples (>600 millions of CDRs) we moved from time domain of communication patterns to the frequency domain, applying fast Fourier transform algorithm for each group of daily time series. Then, we computed second-order features by mathematical functions, such as mean, median, sum, variance, skewness, kurtosis, entropy, characterising the distributions and it's properties. Also we found that only a small set of harmonics in Fourier domain explains the response variable variance for each type of first-order feature space time series, which reduces the computational complexity by a number of orders. Finally, the computed second-order feature space for each spatial had a reduced number of dimensions (>3000), but still represented the temporal and spatial characteristics (i.e. diversity and regularity) of communication patterns.

A feature selection step was performed before the model building. The feature selection was done on a reduced sample of the training data, which was one week long. The metric used was the decrease in Gini index, which is the impurity measure of a decision tree node. This method is proven to be superior than correlation-based measures or information gain criteria. The feature selection step reduced the feature space to 32 dimensions for each of the two models without losing much accuracy.

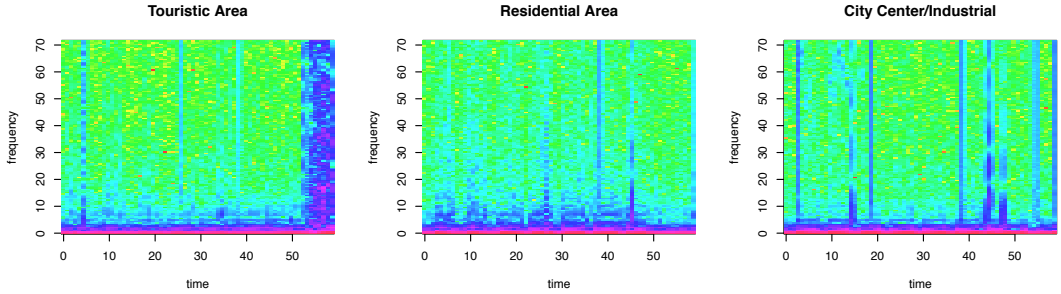


Figure 1: Spectral Characteristics of Typical Response Variables

3 Results and Discussion

Prediction metrics for daily average energy consumption for the next 7 days are 2.43 times better than the baseline, which is the training set mean value (MSE = 325.2679 compared to the baseline MSE = 790.6041). Prediction metrics for daily peak energy consumption model for the next 7 days prediction interval are 59.93 times better than the baseline, which is the training set maximum value (MSE = 601.7531 vs baseline MSE = 36062.7851).

Our results prove that human dynamics, extracted from aggregated and anonymized mobile phone data, are good proxies for modelling energy consumption. The introduced models technically do not account for seasonality on a yearly scale due to the 2-months limitation of the data sample. However, this limitation could easily be solved by training the model on more data.

Our results have several practical implications. Our approach could help to optimize the economy of energy producers/distributors value chain, also acting as an efficient tool for meeting peak electrical energy demand. Hence, could help to reduce total primary energy consumption and ecological footprint including climate change still meeting the people's energy needs.

References

[1] H. Allcott and S. Mullainathan. Behavioral science and energy policy. *Science*, 327:1204–1205, 2010.

[2] Leo Breiman. Random forests-random features. Technical report, Technical Report 567, Department of Statistics, UC Berkeley, 1999. 31, 1999.

[3] Z.J. Kolter and J. Ferreira. A large-scale study on predicting and contextualizing building energy usage. In *Proceedings of the Conference on Artificial Intelligence (AAAI), Special Track on Computational Sustainability and AI*, 2011.

Predicting Crime Hotspots Using Aggregated and Anonymized Data on People Dynamics

Andrey Bogomolov[1], Bruno Lepri[2], Jacopo Staiano[2], Emmanuel Letouze[4], Nuria Oliver[3], Fabio Pianesi[2] and Alex Pentland[1]

[1]MIT Media Lab, USA {abogomol,pentland}@mit.edu

[2]FBK, Italy {lepri,staiano,pianesi}@fbk.eu

[3]Telefonica Research, Spain nuriao@tid.es

[4]University of California, USA eletouze@datapopalliance.org

1 Introduction

Crime is a well-known social problem affecting the quality of life and the economic development of a society. Urbanists and architects have investigated the relationships between people dynamics, urban environment and crime. Urban activist Jane Jacobs emphasized natural surveillance as a key deterrent for crime: as people are moving around an area, they will be eyes on the street able to observe what is going on around them. Hence, Jacobs suggests that high diversity among the population and a high number of visitors contribute to the safety of a given area and lead to less crime [2]. Conversely, Newman proposed an alternative theory arguing that a high mix of people creates the anonymity needed for crime [3]. Thus, according to Newman, low population diversity, low visitors ratio and a high ratio of residents are the features contributing to an areas safety. Several studies have tried to shed light onto these conflicting theories.

Criminologists have also started to investigate in detail significant concentrations of crime at micro levels of geography, regardless of the specific unit of analysis. Research has shown that in what are generally seen as good parts of town there are often streets with strong crime concentrations, and in what are often defined as bad neighborhoods, there are locations relatively free of crime.

In this paper, we propose a novel place-centric and data-driven approach for crime prediction: specifically we investigate the predictive power of people dynamics – derived from a combination of mobile network activity and demographic information – to determine whether a specific geographic area is likely to become a scene of the crime.

2 Methodology and Results

In our model, we used 1 month of aggregated and anonymized mobile data combined with demographics information to predict whether a particular neighborhood in London will be a crime *hotspot* or not in the following month, *i.e.* whether it will have more or less crimes than the median number of crimes per neighborhood in London's metropolitan area. Our model (based on random forests) achieves 68.37% accuracy when using only mobile + demographics information, and 69.54% accuracy when adding census data (borough profiles data). For more detail, see [1].

A spatial visualisation of our results is reported on a map of the London metropolitan area in Figure 2 and compared with a similar visualisation of the ground truth labels in Figure 1. In the maps, green represents low crime (class 0, number of crimes $<$ median number of crimes) and red high crime (class 1, number of crimes $>$ median).

3 Discussion

Our results show that human behavioral and demographic data (at a daily and monthly scale) significantly improve the prediction accuracy when compared to using rich statistical data about a boroughs population (households census, demographics, ethnicity, employment, etc.). The borough profiles data provides a fairly detailed view of the living conditions of a particular area in a city, yet this data is expensive and time-consuming to collect. Hence, this type of data is typically updated with low frequency (*e.g.* every

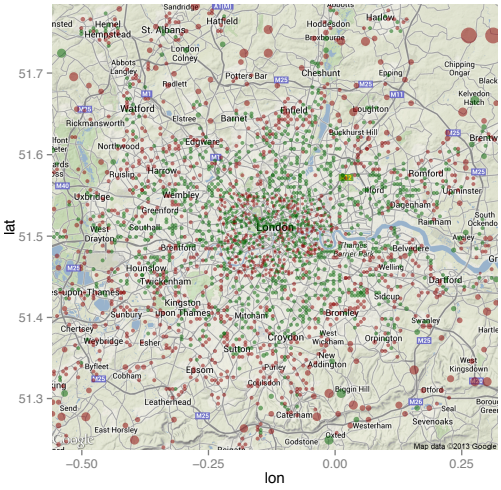


Figure 1: Ground Truth

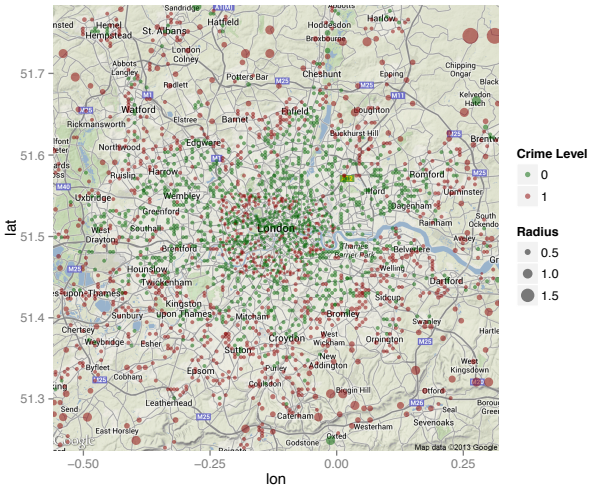


Figure 2: Predicted Crime Hotspots

few years) making difficult the observation of potential changes. Human behavioral data derived from mobile network activity combined with demographics provides significantly finer temporal and spatial resolution.

The features with highest predictive power in our combined model (mobile+borough profiles variables) are derived from the mobile data and not from the official statistics coming from borough profiles.

Moreover, higher-level features extracted over a sequence of days from variables encoding the daily dynamics have more predictive power than features extracted on a monthly basis. This finding points out at the importance of capturing the temporal dynamics of a geographical area in order to predict its levels of crime.

Interestingly, features derived from the percentage of people in a certain cell who are at home both at a daily and monthly basis seem to be of extreme importance. In fact, 11 of the top 20 features are related to the at home variable. Newman approach of defensible space postulates the relevance of a high number of residents in an area to reduce crime. The predictive power of home variables seems to confirm their relevance. However, we found positive associations between the home variables and crime. Hence, our findings do not support Newman thesis, suggesting that an increased ratio of residents is linked to less crime and higher urban safety. It is also interesting to note the role played by the unpredictability of the variables, captured by Shannon entropy features. The entropy-based features in fact seem useful for predicting the crime level of places (8 features out of the top 20 are entropy-based features). In our study, the Shannon entropy captures the predictable structure of a place in terms of the types of people that are in that area over the course of a day. A place with high entropy would have a lot of variety in the types of people visiting it on a daily basis, whereas a place with low entropy would be characterised by regular patterns over time. In this case, the daily unpredictability in patterns related to different age groups, different use (home vs work) and different genders seems to be a good predictor for the crime level in a given area. Our findings support Jacobs theory of natural surveillance [2] that high diversity of functions in a area and high diversity of people (gender-diversity and age-diversity) act as eyes on the street decreasing the number of crimes.

We believe that this work helps deepen our understanding of the relationship between human dynamics, urban characteristics and crime.

References

[1] A. Bogomolov, B. Lepri, J. Staiano, N. Oliver, F. Pianesi, and A. Pentland. Once upon a crime: Towards crime prediction from demographics and mobile data. In *Proc. of ICMI*, pages 427–434. ACM, 2014.

[2] J. Jacobs. *The death and the life of great american cities.* Random House Inc., 1961.

[3] P. Newman. *Defensible space: Crime prevention through urban design.* Macmillian Pub Co, 1972.

Estimating variability in health facility catchment population sizes using mobile phone call records

zu Erbach-Schoenberg, E.[1], Sorichetta, A.[1], Linard, C.[2], Lourenco, C.[1,3], Alegana, V.[1], Bird, T.J.[1], Tatem, A.J.[1,4,5]

[1]University of Southampton, [2]Université Libre de Bruxelles

[3] Clinton Health Access Initiative, [4]Fogarty International Center, [5]Flowminder foundation

1 Introduction

Many health indicators are measured through passive surveillance systems reporting at the level of health facilities. To convert these measures to population–level metrics, estimates of facility catchment population sizes are needed. A range of recent studies have highlighted the dynamic nature of human populations through quantitative analyses, particularly in low income settings [Bengtsson et al., 2011, Bharti et al., 2011, Lu et al., 2013]. However, populations are frequently assumed to be static over the study period due to a lack of time–resolved data. Recent analyses have shown the potential of using anonymised mobile phone call data records (CDRs) to produce accurate and dynamic population distribution maps [Deville et al., 2014]. Here we describe how CDRs can be used to estimate changes in population distribution in a low income setting and how measures relevant for public health applications can be calculated using this data. Using the example of health facility catchment areas in Namibia, we demonstrate that the population size of health facility catchments varies throughout time and that CDRs can be used to improve catchment area population size estimates.

2 Data and Methods

Mobile phone call data records covering a 3.5 year period from October 2010 to April 2014 were provided by the leading mobile phone service provider in Namibia, MTC (Mobile Telecommunications Limited). Following methods outlined in previous studies [Bengtsson et al., 2011, Gonzalez et al., 2008, Tatem et al., 2009, Wesolowski et al., 2012], we used the location of the towers through which an individual's communications were routed to calculate the daily locations for the anonymous subscribers and subsequently estimated the number of daily users for each mobile phone tower.

The user density over the coverage area of each tower (determined by its Voronoi polygon) was spatially extrapolated to census units, to obtain density values at the level of census units. Following the method described in Deville et al. [2014], we used a linear regression on training data to model the relationship between census unit mobile phone user density and observed population density according to the most recent census, conducted in August–September 2011.

Using the obtained parameters, we predicted population densities for each census unit for each month within the period covered by the data, adjusting the estimates to match the total predicted population count to the census counts. We then used a dataset detailing health facility catchment areas, created based on travel time [Alegana et al., 2012], to estimate the size and change in size of the population in each catchment area (see Figure 1 for catchment areas overlayed on a static population distribution estimate provided by the Worldpop project).

To demonstrate the importance of population mobility in the context of disease surveillance, we then applied our monthly population estimates to assess population variability in health-facility catchment areas.

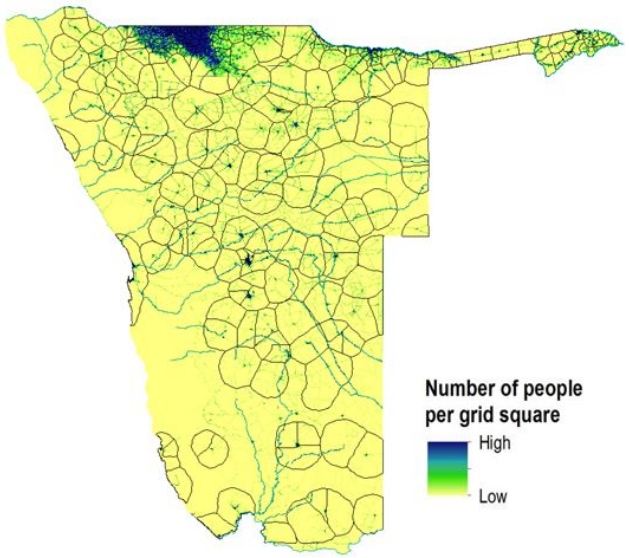


Figure 1: Catchment areas overlayed on static population estimate layer (Worldpop project `http://www.worldpop.org.uk`).

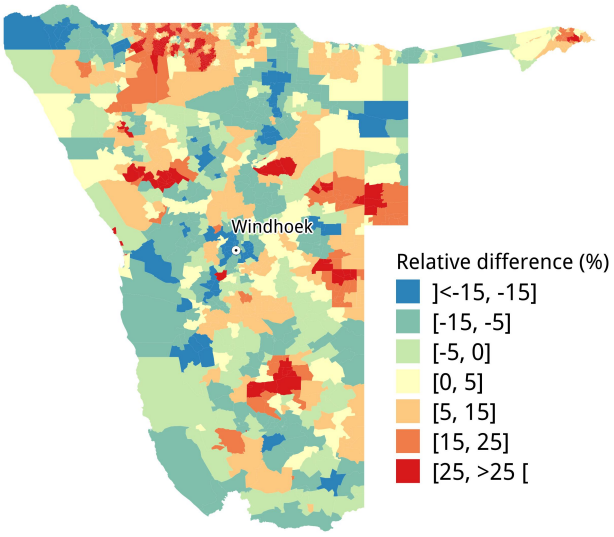


Figure 2: Percent difference in population numbers between November 2011 and December 2011. Red corresponds to a strong increase in population numbers in December, whereas blue corresponds to a decrease. The most substantial change in population distribution at Christmas time is the migration from the capital, Windhoek, to the North and North-East regions of the country.

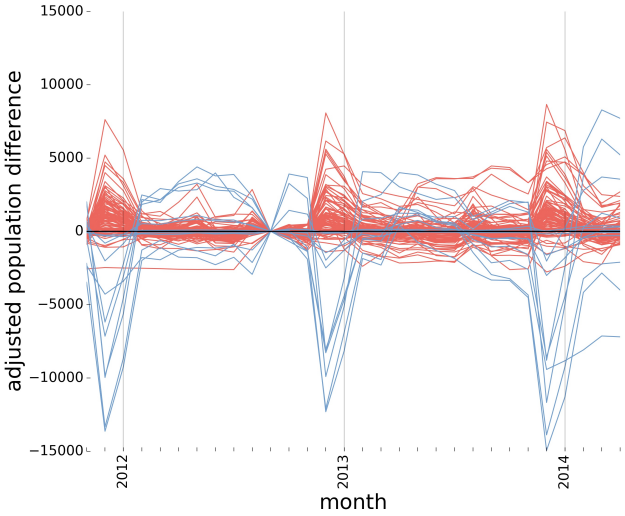


Figure 3: Change of catchment area population counts, showing the difference of the monthly prediction to the estimated based on static census population counts. Catchment area populations for facilities in the North (shown in red) are substantially larger than the static estimate during the holiday period. The opposite can be observed for health facilities in and around Windhoek (blue), showing a marked decrease in catchment population sizes for the same period of time.

3 Results

The population distribution is dynamic with a particularly marked change around the Christmas holiday period. Figure 2 shows the difference in population numbers for each census unit between November 2011 and December 2011 as percent change. We can clearly see a marked decrease in the population number around the capital, Windhoek. At the same time population sizes increase substantially in the North of the country. The opposite dynamic can be observed in the period following the holidays (not shown).

In Figure 3 we show the relative difference of estimated catchment area populations to the population number estimated from the census. We show the difference for health facilities from the North (shown in red) and the Windhoek area (blue). The peaks of relative increase in population numbers in the North around December correspond to the lows for the Windhoek region.

4 Conclusion

We show that even in a sparsely populated country such as Namibia with large distances between population centres, populations are highly dynamic across the year. As a result, catchment populations sizes of health facilities can change substantially through the year, with dynamic estimates differing by up to 15000 people (or up to a 30% difference) from a static estimate. CDRs can be used to assess the seasonality of population density and improve estimates by generating dynamic estimates.

5 Implications

The results have big implications on health metrics that rely on estimates of catchment denominators, with likely substantial under- and overestimations of population prevalences of health outcomes present. Moreover, the period of largest changes in population distribution coincides with the beginning of the malaria transmission season. Therefore, neglecting to take into account seasonality is likely to bias passive surveillance measures. Furthermore, we have seen that the observed changes in population density can be rapid at some times of the year, with substantial differences within the space of a single month. This has implications for the design of targeted intervention studies, intervention planning and the difficulties of reliably quantifying effects.

6 Perspectives

Compared to the method presented in Deville et al. [2014] there are some additional technical details to consider due to the different setting. While in France mobile phone coverage can be assumed to be complete, some areas in Namibia have no coverage. While these are often sparsely populated, using the same method as applied to the France data might lead to estimation errors as a result of substantially overestimating the coverage area of certain towers. Additionally, due to the growing mobile phone market in Namibia and the long period covered by our data set, the set of towers is dynamic due to continuing growth of the mobile phone network.

References

V.A. Alegana, J.A. Wright, U. Pentrina, A.M. Noor, R.W. Snow, and P.M. Atkinson. Spatial modelling of healthcare utilisation for treatment of fever in Namibia. *International Journal of Health Geography*, 11(6):10–1186, 2012.

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

N. Bharti, A.J. Tatem, M.J. Ferrari, R.F. Grais, A. Djibo, and B.T. Grenfell. Explaining seasonal fluctuations of measles in Niger using nighttime lights imagery. *Science*, 334(6061):1424–1427, 2011.

P. Deville, C. Linard, S. Martin, M. Gilbert, F.R. Stevens, A.E. Gaughan, V.D. Blondel, and A.J. Tatem. Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences*, 111(45):15888–15893, 2014.

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

X. Lu, E. Wetter, N. Bharti, A.J. Tatem, and L. Bengtsson. Approaching the limit of predictability in human mobility. *Scientific reports*, 3(2923), 2013.

A.J. Tatem, Y. Qiu, D.L. Smith, O. Sabot, A.S. Ali, and B. Moonen. The use of mobile phone data for the estimation of the travel patterns and imported *Plasmodium falciparum* rates among Zanzibar residents. *Malar J*, 8:287, 2009.

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

Measuring *de facto* populations with Mobile Network Operator's Call Detail Record data

Erki Saluveer[1,2], Rein Ahas[1], Siiri Silm[1], Margus Tiru[1,2]

[1] Department of Geography, University of Tartu, Estonia, rein.ahas@ut.ee, http://mobilitylab.ut.ee/eng/

[2] Positium LBS, spin-off company of Univeristy of Tartu, margus.tiru@positium.ee, http://positium.com/

The aim of this presentation is to introduce a methodology for measuring *de facto* populations based on mobile positioning databases. In the mobile and rapidly changing world, demographic processes must also be monitored more precisely than before. Censuses are usually conducted with a 10-year cycle and population register data is usually published as at the 1st of January, but today, it is necessary to know the number of residents located in an area at every moment of time. The consumers of such information include urban planners, transport managers, marketing companies, managers of emergency situations, and many others.

De jure populations are determined on the basis of an official fact or registration (census, register). *De facto* populations are formed by the people located in a certain area at a certain moment of time. (1) "Population groups or segments of de facto population consist six categories: (a) visitor population; (b) homeless population; (c) seasonal population, which we subdivide into (d) the amenity seeking population and (e) migrant workers and their families; (f), the portion of the Daytime population that consists of residents from elsewhere; and (g) the De Jure population that is "present." (1). For accurate measurement of de facto populations, various registers and sources of data are usually combined (2); they can also be measured within a limited unit of space and at a specific unit of time through observation or survey. They are more difficult to measure in large areas (3,4).

When mapping the distribution of population with the help of secondary data, it is necessary to know the factors and migration processes influencing the population. Distribution of population is most directly influenced by the geography of the building stock and the hierarchy of the population system. De jure populations can be directly connected to the locations of residential buildings. In the case of de facto populations, there are stronger connections with the locations of places of work, educational institutions, services, and various public agencies as well. The population system and the geographical distribution of the population are, for example, also influenced by the hierarchy of the population system and the development cycle of cities. Of population processes, the short-term dynamics of the population, its circadian, weekly, and seasonal cycle, are very important in mapping de facto populations (5,6). Migration processes that are tightly related to demography, economy, and international relations are also an important factor (7,8).

In this presentation, we are using passive mobile positioning, i.e. Mobile Network Operator (MNO) Call Detail Record CDR) and Data Detail Record (DDR) data, to measure *de facto* population. Mobile data has been used for

measuring population in many study projects (9, 10). Only the consent of the mobile operator and ensuring the procedures required for privacy protection are needed. Eurostat has also been most interested in this source of data (11).

In this presentation, we introduce a methodology, which produces de facto population statistics from CDR files. The CDR data include: the anonymous ID connected to the phone number (follows the same phone through history); call activity (any active use of the phone network, calls and text messages in and out, data services) location Cell ID (network cell location based on the location of the antenna); time of the call activity (with the accuracy of a second) (12). The CDR data are distributed unevenly in time and space, but they are collected for a long period of time. Thus, several algorithms for temporal and spatial interpolation of the data are needed. In our research, we have developed important data processing algorithms in the following stages:

a) Temporal interpolation of the data – harmonisation of the temporal distribution of location points.
b) Spatial interpolation – harmonisation of the special distribution of location points.
c) Finding anchor points – determining regularly visited and meaningful places.
d) Determining population segments – segmentation of population groups of different statuses on the basis of characteristics.

We determine the following units of population segments:

a) Permanent residents
b) Temporary residents
c) Domestic visitors
d) Foreign visitors
e) Transit visitors

In the analysis, we used algorithms to calculate the segments of de facto residents for all Estonian local municipalities and larger towns for the years 2012-2014. The calculated population segment figures are compared to the 2011 Census data and the later data will be compared to the number of people on the Census day (1.01.2012).

The results show that the number of permanent residents correlates best with the number of residents in the Census, the relationship is very strong (the correlation coefficient is 0.998). There is also a very strong relationship between the number of residents in the Census and the number of temporary residents and the number of visitors (Fig 1). There is a medium strength relationship between the number of residents in the Census and the number of foreign visitors and the transit segment.

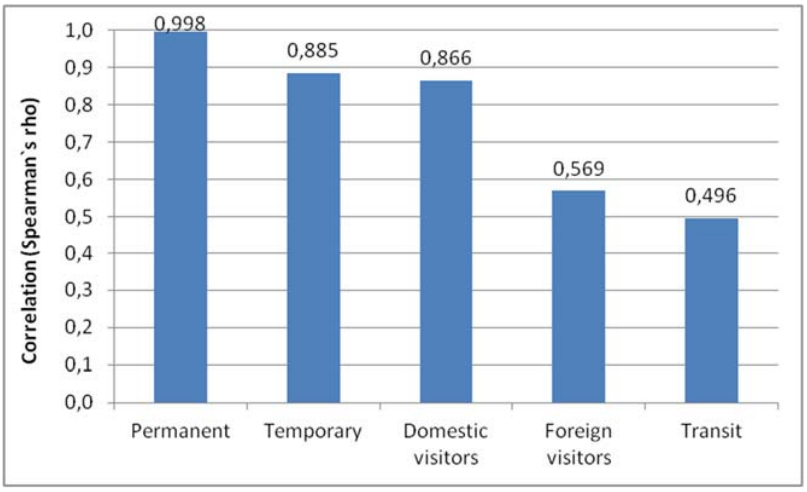


Figure 1. Average correlations of the whole period between population segments and the number of residents in the Census.

On the Census day (1.01.2012), there was a very strong relationship between all segments and the number of residents, the correlation coefficient was higher than 0.9 in the case of all segments. The number of permanent residents correlates best with the number of residents on the census day (correlation coefficient 0.998).

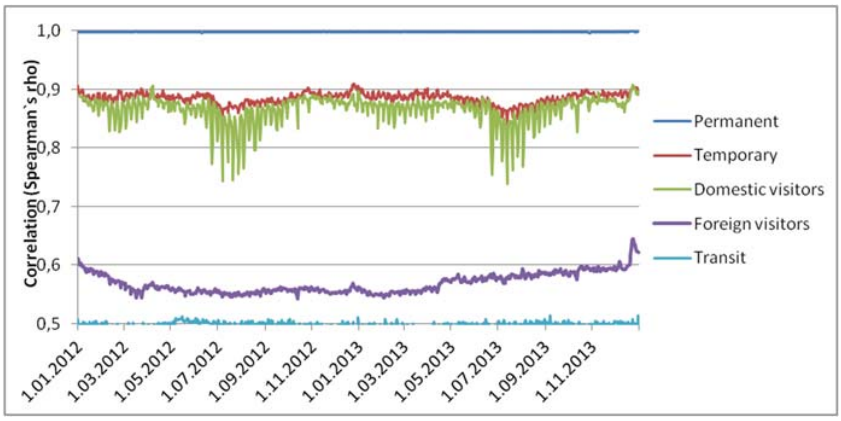


Figure 2. Variation of the correlations between population segments and the number of residents in the Census by days.

Comparison of the figures calculated by using algorithms to the Census data also shows that clear regularities have developed between population segments and certain days of the week. Above all, weekend days and working days are statistically different. In the case of certain segments, there are also great differences between working days (such as Friday). According to the theory of temporary migrations, it is a rhythmic process, a decrease in one place means an increase in another (5,6). The segment of foreign visitors and transit also behaves differently. There are also certain regularities in the variation of monthly population figures, with the summer months of June, July, and August as a holiday and tourism period most distinct. These results relate well with the theory of seasonal migrations and earlier practice. There are also statistically significant differences in the locations and distribution of population segments in December; here, the influences of events are significant in addition to the holidays.

The most complicated issue with the dataset has been distinguishing visits and transit, because it is not possible to register the exact amount of time spent in an area on the basis of CRD and DDR – only the locations of phones are registered. Since the number of phones does not characterise the presence of people with an absolute accuracy, we have used various correction coefficients created on the basis of reference data to generalise the dataset to the general population, and are using databases of censuses, accommodation statistics, road counters, ferry connections, etc.

References

(1)Swanson, D.A., Tayman, J. 2011. On Estimating a De Facto Population and Its Components, Review of Economics & Finance, 1: 17-31.

(2)Sutton, P., Roberts, C., Elvidge, D., & Baugh, K. (2001). Census from heaven: An estimate of the global human population using night-time satellite imagery. International Journal of Remote Sensing, 22, 3061–3076.

(3)UN 1991. Handbook of Vital Statistics Systems and Methods, Volume 1: Legal, Organisational and Technical Aspects, United Nations Studies in Methods, Glossary, Series F, No. 35, United Nations, New York.

(4)Smith, S. K.,1989. “Toward a Methodology for Estimating Temporary Residents”, Journal of the American Statistical Association, 84: 430-436.

(5)Bell, M. and Brown, D. 2006: Who are the visitors? Characteristics of temporary movers in Australia. Population, Space and Place 12(2), 77–146.

(6)Silm,S., Ahas, R., 2010. 'The seasonal variability of population in Estonian municipalities, Environment and Planning A, 42(10) 2527-2546.

(7)Zelinsky, W ., 1971. The hypothesis of the mobility transition. The Geographical Review LX1, 219-249.

(8)Bhaduri, B. (2007). Population distribution during the day. In S. Shekhar & H. Xiong (Eds.), Encyclopedia of GIS, Springer, December 2007 (print edition ISBN 978-0-387-30858-6).

(9)Reades, J. and Calabrese, F. and Sevtsuk, A. and Ratti, C. (2007), ‘Cellular Census: Explorations in Urban Data Collection’, IEEE, Pervasive Computing, Vol. 6, No. 3, pp. 30-38.

(10)Devillea, P., Linard, C., Maryine, S., Gilbert, M., Stevens, F.R., Gaughanf, A.E., Blondel, V., Tatem, A.J. 2014. Dynamic population mapping using mobile phone data, PNAS, 111(45): 15888–15893.

(11)Positium LBS 2014. Feasibility Study on the Use of Mobile Positioning Data for Tourism Statistics, Consolidated Report Eurostat Contract No 30501.2012.001-2012.452, 31p. epp.eurostat.ec.europa.eu/portal/page/portal/tourism/documents/MP_Consolidated%20report.pdf

(12)Ahas, R. Aasa, A., Roose, A., Mark, Ü., Silm, S. 2008. Evaluating passive mobile positioning data for tourism surveys: An Estonian case study. Tourism Management 29(3): 469–486. http://www.sciencedirect.com/science/article/pii/S0261517707001355

(13) Hugo, G.J. 2006 Temporary migration and the labour market in Australia. Australian Geographer 37, 211–31.
(14) Kang, CG, Liu, Y., Ma, XJ., Wu, L., 2012. Towards Estimating Urban Population Distributions from Mobile Call Data, Journal of Urban Technology 19(4): 3-21.

Investigating the relationships between spatial structures and urban characteristics

Marco De Nadai[1], Bruno Lepri[2], Roberto Larcher[3] and Nicu Sebe[1]

[1]University of Trento
[2]Fondazione Bruno Kessler
[3]Telecom Italia - SKIL

1 Introduction

Nowadays, the majority of people live in cities and the urban areas density is generally increasing. Several works show that the size of cities plays a fundamental role in a systematic acceleration of social and economic life [1]. These gains are applied to different quantities including economic output, wages, patents and epidemics and the average increase in these urban quantities, Y in relation to the population size N, is usually described by a super-linear scale-invariant law. Recently, theoretical works suggest that the origin of this super-linear scaling stems from social interactions [4]. Hence, understanding city's dynamics and spatial dynamics is of paramount relevance. Morphological urban analysis have been widely studied in literature. However, this was almost limited to static reports coming from census or remote sensing data.
The recent availability of new large-scale data sets, such as those automatically collected by mobile phone networks, open new possibilities of studying city dynamics.

In this paper, we analyzed the communication networks of 12 Italian cities with more than 150'000 inhabitants in order to investigate the impact of cities' spatial structures and dynamics on different socioeconomic outcomes such as economic growth, innovation, contagious disease rates and crime.

2 Methodology and Results

Our analysis is based on hourly aggregated and anonymous mobile traces provided by an Italian telecommunications operator, which concern 14 Italian cities urban areas during five months (from February 2014 to June 2014).

Given the irregular spatial distribution of the antennas per each technology (GSM, UMTS, etc.), the city area is spatially aggregated in a regular square grid with size α in order to simply compare the city's areas and the cities among each others. The value α has to correctly describe the city's neighbourhoods: a very small value can create cells in which the same quarter is described by more than one cell, while a too big value can hide interesting information and differences between quarters. In order to overcome the arbitrariness limit derived by the α choice, α was set to the mean distance between the nearest neighbourhoods' centroids of Milan ($0.6km$). This means that a square cell of $0.6km^2$ can correctly describe a quarter. The mobile traffic was re-calculated using the square grid, giving the access to the local activity density $\rho(i,t)$ at time t and place (square) i.

Firstly, we investigated the *average people compactness* in the city. This is an important urban metrics which helps to understand how citizens move during the day. For this reason the average distance between phone users in the city D_v is computed thanks to the Venables distance [2]. This distance was weighted with the squared city's area $\sqrt{A}$ in order to compare different cities. From fig. 1 it is possible to see that Milan has a peak around 9, were people are still probably at home or in the commuting system, then it collapses during the day, displaying a spatial concentration of people, probably because they work and/or tourism is more active. Contrarily, Naples shows an almost constant distance, probably meaning that workplaces and residential places are less separated among each other than in Milan, or that Naples is more equally spatial distributed. From the distance D_v it is also possible to study the *dilatation coefficient* which describes an entire city by:

$$\mu = \frac{max_t(D_v(t))}{min_t(D_v(t))}$$

. In figure fig. 1 it is possible to compare the different cities. This describes how the city is spatially organized, showing for example if workplaces, schools and leisure places are concentrated in the same area.

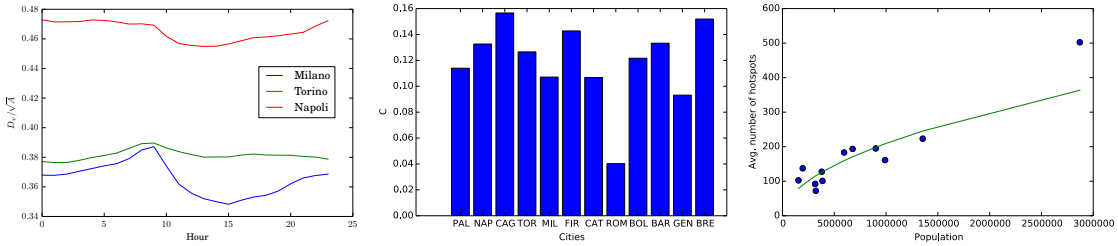


Figure 1: (a). Time evolution of the average distance between phone users in the various cities. (b). Spatial distance between hotspots. (c). Power law fit between the average number of hotspots and cities' population.

In order to look the morphological distribution of the city, it is very useful to look at its *hotspots*. The simplest way to define the hotspots is by choosing a threshold δ and considering hotspots all the points with a density $\rho(i,t) \geq \delta$. This threshold can be chosen by experience or by observing the data, leading to an obvious arbitrariness in the results and to a unique threshold among time t. On the contrary, we used the Loubar threshold $\rho(i,t)_{Lou}$, defined by [2], from the tangent of point $p(1,1)$ of the Lorenz Curve of the density ρ, which expresses the spatial inequality of a city at time t. The interesting features explored by applying this method are here presented.

Louf and Barthelemy [3] showed that in U.S activity centres, as determined from employment data, N_a scales sub-linearly (with $\beta \approx 0.64$) in relation to the population size. More recently, Louail et al. [2] demonstrated that there is a sub-linear relation, with a similar exponent β. between activity (mobile) hotspots and population for Spanish cities. Here, we firstly try to verify the aforementioned sub-linearity relation for Italian cites, then we investigate the associations between activity (mobile) hotspots and social quantities such as criminality, employment.

Interestingly, the power law fit between population and number of hotspots confirms the results obtained with Spanish cities in [2]. Furthemore, we carried out correlational analyses to investigate whether associations between the average number of hotspots and social quantities like GDP, employment, unemployment, patents intensity and criminality exist. For these analysis we employed the non-parametric Spearman's Rho method with a level of significance $p < .05$ (see table 1 reporting Spearman correlation coefficients and the power law fit's exponent β).

Variable	Spearman correlation $[-1,1]$	β
Population	0.96	0.54
GDP	0.60	0.68
Employment	0.83	0.61
Unemployment	0.56	0.40

Table 1: significant Spearman correlation coefficients and power law exponent (β) between various social quantities and the average number of hotspots

References

[1] Luís MA Bettencourt, José Lobo, Dirk Helbing, Christian Kühnert, and Geoffrey B West. Growth, innovation, scaling, and the pace of life in cities. *PNAS*, 104(17):7301–7306, 2007.

[2] Thomas Louail, Maxime Lenormand, Oliva García Cantú, Miguel Picornell, Ricardo Herranz, Enrique Frias-Martinez, José J Ramasco, and Marc Barthelemy. From mobile phone data to the spatial structure of cities. *Scientific Reports*, 2014.

[3] Rémi Louf and Marc Barthelemy. Modeling the polycentric transition of cities. *Physical review letters*, 111(19):198702, 2013.

[4] Wei Pan, Manuel Cebrian, Ghoshal Gourab, Krumme Coco, and Alex Pentland. Urban characteristics attributable to density-driven tie formation. *Nature Communications*, 4(1961), 2013.

Session 3 :: Economies

BEHAVIOR REVEALED IN MOBILE PHONE USAGE PREDICTS LOAN REPAYMENT

DANIEL BJÖRKEGREN*

Many households in developing countries lack formal financial histories, making it difficult for banks to allocate capital, and for potential borrowers to obtain loans. However, many unbanked households have mobile phones, and even prepaid phones generate rich transaction histories. This project shows that behavioral signatures in mobile phone data predict default with accuracy near that of traditional credit scoring methods that rely on financial histories. The method is demonstrated using call records matched to loan outcomes for a sample of borrowers in Haiti.

1. Introduction

Many studies have found that small firms in developing countries have access to opportunities with high returns that, puzzlingly, remain untapped (De Mel et al., 2008; McKenzie and Woodruff, 2008; Banerjee and Duflo, 2014). One reason these opportunities may remain untapped is if potential lenders have difficulty identifying the investments that will be profitable.

Developing country banks that would lend to small firms face several particular challenges. In developed countries, banks have access to robust information on borrower reputation through credit bureaus, which aggregate information on an individual's historical mangement of credit, but many developing country households do not interact with formal institutions that generate these forms of data. As a result, lenders have very little formal information on potential borrowers. This is particularly problematic, as banks who would lend to small or informal businesses may have little recourse if a borrower were to default. Even when banks can rely on institutions like police and courts, it is costly to follow up on small loans.

Traditional microfinance has presented one solution to the repayment problem, but it is not clear that the investments currently selected by microfinance have generated growth for small businesses (see for example Karlan and Zinman, 2011; Banerjee et al., 2014).

This paper introduces a new method to identify profitable investments, using information on potential borrowers that is already being collected by mobile phone networks.

Although unbanked households lack the formal records needed for traditional credit scores, many have maintained a rich history of interaction with a formal institution over an extended period of time—their mobile phone activity, recorded by their operator. Even with prepaid plans, operator records yield rich information about individual behavior and social networks. If indicators derived from this data are predictive of creditworthiness, they can help banks identify profitable opportunities. There are many straightforward indicators of behavior that are plausibly related to loan repayment. For example, a responsible borrower may keep their phone topped up to a minimum threshold so they have credit in case of emergency, whereas one prone to default may allow it to run out and depend on others to call them.

This paper demonstrates that indicators of behavior derived from mobile phone transaction records predict loan repayment with accuracy near that of traditional credit scores.

2. Data

My organizational partner is EFL (Entrepreneurial Finance Lab), which works on alternative credit scoring methods in developing and emerging markets with an emphasis on the unbanked.[1] EFL obtained linked phone and completed loan data for a sample of borrowers in Haiti. First, anonymous records were obtained for a sample of borrowers who took out a small loan from a local bank. These records include basic demographics such as age and gender, the terms of the loan provided, and whether the loan was defaulted on (defined by 90 days of nonpayment).[2] Many of these borrowers also have mobile phone accounts with a large operator in the country. These borrowers were matched to their prepaid phone accounts using encrypted (anonymous) individual identifiers. Mobile phone transaction records (CDR) were obtained for these matched accounts, including metadata for each call, SMS, top up, and data access, for the year of 2012.

Because we aim to predict default based on the information available at the time a loan was granted, only mobile phone transactions that precede the loan date are included. We focus on the bank's small loan product. The data includes 3,131 loans granted between January and September 2012, of which 12.8% ended in default. Borrowers have a median loan size of $184 and term of 5 months (Haiti's per capita GDP was $775.52 in 2012).[3]

*Brown University. E-mail: dan@bjorkegren.com, Web: http://dan.bjorkegren.com

Revision January 12, 2015. Preliminary and incomplete. Thanks to Entrepreneurial Finance Lab for data, and Nathan Eagle, Seema Jayachandran, and Jeff Berens for helpful discussions. This work was supported by the Stanford Institute for Economic Policy Research through the Shultz Fellowship in Economic Policy.

[1] From their website, website, "EFL Global develops credit scoring models for un-banked and thin-file consumers and MSMEs, using many types of alternative data such as psychometrics, mobile phones, social media, GIS, and traditional demographic and financial data. We work with lenders across Latin America, Africa and Asia." http://www.eflglobal.com

[2] Loans that ended in default were oversampled to aid in estimation.

[3] All results reported in US dollars.

3. Method and Results

The goal is to predict the likelihood of default using behavioral features derived from mobile phone usage. The model is estimated using data on loans that have already been completed; these estimates are used to predict whether a loan would end in default based on the information available at the time the loan was granted. Because this sample of individuals did obtain loans, risk is reported among those who were allocated loans based on the bank's scoring method at the time; default risk among unbanked populations may differ, and can be explored in follow up work.[4]

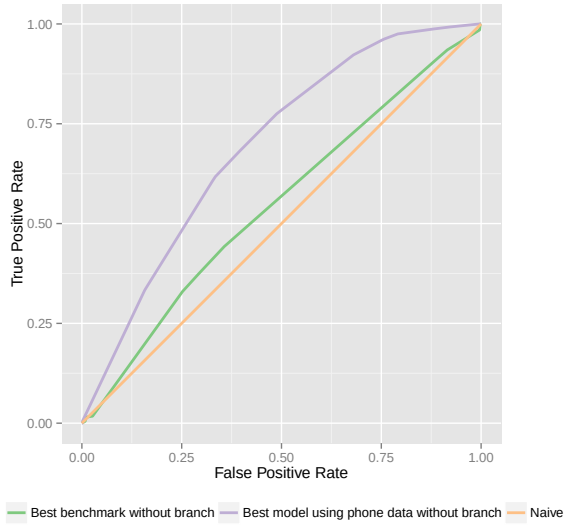
The loan data provides an indicator for whether a particular borrower defaulted on their loan (90 days past due). From the phone data we derive 6,841 features with variation. Features include measures of usage: intensity and distribution over space and time, top up and depletion patterns, mobility, the pattern of handset use, and strength and diversity of social network connections. The performance of these features will be compared against a benchmark using the characteristics that the bank recorded at the time of the loan: gender, age, loan size, and the loan term in days.

A first question is how individual features correlate with default. Benchmark features have very low correlation with default (magnitudes between 0.004 and 0.06), which is reasonable: the bank should already be incorporating the features it observes into the decision of whether to extend credit. Several categories of features from the phone data have highly significant correlations with magnitudes between 0.12 and 0.18, including variation and periodicity of usage, and several measures of mobility.

While individual features are significantly correlated with default, the best prediction method will take into account multiple features at the same time. Because of the large number of potential predictors, including all of them in a simple method like ordinary least squares (OLS) would lead to overfitting. Instead, two types of models are estimated, including a specification of OLS chosen by a model selection procedure (stepwise search using the Bayesian Information Criterion) and random forests. To best measure how the method will perform out of sample, all estimates are computed using cross validation with 5 folds. The data is randomly divided into 5 folds, and the outcomes for each fold are predicted using a model estimated on the omitted folds.

The main result we report is the area under the receiver operating characteristic curve (AUC), which should range from 0.5 to 1.0. The receiver operating characteristic curve (ROC) plots the true positive rate of a classifier against the false positive rate. If the ROC of a classifier has nonconvexities, it is possible to reweight it to obtain the convex hull of the original ROC; thus I also consider the area under the convex hull of the receiver operating characteristic curve (AUC-H). Figure 1 presents the convex hulls of the best benchmark model and the best model using indicators derived from phone data.

Figure 1. Model Performance: Receiver Operating Characteristic Curve



The benchmark models using only demographic and loan data have moderate AUCs ranging from 0.531-0.536. Adding features derived from phone data improves AUCs to the range of 0.660-0.682, and the AUC-H of the best model is 0.690 (stepwise OLS). This performance is near a sample of published results of AUC estimates from other studies that use traditional credit scoring methods with financial data in more developed settings. Baesens et al. (2003) reports top AUCs ranging 0.668-0.758 for different UK samples and 0.776-0.791 using for samples from the Benelux countries. Calabrese and Osmetti (2013) reports a top AUC of 0.723 for small and medium enterprises in Italy.

To account for the possibility of unforeseen shocks, ideally the model would be tested not only on out of sample individuals, but also on out of sample time periods. Since the data spans only a short time this is left for future work. Results so far suggest that the performance of the model without financial data approaches the performance obtained with traditional credit scoring methods in more developed settings. Performance is likely to improve with additional tuning and data (this paper uses an average of 2.5 months of phone data per borrower).

[4]The measures in this paper correspond with a counterfactual of being more selective in extending loans.

References

Baesens, B., T. Van Gestel, S. Viaene, M. Stepanova, J. Suykens, and J. Vanthienen (2003): "Benchmarking state-of-the-art classification algorithms for credit scoring," *Journal of the Operational Research Society*, 54, 627–635.

Banerjee, A., E. Duflo, C. Kinnan, and R. Glennerster (2014): "The Miracle of Microfinance? Evidence from a Randomized Experiment," .

Banerjee, A. V. and E. Duflo (2014): "Do Firms Want to Borrow More? Testing Credit Constraints Using a Directed Lending Program," *The Review of Economic Studies*, 81, 572–607.

Calabrese, R. and S. A. Osmetti (2013): "Modelling small and medium enterprise loan defaults as rare events: the generalized extreme value regression model," *Journal of Applied Statistics*, 40, 1172–1188.

De Mel, S., D. McKenzie, and C. Woodruff (2008): "Returns to Capital in Microenterprises: Evidence from a Field Experiment," *The Quarterly Journal of Economics*, 123, 1329–1372.

Karlan, D. and J. Zinman (2011): "Microcredit in Theory and Practice: Using Randomized Credit Scoring for Impact Evaluation," *Science*, 332, 1278–1284.

McKenzie, D. and C. Woodruff (2008): "Experimental Evidence on Returns to Capital and Access to Finance in Mexico," *The World Bank Economic Review*, 22, 457–482.

Estimating Food Consumption and Poverty Indices with Mobile Phone Data

Adeline Decuyper*, Alex Rutherford†, Amit Wadhwa§, Jean Martin Bauer§, Gautier Krings*‡, Thoralf Gutierrez‡, Vincent D. Blondel*, Miguel A. Luengo-Oroz†
*ICTEAM institute, Université catholique de Louvain, Belgium
†United Nations Global Pulse
‡Real Impact Analytics, Brussels, Belgium
§Vulnerability Analysis and Mapping, World Food Programme

Recent studies have shown the value of mobile phone data to tackle problems related to economic development and humanitarian action. In this research, we assess the suitability of indicators derived from mobile phone data as a proxy for food security indicators. We compare the measures extracted from call detail records and airtime credit purchases to the results of a nationwide household survey conducted at the same time. Results show high correlations ($> .8$) between mobile phone data derived indicators and several relevant food security variables such as expenditure on food or vegetable consumption. This correspondence suggests that, in the future, proxies derived from mobile phone data could be used to provide valuable up-to-date operational information on food security throughout low and middle income countries.

In recent years the explosion in the use of digital services has led to what has come to be called the "Big Data Revolution". This revolution has fundamentally changed how companies are able to understand their users through analysis of data produced passively. However, more recently, as low-cost mobile handsets and internet usage have proliferated in developing countries, reaching 90% coverage in the developing world in 2014 [1], the potential of these signals to transform development and humanitarian action has emerged.

Big Data has the potential to guide policy makers by providing an alternative to traditional data sources such as costly and time-consuming manual surveys [2]. For planning purposes, access to reliable and up-to-date statistical data is essential to humanitarian organizations when deciding where and when help is most needed. Many rich data sources exist with the promise of providing early warning and real-time monitoring of vulnerable populations including remote sensing, social media, remittances and anonymized mobile phone records.

In particular, many recent studies have shown how valuable mobile phone data in the form of Call Detail Records (CDRs) can be used to guide agile development policy and humanitarian action [3], [4], [5]. The Orange D4D challenge [3], using Côte d'Ivoire as a case study, led to a large number of innovative uses of CDR data ranging from mobility modeling for transport optimization [4] to epidemic modeling [5] and network analysis of social communities [6]. The challenge was such a success that other challenges of big data exploration have followed [7], [8], and the results of a second development challenge, this time on data from Senegal, will be presented in April 2015 [8].

Given the importance of mobile devices to people in developing economies for accessing information and economic opportunities, phone usage data represents a clear barometer of a user's socio-economic conditions in the absence or difficulty of collecting official statistics [2], [9], [10]. Many studies have already addressed the question of the relationship between the mobile phone usage and socio-economic levels [11], [10], [12], [9], [13].

The objective of this research is to assess the suitability of metrics derived from mobile phone data, specifically CDRs and airtime credit purchases, as proxies for food security and poverty indicators in a low-income country context. We use a country of central Africa as a case study, and compare at a fine-grained spatial scale features measured from mobile phone data collected in 2012 with the results of a detailed food security survey conducted at the same time. The World Food Programme regularly assesses the situation of nutrition, their reports with updated information are available online [14]. While food access has greatly improved between 2009 and 2012, there is still approximately one household in five that is food insecure in these regions. One of the recommended steps for the way forward to tackle malnutrition in Africa is to obtain better and more frequent information on food security. Therefore, a real-time proxy of relevant trends on food security indicators would be of great interest for regions where household surveys can be time and resource consuming if they are possible at all. In order to complement the information obtained from food security indicators, we also assess the possibility of using the same mobile phone data to map poverty levels, using the results of a survey on non-monetary poverty as a ground truth comparison.

We analyze the correlations between each pair of mobile phone metric and survey variable aggregated over the different administrative divisions of the country. We find very high correlations (> 0.7) between several food security indicators and measures computed from airtime purchases. Generally, the sum of expenses and the maximum top-up value are the mobile variables that correlate best with several of the food indicators, see table I for a few selected correlation coefficients.

TABLE I
SELECTED VARIABLES FROM HOUSEHOLD SURVEYS WITH HIGH CORRELATION WITH THE SUM OF EXPENSES IN MOBILE CONSUMPTION.

survey variable	**cor. sum**	95% **conf. int.**
Share food from production	-0.64	$[-0.71, -0.55]$
Share food from purchases	0.66	$[0.58, 0.73]$
Food consumption score	0.62	$[0.54, 0.7]$
Monthly food expenditure	0.82	$[0.77, 0.86]$
Education expenditure	0.72	$[0.65, 0.78]$
Communication expenditure	0.69	$[0.62, 0.75]$
Total monthly expenditure	0.6	$[0.51, 0.68]$
Per capita income	0.63	$[0.54, 0.7]$
Multidimensional poverty index	-0.75	$[-0.8, -0.69]$

The correlation shown here is the correlation between each indicator, and the sum of expenses on mobile phone airtime credit, along with their 95% confidence intervals. All p-values are below 10^{-15}.

For selected variables, we also test linear and quadratic regression models, using more than one mobile phone derived metric. The highest correlation value of all the models tested in this study corresponds to a quadratic model with the sum of expenses and the average amount of top-up as mobile phone derived indicators that shadows the survey variable measuring the amount of food expenses (0.89), see figure 1.

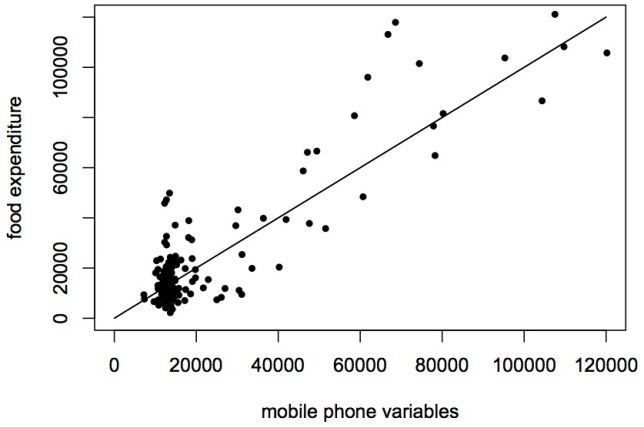


Fig. 1. **Quadratic combination of CDR variables against expenses on food.** Correlation coefficient: 0.89.

Looking in detail at the relations between the consumption of each food item (survey question "how many times have you eaten [item] in the last 7 days?") and mobile phone expenditure (measured by the average of the sum of airtime expenses per user over 6 months), we found different correlation values ranging from no relation to very high correlation between certain food items and airtime expenses. In particular the consumption of vitamin rich vegetables, rice, bread, sugar and fresh meat have a very high correlation (> 0.7) with the airtime purchases. On the other side of the spectrum, broadly cultivated items like cassava and beans have no relation with the expenditure on mobile phones. These results are particularly interesting, as the products from the first group are mainly bought in the market while the products with consumption uncorrelated to top up values are cultivated for personal consumption. Therefore this study is compatible with a new hypothesis: expenditure in mobile phone top up is proportional to the expenditure in food in the markets.

When analyzing the relation between mobile phone variables and poverty indices, we have been able to create a proxy indicator for multidimensional poverty index at the sector level ($> .8$ correlation). Interestingly, we have found that the most important variable to shadow poverty levels is the top up information.

This study is a first step towards demonstrating the utility of mobile phone data, as a novel new data source for food security monitoring. All in all, this analysis shows that simple statistics on top-up data could very accurately provide valuable information about the evolution of food security at a fine grained spatial level. In particular, we have begun to outline which aspects of ground truth food security information are reflected in CDR and mobile phone top-up information. This research illustrates a feasible mechanism whereby mobile data could be shared outside a mobile company at a lower resolution than the individual level, yet still providing information to guide policy.

The full article of this research is available online [15].

REFERENCES

[1] (2014). The world in 2014 : Ict facts and figures. International Telecommunication Union. http://www.itu.int/.
[2] Deville P, Linard C, Martin S, Gilbert M, Stevens F, et al. (2014) Dynamic population mapping using mobile phone data. PNAS 111: 15888-15893.
[3] (2013) Mobile Phone Data for Development - Analysis of mobile phone datasets for the development of Ivory Coast. Orange Data For Development Challenge, D4D. URL http://perso.uclouvain.be/vincent.blondel/netmob/2013/D4D-book.pdf.
[4] Berlingerio M, Calabrese F, Di Lorenzo G, Nair R, Pinelli F, et al. (2013) Allaboard: A system for exploring urban mobility and optimizing public transport using cellphone data. In: Blockeel H, Kersting K, Nijssen S, Železný F, editors, Machine Learning and Knowledge Discovery in Databases, Springer Berlin Heidelberg, volume 8190 of *Lecture Notes in Computer Science*. pp. 663-666.
[5] Lima A, De Domenico M, Pejovic V, Musolesi M (2013) Exploiting cellular data for disease containment and information campaigns strategies in country-wide epidemics. CoRR abs/1306.4534.
[6] Amini A, Kung K, Kang C, Sobolevsky S, Ratti C (2014) The impact of social segregation on human mobility in developing and industrialized regions. EPJ Data Science 3: 6.
[7] Telecom Italia Big Data Challenge. URL http://www.telecomitalia.com/tit/en/bigdatachallenge/contest.html.
[8] de Montjoye Y, Smoreda Z, Trinquart R, Ziemlicki C, Blondel V (2014) D4D-Senegal: The second mobile phone data for development challenge. arXiv preprint arXiv:14074885 .
[9] Frias-Martinez V, Soto V, Virseda J, Frias-Martinez E (2013) Can cell phone traces measure social development? In: Third Conference on the Analysis of Mobile Phone Datasets, NetMob.
[10] Gutierrez T, Krings G, Blondel VD (2013) Evaluating socio-economic state of a country analyzing airtime credit and mobile phone datasets. CoRR abs/1309.4496.
[11] Eagle N, Macy M, Claxton R (2010) Network diversity and economic development. Science 328: 1029-1031.

[12] Frias-Martinez V, Virseda J (2012) On the relationship between socio-economic factors and cell phone usage. In: Proceedings of the Fifth International Conference on Information and Communication Technologies and Development. New York, NY, USA: ACM, ICTD '12, pp. 76–84. doi:10.1145/2160673.2160684. URL http://doi.acm.org/10.1145/2160673.2160684.

[13] Smith-Clarke C, Mashhadi A, Capra L (2014) Poverty on the cheap: Estimating poverty maps using aggregated mobile communication networks. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. New York, NY, USA: ACM, CHI '14, pp. 511–520. doi:10.1145/2556288.2557358. URL http://doi.acm.org/10.1145/2556288.2557358.

[14] WFP. World food programme survey data portal. international household survey network. URL http://nada.vam.wfp.org.

[15] Decuyper A, Rutherford A, Wadhwa A, Bauer J, Krings G, et al. (2014) Estimating food consumption and poverty indices with mobile phone data. CoRR abs/1412.2595.

Mobility and Productivity: Quantifying Urban Economic Activity using Cell Phone Data°

Gabriel E. Kreindler [1,2] and Yuhei Miyauchi [1,2*]

[1] *Department of Economics, Massachusetts Institute of Technology, Cambridge MA*
[2] *LIRNEasia, Colombo, Sri Lanka*

How do daily commuting flows relate to the temporal and spatial distribution of urban economic activity? Simply knowing where people work will not fully capture economic activity, because locations vary in productivity. We show how the entire matrix of commuting flows obtained from cell phone data can be used to estimate productivity, which allows for a more accurate estimation of economic activity. We achieve this by estimating a gravity equation derived from a discrete choice model of job choice. To validate our approach, we use the model to predict mean residential income at commuting origins, which we compare with a new dataset of nighttime lights. We discuss potential applications.

Human mobility and economic activity are intertwined. Previous work has established that international migration flows are responsive to economic activity in the destination country, as well as to distance [1], [2]. Regional and seasonal migration also respond in a similar fashion [3], [4]. These results are suggestive of a relationship between mobility and economic activity at much finer temporal and geographic scales. However, to date this relationship has not been studied quantitatively mainly due to data limitations. This is the goal of our project. We study the link between daily mobility (commuting) flows and economic activity within cities, and quantify it using cell phone data.

Our project has three contributions. First, we set up a theoretical model that links individual home- and job-choice decisions, commuting flows and economic productivity, and show that it can be implemented empirically through a gravity model. Second, we use cell phone data from Sri Lanka to construct a measure of commuting with fine urban spatial and temporal variation, and proceed to estimate the model.[1] Third, we validate the model's performance using a new and improved data source of nighttime lights.

Our work builds on a recent literature on urban sensing [5], and the measure of economic activity may eventually assist policy in urban areas. This detailed data can help improve economic policy [6], as well as benefit urban-planning by making it more responsive to detailed trends in the concentration of economic activity.

Discrete Choice Model and Gravity Equation

In the model, based on [7], workers choose their work destination by trading off the wage offered at each destination, against the distance needed to reach each destination. Naturally, high wages may compensate for the cost of commuting. Other idiosyncratic factors that influence job choice are modelled as random shocks from a pre-specified distribution (Fréchet or type II extreme value). This setup implies that the commuting probability π_{ij} that a worker residing in origin i commutes to destination j satisfies the following gravity equation:

$$\log(\pi_{ij}) = \psi_j + \beta \log(D_{ij}) - \mu_i + \varepsilon_{ij}, \qquad (1)$$

where ψ_j is a log transformation of the wage at j, D_{ij} is a measure of distance between i and j, μ_i captures origin-specific factors, and ε_{ij} is measurement error. Intuitively, ψ_j captures a destination's attractiveness, after controlling for distances to all origin locations, and their respective sizes. The benefit of using an explicit model of workers' decisions is that it allows us to back out other useful economic measures, such as the output and the average residential income at a particular location.

In the full paper, we study several features and extensions of the model, such as the importance of the aggregation level, (i, j) pair specific factors, and worker heterogeneity. The model can also accommodate a variety of assumptions on how workers choose their home location.

Commuting Flows and Model Estimation

Empirically, the model can be mapped to the real world using commuting flows extracted from cell phone data. We use a simple algorithm to construct daily commuting flows from CDR data, building on the literature on this topic [8], [9]. Figure 1 shows the distribution of smoothed commuting flows from a particular origin location (i.e. a given cell tower Voronoi cell) in the greater metropolitan

° The authors are grateful to the LIRNEasia organization for providing access to cell phone data and an excellent working environment, to Sriganesh Lokanathan, research manager at LIRNEasia, whose dedication and relentless efforts made this project possible, and to Danaja Maldeniya for constructive feedback on the project. We acknowledge funding from the International Development Research Centre (IDRC).
* Contact: gek@mit.edu; miyauchi@mit.edu.

[1] Our approach will perform significantly better in *urban* areas, due to higher cell tower density relative to less developed areas.

Colombo area. What is noteworthy is that there is significant variation in commuting flows even after accounting for commuting distance, and this variation appears to capture anecdotal patterns of commuting in the Colombo area.

We estimate equation (1) on commuting volumes between pairs of towers in Sri Lanka. There are 3,047 towers and ≈ 1,7 million pairs of towers in our sample, covering over 300 million commuting trips. We use a linear regression model with two sets of fixed effects (corresponding to origin and destination locations). Using the estimated fixed effects and other coefficients, we construct the residential income and output measures.

Comparison with Nighttime Lights

Having estimated economic productivity (wages) and economic activity (output), exclusively from mobility flows derived from cell phone data, we would ideally proceed to validate these measure using independent wage and output measures. Unfortunately, in our context this type of data is only available aggregated at province level.

Instead, we present results from a validation exercise using a measure Φ_i of mean *residential* income in origin location i, computed using the model. Intuitively, Φ_i captures the average of wages brought "home" by workers who live in i and who work in various other destinations. To assess whether Φ_i contains non-trivial information, we compare it with nighttime lights, which is a recognized measure of residential income [10]–[12]. We use a new version captured by the VIIRS satellites and curated by the Earth Observatory Group (EOG) at NOAA. The VIIRS data has higher spatial and temporal resolution than the older OLS data, and it does not have a saturation point.

Figure 2 shows a graphical comparison: nighttime light VIIRS in the top panel, and the $\log(\Phi_i)$ measure (at the tower cell level) in the bottom panel. The correspondence is generally good, yet there are clear points where the model can be improved (e.g. patterns along roads). In the paper we show using regression analysis that the income measure is informative after controlling for population density (interpolated from the census), and various simple indicators derived from cell phone data.

Ongoing and Future Work

We plan to exploit the time variation in the cell phone data to look at changes in the spatial distribution of economic activity over short and medium time horizons. For example, we can investigate the effects of short term disturbances, such as the transportation restrictions imposed by national events. We can also study whether, in the medium term, economic activity becomes more concentrated around city centers, or whether it actually moves towards city outskirts.

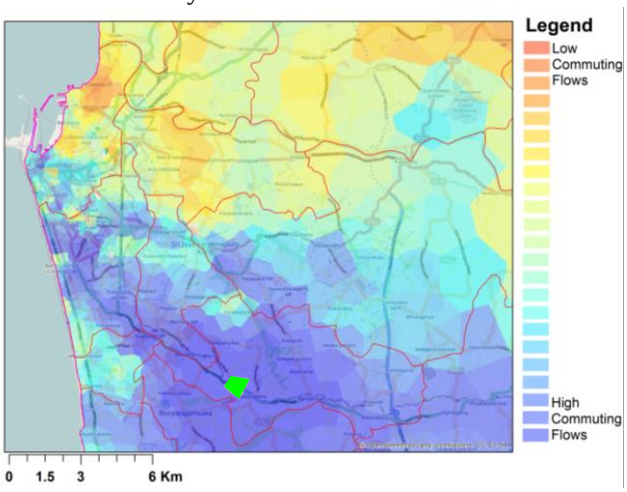


Figure 1. Smoothed commuting flows from a given origin tower cell (indicated by the green cell). Red lines indicate sub-district (divisional secretariat) boundaries.

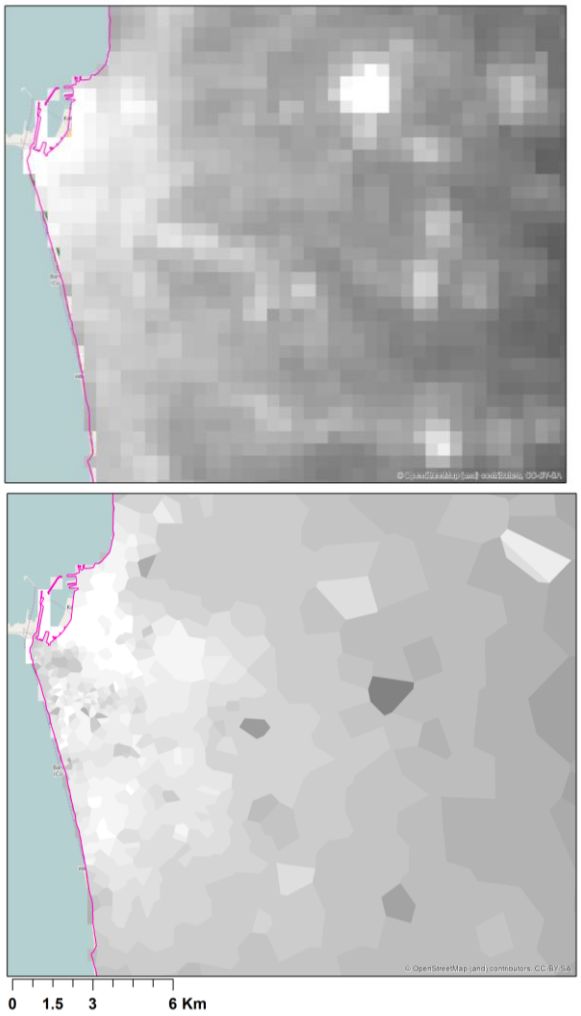


Figure 2. The top panel shows a log transformation of VIIRS nighttime lights. (The bright spot corresponds to an oil refinery.) The lower panel shows the measure of residential income $\log(\Phi_i)$ derived from the model (white corresponds to higher values).

References

[1] A. M. Mayda, "International migration: A panel data analysis of the determinants of bilateral flows," *J. Popul. Econ.*, vol. 23, pp. 1249–1274, 2010.

[2] F. Ortega and G. Peri, "The Effect of Income and Immigration Policies on International Migration," *NBER Work. Pap. Ser.*, no. 18322, 2012.

[3] O. J. Blanchard and F. Katz, "Regional Evolutions," *Brookings Pap. Econ. Act.*, vol. 23, no. 1, pp. 1–76, 1992.

[4] D. Hare, "'Push' versus 'pull' factors in migration outflows and returns: Determinants of migration status and spell duration among China's rural population," *J. Dev. Stud.*, vol. 35, no. January 2015, pp. 45–72, 1999.

[5] F. Calabrese, L. Ferrari, and V. D. Blondel, "Urban Sensing Using Mobile Phone Network Data: A Survey of Research," *ACM Comput. Surv.*, vol. 47, no. 2, pp. 1–20, Nov. 2014.

[6] J. Mitchell, R. J. Smith, M. R. Weale, S. Wright, and E. L. Salazar, "An indicator of monthly GDP and an early estimate of quarterly GDP growth," *Econ. J.*, vol. 115, pp. 108–129, 2005.

[7] G. M. Ahlfeldt, S. J. Redding, D. M. Sturm, C. Fieler, G. Grossman, B. Honore, U. Mueller, S. Kortum, E. Morales, and F. Hall, "THE ECONOMICS OF DENSITY :," *NBER Work. Pap. Ser.*, no. 20354, 2014.

[8] F. Calabrese, G. Di Lorenzo, L. Liu, and C. Ratti, "Estimating Origin-Destination Flows Using Mobile Phone Location Data," *IEEE Pervasive Comput.*, vol. 10, no. 4, pp. 36–44, Apr. 2011.

[9] P. Wang, T. Hunter, A. M. Bayen, K. Schechtner, and M. C. González, "Understanding road usage patterns in urban areas.," *Sci. Rep.*, vol. 2, p. 1001, Jan. 2012.

[10] J. V. Henderson and A. Storeygard, "Measuring Economic Growth from Outer Space," *Am. Econ. Rev.*, vol. 102, no. 2, pp. 994–1028, 2012.

[11] X. Chen and W. D. Nordhaus, "Using luminosity data as a proxy for economic statistics.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 108, no. 21, pp. 8589–8594, 2011.

[12] C. Mellander, K. Stolarick, Z. Matheson, and J. Lobo, "Night-Time Light Data : A Good Proxy Measure for Economic Activity ?," no. 315, p. 33, 2013.

Cignifi Risk Solutions

An application of mobile phone data for Brazil microcredit approval

Data Science Team*

Cignifi Inc.

In the past few decades, microfinance has gained broad attention as a way to promote the financial inclusion of the poor population, especially in developing countries. Microfinance also promotes economic development and increased employment. Popularized through the microcredit movement of the 1970s, a movement that included the founding of the Grameen Bank by Muhammad Yunus in Bangladesh, the system of microfinance has evolved to a system more comprehensive than one focused on just lending. The four main types of financial solutions that are offered are: i) credit; ii) savings; iii) insurance; iv) access to payment alternatives. Microcredit remains the best-known aspect of microfinance, and its development is linked primarily to the use of innovative mechanisms that make it possible to mitigate different market failures that keep the poor outside of the loan market.

Much debate has risen around the mixed outcome of microcredit. Proponents state that it helps the borrowers get out of poverty through the deploying of loans for small businesses and other productive means, while critics have mainly focused on the high interest rates leading the poor into a debt trap. The high interest rates are generally caused by: i) lack of relevant information to make a decision about credit; ii) high transaction cost. To overcome these two obstacles, true innovation is needed, and here we present a case study from Brazil, one utilizing alternative data from mobile telecommunications to gain credit insight and facilitate lending decisions with a much lower transaction cost.

Despite Brazil having taken the first initiative (the UNO project early in the 1970s), the microfinance system development there, after adjusting for the size of the population, is still lagging behind countries like Bangladesh, India or, in Latin America, Peru and Colombia. This leaves a huge gap in the meeting of the potential demand. The Central Bank reports that only 9% of the potential demand for microcredit in the country had been addressed in 2011 [1]. The aforementioned obstacles, notably information asymmetry and transaction costs, need to be surpassed, and this can only be done through innovation.

One of the most promising innovations that occurred recently is mobile payments, or pagamentos móveis. While by official accounts 56% of the adult population has bank

*Contact Qiuyan Xu, qxu@cinifi.com or Adriano A. Massuia, amassuia@cignifi.com for further information.

accounts in the country, the number drops to about 33%, even in an era of economic growth and full employment, when this value is adjusted by taking into consideration the poorest (20% of low-income) [2]. In comparison, the country reached the milestone of 274 million active mobile lines in March 2014, with a density of 135.3 active mobile lines per 100 inhabitants [3], making the use of a mobile payments system one of the most promising initiatives for financial inclusion.

We established a partnership with a Brazilian Telecommunication Operator (name not to be disclosed due to confidentiality agreement) and Microinvestor (Itau Microcredit), and from them we received information related to payment history by contract. The overlap between the information available on the mobile operator and the product information on microcredit restricted the study period to the year 2013. Because of the nature of the data, with its many open contracts with outstanding installment, survival analysis is used to estimate the default risk based on all call detail records (CDR).

Certain CDR variables are found to be significant to differentiate the cr These variables include: i) Numbe with standard price; ii) Number of calls to mobiles from the same carrier, but to different regions; and iii) Number of SMS received. Based on our model, it seems the population can be segmented to different risk profiles through key variables, a process which will allow operational deployment. Hence, creative ways to use CDR can potentially address the issues related to the lack of traditional credit history information and the high transaction cost.

References

[1] BCB *Relatório de Inclusão Financeira* 2011. Disponível em

[2] Demirgüç-Kunt, Asli; Klapper, Leora *Measuring Financial Inclusion: The Global Findex Database* 2012: Policy Research Working Paper 6025, World Bank, Washington, DC.

[3] Teleco *Estatísticas de celulares no Brasil* 2014. Disponível em: Acesso em: 18/04/2014.

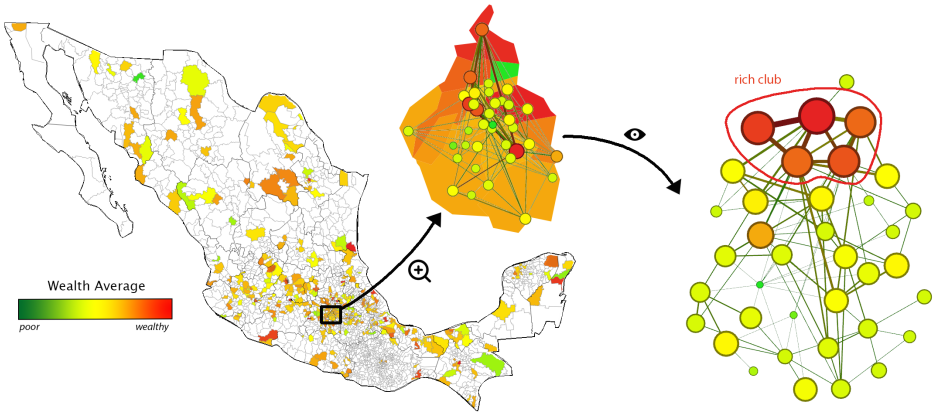
Socioeconomic correlations in communication networks

Yannick Leo[1], Eric Fleury[1], Carlos Sarraute[2], J. Ignacio Alvarez-Hamelin[3], and Márton Karsai[1,@]

[1] LIP, Université de Lyon, UMR 5668 CNRS - ENS Lyon - INRIA - UCB Lyon 1, IXXI
[2] Grandata Labs, Bartolome Cruz 1818 Vicente Lopez. Buenos Aires, Argentina
[3] CONICET, Universidad de Buenos Aires, Av. Paseo Colón 850, C1063ACV Buenos Aires, Argentina
[@] Corresponding author: marton.karsai@ens-lyon.fr

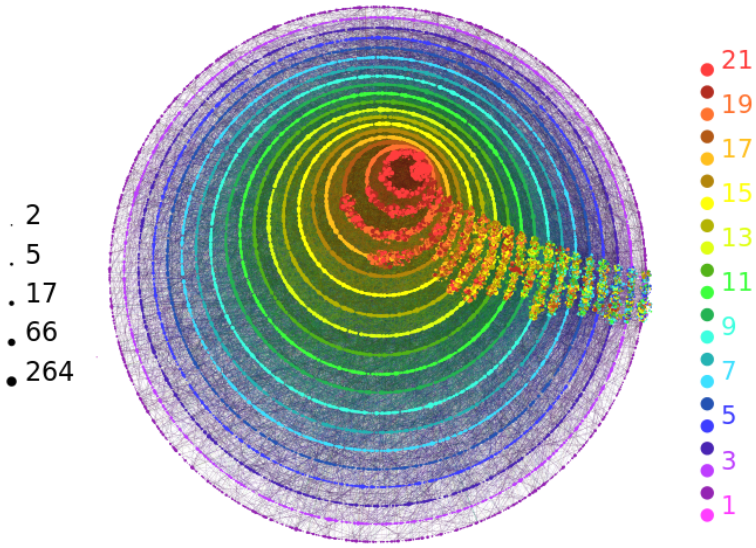
In this work we study the socioeconomic structure of a communication network by combining mobile communication records and bank credit informations of a large number of individuals living in Mexico. We provide empirical evidences about present economic unbalances suggesting not only the distribution of wealth but also the distribution of debts to follow the Pareto principle. Further we study the internal and interconnected structure of socioeconomic groups. Through a weighted core analysis we signal assortative correlations between people regarding their economic capacities, and show the existence of “rich-clubs” indicating present social stratification in the social structure.

Our understanding about the structure and dynamics of social systems has been developed considerably during the last years due to the recent availability of large digital datasets collecting interactions of millions of individuals. One of the most promising direction of studies involves mobile phone communication datasets. As mobile phones became personal items of our everyday life they are optimal to collect information about the location, communication, online, and physical activities of people, and to map out their temporally resolved social-communication network. These advancements give us to the opportunity to study the structure and evolution of very large social networks and signal general rules in human behaviour.



Geographical distribution and average economical capacities of mobile phone users in Mexico (left panel). Sample of users in Mexico City pinned at their most frequented location are depicted together with their social ties extracted from their call detailed records (middle panel). The social network of selected users demonstrates the existence of “rich-clubs” in the social structure where the wealthiest people appear to be connected (right panel). Colours are coding economical capacities, the width of links scales with the wealth of the ending nodes, and the size of nodes reflects the sum of link weights for each individual.

However, although these studies consider the temporal, structural, and spatial characters of human interactions they commonly miss one important dimension regarding the economic status of individuals. Economic capacities of people may largely determine their communication and social behaviour thus the emerging structure of the global social-communication network. Studies combining the social network with economic data could help us better understand spatial, and social segregation, or economic imbalances evolving in the society.



Weighted core decomposition of the largest connected component of the mobile communication network of users with known economic status. Size of vertices represents the number of contacts of individuals in the social graph, while colours decode their economic status ranging from least (magenta) to the most (red) wealthiest ones. The colour of circles is assigned according to the person's core shell determined by [5], while link colours on each end denote the shell of the connected node on the opposite end. Social stratification is evidenced by the narrow shell circles indicating that links connect mostly nodes from neighbouring shells. Components distributed radially on the right hand side are ones, which became disconnected from the largest component during the core decomposition process.

Here we propose a study, which moves along this direction by considering information about the mobile communication, location, and economic capacities of people. We use a communication dataset covering the mobile phone interactions of millions of people who are the clients of a single mobile provider in Mexico [1,2]. The data collects the geo-localised call detailed records (CDRs) for 6 consecutive months. This anonymised dataset is combined with bank credit informations of clients of a bank in the same county. The credit data collects the time, location, and amount of bank card purchases and the monthly evolution of incomes, spendings, and debts of the cell phone operator clients.

Using these informations we quantify the economic status of people to estimate their wealth and debts and infer these measures with the structure and evolution of the social-communication network extracted from the CDRs. We provide empirical evidences about present economic unbalances suggesting not only the distribution of wealth but also the distribution of debts to follow the Pareto principle [3,4]. We categorise individuals into different economic classes to understand the internal and interconnected structure of socioeconomic groups. Based on the economic status measures we perform a core analysis and show assortative correlations between people regarding their economic capacities. Further we provide quantitative evidences about the existence of “rich-clubs” and social stratification in the social structure.

References

[1] C. Sarraute, P. Blanc and J. Burroni, A Study of Age and Gender seen through Mobile Phone Usage Patterns in Mexico. *ASONAM IEEE/ACM* 836-843 (2014).

[2] J. Brea et. al., Harnessing Mobile Phone Social Network Topology to Infer Users Demographic Attributes. *SNAKDD’14 ACM* (2014).

[3] V. Pareto, Le Cours d’E´conomie Politique (Macmillan, London, 1897).

[4] H. Aoyama, et. al.,Pareto’s Law for Income of Individuals and Debt of Bankrupt Companies. *Fractals* **8** 293-300 (2000).

[5] J. I. Alvarez-Hamelin, L. Dall'Asta, A. Barrat, and A. Vespignani, Large scale networks fingerprinting and visualization using the k-core decomposition. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems* **18**, 41-50, Cambridge, MA, MIT Press (2006).

Understanding the Role of Social Networks on Labor Market Outcomes Using a Large Dataset from a Mobile Network

Filipa Reis, Pedro Ferreira
H. John Heinz III College, Carnegie Mellon University

This study uses call detail records from a major mobile carrier to identify job information networks within the neighborhoods of a large European metropolitan area. We start by identifying home and work locations for a random sample of business users using data on the cell towers they use to route calls. We show that the propensity of two workers to work in the same location (cell tower) is higher when they also live in the same location (cell tower). We extend this model by adding information on who calls whom, which allows us to label pairs of workers as likely to know each other – friends in short. We find that the effect of friendship on the propensity to work in the same location is two orders of magnitude greater than that associated with living in the same location. Our findings are robust to alternative specifications of our dependent and independent variables such as the definition and the strength of friendship, the distance between home and work locations and using working for the same company in lieu of (just) working in the same location. These specifications are introduced to reduce our concerns with homophily and reverse causality. We also show that job information networks are more important in less privileged neighborhoods, measured by mortgages costs and unemployment rates. This finding provides some evidence that social networks may not convey enough information about job market opportunities to help underprivileged individuals.

The deliberate use of formal and informal social connections for information access and exchange is a strategy commonly employed by individuals particularly evident during job searches (Ioannides and Loury, 2004). Previous literature has shown that social interactions among people who live in close geographical proximity play a significant role on labor market outcomes (Bayer, Ross, and Topa, 2008). The research methodologies traditionally used to study this phenomenon – social experiments (Jacob, 2005; Katz, Kling, and Liebman, 2001), and survey data (Bayer, Ross, and Topa, 2008) – are only rarely produced, extremely costly to generate and, most importantly, very seldom contain information on actual social networks. The inability to observe who talks to whom may result in confoundedness between social connections and unobserved neighborhood characteristics.

The recent growth in the use of mobile communications and the subsequent availability of datasets on who calls whom, when and from where, opens up avenues for novel research approaches that may complement, or even substitute, the more traditional research methodologies (e.g. Soto, Frias-Martinez, Virseda, Frias-Martinez, 2011; Frias-Martinez and Virseda, 2012; Blumstock, 2012). Using data from mobile communications may help alleviate the above-mentioned concern; though certainly not all communication occurs over mobile phones, it is likely that one's most important social connections show up in Call Detailed Records (CDRs).

In this work, we use eleven months of CDRs from a major European mobile carrier to study how social connections among neighbors and friends affect their propensity to work together. This carrier is the market leader in the corporate segment in the country and in particular in the region we analyze. We start our analysis by inferring home and work location for a random sample of business subscribers. These locations are defined as the most used cell tower during nighttime (between 7 pm and 7 am) and daytime (between 1 pm and 5 pm), respectively. Using the geographical information associated to cell towers (latitude and longitude) we identify the lowest possible statistical region – parishes -- in which cell towers are located. With this information, we classify subscribers who live in the same cell tower as neighbors. We then assess the propensity of pairs of neighbors to work in the same location (defined as working in the same cell tower), compared to the propensity of pairs of people who live in the same parish to do so, interpreting the difference between these two effects as an indication of job information networks among neighbors.

Our results in Table 1 show that, controlling for parish fixed-effects, pairs of neighbors are 33% more likely to work in the same location than pairs of subscribers who just live in the same parish (an increase of 0.002, p<0.001, relative to the baseline propensity of 0.006, p<0.001). We then extend our model by adding information on whether these pairs of individuals use their mobile phones to talk to each other. If they do so, we call them friends. This is where our analysis adds a critical step to what has been previously done in the literature. The latter has been unable to account for the actual social connections among neighbors. We find that the propensity of two neighbors who know each other -- friends -- to work in the same location is two orders of magnitude greater (0.2601, p<0.001) than that observed for neighbors who do not talk to each other (0.0013, p<0.001). These results are robust to a different specification of the dependent variable, namely, an indicator of whether the two neighbors work for the same company (which does not actually entail working in the same location). These results are also robust to different specifications of friendship.

	(1)	(2)	(3)	(4)
dep. var:	**wcell**	**wcell**	**wcell**	**wcell**
neighbors	-0.0002	0.0020***	0.0013***	0.0010***
	(0.0001)	(0.0002)	(0.0002)	(0.0002)
friends			0.2601***	0.2250***
			(0.0075)	(0.0087)
neighbors and friends				0.1075***
				(0.0165)
intercept	0.0061***	-	-	-
	(0.0001)	-	-	-
fixed effects for home parish	no	yes	yes	yes
N	1495157	1495157	1495157	1495157
R-sq	0.000	0.007	0.033	0.034

Standard errors in parentheses
+ p<0.1 * p<0.05 ** p<0.01 *** p<0.001

Table 1. Regression results for 1,495,157 subscriber pairs. Our dependent variable (wcell) indicates whether *i* and *j* work in the same cell tower; *neighbors* indicates whether they live in the same cell. (1) Gives us the baseline propensity for two subscribers who live in the same parish (but not in the same cell) to work in the same location, 0.0061 (p<0.001). In (2), we add parish-level fixed-effects to control for potentially time-invariant unobserved parish characteristics. In (3) we add *friends* indicating whether *i* and *j* call each other, capturing the effect of being friends on the propensity to work in the same location, 0.2601 (p<0.001), controlling for being (or not) neighbors. In (4) the interaction term, *neighbors and friends,* captures how the effect of being friends on the propensity to work in the same location changes for neighbors versus people who are not neighbors. The positive result (0.1075, p<0.001) indicates that being friends increases the propensity to work in the same location across neighbors more than across pairs of people that live further apart.

Finally, we also test whether job information networks among neighbors play a more important role in less privileged neighborhoods as suggested by prior research (Elliot, 1999; Weinberg, Reagan, and Yankow, 2004). Our results confirm that friendships among neighbors living in neighborhoods with lower mortgage costs and /or higher unemployment rates play an important role in determining whether people work in the same location or for the same company. This finding provides some evidence that social connections may not necessarily help improve the employment conditions for disadvantaged individuals and calls for re-allocation programs that may help the poorer find housing in more affluent neighborhoods.

Our research contributes to the body of evidence on the role of social networks on labor market outcomes. Using only observational data, we extend previous analyses (e.g. Bayer et al., 2008) to gain further insight on the role of job information networks at the neighborhood level and how it may vary according to the neighborhood's socioeconomic status. Our work provides some evidence of the potential applications of big data generated by ICTs to explore and analyze social and economic issues.

Bayer P., Ross, S.L., Topa, G. 2008. Place of Work and Place of Residence: Informal Hiring Networks and Labor Market Ouctomes. Journal of Political Economy, Vol. 116 (6), pp. 1150-1196.

Blumenstock, J. 2012 Inferring Patterns of Internal Migration from Mobile Phone Call Records: Evidence from Rwanda. Information Technology and Development 18.

Elliot J. R., 1999. Social Isolation and Labor Market Insulation: Network and Neighborhood Effects on Less-Educated Urban Workers. The Sociological Quarterly, Vol. 40 (2), pp. 199-216.

Frias-Martinez, V., Virseda, J., Rubio, V., Frias-Martinez, E. 2010. Towards large-scale technology impact analyses: automatic residential localization from mobile phone-call data. In Proceedings of the 4th ACM/IEEE International Conference on Information and Communication Technologies and Development (ICTD '10). ACM, New York, NY, USA.

Ioannides Y. M., Loury, L. D. 2004. Job Information Networks, Neighborhood Effects, and Inequality. Journal of Economic Literature, Vol. 42 (4), pp. 1056-1093.

Jacob, B. 2004. Public Housing, Housing Vouchers and Student Achievement: Evidence from the Public Housing Demolitions in Chicago. American Economic Review Vol. 94, pp. 233-58.

Katz, L. F., Kling, J.R., Liebman J.B. 2001. Moving to Opportunity in Boston: Early Results of a Randomized Mobility Experiment. Quarterly Journal of Economics Vol. 116, pp. 607-654.

Soto, V., Frias-Martinez, V., Virseda, J., and Frias-Martinez, E. 2011. Prediction of Socioeconomic Levels using Cell Phone Records. User Modeling, Adaption and Personalization Lecture Notes in Computer Science, Vol. 6787, pp. 377-388.

Weinberg B. A., P. B. Reagan, Yankow, J. J. 2004. Do Neighborhoods Affect Hours Worked? Evidence from Longitudinal Data. Journal of Labor Economics, Vol. 22(4), pp. 891-924.

Session 4 :: Societies (I)

Persistent daily patterns in mobile telephone communication

T. Aledavood[1,*], E. López[2], S.G.B. Roberts[3], F. Reed-Tsochas[2,4], R.I.M. Dunbar[5], E. Moro[6,] J. Saramäki[1]

[1] Dept. of Biomedical Engineering and Computational Science, Aalto University School of Science, Espoo, Finland; [2] CABDyN Complexity Center, Saïd Business School, University of Oxford, UK; [3] Department of Psychology, University of Chester, UK; [4] Department of Sociology, University of Oxford, UK; [5]Department of Experimental Psychology, University of Oxford, UK; [6] Department of Mathematics, Universidad Carlos III de Madrid, Spain; * email: *talayeh.aledavood@aalto.fi*

Human activity patterns are known to follow circadian rhythms[1,2], and this is also reflected in communication activity. We investigate the daily call patterns of 24 students over a period of 18 months[3]. During this period, the students finished high school and went to university or work. We show that while their call patterns are coarsely similar, each individual is different and follows their own rhythm of communication. These individual rhythms are seen to be persistent despite the major changes in the students' circumstances. Using entropy measures, we show that the patterns are partially activity-driven – i.e., around certain times communication is fairly random and shows no preference to specific alters (acquaintances, relatives) – and partially socially driven, such that at given times certain alters are preferred. While there is more variation here than in communication rhythms, it is seen that often, daytime calls are more random and evening/night time calls more alter-specific and focused towards alters with higher ranks (rank of each alter is determined based on fraction of calls received by that alter). We also use survey information on the called alters to determine whether calls to family members or alters of same and opposite gender follow specific rhythms.

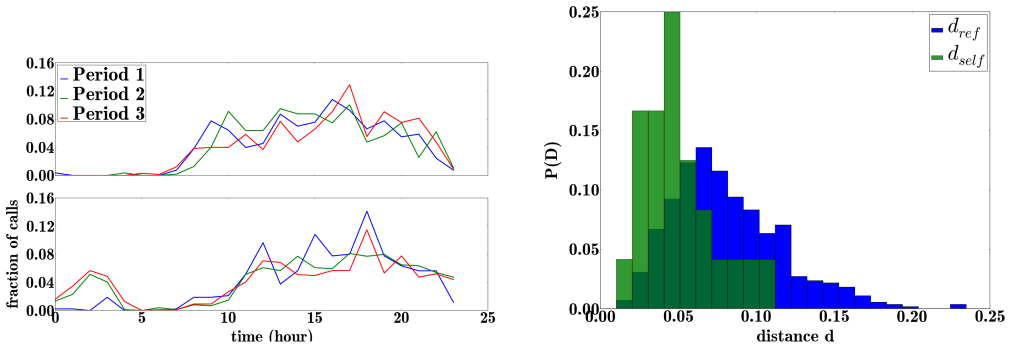


Figure: LEFT: the daily call patterns of two students during three consecutive 6-month time periods. RIGHT: PDF of the Jensen-Shannon divergence measures of the daily patterns of students, where d_{self} measures the distance between the different 6-month patterns of the same student, and d_{ref} the distance of the patterns between students within one period. It is seen that on average, the distance between the patterns of one student is smaller than the reference distance, i.e. the patterns show persistence in time.

References

1. H.-H. Jo, M. Karsai, J. Kertesz, and K. Kaski, *Circadian pattern and burstiness in mobile phone communication*, New Journal of Physics **14**, 013055 (2012)

2. T. Louail, Maxime Lenormand, Oliva García Cantú, Miguel Picornell, Ricardo Herranz, Enrique Frias-Martinez, José J. Ramasco, Marc Barthelemy, *From mobile phone data to the spatial structure of cities*, arXiv: 1401.4540 (2014)
3. Jari Saramäki, E. A. Leicht, Eduardo López, Sam G. B. Roberts, Felix Reed-Tsochas, and Robin I. M. Dunbar, *Persistence of social signatures in human communication*, PNAS **111**, 942 (2014)

Inferring social ties from WiFi scan results

Piotr Sapiezynski, David Kofoed Wind,
Arkadiusz Stopczynski, Sune Lehmann

It has become increasingly popular to study human interactions and mobility through mobile sensors embedded in smartphones [1, 2]. Most commonly, Bluetooth is used as a ground truth proxy for face-to-face interactions because of its short physical range (up to ~10 meters) and relatively low impact on battery of the sensing device. There are a number of reasons why it is crucial to investigate alternative approaches to sensing physical proximity. Firstly, relying on a single channel is prone to error, especially considering frequent hardware and firmware malfunctions of the sensing devices. Furthermore, while Bluetooth data is rarely collected outside of controlled social science experiments, other channels—such as WiFi scans—are often collected during normal operation in many contexts. Such WiFi scan information may be available for hundreds of millions of people: many of the most popular applications available in Google Play market require the permission which allows them to read the list of all visible access points in each scan. This situation has important privacy implications, which we investigate here in context of inferring the social network of users based on their WiFi scan results.

Based on the data collected during a longitudinal study with ~1000 participants, we extend previously suggested approaches to WiFi-based proximity inference: we investigate machine learning methods as well as verify our previous findings [2] by using now-available information with higher temporal resolution. To validate the idea of using WiFi as a proxy for physical proximity, and to explore the possibilities provided by WiFi scan information, we infer social network of users. We train a statistical model exploring various behavioral features (see Figure 1) which, given the WiFi scans of two individuals, estimates the probability that the individuals are connected as friends on Facebook. To evaluate the models, we compare its estimations to the actual online social network of the participants (Figure 2).

References

[1] N. Eagle and A. Pentland. Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing*, 10(4):255–268, 2006.

[2] A. Stopczynski, V. Sekara, P. Sapiezynski, A. Cuttone, MM Madsen, J. Eg Larsen, and S. Lehmann. Measuring large-scale social networks with high resolution. *PLoS ONE*, 9(4):e95978, 04 2014.

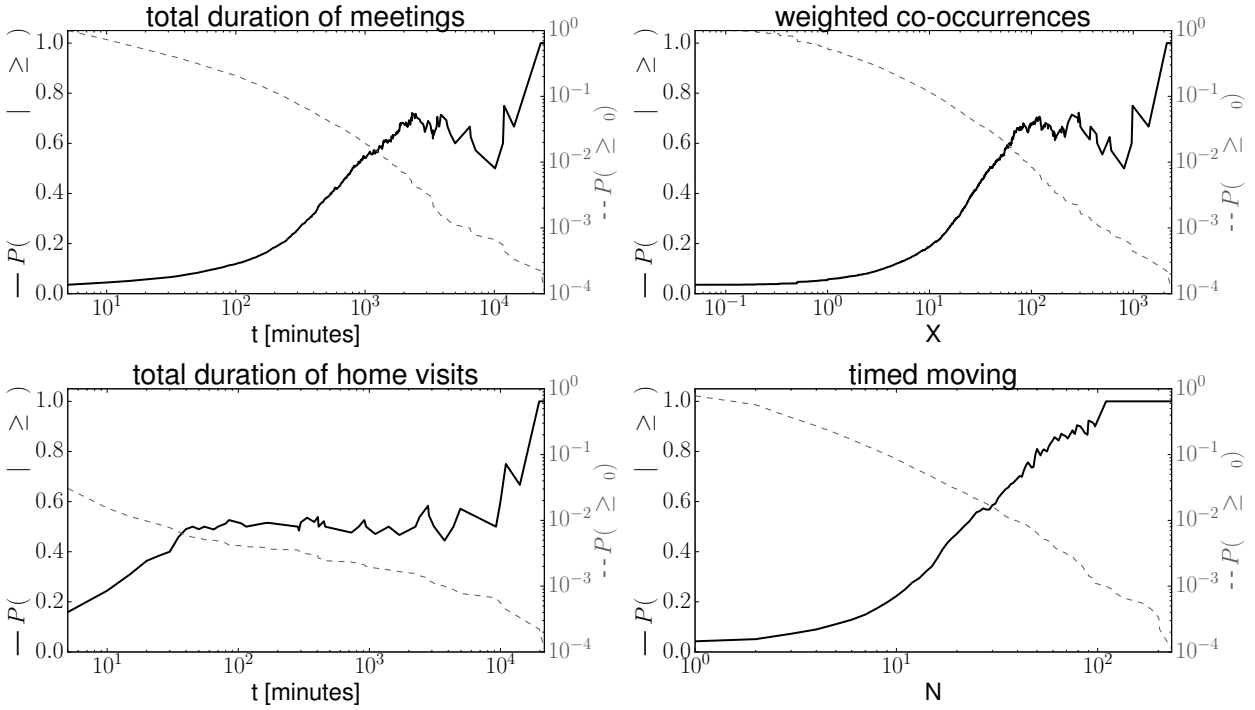


Figure 1: We extract various behavioral features from WiFi scans of the participants' phones: we measure how much time they spend together in total; using the weighted co-occurrences feature we incorporate information about intimacy of meetings: the lower the number of people present at a location, the more important the meeting is; home visits are indicative of friendships, but because of densely populated blocks and dormitories, many false positives are generated; we also observe instances of two people arriving at a location or leaving it at the same time.

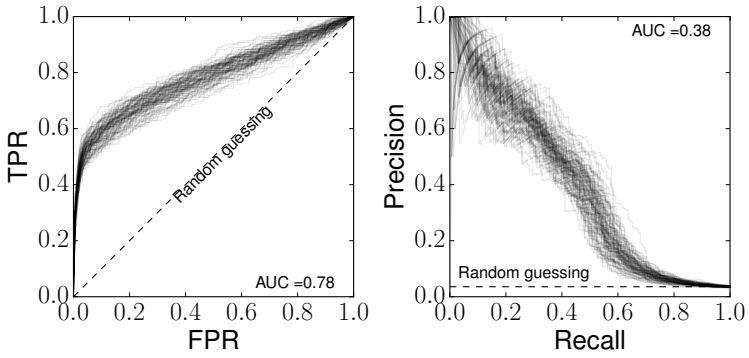


Figure 2: We train 100 instances of Random Forest Classifier using the presented features (among others) to infer friendship links on the online social network. Clearly, physical proximity is a very strong indicator of online friendship links, but there are also many links online which are not reflected in offline relations - a fact highlighted by the saturation of ROC curve at $TPR \approx 0.5$.

On the complementary roles of face-to-face and mediated social interactions

Guy Zyskind
MIT Media Lab

Bruno Lepri
Fondazione Bruno Kessler

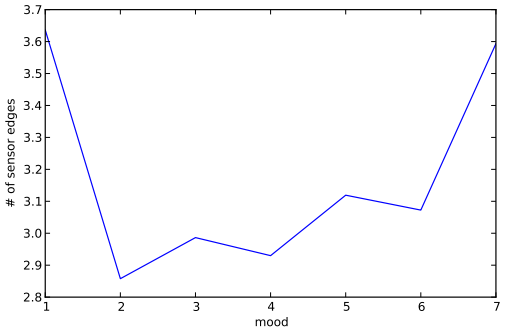
Alex (Sandy) Pentland
MIT Media Lab

Erez Shmueli
Tel-Aviv University

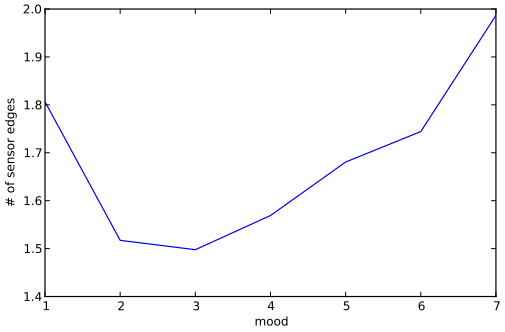
Social interaction patterns have long been a subject of great interest. Several recent studies (e.g., [3, 2]) have examined the influence of face-to-face and mediated (e.g., phone, text and online social networks) communication channels on employees' mood, but since their motivation stemmed from professional productivity, they did not include interactions that occurred outside the working environment. In this paper, we seek to understand how mood is affected by face-to-face and mediated interactions during the entire day, focusing on the relationship between the two types of interactions. Unlike previous studies, our analysis is performed on two distinct high-resolution datasets that were collected over long periods of time. We find that in different moods, people prefer one communication method (either face-to-face or mediated) and that the two types of interactions have complementary roles in practice. Moreover, this relationship becomes specifically strong in cases of extreme mood states – i.e. when people report being inordinately happy or unhappy. Our findings provide initial insights into how mood is affected by different types of social interactions, and can be used in the future as a basis for predicting and influencing individuals' mood based on the pattern of their social interactions.

For our analysis, we used the *Friends and Family* dataset [1] and the *Mobile Territorial Lab* dataset[1]. The *Friends and Family* dataset contains data on $n = 130$ participants that has been collected over the course of 12 months. The *Mobile Territorial Lab* dataset contains data on $n = 70$ participants that has been collected over the course of 6 weeks. In both datasets, call and text logs were collected using a dedicated Android app installed on the mobile devices of the participants. In addition, Bluetooth scans were collected and used as a proxy to face-to-face encounters with other individuals. Finally, mood data was collected via self-reported surveys that were completed by the participants on a daily basis. The self-reported surveys were introduced automatically as part of a mobile application, thus ensuring the reliability of the self-reported results. While the *Friends and Family* dataset measured mood directly on a scale of $1-7$, the *Mobile Territorial Lab* followed the *Positive and Negative Affect Schedule* measure (PANAS) [4]. In order to make the results in the *Mobile Territorial Lab* dataset comparable to those of the *Friends and Family* dataset, we computed the average PA portion of the survey, rounded to the nearest integer, resulting in 5 distinct mood states.

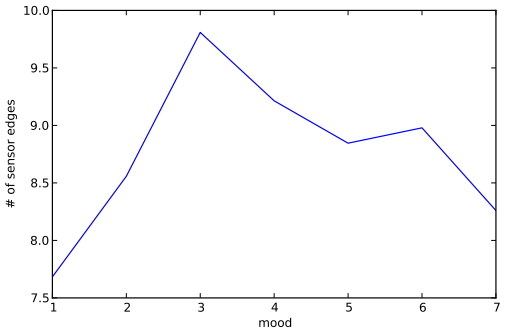
[1] http://www.mobileterritoriallab.eu/



(a) Average number of distinct daily calls.



(b) Average number of distinct daily texts.

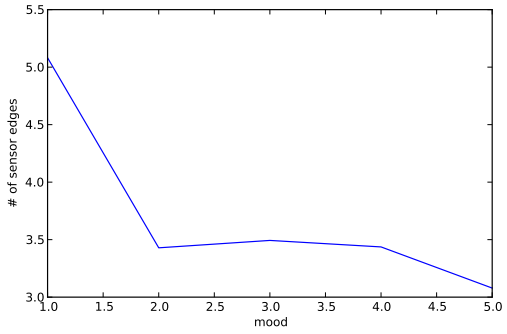


(c) Average number of distinct face-to-face encounters.

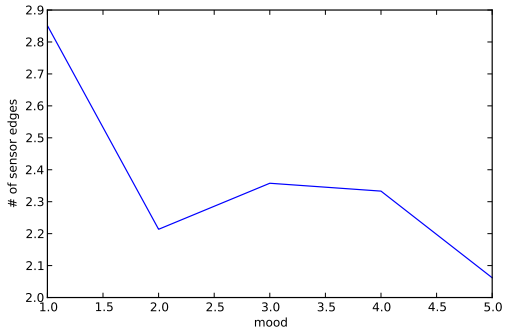
Figure 1: Communication patterns effect on mood in the*Friends and Family* dataset.

Figure 1 shows the average number of distinct interactions depending on the mood, for all individuals and for all

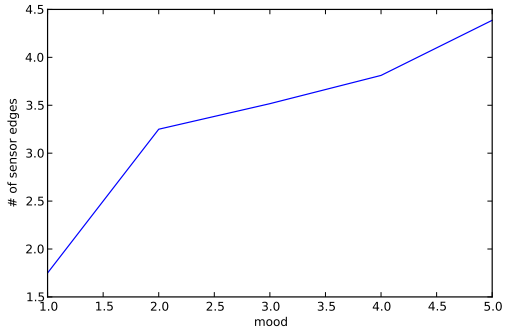
days, in the *Friends and Family* dataset. As can be seen in the figure, calls and texts, both being mediated types of communication, exhibit similar patterns, while face-to-face follows a complementary pattern. More specifically, in extreme situations ($mood = 1$ or $mood = 7$), people prefer to use mediated communication over face-to-face interaction, and the opposite result holds in intermediate mood states ($mood \in [2, 6]$).



(a) Average number of distinct daily calls.



(b) Average number of distinct daily texts.



(c) Average number of distinct face-to-face encounters.

Figure 2: Communication patterns effect on mood in the *Mobile Territorial Lab* dataset.

Figure 2 presents the same analysis for the *Mobile Territorial Lab* dataset, focusing on Positive Affect (PA) questions. As can be seen in the figure, the results are consistent in the sense that the same complementary relationship between mediated communication (i.e., calls and texts) and face-to-face encounters is present. In addition, while not shown here, it is constructive to note that the Negative Affect (NA) questions also followed a consistent pattern.

In conclusion, understanding how mood is affected by different types of social interactions provides a mechanism for inferring well-being. Hence, our results could be used as a first important step in achieving this goal. Further research could examine more closely different types of subjects based on their personality traits (e.g., using the Big Five personality traits model) and see if it affects their communication preferences. Similarly, controlling for the strength of the ties could also reveal interesting results.

REFERENCES

1. Aharony, N., Pan, W., Ip, C., Khayal, I., and Pentland, A. Social fMRI: Investigating and shaping social mechanisms in the real world. *Pervasive and Mobile Computing 7*, 6 (Dec. 2011), 643–659.
2. Counts, S. Understanding Affect in the Workplace via Social Media.
3. Mark, G., Iqbal, S., Czerwinski, M., and Johns, P. Capturing the M ood : Facebook and Face - to - Face Encounters in the Workplace.
4. Watson, D., Clark, L. A., and Tellegen, A. Development and validation of brief measures of positive and negative affect: the panas scales. *Journal of personality and social psychology 54*, 6 (1988), 1063.

Understanding User Attributes from Calling Behavior:

Exploring Call Detail Records through Field Observations and Potential of Estimating User Attributes of Anonymized Call Records

Ayumi Arai,
Hiroshi Kanasugi, Xiaowei Shao, Ryosuke Shibasaki
University of Tokyo
Tokyo, Japan

Apichon Witayangkurn
Asian Institute of Technology
Pathumthani, Thailand

Teerayut Horanont
Thammasat University
Pathumthani, Thail

1. INTRODUCTION

With the rapid spread of mobile phones, analyses of large-scale data such as GPS logs and call detail records (CDRs) have provided detailed descriptions of human mobility. An increasing body of human mobility research is focused on modeling the properties of human mobility patterns, where they are generalized in a quantitative manner. Although these studies succeed in mining mobility patterns [3,4], the outcomes of large-scale data analyses describe the movement of crowds, because the data is anonymized.

In addition to such large-scale datasets, there are other types of data collected through conventional methods. In the field of urban planning and transportation, many empirical studies have attempted to identify the factors that affect human activities and travel patterns, by analyzing data collected through questionnaire surveys. Family and social obligations are often presented as significant factors affecting daily travel and activity behavior [9]. This implies that people's activity patterns are constrained by social ties. This observation is consistent with most previous analysis of human mobility patterns from large-scale datasets, where people routinely visit a limited number of locations [13,14]. Though such conventional approaches may be less efficient in defining human activity and travel patterns in terms of population size and data period length, it can link activity and travel patterns with people's attributes to some extent [12]. To further investigate the properties of user attributes in anonymized data, it is critical to analyze this large-scale data in combination with such secondary data [8].

In fact, emerging studies are attempting to analyze data derived from mobile phones in combination with secondary data. [2] observed that the ratio of shared phone usage and call type, such as incoming and outbound calls, shows significant differences by gender, on average. In addition, the number of usages during a specified length of time was shown to differ according to income level. This observation is consistent with another research project, which analyzed social connectivity through social networking activities [1]. Utilizing sensor data from volunteer mobile users, [10] proposed prediction models based on user demographic attributes.

In this work, we provide a novel approach for extracting features from calling behavior, which can be constructed from anonymized CDRs. This technique can reveal distinctive traits that help identify the demographic attributes of mobile phone users. This study is unique, because we focus on extracting lifestyle traits and routines to identify user attributes. We statistically analyze data collected through a field survey that focused on the demographic attributes, calling behavior, and weekly activity patterns of mobile users. Several key features are empirically derived to analyze user attributes.

Contributions of our work are described below:

- Statistical analysis results are provided for calling behavior based on field survey data. We extract calling behavior traits to differentiate gender that correspond to the patterns of routine activities. To do so, we introduce the concept of weekly activity patterns, which specifies whether the day's call records correspond to a day where the user is engaged in their primary routine.
- Prototypes of calling behaviors are provided. We describe how the prototypes and user attributes are related. Potential of applying our results to estimate user attributes of anonymized call records are discussed.

2. DATA

To understand the hidden properties of CDRs, we conducted a field survey, named the Survey on Patterns of Activity for Comprehensive Explorations of Mobile Phone Users in Dhaka (SPACE). The purpose of SPACE is to understand the calling behavior, characteristics, and lifestyles of mobile phone users. A unique characteristic of the survey data is that it includes actual call records from mobile phone users, as single-day records from 922 handsets. The records specify the call's location type and other basic attributes. The data contain neither mobile phone numbers nor any other information explicitly specifying individuals. Our survey site was Greater Dhaka, which is composed of Dhaka City Cooperation (DCC), selected surrounding municipal areas, and regions outside of the urban area. SPACE was conducted from November 27, 2013 through January 4, 2014. The survey interviewed 810 households, and included 922 mobile phone users.

3. LIFESTYLE AND CALLING BEHAVIOR

To understand activity patterns, it is vital to identify a certain set of locations where people spend the majority of their time. Many studies identify two dominant locations for people as their home and work or school (hereinafter referred to as "*Home*" and "*Primary out-of home location*"), which can explain a significant portion of their activity [13] and thereby their activity patterns and portions of their lifestyles. Therefore, in this section we examine the time distribution of calls initiated from *Home*, *Primary out-of-home location*, or *Others* by analyzing the calling behavior of SPACES's mobile phone users. Because sequential location histories can infer similarities between people [7], we expect the distributions to vary according to differences in user attributes.

This abstract is prepared based on the paper presented at *the 12th International Conference on Advances in Mobile Computing and Multimedia* (MoMM'14), December 08 - 10 2014, Kaohsiung, Taiwan

We extracted calling behavior prototypes according to gender, and examine how their traits differ.

For the analysis provided this section we introduce the concept of weekly activity patterns, which specifies whether mobile users are engaged in their primary routine on the day the call was made. Routines can be any activities users spend the majority of their time on during the day. The routine that the user follows on the highest number of days during the week is considered as their primary routine.

3.1 Weekly Activity Patterns

In this subsection, we describe the weekly activity patterns of the 922 mobile users recorded in the SPACE data. Weekly activity patterns are determined based on the number of weekly routines (primary and non-primary), and the number of days on which the primary routine is followed. By classifying the activity according to a pattern, we examine how activity patterns are linked to calling behavior in the following section. We assume that these patterns can also be partially extracted from anonymized CDRs, which allows us to estimate significant locations such as home and work places. Based on this technique, we consider it possible to reconstruct weekly activity patterns from CDRs, to which we can apply this concept. Table 1 classifies the mobile users into four patterns, according to the following criteria.

Table 1. Weekly activity patterns of 922 mobile users in SPACE data

Pattern	Description of people classified into the pattern	Number of routines	Number of days for the primary routine	%
Group (1)	Those who repeat their primary routine activity every day. Mainly composed of income earners and those who do household tasks.	1	7	56
Group (2)	Those who repeat their primary routine activity every day, except for Friday. Mainly composed of income earners or are students.	2	6	30
Group (3)	Those who repeat their primary routine activity every day except for Friday and Saturday. Mostly composed of public sector employees or students.	2	5	9
Group (4)	Mostly income earners and their activity patterns do not follow Pattern (1), (2), or (3).	1 or more	1 to 6	5

3.2 Calling Behavior by Gender

This subsection compares the calling behavior of males and females by analyzing the call records of the 922 mobile users. First, we examine calling behavior by analyzing call location trends. Figure 1 shows percentages for the number of calls at (a) *Home* and (b) *Primary out-of-home location* among the total number of calls on the primary routine day. As the figure shows, males and females exhibit distinctly different call location trends. Across all patterns, females tend to call predominantly from *Home* while males call from both the *Home* and *Primary out-of-home location*. This indicates that identifying dominant call locations on the primary routine day is important for determining gender. This feature is particularly distinctive among those classified into Group (1), where the primary routine is repeated every day. We conclude that considering the locations of calls and the type of day (primary or non-primary routine day) while examining call records is crucial for extracting gender-wise traits. Assuming that weekly activity patterns can also be reconstructed from the CDRs, we conclude that our findings can be utilized to estimate the gender of the user who produced an anonymized CDR.

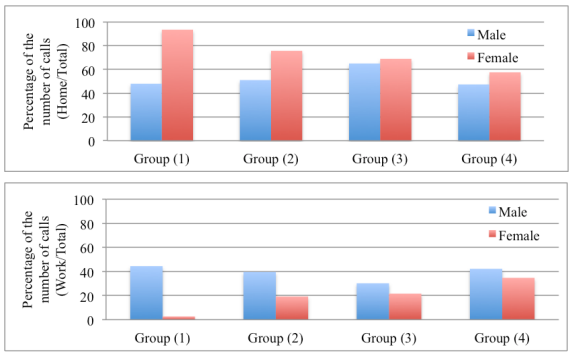


Figure 1. Percentage of calls from (a) *Home* and (b) *Primary out-of-home location* against total calls on the primary routine day by gender (a) Upper (b) Lower

4. PROTOTYPES OF CALLING BEHAVIOR

In this section, we narrowed the time window from a weekly basis to an hourly basis, to understand trends in calling behavior across gender and occupation types. As a result, we aggregated the call times by hour in this section. We extracted parts-based representations of calling behavior by conducting vector quantization against the time and location distribution of call records. This enabled us to cluster the data into mutually exclusive prototypes [5].

4.1 Method for Extracting Prototypes

We employed non-negative matrix factorization (NMF) for vector quantization. NMF was applied to the call records of the 922 mobile users, where the distribution of call records for a single day is expressed as a 72 x 1 matrix. The first set of 24 elements out of 72 consists of hourly counts of call records for *Home*. The first element is the total number of calls from *Hour 0*, occurring between 0:00 and 0:59, and the 24th element denotes *Hour 23*, occurring between 23:00 and 23:59. The next set of 24 elements is structured similarly to the first set, and it accounts for the number of calls from the *Primary out-of home location*. In a similar manner, hourly counts of calls for other locations, which are any locations except for home and work/school, are captured by an additional set of 24 elements. As a result, we obtained 922 sets of 72 x 1 column vectors. To employ NMF, we solved the equation below by following an algorithm, which allows only additive combinations [9]:

$$V \approx WH$$

where we obtained non-negative matrix factors W and H given a non-negative matrix V. Given 922 sets of 72 x 1 column vectors, the vectors were placed in the columns of a 72 x 922 matrix V. This matrix was approximately factorized into a 72 x r matrix W, and an r x 922 matrix H. Here we selected r = 3 to analyze differences in the major features of calling behavior. We then defined the cost function that evaluates the quality of approximation for iterative updates of W and H. We calculated the distance between two non-negative matrices and measured the square of the Euclidian distance [14]. As a result of repeated iterations, we obtained an optimal matrix factorization.

To understand the calling behavior prototypes for males and females, we split the call records obtained from SPACE by gender. Then, we separated the records by primary and non-primary

routine days. We assumed that the calling behavior of the primary routine days was different from that of the non-primary routine days, based on the analysis results described in the previous section.

4.2 Prototypes of the Primary Routine Day

Figures 2(a) and 2(b) show three calling behavior prototypes for males and females, whose call records in the SPACE data fall on their primary routine days. Each prototype is expressed in a 24 x 3 matrix, showing time distributions between zero and 23 hours in the rows, and call locations of *Home*, *Primary out-of home location*, and *Others* in each column from left to right. The intensity of the color indicates the intensity of calls for the specified time band and location. For instance, in each matrix, the cell on the top of the left column represents the intensity of calls for *Hour 0* from home. The percentage indicates how significant each component is for explaining the population's call record trends. For example, Pattern m1, whose percentage is 46% and the greatest among the three, is the most dominant pattern for the calling behavior of males on the primary routine day.

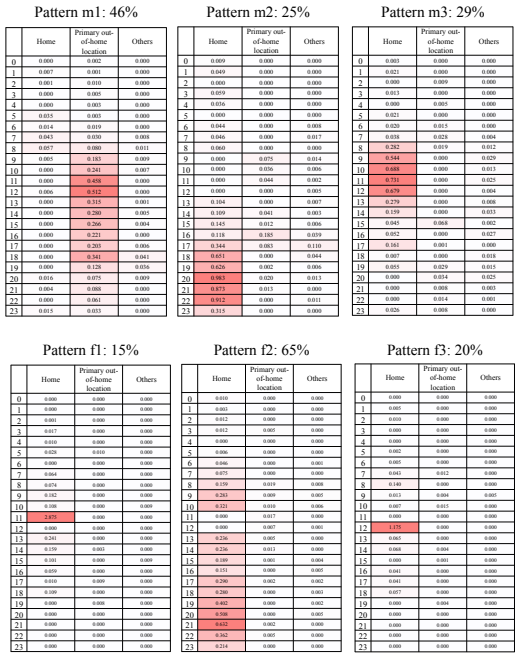


Figure 2. Time distributions of call locations for (a) male and (b) female on the primary routine day (a) Upper (b) Lower

Pattern m1 shows that males tended to call from *Primary out-of home location* around midday, peaking in *Hour 12*. In contrast, the most dominant pattern for females, Pattern f2, shows a peak of *Hour 21* at *Home*. Furthermore, all extracted patterns for females show a higher intensity of calls from *Home*. This trend follows the one exhibited in Figure 2, which indicates that call locations could be a key to distinguishing the gender of users. Although the trend in call locations partially captures gender differences, frequently recorded hours for the remainder of males, Pattern m2 and m3, are apparently difficult to distinguish from those for females, Patterns f1, f2, and f3. For both males and females, intensive calling hours are *Hour 11*, *Hour 12*, *Hour 20*, and *Hour 21* from *Home*. It is still fair to conclude that the location of calls around midday could be a key to identifying the user's gender. That is, males mostly call from *Primary out-of home location* and females call from *Home* around midday on their primary routine days.

Figures 3(a) and 3(b) describe the distribution of the calling patterns illustrated in Figures 2(a) and 2(b) for individual income levels. We split call records for males into four groups based on their individual income level. Call records for females were split into three income groups and a separate non-income group, because the majority of the female users did not earn an income.

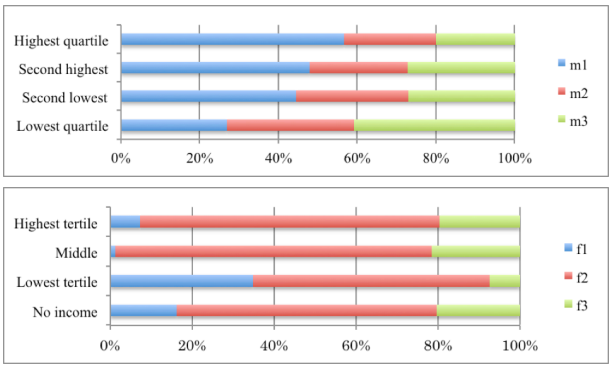


Figure 3. Distribution of individual income levels for three principle patterns for (a) male and (b) female (a) Upper (b) Lower

Among males, we observed that their common patterns varied according to their income level. As described in Figure 2(a), the higher income levels contained larger ratios of Pattern m1. Conversely, the lower individual income levels contained larger ratios of Pattern m3. This result is consistent with our field observations where males in lower-income groups tend to be engaged in self-employed jobs whose work locations tend to be their own home. However, no particular trends were observed among females across the income levels. This result indicates that further examination is necessary to determine the number of clusters when we employ NMF to obtain prototypes. Our findings imply that calling behavior could be a key to understanding the individual income level for males, but not for females.

It is worth noting that the income level used here is not a household income but individual income, which means we assess how much each person earns annually. Incidentally, we do not observe distinctive differences in calling patterns due to different household income levels when we use the household income level for analysis. We assume this implies that calling behavior strongly reflects the characteristics of individuals, rather than those of households.

5. CONCLUSIONS

In this work, we identified calling behavior traits that can distinguish demographic attributes of mobile users. We analyzed aggregated single-day records by focusing on calling behavior traits, which can also be generated from anonymized CDRs. Analysis results suggest that a higher ratio of calls from home can be a key to distinguishing females from males. Specifically for males on the primary routine day, the higher the individual income level, the higher the probability that they would initiate calls from their primary out-of-home locations around midday. The lower the income level of the user, the higher the probability that they would initiate calls from home in the morning. Conversely, there were no distinctive differences in primary routine day calling behavior for females related to their individual income level. Our findings suggested that time of day and call location distribution were keys to extracting differences in calling behavior by gender. That is, identifying types of locations, such as home and primary out-of-home is crucial to estimate the user attributes of anonymized call records.

With our work, we exploited the potential of deriving demographic attributes from anonymized CDRs. Although experiments were performed with a limited number of call records, the techniques developed in this study are capable of extracting gender traits from large-scale CDRs. Experimental results infer that our approach is capable of constructing a demographic attribute prediction model based on anonymized CDRs. Considering the importance of specifying call locations to extract demographic attributes, further studies are necessary to improve existing CDR location labeling methods.

6. ACKNOWLEDGMENTS

We appreciate the telecommunications operator and volunteers who provided data for the research. Part of this work was supported by GRENE-ei, funded by Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT).

7. REFERENCES

[1] Aarthi, S., Bharanidharan, S., Saravanan, M., and Anand, V. 2011. Predicting customer demographics in a mobile social network. In *2011 IEEE International Conference on Advances in Social Networks Analysis and Mining,* 553-554. IEEE.

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

[3] González, M. C., Hidalgo, C. A., and Barabási, A. L. 2008. Understanding individual human mobility patterns. *Nature.* 453, 7196, 779-782.

[4] Isaacman, S., Becker, R., Cáceres, R., Kobourov, S., Martonosi, M. J., Rowland, J., and Varshavsky, A. 2011. Identifying important places in people's lives from cellular network data. In *Pervasive Computing*, 133-151. Springer Berlin Heidelberg.

[5] Lee, D. D. and Seung, H. S. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature.* 401, 788-791.

[6] Lee, D. D. and Seung, H. S. 2000. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing 13.* Cambridge, MIT Press.

[7] Li, Q., Zheng, Y., Xie, X., Chen, Y., Liu, W., and Ma, W. Y. 2008. Mining user similarity based on location history. In *Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 298-307. ACM. New York, NY.

[8] Lu, X., Bengtsson, L., and Holme, P. 2012. Predictability of population displacement after the 2010 Haiti earthquake. *Proceedings of the National Academy of Sciences*, *109*, 29, 11576-11581.

[9] Mo, K., Tan, Ben., Zhong, Erheng., and Yang, Q. 2012. Report of task 3: your phone understands you. Paper presented at Nokia mobile data challenge 2012 workshop, Newcastle, UK, 18-19 June 2012.

[10] Lu, X., and Pas, E. I. 1999. Socio-demographics, activity participation and travel behavior. *Transportation Research Part A: Policy and Practice*, 33(1), 1-18.

[11] Paatero, P. and Tapper, U. 1997. Least squares formulation of robust non-negative factor analysis. *Chemometr. Intell. Lab.* 37, 23-35.

[12] Pas, E. I. 1984. The effect of selected sociodemographic characteristics on daily travel-activity behavior. *Environ. Plann. A*, 16, 5, 571-581.

[13] Song, C., Koren, T. K., Wang, P, and Barabási, A. L. 2010. Modelling the scaling properties of human mobility. *Nat. Phys. 6*, 10, 818-823.

[14] Szell, M., Sinatra, R., Petri, G., Thurner, S., and Latora, V. 2012. Understanding mobility in a social petri dish. *Scientific reports*, 2.

The Evolution of Social Strategies across the Lifespan[1]

Nitesh V. Chawla[†,*], Jie Tang[‡], Yuxiao Dong[†]

† Interdisciplinary Center for Network Science and Applications, Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN, United States of America

‡ Department of Computer Science and Technology, Tsinghua University, Beijing, P. R. China

∗ **Email: nchawla@nd.edu**

Abstract

Users with demographic profiles in social networks offer the potential to understand the social principles that underpin the highly connected world, from individuals, to groups, to societies. In this work [1], we aim to harness the power of computational social science to discover the social strategies in human communication by which people use to fulfill social needs, and to infer users' demographics based on their daily mobile communication behaviors.

We employ a real-world large mobile network comprised of more than 7 million users and over 1 billion communication records (CALL and SMS) as the basis of our study. In this dataset, around 45% of the users are female and 55% are male. In our study, we focus on users aged between 18 and 80 years old.

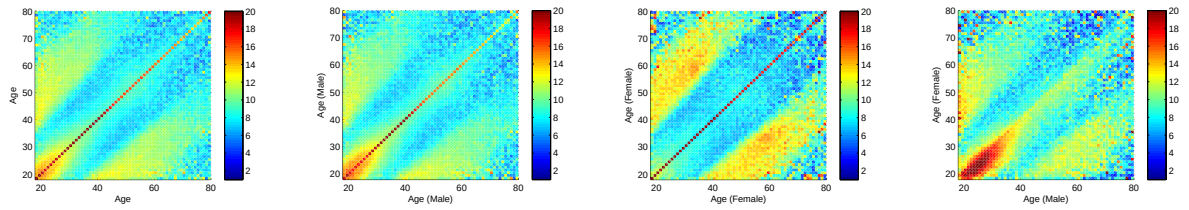
The dynamic social strategies used by people to meet their social needs across their lifetime indicate more complex and crucial social theories than can be revealed by a static view [2]. Our research unveils the significant social strategies across one's lifespan in human communication. Specifically, we investigate the interplay of communication interactions and demographic characteristics in the perspective of micro-level social structures, including social ego, social tie, and social triad.

We found the following social strategies on social ego. First, young people (who have higher degree) are very active in broadening their social circles, while seniors (who have higher clustering coefficients) tend to keep small but more stable connections. Second, people tend to communicate with others of similar age and gender, i.e., age homophily and gender homophily. Third, young people put increasing focus on the same generation and decreasing focus on the older generation, and the middle-age people devote more attention on the younger generation even along with the sacrifice of homophily.

The social strategies on social tie can be summarized as follows. First, frequent cross-generation interactions are maintained to bridge age gaps (Cf. Figure 1). Second, young male maintain more frequent and broader social connections than young females (Cf. Figures 1(b) and 1(c)). Third, opposite-gender interactions are much more frequent than those between young same-gender users (Cf. Figure 1(d)). However, when people become mature, reversely, same-gender interactions are more frequent than those between opposite-gender users.

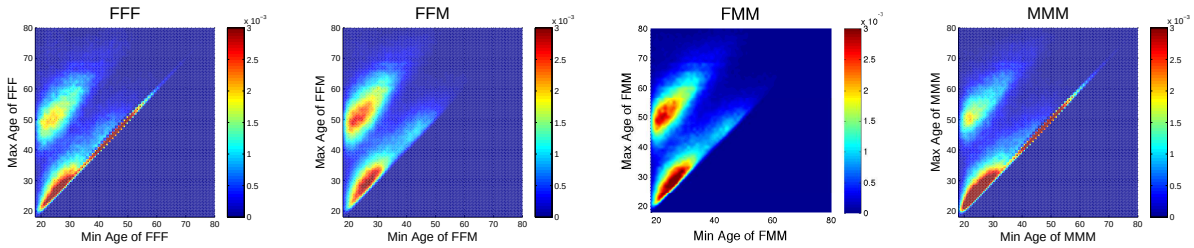
More interestingly, we highlight the social strategies on social triad unveiled from Figure 2.

[1]This work was published in the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining ($KDD'14$) [1] and also is submitted to International Conference on Computational Social Science (IC^2S^2).



(a) #calls per pair (b) #calls per M-M pair (c) #calls per F-F pair (d) #calls per M-F pair

Figure 1. Strength of social tie. XY-axis: age of users with specific gender. The spectrum color represents the number of calls per month. (a), (b), and (c) are symmetric.



(a) FFF demog. dist. (b) FFM demog. dist. (c) FMM demog. dist. (d) MMM demog. dist.

Figure 2. Demographic distribution in social triad. X-axis: minimum age of three users in a triad. Y-axis: maximum age of three users. The spectrum color represents the distributions.

People expand both the same-gender and opposite-gender social circles during the dating active period. However, people's attention to opposite-gender groups quickly disappears after entering into middle-age (Cf. Figures 2(b) and 2(c)), and the insistence and social investment on same-gender social groups lasts for a lifetime (Cf. Figures 2(a) and 2(d)).

Based on these discovered social strategies, we further study to what extent users' demographic information can be inferred by mobile social networks. The objective is to infer users' gender and age simultaneously by leveraging their interrelations. We present a computational model—that is, the *WhoAmI* framework, a Multiple Dependent-Variable Factor Graph model, whereby the social interrelations between users with different demographic profiles can be modeled. On both CALL and SMS networks, the *WhoAmI* method can achieve an accuracy of 80% for predicting users' gender and 73% for their age according to their daily mobile communication patterns, significantly outperforming several alternative data mining methods.

References

1. Y. Dong, Y. Yang, J. Tang, Y. Yang, and N. V. Chawla. Inferring user demographics and social strategies in mobile social networks. In *KDD'14*, pages 15–24. ACM, 2014.

2. V. Palchykov, K. Kaski, J. Kertész, A.-L. Barabási, and R. I. M. Dunbar. Sex differences in intimate relationships. *Scientific Reports*, 2:370, 2012.

Temporal dynamics of intra and inter-city social networks

Alejandro Llorente,[1] Manuel Cebrian,[2] and Esteban Moro[1]

[1] *Departamento de Matemáticas & GISC, Universidad Carlos III de Madrid, 28911 Leganés, Spain*
[2] *National Information and Communications Technology Australia, University of Melbourne, Victoria 3010, Australia*

During recent years, the analysis of different city properties as a function of their size has attracted the attention of economists, sociologists and environmentalists [1–3, 8]. Cities are important economic and cultural centers, and account for significant and increasing part of real-world and online activity in a modern urbanized society. Thus it is important to understand their properties and how they scale with population. While some factors such as wealth creation and innovation have been shown to scale superlinearly with the city size, other quantities, such as infrastructure, scale sublinearly. A possible origin of the non-linear relationship between city societal properties and their size lies in the number of social ties within cities. In recent papers it has been suggested [7] and found [8] that average connectivity of people in larger cities is higher than in smaller cities. As a consequence, contagious disease rates, information diffusion speeds or innovation propagation spreading is higher on larger cities [7], possibly leading to the superlinear scaling of societal properties with city size.

However, this aggregated static analysis of city social connectivity contradicts the highly dynamical nature of social relationships [4]: ties are formed and destroyed at a constant rate in our society and individuals show very different strategies regarding the way they manage their social connectivity, mostly due to time or cognitive limitations [6]. In particular, recent works [5] showed that either individuals keep their relationships constant in time (social keepers) or they vary the people they are connected to (social explorers). On top of that tie persistence is old, with an average of 10% of total ties decaying per month [5]. A highly dynamical evolution of network ties might be misinterpreted with larger social connectivity in the aggregated static picture of social networks. Thus, it is of interest to elucidate whether the aggregated super-scaling of number of ties within cities is also observed instantaneously in time and what is the influence and heterogeneity of social strategies in that super-scaling. Moreover, tie dynamics has a great impact on the way information or disease spread. In general, the bursty nature of tie contacts and their formation/decay dynamics slows down propagation [4] and thus network dynamics within cities will play an important role on the way information/disease spread.

In this work we investigate the dynamical properties of connections within and across regions to characterize and understand the impact of network dynamics in the super-scaling properties of regions and how information spreading is affected by the stability of ties within and across regions. To this end we have analyzed a mobile phone database consisting on 9000 million calls of 20 million users in an industrialized European country during 19 months. Users are geolocated according to their postcode (our regions) billing address and thus we can study the geographical extent of ties within the country. Using the method presented in [5] we were able to detect when ties are created and/or destroyed during our observation period. Each user i is then characterized by his social strategy γ_i, namely the ratio between his social activity n_i (the total number of links created or destroyed in the period) to his social capacity κ_i the number of instantly maintained relationships. Users with $\gamma_i \gg 1$ have a large turnover in their social neighborhood (social explorers), while those with $\gamma_i \ll 1$ keep a stable local network around them (social keepers). Using this definitions we have studied the geographical features of individual's strategies and of ties. Our main results are:

- At the region level, we have corroborated that in our database the aggregated number of ties k_α within a geographical area α does indeed scale super-linearly with those areas population N_α. In fact we get that $k_\alpha \sim N_\alpha^{1.14}$ in line with other works [8]. However, tie dynamics is much more super-linear and the number of ties created or destroyed in region α scales like $n_\alpha \sim N_\alpha^{1}.33$. Our result suggest part of the super-linear character of social connectivity within cities comes from the more dynamical nature of relationships in larger cities.

- At the users' level we found that the explanation for the more dynamical nature of networks within regions is due to the the higher proportion of social explorers in larger regions. In fact we find that the number of social explorers $N_\alpha^{(exp)}$ scales super-linearly with city size $N_\alpha^{(exp)} \sim N_\alpha^{1.27}$ (see figure 2a). Thus, human social strategies are different depending on whether individuals live in small or larger areas, with a higher tendency to become social explorer if we live in larger urban areas.

- Away from region boundaries we have studied also the stability of links as a function of the distance between individuals. As it is well known, the probability to have a link between two persons i and j that are at a geographical distance d_{ij} decays slowly with that distance $P(d_{ij}) \sim d_{ij}^{\eta}$. We also observed that result in our database, but also that the stability of links (the tie lifetime) depends on the geographical distance. Specifically, ties become more and more unstable (short) with increasing d_{ij} up to the geographical scale of $50km$ from

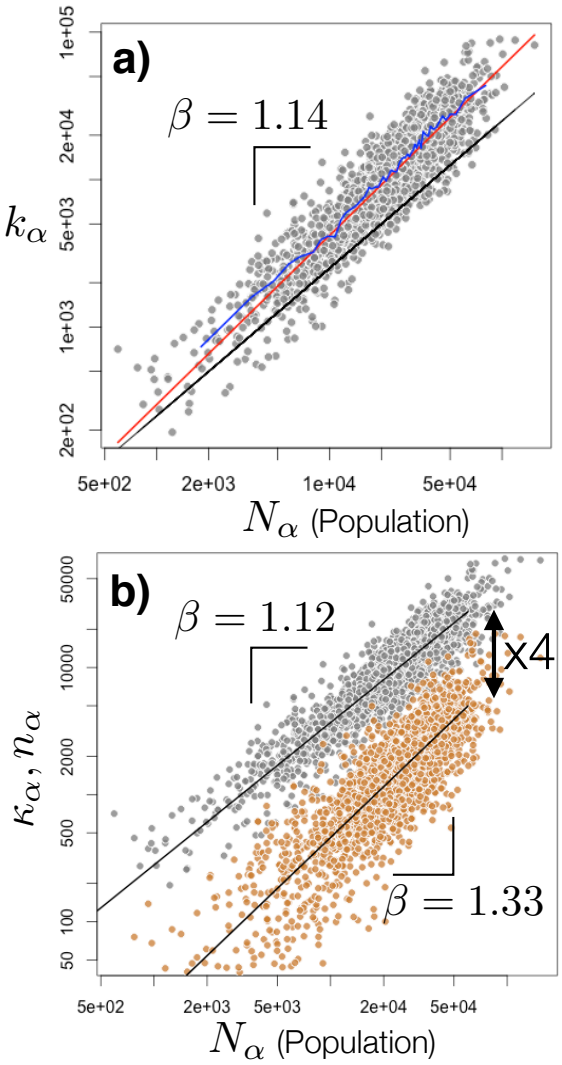


FIG. 1. a) Scaling of the aggregated number of ties k_α within a geographical area α and its population N_α. Each circle represents a postcode in our database. Blue line is an running average over the data, while the red line is a nonlinear fit to $k_\alpha \sim N^\beta$. For comparison we also show the linear relationship $k_\alpha \sim N_\alpha$ (black line). b) Scaling of the aggregated activity n_α (number of ties destroyed/formed within the region, grey symbols) and social capacity κ_α (number of instantly observed ties within the city, orange symbols) as a function of the region population N_α. Black lines are nonlinear fits $\kappa_\alpha, n_\alpha \sim N_\alpha^\beta$.

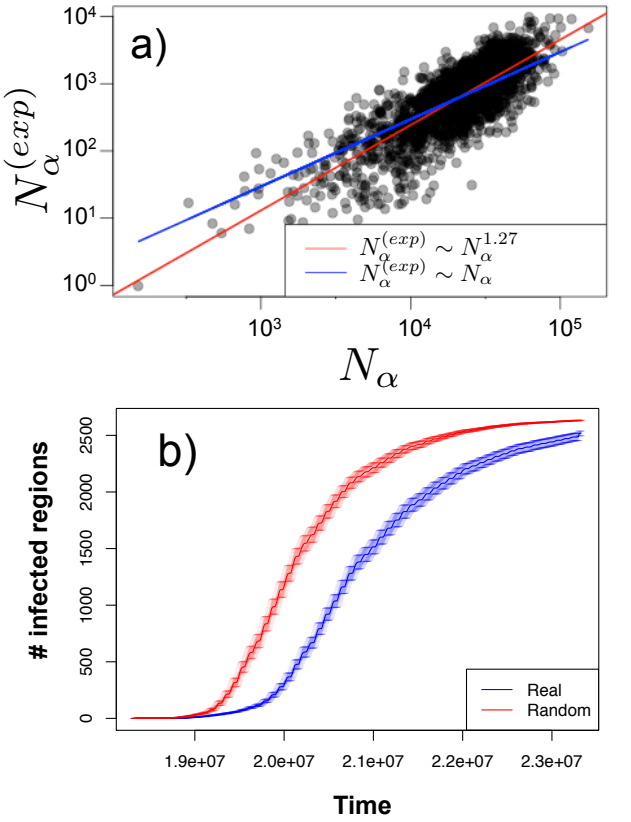


FIG. 2. a) Relationship between the region size and the number of social explorers living there, exhibiting superscaling. b) Number of infected regions as a function of time in the SI process for the real and the randomized data. We consider that a region is infected when a fraction $p = 0.6$ of its inhabitants is infected.

which stability of ties remains constant. Thus, network dynamics is strongly affected by geography: whereas estable links appear at different geographical distances (with higher probability at smaller distances), short duration ones are more likely to appear at higher distances.

- Finally, we have investigated the geographical nature of network dynamics in the process of information diffusion at the country level. The larger proportion of social explorers in larger regions and the fact that long-ranged ties are more unstable might have a paramount impact in the way information spreads at larger distances. To analyze this we have performed simulations of simple contagious models (Susceptible-Infected, SI) on the original call database and compare it to the randomized version of the calls timestamps. Similary to other previous works, an acceleration of the number of infected individuals is provoked by the time randomization as a consequence of the uniformization of the times between interactions, breaking their burstiness. More interestingly, not only the number of total infections is accelerated but also the number of different areas where a breakout has been declared, understood as those regions where the proportion of infected individuals is higher than a threshold p. We conclude that this effect is provoked by the increase of instantaneous relationships within the cities, accelerating the internal infection processes without

slowing the inter-region dissemination.

In conclusion we have found that network dynamics does depend on the geographical scale they happens. In particular we find that larger cities are more dynamical and tend to have larger number of unstable (sort) ties due to the larger proportion of individuals with social exploring strategies. At larger distances we also observe that links are shorter, which yields to a slower infection of different geographical areas. Thus, it is not only that information propagates slower in time because of network dynamics it also does it slowly in geographical terms. Our results suggest that social networks do behave differently depending on city properties: most of network dynamics happens in larger, denser urban areas, while small, countryside areas display less evolution in social connectivity. That rapid network evolution might be behind the existence of more opportunities, connections, meetings, etc. that yield to better performance (in terms of wealth, innovation, etc.). But it also challenges our understanding of how individuals manage their social connectivity at different time and spatial scales.

[1] Michael Batty. The size, scale, and shape of cities. *science*, 319(5864):769–771, 2008.

[2] Luís MA Bettencourt. The origins of scaling in cities. *science*, 340(6139):1438–1441, 2013.

[3] Luís MA Bettencourt, José Lobo, Dirk Helbing, Christian Kühnert, and Geoffrey B West. Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the National Academy of Sciences*, 104(17):7301–7306, 2007.

[4] Petter Holme and Jari Saramäki. Temporal networks. *Physics reports*, 519(3):97–125, 2012.

[5] Giovanna Miritello, Rubén Lara, Manuel Cebrian, and Esteban Moro. Limited communication capacity unveils strategies for human interaction. *Scientific reports*, 3, 2013.

[6] Giovanna Miritello, Esteban Moro, Rubén Lara, Rocío Martínez-López, John Belchamber, Sam GB Roberts, and Robin IM Dunbar. Time as a limited resource: Communication strategy in mobile phone networks. *Social Networks*, 35(1):89–95, 2013.

[7] Wei Pan, Gourab Ghoshal, Coco Krumme, Manuel Cebrian, and Alex Pentland. Urban characteristics attributable to density-driven tie formation. *Nature communications*, 4, 2013.

[8] Markus Schläpfer, Luís MA Bettencourt, Sébastian Grauwin, Mathias Raschke, Rob Claxton, Zbigniew Smoreda, Geoffrey B West, and Carlo Ratti. The scaling of human interactions with city size. *Journal of The Royal Society Interface*, 11(98):20130789, 2014.

Session 5 :: Societies (II)

THE ADOPTION OF NETWORK GOODS: EVIDENCE FROM THE SPREAD OF MOBILE PHONES IN RWANDA

DANIEL BJÖRKEGREN*
BROWN UNIVERSITY

This project develops a method to estimate and simulate the adoption of a network good. I estimate demand for mobile phones as a function of individuals' social networks, coverage, and prices, using transaction data from nearly the entire network of Rwandan mobile phone subscribers over 4.5 years. I then simulate the effects of an adoption subsidy and a requirement to provide coverage in rural areas.

Full version of paper: http://dan.bjorkegren.com

1. Introduction

Improvements in communication, through mobile phones as well as associated services such as mobile money and mobile internet, have the potential to knit even remote villages into the global economy. Although these goods can generate large efficiency gains (Jensen, 2007; Jack and Suri, 2011), their allocations are likely to be inefficient due to network effects. Individuals are unlikely to internalize all the benefits their adoption generates, so adoption is likely to be suboptimal unless the firms operating the network use sophisticated pricing mechanisms. Also, in competitive markets, any single firm will internalize only a small share of the benefits it generates. If instead a market is so concentrated that these benefits are internalized by a small number of firms, the ability of these firms to exert market power raises welfare concerns.

That mobile phones have spread rapidly across varied environments suggests that, in this case, failures by consumers and firms to internalize network effects have not prevented widespread adoption. Yet, because adoption is widespread, even small inefficiencies can have large welfare consequences. Also, there may be many beneficial network goods that we do not observe, because they would diffuse only within a narrow band of policy conditions. The fact that mobile phones have spread makes it possible to observe the network of benefits they generate, and to empirically study policies that may lead to more efficient adoption of other network goods.

Firms and governments use many different policies to guide the provision and adoption of network goods. While theoretical work provides intuition about network effects, there is little empirical work to guide policy choices.[1] Empirical work has been limited for three reasons. It is costly to measure an entire network using traditional data sources. It is also difficult to identify network effects: one individual may adopt after a contact adopts because the contact provides network benefits, or because connected individuals share similar traits or are exposed to similar environments. And even if these two issues are overcome, it is difficult to evaluate policies, which can cause effects to ripple through the entire network. As a result, there remain open questions about how to design policies that better capture the spillover benefits associated with network effects, as well as policies that overcome suboptimal provision arising from high concentrations in industries providing network goods.

In this project, I overcome previous limitations using a new empirical approach and 5.3 billion transaction records from Rwanda's dominant mobile phone operator as the network expanded from 300,000 to 1.5 million subscribers. I estimate a structural model of demand for mobile phones, and then use this model to simulate the effects of two policies.

2. Model

My method has three steps:

First, acknowledging that the utility of owning a mobile phone is derived from its usage, I model the utility of using a phone. I observe every connection between subscribers, as well as the calls placed across each connection. Because 99% of accounts are prepaid and the person placing a call pays for it by the second, a subscriber must value a connection at least as much as the cost of calls placed across it. Because calling prices changed over this period, I can estimate the underlying demand curve for communication across each link, and thus the value of each connection.

Let G be the communication graph. Each individual i has a set of contacts $G_i \subset G$, where a directed link $ij \in G$ indicates that i has a potential desire to call j over the mobile phone network. Let S_t be the subset of nodes subscribing in month t. At each period t, individual i can call any contact j that currently subscribes, $j \in G_i \cap S_t$, to receive utility u_{ijt}. Each month, i draws a communication shock ϵ_{ijt} representing a desire to call contact j. Given the shock, i chooses a total duration $d \geq 0$ for that month, solving:

*E-mail: danbjork@brown.edu, Web: http://dan.bjorkegren.com

Revision January 6, 2015. Preliminary and incomplete. I am grateful to Michael Kremer, Greg Lewis, and Ariel Pakes for guidance and encouragement. Thank you to Nathan Eagle for providing access to the data, computing facilities, and helpful conversations. In Rwanda, I thank the staff of my telecom partner and government agencies. This work was supported by the Stanford Institute for Economic Policy Research through the Shultz Fellowship in Economic Policy.

[1] Early theoretical work includes Rohlfs (1974), Katz and Shapiro (1985), and Farrell and Saloner (1985). Most empirical work on network goods measures the extent of network effects; see for example Saloner and Shepard (1995), Goolsbee and Klenow (2002), and Tucker (2008). The paper closest in spirit to this one is Ryan and Tucker (2012), which estimates the adoption of a videoconferencing system over a small corporate network, and evaluates policies of seeding adoption.

$$u_{ijt} = \max_{d \geq 0} v_{ij}(d, \epsilon_{ijt}) - c_{ijt} d$$

where $v(d, \epsilon)$ represents the benefit of making calls of a total duration of d and c_{ijt} represents the per-second cost.

Second, I model the decision to adopt a mobile phone. The utility of having a phone in a given period is given by the utility of communicating with contacts that have phones: each month i is on the network, he receives expected utility:

$$u_{it} = \sum_{j \in G_i \cap S_t} Eu_{ijt}(p_t, \boldsymbol{\phi}_t) + w \cdot Eu_{jit}(p_t, \boldsymbol{\phi}_t) + \eta_i$$

where u_{ijt} represents calls from i to j (which i pays for), u_{jit} represents calls from j to i (which j pays for), and $w \in \{0, 1\}$ specifies whether recipients value incoming calls. Individual i chooses when to adopt by weighing the discounted stream of these benefits against the declining price of a handset, which is represented by the price index $p_t^{handset}$. Then, i considers the utility of adopting at time τ to be:

$$U_i^\tau = \sum_{t=\tau}^{\infty} \delta^t Eu_{it}(p_t, \boldsymbol{\phi}_t) - \delta^\tau \beta_{price} p_\tau^{handset}$$

I estimate the parameters of this model using maximum likelihood and moment inequalities.

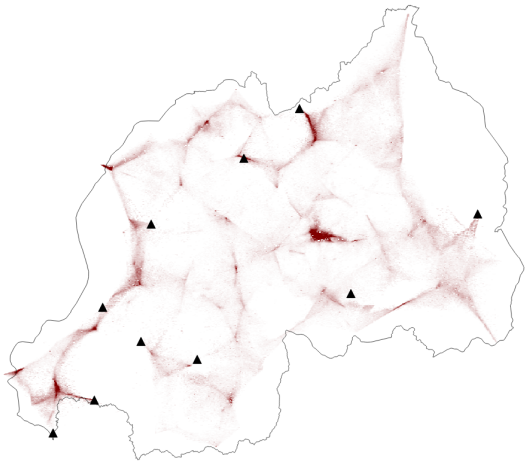
Finally, to evaluate the impact of policies, I use a simulation method that allows each individual to react directly to a policy change, and to each other's responses, capturing effects that ripple through the network and across physical space.

3. Results

I use this approach to answer two sets of policy questions, by simulating how adoption would have proceeded if conditions were less favorable. Because individuals tend not to internalize adoption spillovers, it is common for firms or governments to subsidize adoption of network goods. I analyze a rural adoption subsidy program implemented by the Rwandan government in 2008. I first use phone data to determine how subsidized handsets were ultimately used, and then use the simulation method to determine how the policy affected the entire network. I find that a substantial fraction of the subsidy's impact arises from its impact on nonrecipients—in particular, contacts of recipients account for more than 62% of the effect on revenue. Although the bounds are wide, the subsidy improved welfare, in a low case by $191,108 (representing a social return of 33%), and in a high case by $5.6 million (a social return of 888%).

I also analyze the welfare implications of providing coverage to rural areas. A social planner would expand coverage until the point where building any additional towers would not improve welfare. Firms may stop building before reaching this point: in a competitive market, some of the benefits of expanding coverage will spill over into competitors' networks.[2] And regardless of market structure, firms are unlikely to internalize all of the value generated for consumers. Depending on the shape of private and social benefits from expansion, it may be optimal for a government to require the provision of coverage to areas that are unprofitable to serve. I find that in Rwanda, a government coverage obligation led to the building of a handful of rural towers that were unprofitable for the firm but slightly welfare improving for the country. The impact shifted bounds on welfare in a focal equilibrium upward by roughly $179,381 (0.06%). The impact was also extremely dispersed: over 65% of the gain in consumer surplus accrued to individuals living in areas where coverage was not affected; some of these individuals called in to the covered areas and others were affected indirectly. The impact on revenue by geography is presented in Figure 1: much of the revenue comes from areas far from the towers. Because of this dispersion, it would have been difficult for local communities to raise the funds to build the towers themselves.

Figure 1. Geographic Distribution of Revenue from Rural Expansion



Dropped towers are denoted by triangles; difference in revenue between the baseline and counterfactual in January 2009 is shaded.

4. Conclusion

This project introduces a new method for estimating and simulating the adoption of network goods. I overcome measurement issues that have limited empirical work on network goods using rich new data on the adoption and usage of nearly an entire network of mobile phone users. *(For more details, find the full paper at http://dan.bjorkegren.com)*

[2] A fraction of these benefits can be internalized using interconnection fees, but some will spill into the interiors of competitor networks.

References

Farrell, J. and G. Saloner (1985): "Standardization, Compatibility, and Innovation," *The RAND Journal of Economics*, 16, 70–83.

Goolsbee, A. and P. J. Klenow (2002): "Evidence on Learning and Network Externalities in the Diffusion of Home Computers," *Journal of Law and Economics*, 45, 317–343.

Jack, W. and T. Suri (2011): "Risk Sharing and Transactions Costs: Evidence from Kenya's Mobile Money Revolution," *Working Paper*.

Jensen, R. (2007): "The Digital Provide: Information (Technology), Market Performance, and Welfare in the South Indian Fisheries Sector," *Quarterly Journal of Economics*, 122, 879–924.

Katz, M. L. and C. Shapiro (1985): "Network Externalities, Competition, and Compatibility," *The American Economic Review*, 75, 424–440.

Rohlfs, J. (1974): "A Theory of Interdependent Demand for a Communications Service," *The Bell Journal of Economics and Management Science*, 5, 16–37.

Ryan, S. P. and C. Tucker (2012): "Heterogeneity and the dynamics of technology adoption," *Quantitative Marketing and Economics*, 10, 63–109.

Saloner, G. and A. Shepard (1995): "Adoption of Technologies with Network Effects: An Empirical Examination of the Adoption of Automated Teller Machines," *The RAND Journal of Economics*, 26, 479–501.

Tucker, C. (2008): "Identifying Formal and Informal Influence in Technology Adoption with Network Externalities," *Management Science*, 54, 2024–2038.

The Strength of the Strongest Ties in Collaborative Problem Solving

Yves-Alexandre de Montjoye, PhD Candidate, Media Lab, MIT, Cambridge, MA, USA; **Arkadiusz Stopczynski**, PhD Candidate, DTU Compute, Technical University of Denmark, Copenhagen, Denmark; Erez Shmueli, PhD, Media Lab, MIT, Cambridge, MA, USA; Alex 'Sandy' Pentland, Prof., Media Lab, MIT, Cambridge, MA, USA; Sune Lehmann , DTU Compute, Technical University of Denmark, Copenhagen, Denmark;

Corresponding Author Contact: yva@mit.edu, (617) 324-3842

Keywords: social networks; collaboration; groups; strong ties; complex problem solving

Abstract: Complex problem solving in science, engineering, or business has become a highly collaborative endeavor. Groups of scientists or engineers are collaborating on projects but also using their social networks to gather new ideas and feedback. Here we bridge the literature on group work and information networks by studying groups' problem solving abilities as a function of both their within-group networks but also their members' extended networks. We show that while groups' performance is strongly correlated with its networks of expressive and instrumental ties, only the strongest ties in both networks have an effect on performance, as shown in Figure 1. Both networks of strong ties explain more of the variance than other factors such as measured or self-evaluated technical competencies or personality of the group members. In fact, the inclusion of the network of strong ties renders these factors non-significant in the statistical analysis. Our results have consequences for the organization of groups of scientists, engineers, or other knowledge workers tacking our current most complex problems.

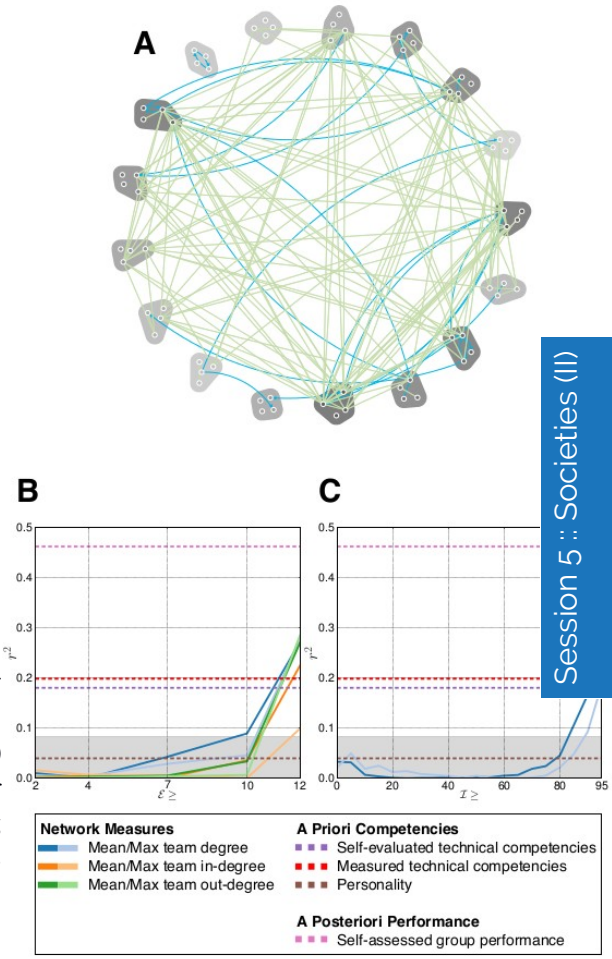


Fig 1 (A) Network of strongest expressive (blue) and instrumental (green) as well as the groups participants where assigned to for the first project. Color saturation is the performance of the groups where darker is higher. (B) Correlation between expressive tie strength and group performance. (C) Correlation between instrumental tie strength and group performance. For both expressive and instrumental ties, the position in the network of strong ties is more important than other a priori characteristics of the group such as self-evaluated and measured proficiency or personality. The gray areas indicates values with $p > 0.05$.

Investigating Social Influence Through Large-Scale Field Experimentation

Johannes Bjelland,[†] Geoffrey Canright,[†] Asif Iqbal,[†] Rich S. Ling,[†‡] Kenth Engø-Monsen,[†] Taimur Qureshi,[†] Christoph Riedl,[§] Pål Roe Sundsøy,[†] David Lazer[§]

[†]Telenor Research, Oslo, Norway; [‡]Nanyang Technological University, Singapore; [§]Northeastern University

The recent emergence of computational social science research methods has enabled us to conduct large-scale, population level experimentation that allow us to estimate causal effects of different policy alternatives. This new experimental paradigm is particularly suited to study patterns of social influence over real-world networks in which social embeddedness and tie strength are often correlated with each other and with homophily. Networks of interaction among individuals provide the primary pathways along which viral contagion spread. This viral information and behavior spreading is relevant to understand fundamental aspects of social influence in the adoption of behavior like smoking or exercising, and economic product adoption. Given limited resources, policy makers and marketers confront a strategic choice: who to initially target with information? Traditional models of contagion suggest that targeting neighboring nodes would be inefficient, because there is the possibility that one node would infect the other virally. More recent work on complex contagion suggests that behavioral confirmation by alters will increase probability of adoption. Here we report on a marketing experiment in which the clustering of targeting is varied, and we find support that clustered targeting is more effective than non-clustered targeting.

Theory. This is the core reasoning behind the "strength of weak ties"—weak ties connect nodes that provide non-redundant information (Granovetter, 1973; Onnela et al., 2007). However, more recent work (Centola & Macy, 2007; Centola, 2010) suggests that the probability of behavioral contagion increases more than linearly with marginal adoption. That is, having two friends that adopt X more than doubles ego's odds of adopting X. This has direct implications for mass efforts for mass behavior change, either by marketers or policymakers: that efforts to change behavior should NOT be randomly disseminated through the population but aimed at clusters of connected individuals.

Methods. Here we report the results of two country-level experiments in which we exposed 19,352 and 27,420 users, through mobile phone text messages, to unique voucher codes that offered them 100MB of free traffic for their data plans. Each user exposed to a code could (a) adopt the code themselves (i.e., redeem the voucher) and (b) pass the code on to their friends. In the experiments we manipulated the number of users within a dyad and closed triangle of "friends" that were exposed to the market offering. That is, in some randomly selected dyads, one randomly selected user was exposed, while in other dyads both were exposed (N=9,318 dyads with one code; N=5,017 dyads with two codes). Conversely, in some randomly selected closed triangles, one randomly selected user, two randomly selected users, or all three users were exposed to the product offering (triangles with one, two, or three codes: 8,408, 4,697, 3,206). We then tracked adoption of voucher codes over two weeks.

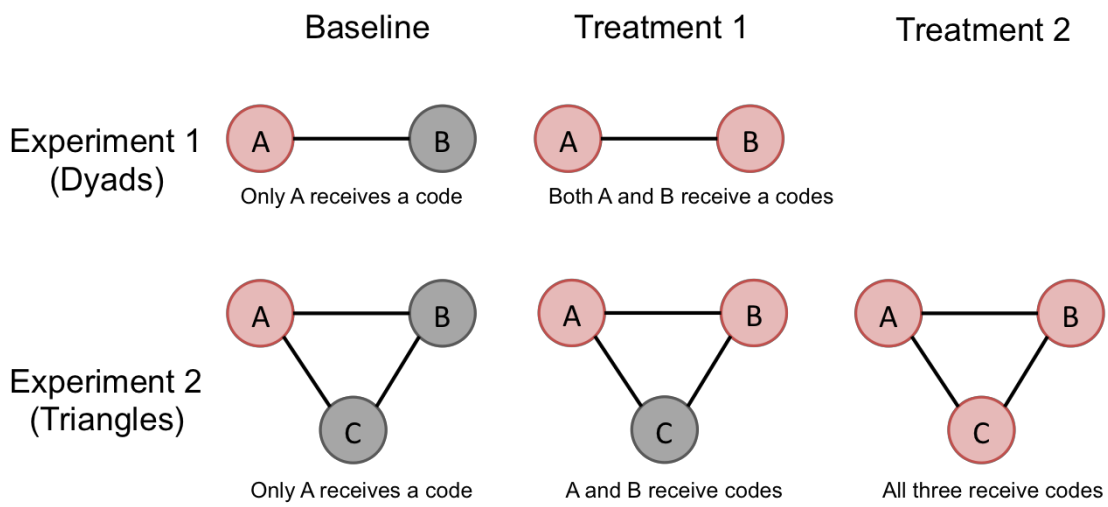


Figure 1. What is the probability of adoption by the focal node (in red)? In the experiment we had two dyad level treatments, and three triad level treatments, varying the number of nodes that received the offer. In experiment 1 probability of focal node adopting was 4.56% in the baseline condition and 4.68% in the treatment condition where both nodes received a code—i.e., employing clustered targeting (not significantly different). In the triangles experiment, the probability of adoption in the baseline condition was 6.28%. In the treatment condition in which two of the three nodes received a code it was 7.18% and in the second treatment condition where all nodes received a code it was 7.08% (3-sample test for equality significant at $p < 0.5$). This represents an increase in adoption probability of 13%-15% from being exposed to clustered treatments.

Results. We find strong causal evidence that increased exposure to the offering through friends significantly increases adoption probability by the focal node in the triad experiment. In dyads, the baseline probability of code adoption remains largely unchanged if a "friend" is equally exposed to a code (adoption probabilities of 4.56% and 4.68%, respectively; not significantly different). However, in the triad experiment, we find that the baseline adoption rate of 6.28% increases to 7.18% and 7.08% if one or two "friends", respectively, have also been exposed to a code (3-sample test for equality significant at $p < 0.5$).

Conclusion. Our results suggest that there is a marginal benefit of exploiting existing ties between individuals to increase behavioral contagion. We show that not only may peer's adopted behavior affect own adoption rates, but simply peer's exposure may already be enough. This has direct implications for efforts to affect mass behavior change and suggests that efforts to change behavior can be more effective if they are aimed at clusters of connected individuals rather than disseminated randomly.

References

Centola, D. (2010). The spread of behavior in an online social network experiment. *Science*, *329*(5996), 1194-1197.

Centola, D., & Macy, M. (2007). Complex contagions and the weakness of long ties1. *American Journal of Sociology*, *113*(3), 702-734.

Granovetter, M. S. (1973). The Strength of Weak Ties. *American Journal of Sociology*, *78*(6), 1360–1380.

Onnela, J-P., et al. (2007). Structure and tie strengths in mobile communication networks. *PNAS*, *104*(18), 7332-7336.

Asymmetric Role of Social Influence in Smartphone Adoption in Large Mobile Networks

Qiwei Han[*†]
qiweih@cmu.edu

Pedro Ferreira[*‡]
pedrof@cmu.edu

João Paulo Costeira[†]
jpc@isr.ist.utl.pt

[*]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
University of Lisbon
Lisbon, 1049-001
Portugal

1. INTRODUCTION

As the number of mobile subscriptions worldwide is reaching nearly 7 billion by the end of 2014 [10], the rapid growth of mobile handset adoption has evolved from communication devices (*a.k.a.* feature phones) with basic telephony service to advanced multifunction computing devices (*a.k.a.* smartphones) that provide one-stop solution to meet popular daily needs. Hence, smartphones have become platforms enabling a multitude of services and applications, through which the business of mobile market is delivered through these devices. Understanding the adoption pattern of this high-technology product is of great interest to not only handset manufacturers and mobile operators, but also to service providers and application developers, and it may also reveal significant managerial and policy implications [11].

Social influences have been acknowledged the important role in determining consumer's decision making process for various mobile communication products and services (e.g. [8, 9, 12, 13]. According to [5], the positive (negative) word-of-mouth message in the mobile phone market can increase (decrease) a company's market share by as much as 10 (20) percent. In particular, with regard to smartphone adoption, both [13] and [17] found positive effect, meaning the propensity for a focal customer to adopt smartphone increases with the number of adoptions in her ego network. However, identifying social influence in large networks, especially in observational studies, still remains as serious concerns to researchers [7], as it is known to confound with endogenous factors such as homophily - the tendency for individuals to choose friends with similar tastes, among others [14]. When correlated behavior between ego and her direct neighbors can be explained by both influence and their inherent similarities, misattribution of homophily to influence may lead to significant overestimation of the latter [3]. Moreover, [19] argued that social influence and homophily are generically confounded and can not be readily separated from each other, because latent homophily may still remain as a component of the estimated influence. Nevertheless, they proposed several constructive suggestions that may help alleviate the issue, one of which our work is motivated: to use community structure as the proxy to control for latent homophily. We intend to estimate the effect of social influence on adoption of iPhone 3G, an iconic smartphone, using dataset from a major wireless carrier operated in one European country (hereinafter EURMO). iPhone 3G is released in July 2008 and EURMO is the sole partner with Apple in this country, so we are able to capture the full cycle of its adoption.

2. DATA

EURMO dataset includes call detail records (CDR) for over 5 million subscribers between August 2008 and June 2009. Subscribers are identified by their anonymized phone numbers. For each call we know the caller and the callee, the timestamp, and the GPS coordinates of the connected cell tower. By aggregating GPS coordinates over the entire period, we can approximate subscriber's home location at municipal level as where they spend most of their days[1]. We further infer the socio-economic indicators (*e.g.* wage) by cross-referencing the latest census. We also have an (incomplete) set of subscriber characteristics such as gender and usage history since their subscription to EURMO, which includes tariff plan, handset and supplementary services (*e.g.* mobile broadband). In our period of analysis, there are 20,570 iPhone 3G adopters with complete profiles. Table 1 lists relevant variables and short descriptions that we extract from EURMO.

We use CDR to construct the social network as an undirected call graph. Specifically, we denote two subscribers to befriend each other if they exchange at least one call in the same calendar month. The mutual relationship between subscribers ensure that we preclude communications that are unlikely to represent the social ties, such as customer services and PBX machines. The resulting network consists of 5,535,388 subscribers and 66,717,468 edges with mean, standard deviation, and median of degree being 24.1, 25.7 and 16, respectively.

3. SUBPOPULATION EXTRACTION

Recent researches on statistical properties of real world social networks provide evidence of the existence of community structure within the network [6, 16]. These findings does not only validate the theoretic role of homophily and influence in tie formation, but also provide several important insights into the problem of community structure inference when we only observe the resulting network as follows. First, uncovered communities should exhibit real social meaning, as individuals in the same community have some natural affinity for each other or some fundamental characteristics in common. Meanwhile, they should be more likely to connect to each other than those who belong to different communities. Hence community discovery method should consider

[1]The municipal location is defined as Nomenclature of Units for Territorial Statistics (NUTS) III, which is a geocode standard across European countries by Eurostat for statistical purposes

Variable	Type	Description
gender	categorical	Self-reported gender (male, female, unknown)
wage	categorical	Inferred wage level (very low, low, average, high, very high)
prepaid	binary	prepaid tariff plan (yes, no)
phone_technology	categorical	handset technology (2G, 2.5G, 3G, 3.5G, other)
mobile_internet	binary	mobile broadband (yes, no)
phone_age	continuous	age of currently owned handset (year)
tenure	continuous	tenure since subscription (year)
region	categorical	home location at municipal level

Table 1: List of covariates extracted from EURMO

two different sources of information together, *i.e.* individual characteristics and social connections among them. Second, in many actual networks, individuals may belong to multiple overlapping communities [1, 15], *e.g.* families, coworkers and friends. This is also aligned with the notion of homophily across different social dimensions. Third, as also noted in [19], misspecification of community structure (*e.g.* simple modular and/or disjoint structure) may even worsen the problem and lead to biased model estimation. Fourth, as the complexity of network structure grows exponentially with the size, computational costs still remain challenging for the analyses on large scale networks. Therefore, extracting subpopulation via community discovery does not only significantly reduce group-level heterogeneity that may potentially confound the result but also help lessen the computational cost [22].

We employ the method of discovering Communities from Edge Structure and Node Attributes (CESNA) for our purpose [21]. CESNA is a community discovery algorithm that consider both node attributes and network structure as well as the interactions between these two sources of information. It can detect overlapping communities with high accuracy and scalability over many existing community detection algorithms, particularly on large scale networks. For sake of space, details beyond the mechanics of CESNA can be found in [21], and we only note the following implementation procedures: i) for each iPhone 3G adopter, we construct the ego-network that contains adopter and their direct neighbors, similarly as [17]; ii) for each subscriber in the ego-network, we extract a list of 0-1 valued covariates specified in table 1 that represent pluralistic homophily including gender and wage (socio-demographic homophily); tariff plan, phone technology and mobile broadband (contextual homophily); and home location (spatial homophily); iii) we apply CESNA on each ego-network using both node and edge information with the optimal number of communities identified through cross-validation; iv) we remove duplicated and nested communities and only retain communities that contain iPhone 3G adopters. As a result, we obtain 11,454 communities with 202,743 subscribers, 14,685 of which are iPhone 3G adopters. After detecting community structure from networks, we are able to "naturally" extract subpopulations with half of the original network size which include over 70% of adopters and their cohesive groups of neighbors with whom are both similar and strongly connected.

4. CORE-PERIPHERY STRUCTURE

From the extracted subpopulation, we observe that over 70% of subscribers belong to only one community and nearly 90% of those belong to two, while only about 5% of subscribers belong to more than 5 communities. This is consistent with the findings suggested in [18, 20] that the intersection of overlapping communities may reveal *core-periphery structure* which complements current views of network organizations. In general, core nodes refer to set of central nodes that are connected to other core nodes as well as peripheral nodes, while peripheral nodes, by contrast, are only loosely connected to the core nodes but not to each other [4]. In this sense, following the measure proposed in [20] that unify both organizing principles of the network, we validate the existence of core-periphery structure in our subpopulation (see Fig. 1).

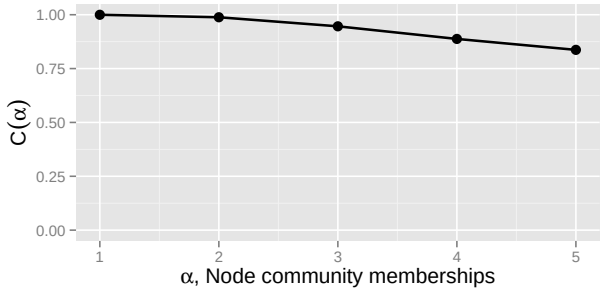


Figure 1: The fraction of nodes $C(\alpha)$ in the largest connected component of the induced subgraph on the nodes who belong to at least a communities. A high $C(\alpha)$ means that there is a single dominant core.

We then define a subscriber as *core* node if she belongs to at least five communities and as *periphery* node if otherwise. Among 9,194 core nodes, 5,548 (60%) are iPhone 3G adopters, whereas only 5% of peripheral nodes are adopters. This provides us with extra implications that iPhone 3G adopters tend to form densely connected groups that have the most shared properties with others thorough overlapping communities.

5. MODEL AND RESULTS

We describe our reduced form latent utility model as follows (index i are dropped for simplicity):

$$\begin{aligned} Y_t &= 1\{U_t > 0\} \\ U_t &= \alpha + \beta \cdot X + \gamma \cdot Z_t + \delta \cdot Core \\ &\quad + \mu_1 \cdot Core\_Friend\_Adopt_{t-1} \\ &\quad + \rho_1 \cdot Core \cdot Core\_Friend\_Adopt_{t-1} \\ &\quad + \mu_2 \cdot Peri\_Friend\_Adopt_{t-1} \\ &\quad + \rho_2 \cdot Core \cdot Peri\_Friend\_Adopt_{t-1} + \epsilon_t \end{aligned} \tag{1}$$

where X and Z_t are time invariant and time variant subscriber-specific characteristics listed in table 1. Dummy variable $Core$ indicates the incremental *network position effect* when subscriber is deemed as core. For social influence, we include the *main effects* of number of core friend adopters $Core\_Friend\_Adopt_{t-1}$ and number of peripheral friend adopters $Peri\_Friend\_Adopt_{t-1}$ and the *interaction effects* between number of friend adopters and subscriber's network position. A significant coefficient for the latter terms captures

the social influence from core friend adopters and peripheral friend adopters, relative to focal subscriber's network position. Besides, we also introduce month and location fixed effects to control for heterogeneity across time and region.

We organize the subpopulation data into a panel where each individual is a subscriber and each period is a calendar month and observations after the first adoption need to be removed from the sample. We use the resulting sample (size=2,116,855) to empirically estimate the social influence on iPhone 3G adoption. Table 2 summarizes the estimates of covariates of interest from equation 1 using Probit model. The full regression table is available upon request.

Dependent Variable: $Adopted_t$	Probit
$Core(\delta)$	1.522^{***}
	(0.021)
$Core\_Friend\_Adopt_{t-1}(\mu_1)$	0.363^{***}
	(0.008)
$Peri\_Friend\_Adopt_{t-1}(\mu_2)$	0.187^{***}
	(0.008)
$Core \cdot Core\_Friend\_Adopt_{t-1}(\rho_1)$	0.092^{***}
	(0.01)
$Core \cdot Peri\_Friend\_Adopt_{t-1}(\rho_2)$	0.244^{***}
	(0.013)

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 2: Regression results of Probit Model

By computing the average partial effect using delta method, we find that for core subscribers, their propensity to adopt iPhone 3G is over 10% higher than peripheral subscribers, indicating that subscribers who are in the higher network position (maybe also social status) are more likely to adopt the smartphone. Meanwhile, having one more core friend adopter and one more peripheral friend adopter will increase the focal subscriber's adoption probability by 0.46% and 0.25%, respectively. However, interpreting the coefficients of interaction terms are far more straightforward [2], as the interaction effect needs to be calculated as the cross-partial derivatives. We find *asymmetric* interaction effect on core and peripheral subscribers when having core and peripheral adopter friends. Specifically, on average with one more core (peripheral) adopter friend, the changes between core and peripheral subscribers is about -1.52% (-1.76%). The implications of this finding are twofold: 1) peripheral subscribers are more likely to get influenced than core subscribers; 2) core subscribers are slightly more likely to be get influenced by a core adopter friend rather than a peripheral adopter friend.

6. REFERENCES

[1] Ahn, Y.-Y., Bagrow, J., and Lehmann, S. Link communities reveal multiscale complexity in networks. *Nature 466* (2010), 761–764.

[2] Ai, C., and Norton, E. Interaction terms in logit and probit models. *Economic Letters 80*, 1 (2003), 123–129.

[3] Aral, S., Muchnik, L., and Sundararajan, A. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of National Academy of Science 106*, 51 (2009), 21544–21549.

[4] Borgatti, S., and Everett, M. Models of core/periphery structures. *Social Networks 21*, 4 (2000), 375–395.

[5] Bughin, J., Doogan, J., and Vetvik, O. A new way to measure word-of-mouth marketing. *Mckinsey Quarterly* (April 2010), 1–9.

[6] Girvan, M., and Newman, M. Community structure in social and biological networks. *Proceedings of National Academy of Sciences 99* (2002), 7821–7826.

[7] Goldsmith-Pinkham, P., and Imbens, G. Social networks and the identification of peer effects. *Journal of Business & Economic Statistics 31*, 3 (2013), 253–264.

[8] Han, Q., and Ferreira, P. Role of peer influence in churn in wireless networks. In *Proceedings of 7th International Conference on Social Computing, SocialCom 2014, Beijing, China* (2014), pp. 176–183.

[9] Hill, S., Provost, F., and Volinsky, C. Network-based marketing: Identifying likely adopters via consumer networks. *Statistical Science 21*, 2 (2006), 256–276.

[10] ITU. The World in 2014: ICT Facts and Figures, 2014.

[11] Kauffman, R., and Techatassanasoontornb, A. Understanding early diffusion of digital wireless phones. *Telecommunications Policy 33*, 8 (2009), 432–450.

[12] Ma, L., Krishnan, R., and Montgomery, A. Latent homophily or social influence? an empirical analysis of purchase within a social network. *Management Science 57*, 9 (2014), 1623–1639.

[13] Matos, M., Ferreira, P., and Krackhardt, D. Peer influence in the diffusion of the iPhone 3G over a very large social network. *MIS Quarterly 38*, 4 (2014), 1103–1133.

[14] McPherson, M., Smith-Lovin, L., and Cook, J. M. Birds of a feather: Homophily in social networks. *Annual Review of Sociology 27* (2001), 415–444.

[15] Palla, G., Derenyi, I., Farkas, I., and Vicsek, T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature 435* (2005), 814–818.

[16] Park, J., and Barbasi, A.-L. Distribution of node characteristics in complex networks. *Proceedings of National Academy of Sciences 106*, 46 (2007), 17916–17920.

[17] Risselada, H., Verhoef, P., and Bijmolt, T. Dynamic effects of social influence and direct marketing on the adoption of high-technology products. *Journal of Marketing 78* (March 2014), 52–68.

[18] Rombach, M. P., Porter, M., Fowler, J., and Mucha, P. Core-periphery structure in networks. *SIAM Journal of Applied Math 74*, 1 (2014), 167–190.

[19] Shalizi, C. R., and Thomas, A. C. Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods & Research 40*, 2 (May 2011), 211–239.

[20] Yang, J., and Leskovec, J. Overlapping communities explain core-periphery organization of networks. *Proceedings of the IEEE 102*, 12 (December 2014), 1892–1902.

[21] Yang, J., McAuley, J., and Leskovec, J. Community detection in networks with node attributes. In *Proceedings of the IEEE 13rd International Conference on Data Mining, Dallas TX, USA* (2013), pp. 1151–1156.

[22] Zhang, B., Krackhardt, D., Krishnan, R., and Doreian, P. An effective and efficient subpopulation extraction method in large social networks. In *Proceedings of the 32rd International Conference on Information Systems, Shanghai, China* (2011).

Using Mobile Phone Data to Predict the Spatial Spread of Cholera

Linus Bengtsson[a,b], Jean Gaudart[c], Xin Lu[d,a,b], Sandra Moore[e], Erik Wetter[b,f], Kankoe Sallah[c], Stanislas Rebaudet[e] and Renaud Piarroux[e]

Corresponding author: Linus Bengtsson (linus.bengtsson@flowminder.org)

[a]*Department of Public Health Sciences. Karolinska Institutet, Stockholm, Sweden;*
[b]*Flowminder Foundation, Stockholm, Sweden;*
[c]*Aix-Marseille University, UMR 912 SESSTIM (INSERM-IRD-AMU), Marseille, France;*
[d]*College of Information System and Management, National University of Defence Technology, Changsha, China;*
[e]*Aix-Marseille University, UMR MD 3, Marseille, France;*
[f]*Stockholm School of Economics, Stockholm, Sweden;*

Effective response to infectious disease epidemics requires focused control measures in areas predicted to be at high risk of new outbreaks. We aimed to test whether mobile operator data could predict the early spatial evolution of the 2010 Haiti cholera epidemic. Daily case data were analysed for 78 study areas from October 16 to December 16, 2010. Movements of 2.9 million anonymous mobile phone SIM cards were used to create a national mobility network. Two gravity models of population mobility were implemented for comparison. Both were optimized based on the complete retrospective epidemic data, available only after the end of the epidemic spread. Risk of an area experiencing an outbreak within seven days showed strong dose-response relationship with the mobile phone-based infectious pressure estimates. The mobile phone-based model performed better (AUC 0.79) than the retrospectively optimized gravity models (AUC 0.66 and 0.74, respectively). Infectious pressure at outbreak onset was significantly correlated with reported cholera cases during the first ten days of the epidemic ($p<0.05$). Mobile operator data is a highly promising data source for improving preparedness and response efforts during cholera outbreaks. Findings may be particularly important for containment efforts of emerging infectious diseases, including high-mortality influenza strains.

Introduction: Re-occurring infectious disease outbreaks due to cholera, measles and other preventable infectious diseases contribute to a major disease burden affecting low- and middle-income countries.[1,2] Concurrently, outbreaks of new infectious diseases with pandemic potential pose a considerable threat to human life and development.[3,4] Response to, and ideally containment of,[5] an infectious disease outbreak can be greatly improved if health care response and outbreak control measures can be focused to areas predicted to be at the highest risk of experiencing new outbreaks.[6,7] Accurate models of the geographic distribution of epidemic risk could significantly enhance the population-level effects of interventions implemented to control the spread of transmittable diseases.[8] Considerable progress has been made in predicting temporal evolution of epidemics once outbreaks have progressed beyond a small initial group of cases.[7,9] However, predicting spatial transmission routes of epidemics has proven to be remarkably difficult, due to the importance of rare, long-distance transmission events,[10] limited data on population mobility, unknown population immunity levels,[9] low sensitivity and specificity of case reports[11] and limited access to accurate and spatiotemporally resolved case data.[12]

Empirical data has provided key insight into the spatial spread of measles in England[13] and Niger[14] as well as into influenza spread in the USA and Europe.[11,12,15] While population mobility plays a key role in such modelling studies,[10,16] it has not been possible, until now, to study detailed and concurrent data on both population mobility and spatiotemporal distribution of cases. Instead, empirical studies have used either models of population mobility, preferentially gravity models,[17] or census data on work-home commuting as proxies for total mobility during outbreaks.[11] It is also not clear how to choose and properly parameterize mobility models across contexts in new outbreaks. This is especially problematic during the critical early outbreak phases, when interventions have the greatest effect, but limited data are available to fit transmission models. Mobile operator data is a promising source of national mobility patterns and has notably been used as in malaria modelling studies.[18,19] However, the extent to which this type of data accurately reflect movements of infectious

persons and its utility in predicting spatial spread of infectious agents have not been evaluated.

The largest cholera epidemic to strike a single country in recent history was the 2010 Haitian outbreak.[20] The 2010 Haiti cholera epidemic provides a unique opportunity to explore the influence of population mobility on the spatial evolution of a large-scale cholera outbreak. First, cholera had not previously affected the country for at least a century, thereby rendering epidemic development unbiased by differential population immunity. Second, the circumstances and location of the onset of the cholera epidemic are well understood.[21-23] Third, daily case reporting based on WHO criteria was initiated very early throughout the country, and the notification system was highly effective.[20-23]

Methods: The daily case reports per health facility enabled us to determine daily case numbers per commune while the epidemic spread throughout the country (October 14 and 64 days onwards). We defined the end of the study period as December 16, when the peak of the epidemic was reached and all but one commune had reported at least one case. In 62 communes out of 140 communes, including the eight communes within the Port-au-Prince metropolitan area, there may have been patients who sought healthcare in neighbouring communes.
Mobile phone data: The analysed anonymous mobile phone data consisted of the last outgoing call or text message each day from October 15 to December 19, 2010 for all 2.9 million users belonging to the largest mobile operator, Digicel Haiti.[24]

We used the mobile phone data to construct a mobility matrix M^{phone}, with elements m_{ij}^{phone}, indicating the average daily proportion of mobile phones relocating from study area i to j, comparing their last registered location on day t with their last registered location on day *t-1*. The mobility network built on the basis of M^{phone} displays strong connectivity both between Port-au-Prince and large parts of the country as well as between other urban areas and their surrounding countryside (Fig. 1). We calculated the infectious pressure $P_j(t)$, sustained by each study area j during the period from October 21 (from seven days after the disease onset of the first case in Haiti) to December 16, according to Eq. 1, in which $c_i(t)$ is the number of reported cases in study area i on day t:

$$P_j(t) = \sum_{i,i\neq j}^{n}\left[m_{ij}^{phone}\sum_{k=1}^{7}c_i(t-k)\right] \qquad \text{(Eq. 1)}$$

For comparative purposes, we implemented a gravity model of population mobility.

In the absence of detailed mobility data, parameter values for the gravity model are unknown and needs to be assigned. We chose to optimize the model based on the retrospective case data from the complete study period. Note that this optimisation thus could not have been performed until after the spatial spread of the epidemic was complete. Our comparison model thus performs better than a model that could have been developed during the epidemic. We produced two separate optimisations. In the first we optimised the gravity model by choosing values for μ_i and δ (0.154 and 122 respectively), which minimised the residual sum of squares between reported daily cholera cases in each study area and the estimated pressure from the gravity model.[25] In the second we chose parameter values (0.158 and 3.5 respectively) that maximised the area under the curve (AUC), among all possible ROC curves. (P^{grav1} and P^{grav2}, respectively).

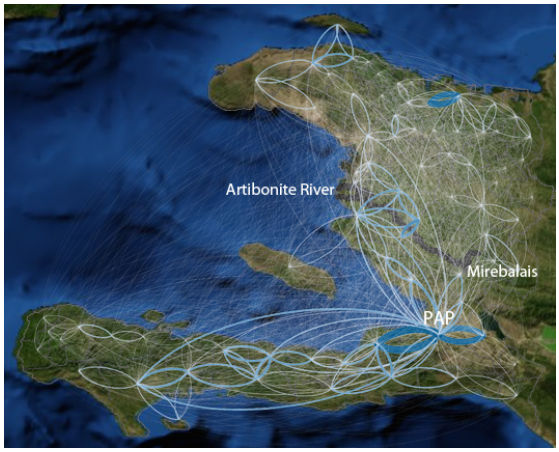


Figure 1: Mobile phone mobility network. The average absolute number of mobile phones moving between the study areas (October 15 to December 19, 2010). The original outbreak location (Mirebelais), the Artibonite River (dark blue) and Port-au-Prince (PAP) are depicted.

Results: The risk of a study area experiencing a new outbreak correlated closely with the infectious pressure. Over a pressure level of 22

(P^{phone}), all areas (six study areas) experienced outbreaks within seven days (see S5 for P^{grav1}, P^{grav2} and sensitivity analyses).

Building upon this strong correlation between infectious pressure and outbreak risk, we created a binary test to predict an outbreak occurring within the upcoming seven days, based solely on thresholds of infectious pressure. We predicted an outbreak to occur at a given pressure threshold) and plotted the corresponding sensitivity and specificity of each threshold (Fig. 2b). We compared the model based on the mobile operator mobility data (P^{phone}) with the gravity models (P^{grav1} and P^{grav2}), for which P^{grav2} was optimised specifically to yield the maximum possible area under the curve (AUC) in this analysis. Comparing these ROC curves, the P^{phone} model clearly performs better than the P^{grav1} model and slightly better than the P^{grav2} model, yielding a higher specificity for a given level of sensitivity. Note that both gravity models rely on parameter optimisations that could not have been performed until the epidemic spread was completed.

Discussion: Our results show that the risk of epidemic onset of cholera in a given area and the initial intensity of local outbreaks could have been anticipated during the early days of the Haitian epidemic using case reports and the mobility patterns of mobile phones. We show that the specificity and sensitivity of predictions of epidemic spread was improved or comparable to currently available optimized mobility models. Most importantly, the predictions based on the mobile operator data did not rely on retrospective optimization of parameter models and could thus be available from the start of an outbreak. This is important as gravity model parameters are highly context specific.[26,27] Although this study focuses only on the influence of human mobility, future mobile phone based models focusing on cholera may benefit from including data on spatial distributions of access to water and sanitation,[28] bacterial transmission via waterways,[25] agricultural practices,[29] differential population immunity levels[30] and interactions between infectiousness and mobility and between infectiousness and phone use.

These results indicate that outbreak preparedness and response to epidemic agents, such as cholera, can be enhanced. The findings may have particular importance for improving early containment efforts of emerging infectious diseases, such as high mortality strains of pandemic influenza, and the response to vaccine-preventable diseases, such as measles, in low-income settings.

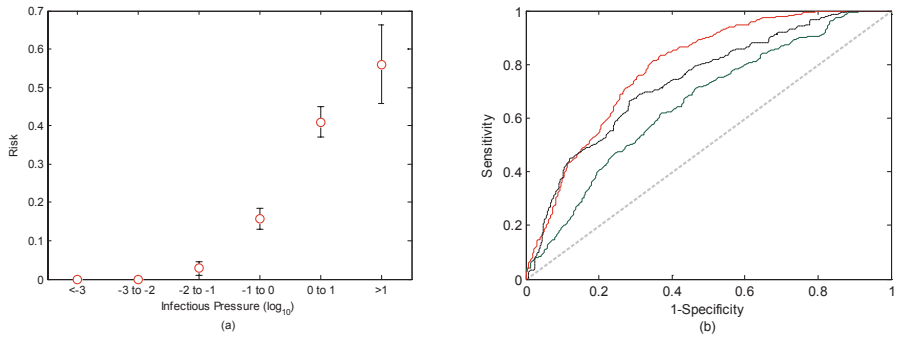


Figure 2: a) Relationship between infectious pressure, calculated from the mobile phone data and the risk of areas experiencing a new outbreak within seven days. b) ROC curve for predicting outbreak at increasing thresholds of infectious pressure (red: P^{phone}, green: P^{grav1}, black: P^{grav2}).

References

1 Black, R. E. *et al.* Global, regional, and national causes of child mortality in 2008: a systematic analysis. *Lancet* **375**, 1969-1987, doi:10.1016/S0140-6736(10)60549-1 (2010).

2 Murray, C. J. *et al.* Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *The Lancet* **380**, 2197-2223 (2013).

3 Murray, C. J., Lopez, A. D., Chin, B., Feehan, D. & Hill, K. H. Estimation of potential global pandemic influenza mortality on the basis of vital registry data from the 1918-20 pandemic: a quantitative analysis. *Lancet* **368**, 2211-2218

4 Smith, R. D., Keogh-Brown, M. R., Barnett, T. & Tait, J. The economy-wide impact of pandemic.

5 Longini, I. M., Jr. *et al.* Containing pandemic influenza at the source. *Science* **309**, 1083-1087,

6 Ferguson, N. M. *et al.* Strategies for containing an emerging influenza pandemic in Southeast Asia. *Nature* **437**, 209-214, doi:10.1038/nature04017 (2005).

7 Anderson, R. M. & May, R. M. *Infectious diseases of humans : dynamics and control*. (Oxford University Press, 1991).
8 Tuite, A. R. *et al.* Cholera epidemic in Haiti, 2010: using a transmission model to explain spatial spread of disease and identify optimal control interventions. *Ann Intern Med* **154**, 593-601
9 Grassly, N. C. & Fraser, C. Mathematical models of infectious disease transmission. *Nature Reviews Microbiology* **6**, 477-487 (2008).
10 Riley, S. Large-scale spatial-transmission models of infectious disease. *Science* **316**, 1298-1301,
11 Viboud, C. *et al.* Synchrony, waves, and spatial hierarchies in the spread of influenza. *Science* **312**, 447-451
12 Eggo, R. M., Cauchemez, S. & Ferguson, N. M. Spatial dynamics of the 1918 influenza pandemic in England, Wales and the United States. *J R Soc Interface* **8**, 233-243, doi:10.1098/rsif.2010.0216 (2011).
13 Grenfell, B. T., Bjornstad, O. N. & Kappey, J. Travelling waves and spatial hierarchies in measles epidemics. *Nature* **414**, 716-723
14 Bharti, N. *et al.* Explaining seasonal fluctuations of measles in Niger using nighttime lights imagery. *Science* **334**, 1424-1427 (2011).
15 Crepey, P. & Barthelemy, M. Detecting robust patterns in the spread of epidemics: a case study of influenza in the United States and France. *Am J Epidemiol* **166**, 1244-1251 (2007).
16 Balcan, D. *et al.* Multiscale mobility networks and the spatial spreading of infectious diseases. *PNAS* **106**, 21484-21489, (2009).
17 Truscott, J. & Ferguson, N. M. Evaluating the adequacy of gravity models as a description of human mobility for epidemic modelling. *PLoS Comput Biol* **8**, e1002699 (2012).
18 Wesolowski, A. *et al.* Quantifying the impact of human mobility on malaria. *Science* **338**, 267-270, doi:10.1126/science.1223467 (2012).
19 Le Menach, A. *et al.* Travel risk, malaria importation and malaria transmission in Zanzibar. *Sci Rep* **1**, 93 (2011).
20 Barzilay, E. J. *et al.* Cholera surveillance during the Haiti epidemic--the first 2 years. *The New England journal of medicine* **368**, 599-609 (2013).
21 Piarroux, R. *et al.* Understanding the cholera epidemic, Haiti. *Emerg Infect Dis* **17**, 1161-1168, doi:10.3201/eid1707.110059 (2011).
22 Frerichs, R. R., Keim, P. S., Barrais, R. & Piarroux, R. Nepalese origin of cholera epidemic in Haiti. *Clin Microbiol Infect* **18**, E158-163 (2012).
23 Frerichs, R. R., Boncy, J., Barrais, R., Keim, P. S. & Piarroux, R. Source attribution of 2010 cholera epidemic in Haiti. *Proceedings of the National Academy of Sciences of the United States of America* **109**, E3208; author reply E3209 (2012).
24 Lu, X., Bengtsson, L. & Holme, P. Predictability of population displacement after the 2010 Haiti earthquake. *PNAS* **109**, 11576-11581 (2012).
25 Rinaldo, A. *et al.* Reassessment of the 2010-2011 Haiti cholera outbreak and rainfall-driven multiseason projections. *PNAS* **109**, 6602-6607, (2012).
26 Garske, T. *et al.* Travel patterns in China. *PLoS One* **6**, e16364 (2011).
27 Alberto Rubio, V. F.-M., Enrique Frias-Martinez and Nuria Oliver. Human Mobility in Advanced and Developing Economies: A Comparative Analysis. *Telefonica Research, Madrid, Spain*.
28 Waldman, R. J., Mintz, E. D. & Papowitz, H. E. The cure for cholera--improving access to safe water and sanitation. *N Engl J Med* **368**, 592-594, doi:10.1056/NEJMp1214179 (2013).
29 Gaudart, J. *et al.* Spatio-temporal dynamics of cholera during the first year of the epidemic in Haiti. *PLoS Negl Trop Dis* **7**, e2145 (2013).
30 Chao, D. L., Halloran, M. E. & Longini, I. M. Vaccination strategies for epidemic cholera in Haiti with implications for the developing world. *PNAS* **108**, 7081-7085 (2011).

Micro Dynamics of Social Interactions: Quantifying Human Life

Vedran Sekara[1], Arkadiusz Stopczynski[1,2] & Sune Lehmann[1,3]

[1]*Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kgs. Lyngby, Denmark,*
[2]*Media Lab, Massachusetts Institute of Technology, Cambridge, MA, USA,*
[3]*The Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark.*

Community structure is a key property in many real world social networks and is thought to play a crucial role in both network function[1] and resilience[2]. Due to frequent changes in the communication and activity patterns of individuals[3] the underlying network topology is under constant evolution[4], but our understanding of the basic features that facilitate this evolution is quite limited[5,6]. Based on a dataset of minute-to-minute social interactions between approximately 1 000 densely connected individuals, spanning more than a year[7], we study the micro dynamics of an entire social system. Slicing temporal data into *micro*-snapshots we show that isolated motifs directly appear (Fig. 1a-c)—thus conventional community detection heuristics become redundant. By tracking the evolution of motifs across time, we demonstrate a new method that identifies social gatherings. On the shortest time-scale, we find that gatherings are fluid, with members coming and going, but organized via a stable core of individuals that are persistent throughout the entirety of meetings. Some cores exhibit a pattern of recurring meetings across weeks and months, each with varying degrees of regularity. In terms of network structure, we find that cores are highly overlapping, and large cores contain rich inner structure with hierarchically nested sub-cores. From an individual perspective, these cores can be used to summarize social contexts, where each person can belong to multiple cores (Fig. 1d). In Fig. 1e we observe the temporal patterns of core participation from late January to late April for the ego-network shown in Fig. 1d—displaying complex participation patterns where regularity is mixed with randomness. Applying cores as a vocabulary of states we summarize human social life by drawing on previous work that has studied the predictability of individuals' trajectories[8]. We show that in analogy to human mobility, social contexts can be predicted with high precision based on past behavior. Combining social behavior of each individual with his/her time resolved geographic mobility we can in fact describe a persons full life. Results show that there is no overall correlation between social and location predictability, meaning that you can be highly predictable with respect to location but not with respect to your social context or vice versa. However, in real life, we have varying degrees of predictability, at night we have low entropy because we sleep in the same location, while during days and evenings our entropy is higher—because we occupy more states. Illustrating that predictability has a inseparable temporal component. During periods when the ability to predict location based on past behavior is systematically lowered, social contexts provide a new way to predict location for individuals.

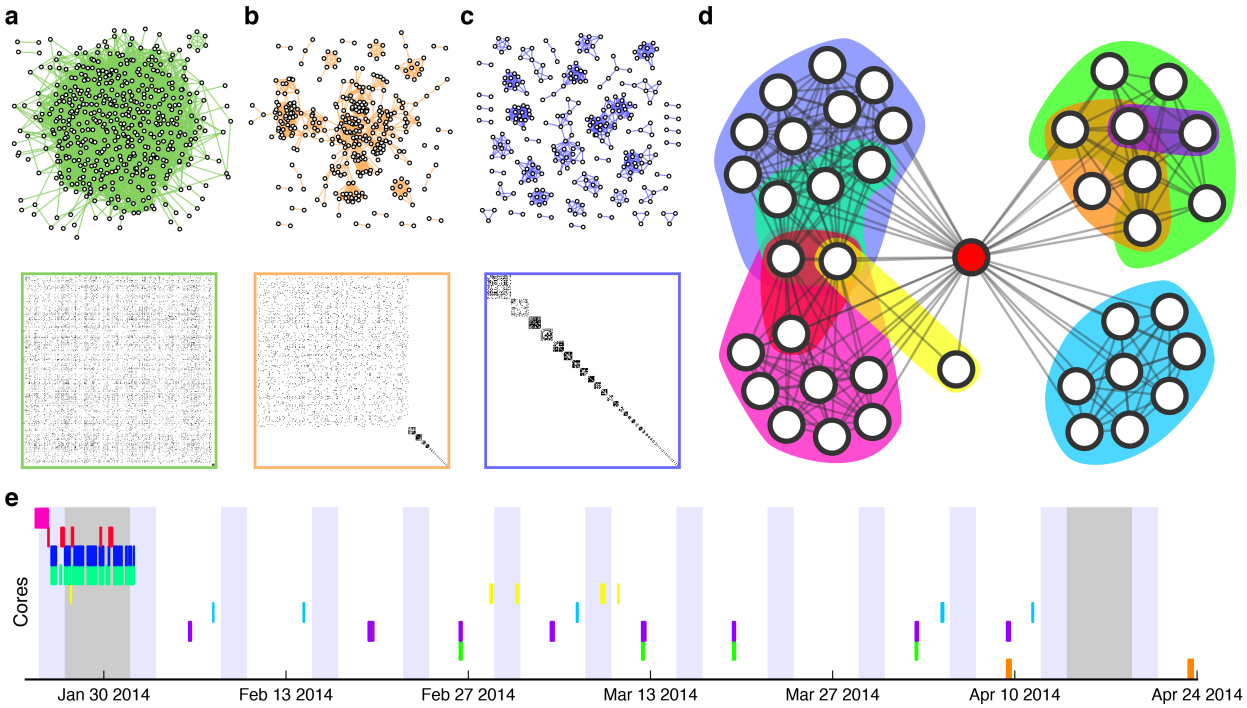


Figure 1: **Dynamics of social behaviour. a-c** Network slices obtained by slicing the social dynamics using varying temporal windows (1 day, 1 hour, and 5 minutes). Below, adjacency matrices, colored in agreement with networks, and sorted according to the largest component. **d**, Ego-centric view of the community structure for a single individual (red node) that belongs 9 temporally distinct groups. In the aggregated sense communities are both overlapping and hierarchically organized. Only frequently occurring communities (appearing on average at least one per month) are shown. **e**, Temporal dynamics of each individual community shown in panel d, colored accordingly, displaying some degree of order but also randomness.

References

1. Fortunato, S. Community detection in graphs. *Physics Reports* **486**, 75–174 (2010).
2. Albert, R., Jeong, H. & Barabási, A.-L. Error and attack tolerance of complex networks. *Nature* **406**, 378–382 (2000).
3. Cattuto, C. *et al.* Dynamics of person-to-person interactions from distributed rfid sensor networks. *PloS one* **5**, e11596 (2010).
4. Holme, P. & Saramäki, J. Temporal networks. *Physics reports* **519**, 97–125 (2012).
5. Kossinets, G. & Watts, D. J. Empirical analysis of an evolving social network. *Science* **311**, 88–90 (2006).
6. Palla, G., Barabási, A.-L. & Vicsek, T. Quantifying social group evolution. *Nature* **446**, 664–667 (2007).
7. Stopczynski, A. *et al.* Measuring large-scale social networks with high resolution. *PLoS ONE* **9**, e95978 (2014).
8. Song, C., Qu, Z., Blumm, N. & Barabási, A.-L. Limits of predictability in human mobility. *Science* **327**, 1018–1021 (2010).

Session 6 :: Cities (II)

NEIGHBORHOOD AND NETWORK SEGREGATION: ETHNIC HOMOPHILY IN A 'SILENTLY SEPARATE' SOCIETY*

Joshua Blumenstock, University of Washington

Ott Toomet, Tartu University

Rein Ahas

Erki Saluveer

December 31, 2014

Abstract

We examine the relationship between geography and ethnic homophily in Estonia, a linguistically divided country. Analyzing the physical locations and cellular communications of tens of thousands of individuals, we document a strong relationship between the ethnic concentration of an individual's geographic neighborhood and the ethnic composition of the people with whom he interacts. The empirical evidence is consistent with a theoretical model in which individuals prefer to form ties with others living close by and of the same ethnicity. Exploiting variation in the data caused by migrants and quasi-exogenous settlement patterns, we find suggestive evidence that the ethnic composition of geographic neighborhoods has a causal influence on the ethnic structure of social networks.

Keywords: residential segregation, homophily, social segregation, minorities

JEL codes: J15, J61

1 Introduction

Ethnic segregation is a prominent feature of most contemporary and historical societies. Such fractionalization has been tied to patterns of economic development and growth, investment in human capital, the efficiency of inefficient labor markets, violence and corruption, as well

*The authors are grateful for thoughtful comments from Mark Ellis, Ira Gang, Matthew Jackson, Štěpán Jurajda, Ramona Angelescu-Naqvi, Xu Tan, and participants of Tartu 2014 Xmas seminar. We also thank Siiri Silm for help with geodata analysis. We gratefully acknowledge financial support from GDN RRC and ESRC and Estonian Science Foundation grants IUT2-17, IUT20-49, and 9247. All errors are our own.

as broader patterns of inequality, prejudice, and discrimination (cf. Easterly and Levine 1997, Collier 1998, Cutler and Glaeser 2007, Bayard, Hellerstein, Neumark, and Troske 1999, Miguel and Gugerty 2005).

Ethnic homophily, or an individual's preference for co-ethnics, plays an important role in shaping patterns of ethnic segregation. When individuals prefer to associate with others of the same type, this influences where people choose to live, where people choose to work, and with whom they choose to interact (Massey 1985). At the same time, the reverse process may also obtain: when people are physically surrounded by others of their own type, they may choose to associate with them to a higher degree The relationship between neighborhood segregation, as defined by the structure of geographic communities, and network segregation, as determined by actual patterns of interaction, is thus theoretically ambiguous. Empirical adjudication is similarly complicated by the difficulty of separately observing both types of segregation on a single population.

Here, we study the relationship between neighborhood segregation and network segregation using a novel dataset that allows us to disentangle the ethnic composition of an individual's physical surroundings from the ethnic composition of people with whom he interacts. The context for this study is Estonia, a country with a long and complex history of ethnic strife and resettlement. Prior to World War II, roughly 94 percent of the Estonian population was ethnically Estonian; however, the Estonia's incorporation into the USSR created a large influx of Russian immigrants into Estonia, and by 1989 roughly 39 percent of the Estonian population was ethnically Russian. Due to Stalin's brutal regime, and the anti-Russian backlash that followed Estonian independence, strong feelings of animosity exist between the two groups. Modern-day Estonian society has been described by Heidmets (1998) as one of "silent separation" where the two ethnic groups occupy the same physical spaces but rarely interact.

Empirically, we analyze a large dataset of anonymized mobile phone communications data that allows us to observe, for tens of thousands of individuals, all locations inhabited be each individual over a five-year period, as well as all phone-based interactions with other individuals in the dataset. Critical to our current analysis, we also observe the language spoken by each mobile phone subscriber. In Estonia, linguistic preference remains a core component of ethnic identity, with most ethnic Russians choosing to speak the Russian language, and most ethnic Estonians choosing to speak the Estonian language (Tammaru 2001). Using these data, we are thus able to separately measure, for each individual, the extent to which she is physically surrounded by coethnics, and the ethnic composition of her social network.

We observe a strong and robust relationship between neighborhood and network homophily. In other words, while the average individual is more likely to interact with coethnics than others, those individuals who are physically surrounded by coethnics are even more likely to interact with coethnics. This relationship exceeds what would be expected under a naive model of geographically constrained random attachment, where individuals randomly interact with those in their geographic network irrespective of ethnicity. This effect persists even when controlling for a range of demographic characteristics.

We further find suggestive evidence that the ethnic composition of an individual's network exerts a causal influence on the social connections formed by the same individual. In particular, when we study the homophilic tendencies of migrants,[1] we find that they are *less* sensitive to their physical surroundings; while they still interact more with coethnics the more they are surrounded by coethnics, this relationship is less pronounced than it is in the population of people who never migrate. It appears that this differential effect for migrants is primarily driven by the fact that migrants remain connected to their place of origin, and do not immediately form connections in their new neighborhood.

The above results are consistent with a simple model of social preferences where geographically close coethnic ties are most likely to form. If interethnic ties are possible, the number of interethnic friends will depend on the "local" ethnic composition. The relationship is weaker for migrants who have just recently arrived in the region and have not yet had time to adapt their networks to the new local environment.

In analyzing the behavior of migrants, we cannot rule out the possibility that individuals with different homophilic tendencies select into migration. However, we exploit several plausibly exogenous sources of variation to help assuage such concerns of endogeneity. First, we use the high-resolution mobility data to infer the exact date of migration, and look separately at homophily for migrants of different cohorts. Here, we find that recent migrants are indeed more likely to be in contact with their origin community than migrants who have lived in the destination community for several years; these recent migrants are also less likely to interact with individuals in their new neighborhood.

As a second robustness test, we analyze the relationship on Soviet-era housing estates. As the housing market was virtually non-existent during the communist period, the residential location was largely exogenous and the segregation preferences play little role accordingly. The find the

[1] Migrants are identified in our data based on the set of geolocated mobile phone towers used to route their calls. Migrants constitute roughly 10% of the sample population, and our results are robust to a variety of plausible ways of classifying migrants.

effect being almost as strong on these estates as in the full sample, suggesting that neighborhoods that play a substantially stronger role in determining the social networks.

Our paper thus documents the strong relationship between neighborhood and network segregation, and provides suggestive evidence on causality. Taken further, these results imply that physical integration may lead to social integration.

The remainder of the paper is organized as follows: Section 2 describes the background and institutions of Estonia, the country we analyze. Section 3 provides a simple theoretical framework for interpreting our results. The data and the empirical approach are described in greater detail in Section 4, and Section 5 discusses the results. In Section 6 we discuss and interpret the results in greater detail before concluding.

2 Estonia: Background and Context

The focus of our empirical analysis is Estonia, a country uniquely suited for the empirical analysis of ethnic segregation. Before World War II, roughly 94 percent of the population was ethnic Estonians, with the remainder largely comprised of ethnic Russians (Katus 1990). During WWII, Estonia was incorporated into the Soviet Union, and fell under the brutal Stalinist regime for nearly a decade. As part of the post-war reconstruction and industrialization effort, it experienced large-scale immigration from other parts of the Soviet Union, mainly from Russia. This process resulted in an increase of the population of the country to 1.57 million by 1989, 39% of whom were ethnic minorities (Tammaru and Kulu 2003).

Most of the migrants from the Soviet Union were Russian-speaking, and were regarded by the ethnic Estonian population as the "Russians" and associated with the harsh regime. Relations between the two dominant ethnic groups deteriorated rapidly, and by the 1970s the society was sharply divied along linguistic lines. Estonians and Russians attended different schools, worked in different establishments and followed different media outlets (Kalmus and Pavelson 2002, Vihalemm 2010). Russian language functioned as the *lingua franca* of the Soviet Union and most Estonians possessed a working knowledge of the language. However, the command of the Estonian language was poor among the minorities (Kulu and Tammaru 2004). Despite the official Soviet policy, Estonians never considered themselves a part of the Soviet nation, and distinguished clearly between in-group (i.e. "Estonians") and out-group (i.e. "Russian") members. This linguistically divided society where two ethnic communities live rather parallel lives has thus been characterized as a"silently separated society" (Heidmets 1998).

After the fall of the USSR, Estonia began a nation-building that has been widely regarded as discriminatory towards ethnic Russians (Pettai 2002). First, the newly elected parliament granted citizenship only to nationals of the pre-WWII republic and to their offspring (Everly 1997). As a result, a sizeable part of the current minority population does not have Estonian citizenship. Second, the Estonian language was made the sole official language of the country, causing a gradual deterioration of Russian language skills among Estonians, particularly among the younger generation. However, a large percentage of Russians are still not able to communicate in Estonian (Kulu and Tammaru 2004). For this reason there is no universally shared language in the country today. The shift in the roles of the languages was also accompanied by relative deterioration of the economic position of Russian speakers (Leping and Toomet 2008). In this way the historic animosity between the two language groups, the high levels of segregation in many important spheres, and the lack of a *lingua franca* contribute to the low number of interethnic contacts and general lack of social integration today. While the attitudes toward the other ethnic groups have been improving through the previous decade, interethnic engagement is still relatively infrequent (Lauristin, Uus, and Seppel 2011). The tensions do occasionally rise to the surface as, for instance, during the large-scale riots in Tallinn in the spring of 2007.[2]

The fall of the Soviet regime also had a signifant impact on patterns of migration and settlement. The Russian immigration of the Soviet era come to a rapid halt, while both urbanization and sub-urbanization gathered momentum. The main mechanism that shaped the ethnic composition of the urban neighborhoods, including in the capital Tallinn, is related to historic immigration and residential construction. Between 1950 and 1989, the population of the country rose by more than 40%, from 1,097,000 to 1,565,000, mainly through immigration from elsewhere in the USSR (Tammaru 2001). In the absence of a housing market, immigrants were usually granted flats in newly built, standardized, high-rise housing estates (Kährik and Tammaru 2010) which nowadays provide accommodation for a large part of the total population. These are often dominated by ethnic Russians, whereas Estonians are over-represented in pre-WWII (and also in the small post-1991) housing stock, and also in detached houses. In this way, the current ethnic composition across urban neighborhoods largely reflects the immigration patterns during the construction periods, rather than factors such as socio-economic status. Recent suburbanization, and the fact that a substantial part of the immigrant population left after the collapse of Soviet

[2]The riots were caused by the relocation of a Soviet World War II monument, popularly referred to as the "Bronze Soldier", from central Tallinn to a military cemetery. From the perspective of ethnic Estonians, the monument was considered to glorify oppressive Soviet rule, while for the Russian-speaking population it was a symbol of victory over the Nazis in the "Great Patriotic War." See Schultze (2011).

Union, has not radically changed this picture (Hess, Tammaru, and Leetmaa 2012).

3 Theoretical Framework

We develop a simple model that includes two different groups of people, two regions, and migration between these regions. The model allows us to describe the expected number of ties within and across groups, and illustrates two important results: first, homophily, or the preference for individuals of the same group, is positively correlated with the neighborhood ethnic composition; and second, this relationship is stronger for people who remain in a single location and weaker for people who migrate. We formalize these results as propositions below and explain the intuition.

We consider a world containing two regions, A and B, and two (ethnic) groups, 0 and 1. The population in region A is n^A and in region B it is n^B. There are $n_0^A = \pi^A n^A$ group-0 members and $n_1^A = (1-\pi^A)n^A$ group-1 members in region A where π^A is the fraction of group-0 members in region A. The expressions for region B are analogous.

Assume that ties between two individuals, located at geographic distance d^g and "ethnic distance" d^e is created by a Poisson process with intensity

$$\mu = \phi(d^g) \cdot \chi(d^e). \tag{1}$$

The first term $\phi(\cdot)$ describes how the intensity depends on the geographic distance and $\chi(\cdot)$ describes the dependence on "ethnic distance", where $d^e = 0$ for members of the same group and $d^e = 1$ for members of the different group. As we only have two regions and two groups, we can label the corresponding function values as ϕ^0 and ϕ^1 for local ties and distant ties, and χ^0 and χ^1 for in-group and out-group ties. We assume "short" ties arise more easily: $\phi^0 > \phi^1$ and $\chi^0 > \chi^1$. Ties are destroyed with Poisson process with intensity δ, independent of their "length". Assuming that the number of actual ties is much smaller than the number of potential ties, we have the following expression for the expected number of individual ties, ν. For instance, for a group 0 member in region A we have:

$$\frac{\mathrm{d}\,\nu}{\mathrm{d}\,t} = \phi^0(\chi^0 n_0^A + \chi^1 n_1^A) + \phi^1(\chi^0 n_0^B + \chi^1 n_1^B) - \delta\nu. \tag{2}$$

The first term describes creation of new local ties, the second term that of distant ties, and the last term the destruction of ties.

If people who remain in one region have lived there long enough, their expected number of ties correspond to these in the steady-state, ν^* and homophily (for group 0) $h^* = \nu_0^*/(\nu_0^* + \nu_1^*)$.[3] [4] Here ν_0^* and ν_1^* are ties to group 0 and group 1, i.e. in-group and out-group ties for group 0.

Proposition 1. *Homophily in social ties is positively related to the ethnic composition of local geographic neighborhoods:* $\frac{\partial h^*}{\partial \pi^A} > 0$.

Proof. See Appendix A. □

Intuitively, as our model explicitly allows for a greater likelihood of local tie formation, the local population composition influences substantially the actual homophily.

As migrants do not posses the the steady-state equilibrium networks, we solve the dynamic equation (2) and have the following result:

Proposition 2. *Conditional on the average level of neighborhood homophily, the homophily of migrants is less sensitive to the neighborhood composition than that of non-migrants.*

Proof. See Appendix A. □

This effect exists because local network ties develop over time. In-migrants, arriving from neighborhoods of different ethnic composition, have only partially adapted to the ethnic composition of their new neighborhoods.

4 Data and Empirical Approach

To analyze the relationship between neighborhood and network segregations, we exploit a large set of data on mobile phone use in Estonia. This dataset permits us to simultaneously observe the locations of thousands of individuals over a period of several years, the ethnicity of those individuals, as well as the extent to which those individuals interacts with others of the same or different ethnicity.

[3] Homophily is a measure of exposure dimension of segregation (Massey and Denton 1988). Here we focus on homophily based on the percentage of contacts in an individual's network, but our empirical results are robust to other common definitions of homophily.

[4] Empirically we observe $\mathbb{E}\, h^* = \mathbb{E}[\nu_0^*/(\nu_0^* + \nu_1^*)]$ instead of $\mathbb{E}\, \nu_0^*/(\mathbb{E}\, \nu_0^* + \mathbb{E}\, \nu_1^*)$. However, under mild assumptions these two expressions are equal. See Appendix A.

4.1 Data

We employ cellphone usage data from the largest mobile service provider in Estonia, EMT, which has roughly 60% market share. We obtained two related datasets from this operator.

Passive Positioning Data: The first dataset contains the locations of each individual over the period from 2007–2012. As is typical for such positioning data, we do not observe the actual location but rather the Cell Global Identity (CGI), i.e. the network antenna which processed the outgoing call.[5] This gives us a spatial resolution of a few hundred meters in dense urban environments, and up to five kilometers in rural areas. The data include the time of each call activity and the corresponding location (CGI). Every network user (as identified by a SIM card with a unique phone number) is assigned a random identification tag, making it possible to track the same user over time.[6] Based on timing, location and regularity of the calls, we attach a place of residence to each cellphone (Ahas, Silm, Järv, Saluveer, and Tiru 2010). We focus on yearly modal place of residence in order to avoid places that are too unstable, or seasonal migration.

Call Graph: The second dataset contains a complete 10-day call graph, which allows us to observe in a fixed window who is communicating with whom. This call data records (CDR) are similar to the passive positioning data, but for each call or SMS event we also observe the ID of the second party.

For each subscriber in our dataset, we additionally observe whether the subscriber prefers the Estonian or Russian language. Since the correlation between ethnicity and language is almost complete (Kulu and Tammaru 2004) we use language as a proxy for ethnic background. All of these data use shared anonymized identifiers which allow us to link long-term location information to the network communication data, and in this way to relate the segregation in communication network to segregation in space.

We perform our empirical analysis on a random sample of 48,781 individual mobile phone subscribers. Of these, 42,604 are Estonian and 6,178 are Russian; 46,835 have a known residence location, and 18,716 live in the metropolitan area.

[5] In a cellular network, a "cell" roughly corresponds to an area where all the network traffic goes through a single antenna. Usually, several antennas are located in one transmission tower and are oriented in different directions. We know the location of the transmission towers and the direction of the antennas. Based on this information, we can construct "typical" cell boundaries; however, the actual boundaries may fluctuate due to network load, obstacles and noise.

[6] The individuals and real phone numbers cannot be identified using the tag in our data. The collection, storage, and processing of the data complies with all European Union requirements regarding the protection of personal data (European Commission 2002). Approval was also obtained from the Estonian Data Protection Inspectorate and the University of Washington Human Subjects Division.

4.2 Empirical Approach

In our empirical analysis, we examine the relationship between the ethnic composition of an individual's immediate physical neigborhood, and the ethnic composition of his call graph. The passive positioning data allows us to determine where individuals live, which in turn makes it possible to observe the physical neighborhood. We will begin by defining physical neighborhood along political boundaries, but our results are robust to several alternative definitions.[7] Similarly, we will initially measure the ethnic composition of the call graph as the fraction of contacts of the same ethnicity, but our results obtain when we define network homophily as the fraction of communication events (i.e., the weighted call graph).

Below, we will separately analyze the relationship between neighborhood and social network homophily, as well as the extent to which people are connected to current and historical regions of residence. In both cases we show the nonparametric relationships graphically, then test the statistical relationship in a regression specification.

The basic regression equation for estimating the relationship between social network homophily h_i and neighborhood own-group percentage P_i for individual i is

$$h_i = \alpha_0 + \beta E_i + \gamma P_i + \eta P_i \cdot M_i + \epsilon_i \tag{3}$$

where E_i indicates the ethnicity of individual i, and M_i indicates whether i is a migrant.[8] Note that as P_i is defined at region level, we cluster the standard errors within regions.

Later, we will also introduce several variants on model (). First, to analyze the geographic structure of networks in greater detail, we will split the connections into local and distant ones, depending on whether these cross a county border. For migrants, we will also separately analyze the extent to which their current social network is comprised of people residing in their current location, or the location from which they migrated. Additionally, to assess the robustness of our results, we will restrict our sample to specific types of individuals, for instance those who were

[7] We perform our analysis using different types of spatial units. Calculation of neighborhood homophily is based on city tracts inside of the capital city. These are spatial units, based on access roads and housing type. Elsewhere in the metropolitan area we rely on municipalities, the area contains 18 suburban municipalities. Finally, outside of the metropolitan area we use counties, there are 14 of these outside of the metropolitan area. Counties are of roughly equal size (though of very unequal population) and broadly correspond to commute-to-work area around an urban center. We choose such an approach to account for different population density and also to take into account the uneven distribution of network antennas. Finally, we analyze migration and geographic tie distance at county level.

[8] As people show a heterogeneous pattern of spatial mobility, we use several definitions of migrants. The strictest definition requires a valid residence region for all 6 years, and only a single move during this period. This gives us 2,614 migrants. The most flexible definition allows up to three missing yearly locations and up to two moves (we analyze the last of these). This gives us 6,592 migrants. All our central results are robust with respect to the definition of migrants.

likely to be assigned to their current place of residence through a Soviet-era natural experiment. Finally, we will also disaggregate migration by cohort, to determine whether the effects of migration are different for recent migrants. Formally, if b is the proportion of given type connections and YSM is years since migration, we estimate:

$$b_i = \alpha_0 + \beta \cdot YSM_i + \epsilon_i. \tag{4}$$

5 Results

5.1 Neighborhood and Network Homophily

We start by presenting the relationship between the regional homophily and the average network homophily for the residents of these neighborhoods. Figure 1 shows the nonparametric relationship between geographic neighbordhood composition and observed homophily in the call graph. Each point indicates the average values for a municipality, with each municipality appearing once for Estonians (hollow circles) and once for Russians (filled circles). The horizontal axis indicates the proportion of a given ethnicity in the municipality and the vertical axis indicates the average homophily of that ethnicity in that region. We see a clear positive relationship for both Estonian speakers and Russian speakers. We also see that while the slope for both groups is similar, Russian speakers are substantially less homophilous.

Next we estimate the same relationship at the individual level, using variants of model (), described above. Table 1 presents four different specifications, where in addition to neighborhood composition we add different combinations of migrant and minority status. All models confirm the visual impression that network composition is strongly related to that of neighborhoods. In the first column, we simply regress individual network homophily h_i on the ethnic composition of the neighborhood P_i. The estimates indicate that a 10 percentage point increase in co-ethnics in a geographic neighborhood corresponds to a 3.5 percentage point increase in co-ethnics in the call network (column 1). This figure is highly significant and robust to the inclusion of several control variables (columns 2-4).

In Column 2 of Table 1, we note that migrants are in general more homophilous than non migrants (by 12 percentage points), but that critically the relationship between the neighborhood and the network is weaker. For migrants, a 10 percentage point incresae in co-ethnic share is only associated with a $10 \times (0.36 - 0.12) = 2.4$ percentage point increase in co-ethnics in the call network.

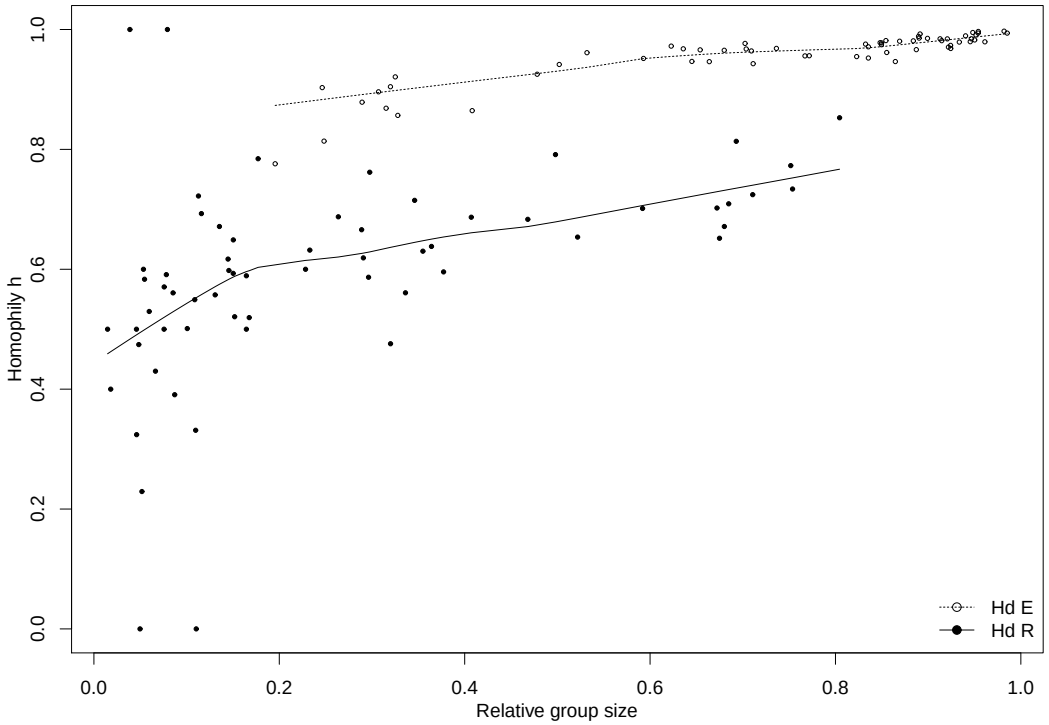


Figure 1: Average homophily as a function of group size across regions. Empty circles denote Estonian-speakers, filled circles Russian speakers. The dashed and solid lines are corresponding smoothed averages.

Model 3 adds controls for ethnicity. We see that while Russians are generally less homophilous than Estonians, they are more sensitive to the neighborhood environment: here, 10 percentage points more co-ethnics in the neighborhood corresponds to $10 \times (0.20 + 0.22) = 4.4$ percentage points higher share in call network. The lower average homophily levels are presumably related to smaller country-wide numbers of Russian speakers while the figure suggests that the stronger correlation is related to neighborhoods with very low group size. The housing type may also play a role as that Russian speakers are overrepresented in Soviet-era high-rise estates.

The above results are evidence of the strong relationship between neighborhood and network segregation. Whether this relationship is causal, however, is not clear. The model presented in Section 3 assumed that neighborhood composition would influence tie formation, and these results are consistent with the two Propositions from that model. Below, we provide additional empirical evidence that appears to indicate there is indeed a casual effect of physical segregation on network homophily.

Outcome: individual homophily h (percentage of co-ethnic contacts in call network)

	Model 1	Model 2	Model 3	Model 4
% coethnics	0.35***	0.36***	0.20***	0.21***
	(0.00)	(0.00)	(0.00)	(0.01)
Migrant		0.12***		0.10***
		(0.02)		(0.02)
Migrant×% coethnics		−0.12***		−0.11***
		(0.02)		(0.02)
Russian			−0.32***	0.34***
			(0.01)	(0.00)
Russian×% coethnics			0.22***	0.22***
			(0.01)	(0.01)
Migrant×Russian				−0.04
				(0.03)
Migrant×Russian×% coethnics				0.02
				(0.06)
Intercept	0.67***	0.66***	0.80***	0.68***
	(0.00)	(0.00)	(0.00)	(0.00)
R^2	0.14	0.14	0.22	0.22
Num. obs.	40819	40819	40819	40819

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 1: Full sample results: relationship between individual homophily and neighborhood ethnic composition.

5.2 Soviet-Era Neighborhoods as a Natural Experiment

If the decision to migrate were exogenous, the fact that the network structure of migrants is less strongly correlated with the geographic struture of their current surroundings could be interpreted as evidence of a causal relationship in which physical surroundings determine network structure. However, since migration and patterns of settlement are not generally exogenous, it may also be the case that people select into migration, and in partcular that people who care less about their physical surroundings are the ones who choose to migrate. To disentangle these two possibilities, we examine the same relationship on a sample of individuals for whom the current location choice is more plausibly exogenous.

As discussed in Section 2, the constant shortage of housing and lack of choice in the housing market created quasi-exogenous variation in patterns of settlement. During the period of highest Russian immigration, 1970s and 80s, individuals had little choice over where to live in cities. We treat this as a natural experiment and limit our analysis here to neighborhoods that are dominated by Soviet-era housing estates. We focus on the capital city Tallinn only. Note that

our setup does not constitute a perfect experiment as we do not know who in our sample did actually live in these neighborhoods during the Soviet period. However, we exclude all the migrants into these areas we are able to identify in the sample.

Outcome: individual homophily h (percentage of co-ethnic contacts in call network)

	Model 1	Model 2	Model 3
% coethnics	0.05	0.20^{***}	0.17^{***}
	(0.04)	(0.04)	(0.04)
Russian		-0.25^{***}	-0.33^{***}
		(0.01)	(0.05)
Russian×% coethnics			0.13
			(0.09)
Intercept	0.85^{***}	0.83^{***}	0.85^{***}
	(0.02)	(0.02)	(0.02)
R^2	0.00	0.16	0.16
Num. obs.	2362	2362	2362

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 2: Soviet-era neighborhoods: relationship between individual homophily and neighborhood ethnic composition.

The results (Table 2) are rather similar to the estimates in the full sample. In particular, the estimate for co-ethnics, 0.17, in Model 3 is statistically indistinguishable to that of Model 3 in the full sample, 0.20 (in Table 1). However, as the sample is smaller, the standard errors are correspondingly larger. The fact that in this sample, who were plausibly exogenously settled into their currently location, we see the same strong relationship between neighborhood and network segregation, lends support to the causal nature of the relationship, i.e., that social networks are at least in part determined through the residential neighborhoods, and that a portion of network segregation is caused by geographic segregation.

5.3 Local and Distant Ties

The homophily-related predictions in Section 3 are based on the fact that local ties are created more easily than distant ones. Unfortunately, as our observations of network structure are based on a single 10-day period, we cannot directly observe the process of tie formation. However, we do observe the geographic structure of ties, which allows us to make two observations that are consistent with our model of tie formation. First, for for migrants, we observe that distant ties are primarily linked to their former place of residence: most of the distant ties were formed when

the distant place still was "local". Second, when we separate the migrants into cohorts by time since migration (we can observe migration 1-6 years ago), we observe an increasing number of local ties and a decreasing number of ties to the previous home region.

Specifically, we split the connections to local and distant ones based on counties (Estonia consist of 15 counties, roughly similar in terms of area but very unevenly populated). We compare the number of ties in the current county of residence, in the previous county of residence (for migrants only), and in all other counties. We select residents of the metropolitan area as of 2012. Table 3 presents their average number of contacts (based on the 10-day callgraph) in selected counties: the Metro area, Ida-Viru (code 44), Pärnu (67), and Tartu (78).[9] The rows correspond to the previous (2011) residence: *Metro* are those who were living in the metropolitan area in 2011 as well, i.e. "stayers"; 44 are those who lived in Ida-Viru and hence they are recent migrants to the metro area, and analogously for the other rows. Columns represent the county of contact.[10] In case of Estonian-Estonian ties (left panel), we see that those who have been in the metropolitan area both for 2011 and 2012 (row labeled "Metro") clearly posses the largest number of the connections in that area. The average number of connections in other counties (columns labeled "44", "67" and "78") is very small. For movers (rows labeled "44", "67" and "78") the picture is different. All of them possess a substantial number of connections in the metro area (after all, they are living there as of 2012) while the number of contacts to their previous county of residence is also relatively large (left panel, main diagonal). However, the number of contacts in the other counties is negligible, exactly as in case of those who never migrate. The Russian-Russian ties (right panel) paint a similar picture. There are too few observations for any inference on interethnic ties (not shown).

To summarize, Table 3 strongly suggests that ties form locally. People who never migrate are almost exclusively connected to their current county while migrants have a substantial number of connections to their previous county.

5.4 Evidence on Tie Creation and Destruction

Analyzing differences in tie structure by year of migration allows us to indirectly test the theory of tie formation posited above. Figure 2 shows the relationship between migration year and the geographic structure of the current social network. The figure indicates that the number of contacts in the current county (circles) is lower for the recent migrants while the number

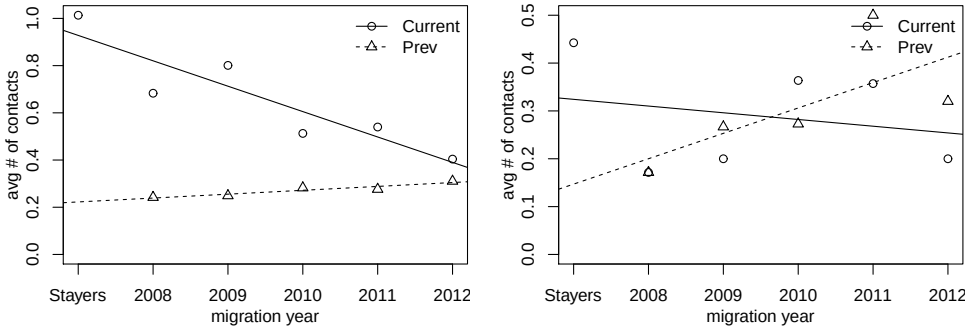
[9] The results for other counties are qualitatively similar.

[10] The numbers are low because we do not observe valid county of residence for the contacts outside of the sample.

Residence 2011	connections to Metro	44	67	78	connections to Metro	44	67	78
	Estonian-Estonian				Russian-Russian			
Metro	0.91	0.01	0.03	0.05	0.45	0.03	0.00	0.00
44	0.35	0.28	0.02	0.02	0.20	0.39	0.00	0.02
67	0.45	0.00	0.32	0.06	0.00	0.00	0.17	0.00
78	0.72	0.00	0.01	0.39	0.14	0.00	0.00	0.29

Notes: Residents of the metropolitan (capital) area 2012 depending on their 2011 residence (in rows), and their number of contacts (degree) in columns. The county codes are 44 = Ida Viru; 67 = Pärnu; 78 = Tartu.

Table 3: Number of contacts in the current residence county, previous residence county, and other counties. Estonian-Estonian and Russian-Russian ties.



(a) Estonian-Estonian

(b) Russian-Russian

Figure 2: Average number of contacts to the current and previous county of residence. The lines represents the corresponding linear fit (with stayers excluded).

of contacts in the previous county (triangles) is decreasing. The figure for Estonian-Estonian ties (left panel) is less noisy and suggests that the ties to the former county are rather resilient and may be related to family or others in-kin (Phithakkitnukoon, Calabrese, Smoreda, and Ratti 2011).

Table 4 gives similar results using individual-level regressions. We estimate the percentage of connections to the current and previous county of residence as a function of years since migration. The table indicates that the share of contacts in the former place of residence falls by about 4 percentage points per year and are replaced by a corresponding growth in connections in the current place of residence.

In summary, our contact distance analysis strongly suggests that ties arise locally over time and also fade away over time when individuals move elsewhere. These outcomes fit to our theoretical framework and suggest that neighborhood population composition is an important determinant of social networks.

Outcome: percentage of contacts in the region		
	Region:	
	Home 2007	Home 2012
Years since migration	-0.04^{***}	0.04^{***}
	(0.01)	(0.01)
Intercept	0.40^{***}	0.44^{***}
	(0.03)	(0.03)
R^2	0.01	0.01
Num. obs.	1203	1203

Notes: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 4: 2007 and 2012 connections percentage, by migrant status and ethnicity

6 Discussion

Our theoretical framework suggests that a number of our findings are compatible with the causality running from neighborhoods to networks. The ethnic composition in residential neighborhoods is related to that in social networks, and the relationship is stronger for stayers and weaker for migrants. In addition, the relationship is similar in Soviet-era high-rise estates, the neighborhoods that were populated in a period with little residential choice. We also show that cellular calls are mostly connecting individuals living in the same commuting area, and if stretching a longer distance, these are likely connections to the former place of residence. All these outcomes suggest that ties typically arise between individuals who are living close to each other. Most importantly, the trade-off between ethnic and geographic distance results in out-group ties, likelihood of which increases along the number of out-groups living in the same neighborhood.

Our central outcome, the positive correlation between network and neighborhood composition can be explained in other ways as well. First, networks may influence neighborhood choice, in particular individuals with certain amount of inter-ethnic contacts may choose to live in a correspondingly mixed environment. However, in this case one expects migrants to be equally sensitive to the neighborhood composition than the non-migrants, and second, one also expects the relationship to be substantially weaker in the Soviet-era neighborhoods. Neither of these predictions is true. Unfortunately, the current cross-sectional network data does not allow to test this hypothesis directly.

Alternatively, both network and neighborhood composition may be jointly determined by a third variable, such as segregation preferences. If this is true, we would not expect the geographic

distance to be a strong determinant of networks, and in addition, we would also expect the relationship to be much weaker in Soviet-era neighborhoods.

7 Conclusion

We use mobile telecommunication data to analyze the relationship between physical segregation and social network network. The data originates from Estonia, a linguistically divided country where the relationship between the corresponding ethnic groups has been characterized by distrust and animosity. These data allow us to compute network homophily and analyze its relationship with the place of residence, separately for people who migrate and those who do not. We document a strong positive relationship between ethnic composition in the residential neighborhood and network homophily. We also show that communication networks are largely local, except for migrants who possess a substantial number of contacts in their previous place of residence. As more time passes since the date of migration, migrants have more contacts in their new location and fewer contacts in their previous location.

A simple theoretical framework suggests that all these outcomes can be explained by neighborhood ethnic composition being an important determinant of social network homophily. While we cannot conclusive rule out alternative explanations, we test several additional specifications that are consistent with this causal relationship.

More speculatively, our results suggest that physical integration might help generate social integration. By contrast, ethnic or racial ghettos may harm social integration, even if such separation might increase the efficiency of local labor markets (Edin, Fredriksson, and Åslund 2003, Damm 2009). Of course, our context, where two well-established ethnic groups share similar socioeconomic and cultural backgrounds, may not be generalizable to contexts where the ethnic groups are separated by much more than just language. Nonetheless, we believe these results provide empirical support for policies designed to promote physical integration.

A Proofs

Expected Homophily The observed communication process does not include the complete network data. A number of links are missing, either because they connect to out-of-sample peers, or because no calls were made during our 10 days of network sampling. Here we show that under independent link sampling, $\mathbb{E}\, h^* = \mathbb{E}[\nu_0^*/(\nu_0^* + \nu_1^*)] = \mathbb{E}\,\nu_0^*/\,\mathbb{E}(\nu_0^* + \nu_1^*)$.

Look at individual i. Assume that in the complete network she has degree $N_i = S_i + D_i$ (communication links to different alters), comprising of S_i same type links and D_i different type links. Assume that the alters are observed independently with probability p. Hence, the observed number of contacts $\tilde{N}_i$ is a random variable and can be expressed as a sum of N_i realizations of independent Bernoulli random variables $A_{ij} \sim \text{Bernoulli}(p)$ where $A_{ij} = 1$ denotes that the alter j of individual i is observed. The observed number of same type alters is $\tilde{S}_i = \sum_{j\in\mathfrak{S}} A_{ij}^S$ and that of different type alters $\tilde{D}_i = \sum_{j\in\mathfrak{D}} A_{ij}^D$ ($\mathfrak{S}$ and $\mathfrak{D}$ denote the set of same and different type friends of i). The true homophily is $H_i = S_i/N_i$ while we observe

$$\tilde{H}_i = \frac{\tilde{S}_i}{\tilde{N}_i} = \frac{\sum_{i\in\mathfrak{S}} A_i^S}{\sum_{i\in\mathfrak{S}} A_i^S + \sum_{i\in\mathfrak{D}} A_i^D}. \tag{5}$$

As A^S and A^D have equal i.i.d distribution, we can apply lemma 3 by Heijmans (1999) and conclude that

$$\mathbb{E}\left[\tilde{H}_i | \tilde{N}_i > 0\right] = \frac{\mathbb{E}\,\tilde{S}_i}{\mathbb{E}\,\tilde{S}_i + \mathbb{E}\,\tilde{D}_i} = \frac{S_i}{S_i + D_i}. \tag{6}$$

Accordingly, we can easily base our inference on the individual homophily on the observed ties in the data.

Propositions

Proposition 1 For stayers, individuals who spend long time in one region, we observe the expected equilibrium number of ties. For group 0:

$$\nu^* = \frac{1}{\delta}\left[\phi^0(\chi^0 n_0^A + \chi^1 n_1^A) + \phi^1(\chi^0 n_0^B + \chi^1 n_1^B)\right] = \tag{7}$$

$$= \frac{1}{\delta}\left\{\phi^0[\chi^0\pi^A + \chi^1(1-\pi^A)]n^A + \phi^1(\chi^0\pi^B + \chi^1(1-\pi^B)]n^B\right\}. \tag{8}$$

This relationship can be used to express the individual network homophily in equilibrium:

$$h^* = \frac{\nu_0^*}{\nu^*} = \frac{\nu_0^*}{\nu_0^* + \nu_1^*} = \frac{\nu_0^{A*} + \nu_0^{B*}}{\nu_0^{A*} + \nu_1^{A*} + \nu_0^{B*} + \nu_1^{B*}} = \\ = \frac{\frac{1}{\delta}\left(\phi^0\chi^0\pi^A n^A + \phi^1\chi^0\pi^B n^B\right)}{\frac{1}{\delta}\left\{\phi^0[\chi^0\pi^A + \chi^1(1-\pi^A)]n^A + \phi^1(\chi^0\pi^B + \chi^1(1-\pi^B)]n^B\right\}} = \\ = \frac{\phi^0\chi^0\pi^A n^A + \phi^1\chi^0\pi^B n^B}{\phi^0[\chi^0\pi^A + \chi^1(1-\pi^A)]n^A + \phi^1[\chi^0\pi^B + \chi^1(1-\pi^B)]n^B}, \quad (9)$$

where ν_0 and ν_1 are ties to group 0 and 1 respectively, ν^A and ν^B are ties to regions A and B, and * denotes the corresponding equilibrium values.

Look at the region A with the local ethnic composition π^A. As the number of ties in region B, ν^{B*}, does not depend on π^A, we have

$$\frac{\partial h^*}{\partial \pi^A} = \frac{\partial}{\partial \pi^A}\left(\frac{\nu_0^{A*} + \nu_0^{B*}}{\nu_0^{A*} + \nu_1^{A*} + \nu_0^{B*} + \nu_1^{B*}}\right) = \\ = \frac{\frac{\partial \nu_0^{A*}}{\partial \pi^A}}{\nu^*} - \frac{\nu_0^*}{(\nu^*)^2}\frac{\partial \nu^A}{\partial \pi^A} = \frac{1}{\delta}\left[\frac{\phi^0}{\nu^*}\chi^0 n^A - \frac{\phi^0}{\nu^*}h(\chi^0 - \chi^1)n^A\right]. \quad (10)$$

The first term in brackets describes the growth of ν_0^* while π^A grows, the second one the corresponding growth of ν^*. As $\chi^0 > \chi^1 > 0$ and $0 \le h \le 1$, the derivative is positive. It increases in the local interaction rate ϕ^0 and in the local population size.

Proposition 2 Look at movers from the region A to B. Assume they initially posses the ties, corresponding to the equilibrium in A. At time $t = 0$ they relocate to B. As ties are neither created nor destroyed instantly, we have at the moment of move $\nu_0^B(0) = \frac{1}{\delta}\phi^1\chi^0\pi^B n^B$ which is not the equilibrium value. Solving the differential equation (2) we find

$$\nu_0^B(t) = \frac{1}{\delta}\phi^0\chi^0\pi^B n^B + \mathrm{e}^{-\delta t}\left[\frac{1}{\delta}\phi^1\chi^0\pi^B n^B - \frac{1}{\delta}\phi^0\chi^0\pi^B n^B\right] = \\ = \frac{1}{\delta}\left[\phi^0 + \mathrm{e}^{-\delta t}(\phi^1 - \phi^0)\right]\chi^0\pi^B n^B \equiv \frac{1}{\delta}\phi_{10}(t)\chi^0\pi^B n^B. \quad (11)$$

Analogous expressions for the other types of ties will be

$$\begin{aligned} \nu_0^A(t) &= \frac{1}{\delta}\phi_{01}(t)\chi^0\pi^A n^A \qquad \nu_1^A(t) = \frac{1}{\delta}\phi_{01}(t)\chi^0(1-\pi^A)n^A \\ \nu_1^B(t) &= \frac{1}{\delta}\phi_{10}(t)\chi^0(1-\pi^B)n^B, \end{aligned} \quad (12)$$

where $\phi_{01}(t) = \phi^1 + \mathrm{e}^{-\delta t}(\phi^0 - \phi^1)$. Note that $\phi^0 > \phi_{01}(t) > \phi^1$ and $\phi^0 > \phi_{10}(t) > \phi^1 \quad \forall t > 0$.

The relationship between migrant's homophily and the new local ethnic composition, $\frac{\partial h(t)}{\partial \pi^B}$, is

$$\frac{\partial h(t)}{\partial \pi^B} = \frac{\partial}{\partial \pi^B}\left(\frac{\nu_0^B(t) + \nu_0^A(t)}{\nu_0^B(t) + \nu_1^B(t) + \nu_0^A(t) + \nu_1^A(t)}\right) = \\ = \frac{\frac{\partial \nu_0^B(t)}{\partial \pi^B}}{\nu(t)} - \frac{h(t)}{\nu(t)}\frac{\partial \nu(t)}{\partial \pi^B} = \frac{1}{\delta}\left[\frac{\phi_{10}(t)}{\nu(t)}\chi^0 n^B - \frac{\phi_{10}(t)}{\nu(t)}h(t)(\chi^0 - \chi^1)n^B\right]. \quad (13)$$

This is positive, by similar argumentation as used for the steady-state equilibrium.

Next, we show that $\phi^0/\nu^* > \phi_{10}(t)/\nu(t)$. It is determined by the sign of

$$\delta \cdot \left[\phi^0 \nu(t) - \phi_{10}(t)\nu^*\right] = \\ = \phi^0\left[\phi_{01}(t)\chi^0 n_0^A + \phi_{01}(t)\chi^1 n_1^A + \phi_{10}(t)\chi^0 n_0^B + \phi_{10}(t)\chi^1 n_1^B\right] - \\ - \phi_{10}(t)\left[\phi^1\chi^0 n_0^A + \phi^1\chi^1 n_1^A + \phi^0\chi^0 n_0^B + \phi^0\chi^1 n_1^B\right] = \\ = \phi_{01}(t)\phi^0(\chi^0 n_0^A + \chi^1 n_1^A) + \phi_{10}(t)\phi^0(\chi^0 n_0^B + \chi^1 n_1^B) - \\ - \phi_{10}(t)\phi^1(\chi^0 n_0^A + \chi^1 n_1^A) - \phi_{10}(t)\phi^0(\chi^0 n_0^B + \chi^1 n_1^B) = \\ = \mathrm{e}^{-\delta t}((\phi^0)^2 - (\phi^1)^2) > 0 \quad (14)$$

where we have used the definition of $\phi_{10}(t)$ and $\phi_{01}(t)$. Accordingly, if homophily levels are comparable, $h(t) = h^*$, then the $\frac{\partial}{\partial \pi^B}h^* > \frac{\partial}{\partial \pi^B}h(t)$ as $\phi^0 > \phi_{10}(t)$. Intuitively, the stayers' networks are primarily determined by the local population composition while the other regions weight more in the movers' networks.

References

Ahas, R., A. Aasa, A. Roose, Ü. Mark, and S. Silm (2008): "Evaluating passive mobile positioning data for tourism surveys: An Estonian case study," *Tourism Management*, 29(3), 469–486.

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.

Collier, P. (1998): "The political economy of ethnicity," in *Annual World Bank Conference on Development Economics*, pp. 387–405.

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

Damm, A. P. (2009): "Ethnic Enclaves and Immigrant Labor Market Outcomes: Quasi-Experimental Evidence," *Journal of Labor Economics*, 27(2), pp. 281–314.

Easterly, W., and R. Levine (1997): "Africa's growth tragedy: policies and ethnic divisions," *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.

European Commission (2002): "Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications)," *Official Journal of European Communities*, L201, 37–47.

Everly, R. (1997): "Ethnic assimilation or ethnic diversity? Integration and Estonia's citizenship law," in *The Integration of non-Estonians into Estonian Society: History, problems and trends*, ed. by A. Kirch, pp. 106–121. Estonian Academy Publishers, Tallinn.

Heidmets, M. (1998): "The Russian Minority: Dilemmas for Estonia," *Trames*, 3(2), 264–72.

Heijmans, R. (1999): "When does the expectation of a ratio equal the ratio of expectations?," *Statistical Papers*, 40(1), 107–115.

Hess, D. B., T. Tammaru, and K. Leetmaa (2012): "Ethnic differences in housing in post-Soviet Estonia," *Cities*, 29(5), 327–333.

Kalmus, V., and M. Pavelson (2002): "Schools in Estonia as Institutional Actors and as a Field of Socialisation," in *The Challenge of the Russian Minority: Emerging Multicultural*

Democracy in Estonia, ed. by M. Lauristin, and M. Heidmets, pp. 227 – 236. Tartu University Press, Tartu.

Katus, K. (1990): "Demographic trends in Estonia throughout the centuries," *Yearbook of Population Research in Finland*, 28, 50–66.

Kulu, H., and T. Tammaru (2004): "Diverging views on integration in Estonia, determinants of Estonian language skills among ethnic minorities," *Journal of Baltic Studies*, 35, 378–401.

Kährik, A., and T. Tammaru (2010): "Population composition in new suburban settlements of the Tallinn metropolita area," *Urban Studies*, 45, 1055–1078.

Lauristin, M., M. Uus, and K. Seppel (2011): "Kodakondsus, kodanikuühiskond ja rahvussuhted (Citizenship, civil society and ethnic relations)," in *Integratsioonimonitor 2011*, pp. 9–50. Kultuuriministeerium, Tallinn, EE.

Leping, K.-O., and O. Toomet (2008): "Emerging ethnic wage gap: Estonia during political and economic transition," *Journal of Comparative Economics*, 36(4), 599–619.

Massey, D. S., and N. A. Denton (1988): "The Dimensions of Residential Segregation," *Social Forces*, 67(2), pp. 281–315.

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

Pettai, I. (2002): "Mutual tolerance of Estonians and non-Estonians (in Estonian)," in *Estonia and Estonians in comparative perspective (in Estonian)*, pp. 213–233. Tartu University Press.

Phithakkitnukoon, S., F. Calabrese, Z. Smoreda, and C. Ratti (2011): "Out of Sight Out of Mind–How Our Mobile Social Network Changes during Migration," in *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pp. 515 –520.

Schultze, J. (2011): "Contact and crisis in interethnic relations," in *The Russian Second Generation in Tallinn and Kohtla-Järve: The TIES Study in Estonia*, ed. by R. Vetik, and J. Helemäe, pp. 165–182. Amsterdam University Press, Amsterdam, NL.

Tammaru, T. (2001): "Suburban growth and suburbanisation under central planning: The case of Soviet Estonia," *Urban Studies*, 38(8), 1341–1357.

Tammaru, T., and H. Kulu (2003): "The ethnic minorities of Estonia: changing size, location, and composition," *Eurasian Geography and Economics*, 44(2), 105–120.

Vihalemm, T. (2010): "To learn or not to learn? Dilemmas of linguistic integration of Russians in Estonia," *Russian Minorities in the Baltic States*, 2, 74–98.

Untangling the effects of residential segregation on individual mobility

Suma Desu[1], Lauren Alexander[2], Marta González[1,2]

[1]Center for Computational Engineering, Massachusetts Institute of Technology, 77 Massachusetts Ave., Cambridge, MA 02139, USA,

[2]Department of Civil Engineering, , Massachusetts Institute of Technology, 77 Massachusetts Ave., Cambridge, MA 02139, USA,

January 30, 2015

1 * Introduction

Physicists and sociologists have longstanding interests in identifying the mechanisms by which individual dynamics lead to collective outcomes. In the social domain, tipping point or threshold models provide one useful framework for connecting the actions of individuals to population processes [1, 2]. In the physics domain mobility models can account for individual preferences that result in population level commuting fluxes at multiple scales [3, 4]. Behavior models, such as models of social interaction, have demonstrated great potential for understanding the dynamics of both residential mobility and segregation by race and ethnicity. In his work [1] , Thomas Schelling laid the conceptual groundwork for understanding the relationship between individual preferences and the evolution of neighborhood compositions. He demonstrated that minute racial preferences in individual residential choices could result in aggregate patterns of residential segregation over time.

Residential choices are often the result of careful planning and consideration, thus the importance of individual preference in deciphering patterns of segregation is well established. On the other hand, daily mobility choices, such as choosing a grocery store, are often ephemeral. Although these daily decisions might seem random, individual mobility data has revealed to be highly non-random, governed by simple laws, and greatly predictable [3, 5, 6]. Accordingly we can use individual trajectories as a proxy to understand the environments to which individuals are exposed.

In this work, we use call detail records (CDRs) generated over multiple months, which are treated and validated, to create average daily origin destination (OD) networks. We use these networks to study the difference in daily mobility choices arising from residentially segregated areas.This represents the first data driven study to investigate whether racial preferences can be discerned from individual mobility choices. If so, results would indicate that not only do we organize residentially by our socio-economic demographics, but also move according to these as well. With the rapid migration into urban areas, understanding how social decisions influence mobility has consequences in many domains such as epidemiology and urban planning. But perhaps more interestingly, this work also servers as a platform to link individual mobility data with social theories that posit racial isolation and concentrated disadvantage heighten exposure to criminality and reduce access to resources and opportunities [7, 8].

2 Methods

Our CDR data set contains more than 8 billion anonymized mobile phone records, obtained from several phone providers, covering 2 million users over two months in the spring of 2010. For each record we are given the following information: an anonymized user identification number, the latitude and longitude of the record, and finally the timestamp at the moment of phone activity, these activities encompass calls, text messing, and web browsing. Typical cell record datasets are given with respect to cell towers, in our case, the provider estimates the location of each record using a triangulation scheme, resulting in location accuracy of 200-300m. Using the stay-point extraction method detailed in [9], we identify 'stay locations' which are the locations where users engage in some activity. Then using the same methodology developed in [10] we assign labels to the stay regions, thereby generating trips by purpose. The stay process yields a timestamp and duration for each stay location, using these and frequency of visits we are able to assign an activity type of 'home', 'work', or 'other' to each stay location for each user. To capture only those users whose home location is adequately represented in the CDR data, we filter out users with less than 8 records. We further filter census tracts that do not have enough residents identified from the CDR data. Next we upscale user trips to represent population level travel patterns by calculating expansion factors for each tract as the ratio of resi-

*corresponding author: sdesu@mit.edu

dents as identified in the CDR data and the census population. We validate these ODs against the Massachusetts Travel Survey at various scales of aggregation; in 1 we show that the accuracy of our data improves at sparser scales.

Residential segregation metrics are also particularly sensitive to which scale is used. The reason is segregation measures using census tract data tend to report higher values for cities with high population densities, where census tracts are smaller and cover only one neighborhood, as compared to smaller, sparser cities where census tracts can cover many neighborhoods. This bias is reduced at smaller levels of spatial aggregation. Consequently, it is important to be careful when comparing the level of residential segregation over larger spatial areas. Moreover, it is advisable to carry out a multi-scale analysis. In [11] a methodology to compute residential segregation on various scales is presented but it has not been widely used in the subsequent literature. Here we creating a segregation profile to measure how scale affects segregation in our study region. To ensure adequate correlation and small enough clusters to capture variation in racial composition we choose to use 350 k-means clusters as the local environments of our users.

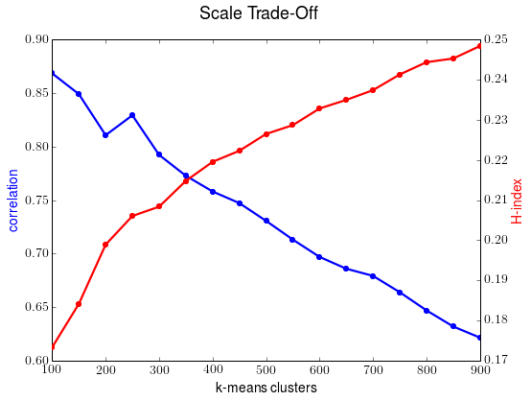


Figure 1: the correlation increases with higher levels of aggregation but the H-index decreases

Using these 350 clusters we construct OD-networks to measure aggregate patterns arising from residentially segregated areas. Since our goal is quantify the effects of residential segregation we assign all users the dominant race of their home tract. Then we construct the race networks 2 as follows: each node represents one particular cluster, and edges are present between nodes if a trip has occurred between the two clusters. The weights on the edges t_{ij}^{r} are the sum of daily flows from $i \rightarrow j$, hence the network is directed. We focus on average daily home other networks because we believe these trips represent individual choices and preserve individuals' residences.

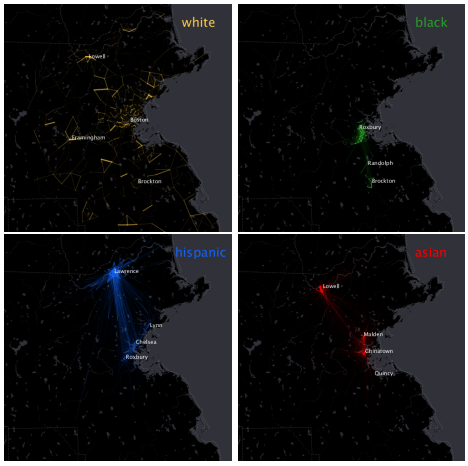


Figure 2: Home Other Networks, we will use this color scheme throughout

Using the static racial composition in 3 as a reference, we see in 2 some long range connections in minority networks are between other areas that have higher concentrations of the same respective minority.

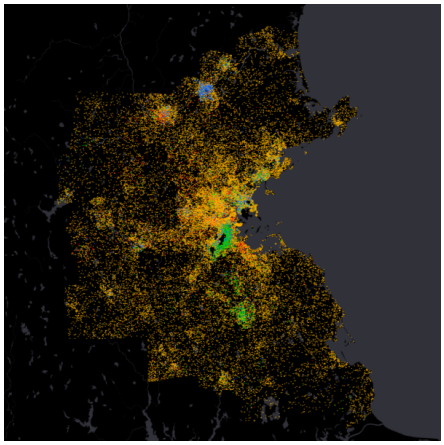


Figure 3: Every dot represents 100 people of a particular race. For each census tract, dots are generated and labeled with races according to their respective racial population distribution in the census. For the visualization dots are uniformly distributed within the borders of corresponding census tracts and colored by the race they represent. We are left with a visual approximate of the racial composition of our study area. This map shows our study region is predominately white and fairly segregation.

3 Metrics

We reformulate traditional measures of segregation, namely entropy and exposure to include dynamic properties of individual mobility. These measures are often cited as proxies for understanding the local environments of minorities, but a more accurate understanding would incorporate data on where these individuals go throughout the day. Sociologists use entropy to measure diversity of racial distributions in locations and this value is the location's contribution to the information-theory based H-index. Instead of measuring the static distributions of locations, we measure the diversity of visits to each location formulated as follows:

Let L be the set of all locations. Firstly, for each location $i \in L$: we compute s_i^r, the incoming strength in the OD-network defined by race r. Next, we normalize by total incoming strength in r's network , s_{total}^r, this gives us the overall importance of location i in each race network. Because we have many more white user's than minorities normalization ensures we weight each network equally. Finally we compute the entropy, E_i , of these normalized values.

$$s_i^r = \sum_{j \in L} t_{ji}^r \tag{1}$$

$$E_i = \sum_{r \in R} \frac{s_i^r}{s_{total}^r} \log_4 \frac{s_i^r}{s_{total}^r} \tag{2}$$

In 4 Downtown Boston and South Cambridge are shown to be important locations in all race networks but there is not much diversity elsewhere.

Exposure is traditionally a measure used to capture the average percentage of race m present in the local environments of race n, denoted ${}_nP_m$. Here we measure the exposure of individual's destinations as a function of distance. To measure the distance between location i and location j we use the haversine function, $h(i,j)$. Most trips in our OD networks occur between 0 - 5 km with very few above 25km, so we measure exposure in the following intervals $d = [0,1,5,10,25,\infty]$, shown in the left hand side of .

$$s_i^r = \sum_{j \in L} t_{ij}^r \qquad \text{if } d_k \leq h(i,j) < d_{k+1} \tag{3}$$

$${}_nP_m = \sum_{i \in L} \frac{s_i^n}{s_{total}^n} * \pi_{im} \tag{4}$$

Threshold models used in segregation literature often cite that there is tolerance parameter governing residential movement. Inspired by this, we create probability of visitation functions with respect to the racial composition of local environments. What we see in the right hand side of seems intuitive, each race has a higher probability of visiting a cluster that has a larger proportion of their own race.

4 Model

From 5 we see a stark pattern emerge: even during daily travel races are likely to visit locations that are dominated by their own race. To untangle whether these are effects of individual preference or simply spatial segregation we test two models. Both are rank based variants of the radiation model in [4]. In both models we predict t_{ij}^r using the outgoing strength s_i^r as production. In the race blind model we use total population as attraction and in the race aware model we only use the population of race r to predict t_{ij}^r. These models allow us to account for the distance decay in trip lengths in addition to discerning if individual's mobility is race blind or race aware.

References

[1] Thomas C Schelling. Dynamic models of segregation. *Journal of mathematical sociology*, 1(2):143–186, 1971.

[2] Mark Granovetter and Roland Soong. Threshold models of diffusion and collective behavior. *Journal of Mathematical sociology*, 9(3):165–179, 1983.

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

[4] Filippo Simini, Marta C González, Amos Maritan, and Albert-László Barabási. A universal model for mobility and migration patterns. *Nature*, 484(7392):96–100, 2012.

[5] Christian M Schneider, Vitaly Belik, Thomas Couronné, Zbigniew Smoreda, and Marta C González. Unravelling daily human mobility motifs. *Journal of The Royal Society Interface*, 10(84):20130246, 2013.

[6] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

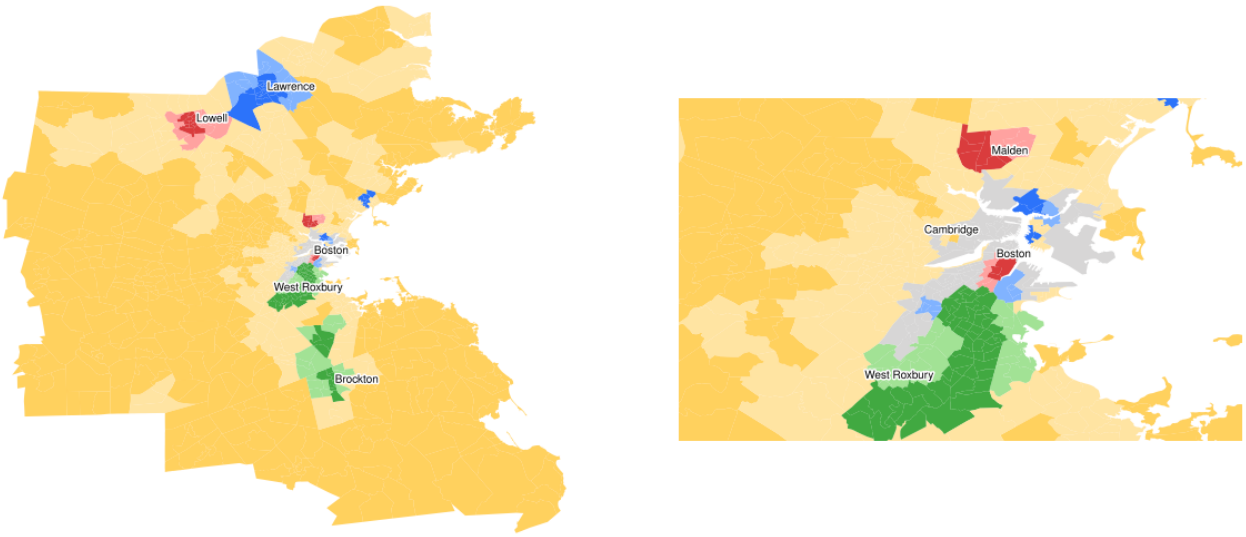


Figure 4: Location Visitation Diversity. Darker colors represent lower diversity in visitation ex. dark green means that location is mostly important in the black network, light green means mostly important in the black network but some diversity in visitation. Grey represents high diversity of visitations.

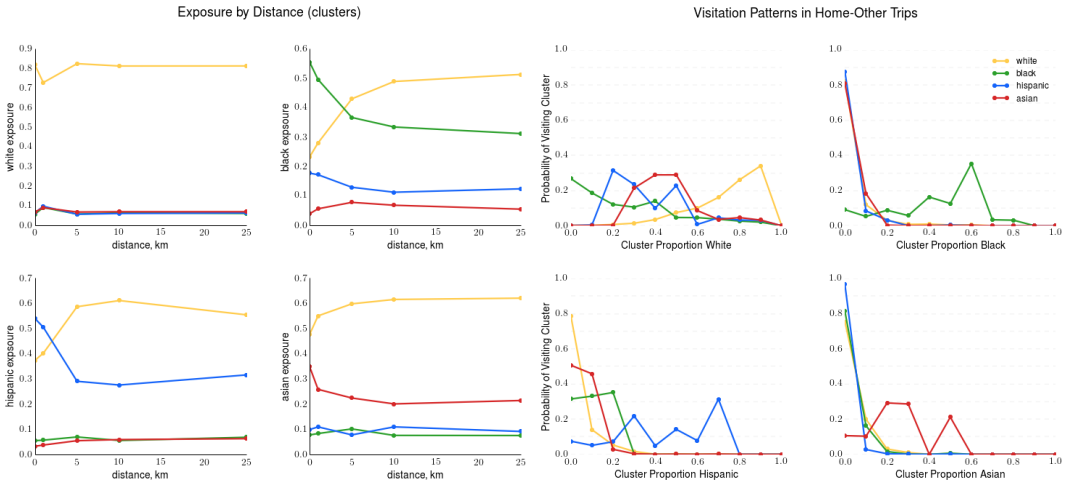


Figure 5: Left: each subplot represents the exposure for one race, in the top left we measure white exposure, the yellow line represents the exposure of white to white and the black line represents the exposure of white to black. etc. For all minorities as distance increases exposure to white increases and exposure to their own race decreases but remains higher than exposure to any other minority group Right: utility functions for each race

[7] Robert J Sampson, Jeffrey D Morenoff, and Thomas Gannon-Rowley. Assessing" neighborhood effects": Social processes and new directions in research. *Annual review of sociology*, pages 443–478, 2002.

[8] Robert D Dietz. The estimation of neighborhood effects in the social sciences: An interdisciplinary approach. *Social Science Research*, 31(4):539–575, 2002.

[9] Shan Jiang, Gaston A Fiore, Yingxiang Yang, Joseph Ferreira Jr, Emilio Frazzoli, and Marta C González. A review of urban computing for mobile phone traces: current methods, challenges and opportunities. In *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*, page 2. ACM, 2013.

[10] Lauren Alexander, Shan Jiang, Mikel Murga, and Marta C. González. Validation of origin-destination trips by purpose and time of day inferred from mobile phone data. *submitted to Transportation Research Part C*, 2014.

[11] Sean F Reardon and David OSullivan. Measures of spatial segregation. *Sociological methodology*, 34(1):121–162, 2004.

Spatial and Social Homophily at a Massive Religious Gathering

Ian Barnett[1]
Tarun Khanna[2]
Jukka-Pekka Onnela[1]

The Kumbh is a religious Hindu festival that has been celebrated for centuries. The 2013 Kumbh Mela, a grander form of the annual Kumbh, had an estimated 40 to 100 million visitors (Sugden, 2013). These estimates, if correct, would make it the largest gathering of people in human history. We used call detail records (CDRs) from one of India's largest cellular service providers to study the spatial and social dynamics of Kumbh attendants. Here, we focused our investigation on homophily, the common finding that individuals who frequently interact with one another tend to be similar in their attributes. What separates our investigation from previous studies in this area is that we examine spatial and social components of homophily simultaneously and, further, do so in the context of a massive gathering of millions of people.

Our data consisted of 390 million communication events among a sample of approximately 4 million people whose communication activities were observed for a three-month period from January 1 to March 31, 2013, a time period that encompassed the 2013 Kumbh Mela. The festival took place within a designated venue that was served by approximately 200 cell towers, which divided the Kumbh area into 200 small spatial regions (technically known as a Voronoi tessellation where each Voronoi cell is centered on a cell tower). We examined communication events that were transmitted by any one of these towers with known spatial locations. Past investigations of homophily have focused on individual covariates, such as age, sex, race, and ethnicity. India is a culturally diverse nation with many languages and customs that vary widely across its many states, and here we investigated individual-level homophily in terms of callers' regions, available as their phone area codes. We defined spatial homophily as a tendency for individuals from a given region of India to be spatially co-located, while at the Kumbh, with other individuals from the same region over and above that expected by chance alone. Similarly, we defined network homophily as a statistical excess of social ties, while at the Kumbh, here measured as phone calls and text messages, between people who shared an area code.

We compared how spatially and socially homophilous the different regions of India were with respect to one another. Overall, we found many of the regions to exhibit relatively strong levels of both types of homophily. Further, we found that states with smaller representation at the Kumbh tended to show significantly stronger spatial and social homophily than the states with greater Kumbh representation (Figure 1). In other words, smaller groups at the Kumbh appeared to display greater cliquishness, both spatially and socially. This phenomenon of minority groups being more homophilous than larger ones has been observed in studies of race and social structure in school settings (Gonzalez, 2007; Vermeij, 2009; Currarini, 2009), but these earlier findings tend to be based on relatively small interaction contexts that are long-lasting. In contrast, the Kumbh is a massive event in which most attendees stay a relatively short period of time, seldom more than a couple of weeks. In this very different type of behavioral context we find that homophilous tendencies persist.

Our findings could be helpful in understanding the dynamics of organic or unplanned displacements of populations. It would be helpful to be able to identify the social factors that lead people to seek out others like them in these types of settings, because these behavioral tendencies can lead to unpredictable and devastating outcomes, such as human stampedes. More broadly, we anticipate that the present work could be used in the context of disasters, like earthquakes and floods, when large numbers of people are forced out of their homes.

[1] Department of Biostatistics, Harvard School of Public Health, Boston, MA
[2] Graduate School of Business Administration, Harvard University, Boston, MA

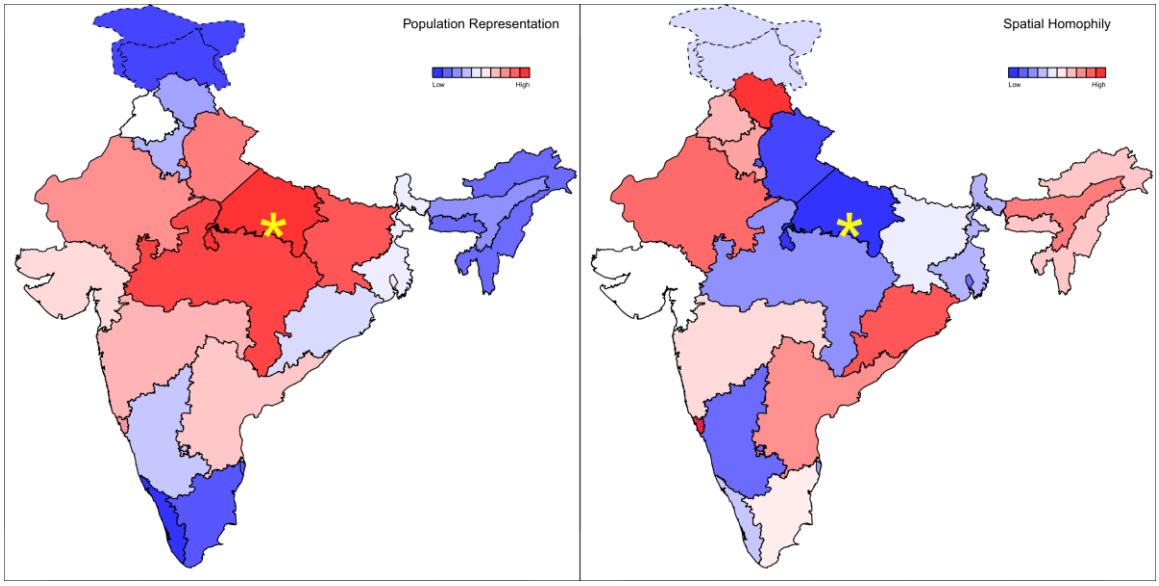


Figure 1: We have divided India into 23 different states, ranked first by representation at the Kumbh (left panel) and then by degree of spatial homophily (right panel). The heat map colors correspond to the rankings. The yellow star is the city of Allahabad, the location of the 2013 Kumbh Mela. The near inversion of colors when comparing the two panels demonstrates a clear negative association between state representation and spatial homophily.

References

Currarini, S. a. (2009). An economic model of friendship: Homophily, minorities, and segregation. *Econometrica, 77*(4), 1003--1045.

Gonzalez, M. a. (2007). Community structure and ethnic preferences in school friendship networks. *Physica A: Statistical mechanics and its applications, 379*(1), 307--316.

Sugden, J. (2013). How the Kumbh Mela Crowds Are Counted. *Wall Street Journal*.

Vermeij, L. a. (2009). Ethnic segregation in context: Social discrimination among native Dutch pupils and their ethnic minority classmates. *Social Networks, 31*(4), 230--239.

Predicting Gender from Mobile Phone Metadata

Eaman Jahani[1], Pål Roe Sundsøy[2], Johannes Bjelland[2], Asif Iqbal[2], Alex Pentland[1] and Yves-Alexandre de Montjoye[1]

[1]*MIT Media Lab, Cambridge, USA*
[2]*Telenor Group, Oslo, Norway*
Email: eaman,pentland,yva@mit.edu
pal-row.sundsoy,johannes.bjelland,asif.iqbal@telenor.com

Keywords: Metadata, Gender Prediction, CDR, Development, Behavioral Indicators

Over the past decade, international organizations, research institutes and governments in rich countries have heavily invested in data collection for evaluating their performance in development areas such as child mortality, spread of diseases and poverty. Novel and scalable analysis techniques have been specifically introduced for the better usage of big data in policy making. As rich countries collect and analyze more and more data from a broad range of applications, the poor and developing countries have lagged behind. Many countries in the developing world have not conducted a census in many years due to infrastructure and resource limitations and market inefficiencies [1]. For example, Afghanistan is still using census data from 1979. The lack of reliable government statistics and census data hinders development and evaluation of policies in countries which need it the most. Given the vital role of census and demographic data in measuring social and economic development, their shortage needs to be addressed.

Researchers have noted that the emergence of the data-driven "computation social science" and the large scale adoption of new technologies offer an unprecedented opportunity to better understand individual behavior and societies at scale [2]. The rise of communication technologies, and in particular the high penetration rate of mobile phones across the globe (90% in the developing countries), has the potential to address the scarcity of development data. While many economic, demographic and population data are missing in developing countries, the abundance of metadata collected by mobile phone operators can fill this gap, thus enable a more informed policy making process [3].

Many international organization, such as UN and world bank, have already started tapping into the great potential of big data for development. A report recently published by the UN data revolution group calls for greater embedding of ubiquitous data collection and analysis into its 2015 sustainable development action plan [4] . In addition, the big data and people project alliance (Data-Pop) has focused on developing methods for collecting and analyzing indicators of human welfare, such as poverty, inequality, human mobility, economic statistics and inflation patterns, mostly in underdeveloped and conflict regions [5]. Recently, a series of research studies have investigated the applicability, accuracy and cost-effectiveness of geographical and temporal aggregation of communication logs in novel applications such as census prediction. Deville showed how mobile phone metadata can be used to generate detailed maps of population distribution on a country-wide scale [6]. Using mobility and communication patterns obtained from mobile phone metadata, Enns constructed a geographic network of regions with low and high malaria transmission and proposed a more cost-effective approach to the malaria control effort [7]. Similarly, researchers have utilized mobile phone mobility and Call Detail Records (CDR) for mapping and containing the spread of Ebola [8].

There are two main challenges with the application of call and mobility data for development purposes. First, while the CDRs in their raw form provide us with precise information on who called whom from where and for how long, they provide very little insights about the individual phone users and their habits. This is especially important in development policy-making which heavily relies on temporal and geographic aggregates of demographic and behavioral data. Second, the lack of subscribers demographic information is even more pronounced in developing countries where this data has the most potential but the vast majority of mobile phones are prepaid. Here, we address these challenges by extracting insightful behavioral indicators from the raw CDRs which can be used for prediction of individual demographic characteristics. Our behavioral indicators range from

radius of gyration to average response delay or the diversity of contacts.

The motivation of this study is to investigate the extent to which our behavioral indicators can replace the expensive and missing census data. Gender is a basic, yet an essential, demographic variable in census data. In this paper, we focus on how behavioral indicators derived from standard mobile phone meta data, CDRs, can be used as input features to machine learning techniques for predicting the gender of the mobile phone user. By combining large-scale CDR meta data with the mobile phone subscriber information, we evaluated the prediction power of these behavioral indicators across two different countries, developed and developing.

Several recent studies have investigated the application of mobile phone usage data to infer the demographic structure of a population. Different variables including age, gender, personality traits and socioeconomic status have been the focus of these studies. The prediction approaches range from network-based collaborative filtering to standard machine learning algorithms. Herrera-Yagüe introduced a method which exploits the homophily in the network connections and uses the available demographic information of a user's contacts along with her SMS, call usage pattern and ego-network structure as the input features to several machine learning algorithms [9]. Brea proposed a graph-based algorithm for predicting the age of mobile phone users. Given a seed of known user ages, the algorithm extracts topological information from the network to infer the user's age group. Brea also discovered that the topology of the seed set relative to the prediction node plays an important role in performance of the algorithm [10]. These methods require the whole communication network around each user along with the information about her neighbors to be available before making any prediction. In contrast to these approaches, our method solely relies on the activities of individual mobile phone users for prediction.

Sarraute employed several individual-based "behavioral variables" and network-based "social variables" as input features to a double-layered learning algorithm called "Population Pyramid Scaling". The algorithm admits a hyper parameter that determines the proportion of nodes to predict, thus adjusting the trade-off between precision and recall [11]. The nature of these "behavioral" and the "social" variables extracted from mobile phone usage patterns is similar to ours, with the difference that we greatly expanded the set of input features to cover new behavioral categories. In contrast to Sarraute's work, we did not limit our testing to the more predictable users and evaluated our model on all of the test data.

We developed a range of novel indicators capturing different aspects of individual habits and behaviors. The complete set of behavioral indicators consisted of 94 features that are extracted from texting, calling and mobility patterns of individual users. We developed a software package which can easily compute these variables from raw carrier logs. We believe these indicators can account for a diverse set of behavioral patterns by only considering the activities of individual users, and not of their neighbors. The indicators can be roughly divided into five different categories:

1. Basic phone usage: Some examples are number of calls, number of texts or number of active days per week.
2. Active user behavior: These indicators aim to capture user's level of proactiveness. Some examples are the percentage of call or text interactions initiated by the user or the delay time in answering a text. We consider a text from user A to user B as a response if it is sent within one hour of receiving a text from user B. The response delay is the expected time it takes for a user to answer a text. Text response rate is another example of user active behavior.
3. Diversity: Entropy or diversity is a quantitative measure capturing how many different values a random variable can take and how evenly they are distributed across the range of all possible values. For example, as a user interacts more evenly with a larger number of contacts, her entropy of contacts increases. In this work, we considered the entropy of call, text, call and text contacts and the unique places a user visits.
4. Regularity: Some examples are weekly deviations in temporal calling pattern, call and text inter-time or the times a user comes back home.
5. Location: These features aim to capture user's mobility pattern. Some examples are the absolute number of different places a user has sent or received a call or text from or the distance between these places. Another example is the mean radius of gyration which is the radius of the smallest circle that contains all the places a user has visited in a day.

Our data set consists of a random sample of more than 550,000 subscribers in a European country and 40,000 subscribers in a South Asian country, whose raw CDRs were collected for a period of 3 months. Longer periods of user data result in more accurate representation of behavioral indicators since the features are aggregated weekly. The dataset is anonymized by encrypting the phone numbers of the callee and the caller in the raw CDRs. An explorative study of the features reveals stark differences and yet remarkable similarities between the two countries with different culture and economic status. For example, figure 1 illustrates the kernel density of mean duration of calls across these two countries along with the same distribution per gender in each country. While the average duration of calls in the European country is 200 minutes longer than the South Asian country, women in

Algorithm	Best Configuration	European Accuracy	South Asian Accuracy
Linear SVM	Penalty = L2, Loss = L2, Cost = 10	71.9%	73 %
SVM with Polynomial Kernel	Degree = 2, Cost = 10 Kernel Coefficient = 0.01	74%	74.9%
SVM with RBF Kernel	Cost = 100, Kernel Coefficient = 0.1	74.1%	74.8%
Random Forest	Split criterion = Entropy, Min samples for split = 5, Features searched for best split = 50%, Number of trees = 400	73.2%	78%

TABLE 1: Gender Prediction Accuracy with Best Classifier Configuration

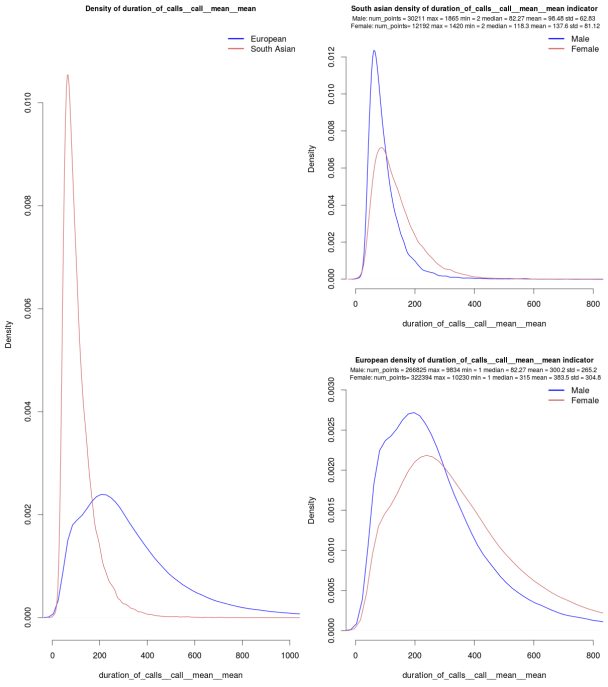


FIGURE 1: The mean duration of calls per country (left), mean duration of calls per gender in the south Asian country (top right) and mean duration of calls per gender in the European country (bottom right)

Herra-Yagüe [9]	Sarraute [11]	Our European Data
65.3%	66.3%	74.1%

TABLE 2: Comparison of Previous Gender Prediction Accuracies with our Worst Performance

both countries are expected to have approximately 30% longer call duration compared to men. This similarity is in fact confirmed by our final machine learning model where mean duration of calls appears as one of the strongest predictors of gender in both countries.

We tested the prediction power of our behavioral indicators on users who were active on at least 3 days of the week and were able to correctly predict gender with 74.1% accuracy in the *worst-case scenario*. We also observed that the strongest predictors of gender are mostly the same across these two countries, despite cultural, social, geographical and economic differences. In particular, we observed that diversity of contacts and places visited is strongly correlated with gender in both of these countries. We examined SVM with polynomial and radial basis function kernels, linear SVM and random forest and concluded that SVM with RBF kernel and random forest yield the highest level of accuracy. To achieve equal precision and recall for both genders in presence of data imbalance, we modified the penalty term for each gender inversely proportional to its relative frequency in the training set. Furthermore, our input features were standardized to the [0,1] interval in the case of SVMs. Table 1 summarizes the accuracy results from various learning models we tested. Our models were tuned through a grid search of stratified 5-fold cross validation scenarios. The accuracy numbers reported in table 1 are from unseen test data held separate from the cross validation set.

Table 2 compares our worst case accuracy (from the European data set) with equivalent accuracy metrics from [9, 11]. Herra-Yagüe [9] accuracy is taken from the best model which uses both isolated link and ego-network features. The accuracy from Sarraute [11] is from the case where a prediction is generated for all nodes (i.e. $q = 1$). We attribute the substantial improvement in our prediction accuracy to the expanded set of learning features. In fact, a close examination of the learned models reveals that some of the most predictive features in both countries are ones that are missing from these previous studies (e.g. interactions per contact, entropy of contacts and places, percent initiated interactions).

These results show that behavioral indicators directly derived from mobile phone metadata can be combined with standard machine learning techniques for accurate and robust prediction of gender. Further work is required to investigate the applicability of these indicators in prediction of other demographic variables such as age and socio-economic status. We believe that combining our individual behavioral features with information obtained from the ego-networks (such as neighbor similarity) can result in even higher accuracy.

REFERENCES

[1] Jerven, M. (2013) *Poor Numbers: How We Are Misled by African Development Statistics and What to Do about It.* Cornell University Press.

[2] Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabsi, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., and Van Alstyne, M. (2009) Computational social science. *Science*, **323**, 721–723.

[3] de Montjoye, Y.-A., Kendall, J., and Kerry, C. (2014) Enabling humanitarian use of mobile phone data. *Brookings, Issues in Technology Innovation*, **26**.

[4] Data Revolution Group (2014) *A World that Counts. Mobilising the Data Revolution for Sustainable Development.* Prepared by United Nations Independent Expert Advisory Group on Data Revolution.

[5] (2015). Data-pop alliance. www.datapopalliance.org. Accessed: 01-20-2015.

[6] Deville, P., Linard, C., Martin, S., Gilbert, M., Stevens, F. R., Gaughan, A. E., Blondel, V. D., and Tatem, A. J. (2014) Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences*, **111**, 15888–15893.

[7] Enns, E. A. and Amuasi, J. H. (2013) Human mobility and communication patterns in côte d'ivoire: A network perspective for malaria control. *Mobile Phone Data for Development: Analysis of Mobile Phone Datasets for the Development of Ivory Coast*, **1**.

[8] Wesolowski, A., Buckee, C., Bengtsson, L., Wetter, E., Lu, X., and Tatem, A. (2014) Commentary: Containing the ebola outbreak the potential and challenge of mobile network data. *PLOS Currents Outbreaks*, **1**.

[9] Herrera-Yagüe, C. and Zufiria, P. J. (2012) Prediction of telephone user attributes based on network neighborhood information. *Proceedings of the 8th International Conference on Machine Learning and Data Mining in Pattern Recognition*, Berlin, Heidelberg MLDM'12, pp. 645–659. Springer-Verlag.

[10] Brea, J., Burroni, J., Minnoni, M., and Sarraute, C. (2014) Harnessing mobile phone social network topology to infer users demographic attributes. *Proceedings of the 8th Workshop on Social Network Mining and Analysis*, New York, NY, USA SNAKDD'14, pp. 1:1–1:9. ACM.

[11] Sarraute, C., Blanc, P., and Burroni, J. (2014) A study of age and gender seen through mobile phone usage patterns in mexico. *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*, Aug, pp. 836–843.

Session 7 :: Crowds (II)

SPATIOTEMPORAL DETECTION OF UNUSUAL HUMAN POPULATION BEHAVIOR USING MOBILE PHONE DATA

ADRIAN DOBRA, NATHALIE E. WILLIAMS, AND NATHAN EAGLE

Abstract. With the aim to contribute to humanitarian response to disasters and violent events, scientists have proposed the development of analytical tools that could identify emergency events in real-time, using mobile phone data. The assumption is that dramatic and discrete changes in behavior, measured with mobile phone data, will indicate extreme events. In this study, we propose an efficient system for spatiotemporal detection of behavioral anomalies from mobile phone data and compare sites with behavioral anomalies to an extensive database of emergency and non-emergency events in Rwanda. Our methodology successfully captures anomalous behavioral patterns associated with a broad range of events, from religious and official holidays to earthquakes, floods, violence against civilians and protests. Our results suggest that human behavioral responses to extreme events are complex and multi-dimensional, including extreme increases and decreases in both calling and movement behaviors. We also find significant temporal and spatial variance in responses to extreme events. Our behavioral anomaly detection system and extensive discussion of results are a significant contribution to the long-term project of creating an effective real-time event detection system with mobile phone data and we discuss the implications of our findings for future research to this end. The complete version of this paper will appear in PLoS ONE and it is available online [1].
KEYWORDS: Big data, call detail record, emergency events, human mobility

1. Introduction

Discrete emergency events, such as terrorist attacks and natural disasters, occur frequently around the globe and regularly cause massive destruction. With the ultimate aim to decrease the humanitarian toll of post-event disasters, scientists have recently begun to understand that several relatively new sources of organically collected data, such as cell phone records, internet blogs, and Twitter, could provide real time or very quick identification of emergency events. While these data are continuously collected by service providers and could ostensibly be made available, the tools for using such data for real-time event identification are still under construction. The broad purpose of this article is to contribute to the long-term goal of development of analytical tools for using mobile phone data to identify emergency events in real time. This can ultimately contribute to quicker humanitarian response and decreases in the severity of disasters. Specifically, we create a system for identifying anomalies in human behavior as manifested in mobile phone data, and discuss the correspondence between these anomalies and actual emergency and non-emergency events that might have caused them.

Previous research has demonstrated that such analytical tools might be possible, by showing that natural and man-made emergency events, such as earthquakes or bombings, can be seen in dramatic increases in calling and mobility behaviors. These studies are able to show such anomalies by comparing behaviors to events where the character, time, and place of these events are already known. We build on these studies, but develop a blind system that is closer in nature to an actual event detection system. Instead of starting with the time and location of an event, then looking for anomalous calling behavior, we develop a behavioral anomaly detection system that identifies days with unusual calling or mobility behavior, as well as the location and geographic extent of these disruptions. Our detection system is scalable as it is able to efficiently process years of country-wide mobile phone records.

For illustration we use mobile phone records from a single cellular services provider from Rwanda. We connect the identified anomalous days and locations with extensive records of violent and political events and natural disasters. Results of this exercise reveal that some days

with anomalous increases in calling and mobility behavior match well with several different kinds of events. In other cases, days with *decreases* in calling and/or mobility match with events. These cases were surprisingly more numerous than events matched with increases in calling and mobility. In still other cases, we do not find good event matches for days with anomalous behavior and we also find cases where emergency events occurred without resulting in anomalous behavior that our system could detect. Notably, we learn as much from the unmatched events and behavioral anomalies as from the matched cases.

We argue that further quantitative and qualitative research into the exact and possibly multi-dimensional nature of human response to emergency events is needed. In this regard, our careful analysis of both the matched events and the events and instances of anomalous behavior that do not match reveal some key insights into further developments needed to better understand human response to emergency events. In fact, it is this outcome, namely the demonstration that human behavioral responses to emergency events are much more complex than previously assumed, that is the most important contribution of this paper. Future research must address this complexity and can benefit from using existing social and psychological theories of behavioral response to threat. We conclude this article by setting out a clear pathway of research aimed at the goal of a creating an effective system of identifying emergency events in real-time (or close to real time) from mobile phone data.

2. Results

Our anomalous behavior detection system identified many days with unusual calling and movement behavior across multiple sites. Here we describe some of these disturbances in which the largest spatial clusters of sites were identified, match some of them with key events that occurred in Rwanda, and discuss what we learn from each matched and unmatched event.

Violence against civilians – September 16, 2005. Our system identified a spatial cluster of four sites with higher than usual call volume and movement frequency on September 17, 2005. These sites are located in the vicinity of the city of Goma, along the Rwanda-DRC border and near the Rwanda-Uganda border. We find no event, emergency or otherwise that occurred on September 17, 2005. However, a violent event, reported by ACLED, occurred in the same area on September 16, 2005. The distance between the event's reported location (latitude -1.67, longitude 29.22) and the centroid of the closest site in the spatial cluster is 1.7 km. Radio France reported that this event involved five armed attacks in which 1 person was killed, 7 others were wounded, and was part of the general trend of violence against civilians.

Violence against civilians – January 15, 2006. We identify 16 sites (eight in the Kigali area) with lower than usual call volume on January 15, 2006. Eight of these sites also had lower than usual movement frequency. A violent event was recorded by ACLED in Kigali on the same day. The distance between the event's reported location (latitude -1.96, longitude 30.04) and the centroid of the closest site with unusual call volume is 3.8 km. Reporters Without Borders and Amnesty International report that four armed men invaded and ransacked the home of Bonaventure Bizumuremyi, the editor of the independent fortnightly Rwandan newspaper, Umuco. Mr. Bizumuremyi was the target of intimidation and harassment, demanding that he stop publishing articles criticizing the ruling Rwandan Patriotic Front (RPF). The armed forces of the RPF forced many independent journalists and human rights activists to leave Rwanda during that same general time period by intimidation, harassment or arbitrary arrest.

Protest – November 25, 2006. Our system identified seven sites with unusually low call volume and movement frequency, and seven additional sites with unusually low movement frequency on November 26, 2006. One of these sites is far from the other 13 and belongs to a separate spatial cluster. This suggests that there were two events on this day, one that created the anomaly in a single site and another that created behavioral anomalies in the remaining 13 sites. The 13 cluster sites are in Kigali and slightly to the east of the city. The single separate site is in the southern part of Rwanda on the Burundi border. We do not have record of any events that occurred on November 26, 2006 in these areas. However, a large protest was recorded on November 25, 2006 in Kigali. The distance between the event's reported location (latitude

-1.96, longitude 30.04) and the centroid of the closest site with unusual call volume (movement frequency) is 3.38 (1.83) km. The New Times of Rwanda reports that 15,000 demonstrators flooded the streets of Rwandas capital Kigali in protest of France's role in the Rwandan genocide, and their call for the arrest and trial of the Rwandan President Paul Kagame.

Protests – November 19, 2008. Our system identified 45 sites in and extending well beyond Kigali that had unusual low call volume and movement frequency. Four additional sites recorded unusually low call volume. The sites are grouped in a large spatial cluster which is indicative of a major common cause of the disturbances at all these sites. On the same day, reports indicate that tens of thousands of Rwandans participated in a series of protests over the arrest in Germany of Rose Kabuye, a prominent Rwandan military and political figure, on alleged involvement in the plane crash that led to the 1994 genocide. The distance between the reported location of the protests (latitude -1.96, longitude 30.04) and the centroid of the closest site with unusual call volume and movement frequency is 1.8 km.

Floods – September 19, 2007. Our system identified 53 sites that had unusually low call volume and movement frequency on September 19, 2007. During the previous week, starting on September 19, torrential rains in the northwest of the country led to severe floods, leaving 15 people dead, 7000 people homeless and displaced, and more than 1000 houses uninhabitable. Floods also contaminated clean water supplies and decimated field crops, leading to concerns about waterborne diseases and food insecurity in the area. On September 18, floods dramatically swept away 42 homes and forced families to evacuate in the middle of the night. The following day, September 19, is when we find behavioral disruptions of decreased calling and movement. Notably, the behavioral anomalies occur across the country, instead of concentrated in the area most affected by flooding. In this case, the date of the behavioral disruption suggests a good match with the flooding event, but the spatial range of behavioral reaction decreases our confidence that the dramatic floods created the dramatic behavioral anomalies. It is possible that other areas of the country were also affected by flooding, that roads were damaged or transportation infrastructure was disrupted, or that families were busy rebuilding homes and crops that were destroyed by the rains. All of these possibilities are plausible explanations for decreased mobility and calling. However, further qualitative and quantitative research on behavioral reactions to similar flood disasters will be necessary to understand if and exactly how people change their communication and movement in response to natural disasters. The contribution of this case study is an indication that reactions to flood disasters might be much more complicated than we currently understand.

Christmas Eve – December 24, 2007 and 2008 . We identified 26 sites with unusually high call and movement frequency on December 24, 2007 and 59 such sites on December 24, 2008. Still more sites recorded only higher than usual call volume (21 in 2007 and 17 in 2008), or only higher than usual movement frequency (1 in 2007 and 2 in 2008). Given that about 90% of Rwandans identify as Christians, it is not surprising that we find behavioral anomalies on Christmas Eve in 2007 and 2008. We expect that people called and visited their families to celebrate the holiday, resulting in the increases we find in both behaviors. The particular features of the behavioral anomaly we find on these two days (large spatial extent, higher than usual calls and mobility) match well the characteristics of this major planned religious event.

New Year's Eve and New Year's Day – January 1 and December 31, 2008, and January 1, 2009. Our system identified more than 20 sites spread throughout Rwanda with unusually high call and movement frequency on each of January 1, 2008, December 31, 2008, and January 1, 2009. Given that New Year's is a national holiday that affects all people in Rwanda (regardless of religion) and given the wide spread of the behavioral anomalies we find, we believe that these anomalies are due to this holiday. Just as with Christmas, it is likely that Rwandans call and visit family and friends more often on New Year's Eve and Day.

International treaty – November 9, 2007. Behavioral anomalies were identified over a large area of Rwanda on November 9, 2007: 52 sites recorded unusually high call volume and movement frequency, three additional sites recorded unusually high call volume and one other site recorded unusually high movement frequency. One political event might explain

this anomalous behavior: on that day, the governments of the Republic of Rwanda and of the Democratic Republic of Congo (DRC) signed the "Nairobi Communiqué" which defined a joint approach to end the threat to peace and stability in both countries and in the Great Lakes region posed by the Rwandan armed groups on Congolese territory. It is plausible that people made more calls to spread information and discuss this major treaty, but it is unclear why such as event would cause increased mobility. We do not find any other event that could plausibly have caused a nationwide response such as this.

Major unknown event – April 24 and 25, 2008. Our system identified unusually low call volume and movement frequency in 61 sites on April 24, 2008 and in 53 sites on the next day. On both days additional sites recorded unusually low call or movement frequency. We have been unable to find an event on or just before these days that could explain anomalous human behavior that lasted at least two consecutive days, affected almost the entire country and led to a significant decrease in the routine behaviors in Rwanda.

Commemoration of the genocide against the Tutsi – April 7 and 8, 2007, and April 7 and 8, 2008. Our system identified 26 sites with unusually low call volume and movement frequency on April 7, 2007 and 24 such sites on April 7, 2008. Our system also found a smaller number of sites with unusually low call volume and movement frequency on April 8, 2007 and 2008. April 7 is an official annual Rwandan holiday which marks the start date of the 1994 genocide. It is a planned event which affects most Rwandans. The behavioral anomalies spread across the country on these days for two years in a row suggest that the remembrance day could be the cause of decreased call volume and mobility frequency.

3. Discussion

In this paper, we contribute to the process of creating a system of detecting emergency events using mobile phone data. An effective event detection system could make significant contributions to humanitarian response and reducing the toll of disasters on human well-being. Towards this end, we develop a method for using mobile phone data to identify days with anomalous calling and mobility behavior, including days with high call volume and/or mobility, and low call volume and/or mobility. Our method also identifies the location of these anomalies and the geographical spread of the disturbances. We compare the days we identify with anomalous behaviors to a database of emergency and non-emergency events. Some days and places with behavioral anomalies match well with events and others do not. We learn from both cases.

Our analysis makes clear that detecting dramatic behavioral anomalies is only part of the work required to create an effective system of emergency event detection. The remaining work that is necessary is serious social-behavioral analysis of the exact types of behaviors that can be expected after different kinds of events and the exact time scales on which they occur. This will require intensive qualitative as well as quantitative analysis. It is only through a thorough understanding of these underlying differential behavioral patterns that an effective detection system can be developed.

References

[1] Dobra A, Williams N, Eagle, N (2015) Spatiotemporal Detection of Unusual Human Population Behavior Using Mobile Phone Data. PLoS ONE to appear. Available at http://arxiv.org/abs/1411.6179

Department of Statistics, Department of Biobehavioral Nursing and Health Systems, Center for Statistics and the Social Sciences and Center for Studies in Demography and Ecology, University of Washington, Box 354322, Seattle, WA 98195
E-mail address: adobra@uw.edu

Department of Sociology and Jackson School of International Studies, University of Washington, Box 353340,Seattle, WA 98195
E-mail address: natw@uw.edu

Department of Epidemiology, Harvard University, Boston, MA 02115
E-mail address: nathan@mit.edu

Estimating Attendance From Cellular Network Data

Marco Mamei
Dipartimento di Scienze e Metodi dell'Ingegneria
University of Modena and Reggio Emilia, Italy
marco.mamei@unimore.it

Massimo Colonna
Engineering & Tilab
Telecom Italia, Italy
massimo.colonna @telecomitalia.it

1. INTRODUCTION

We present a methodology to estimate the number of attendees to events happening in the city from cellular network data. In this work we used anonymized Call Detail Record (CDR) comprising data on where and when users access the cellular network. Our approach is based on two key ideas: (1) we identify the network cells associated to the event location. (2) We verify the attendance of each user, as a measure of whether (s)he generates CDR data during the event, but not at other times. We evaluate our approach to estimate the number of attendees to a number of events ranging from football matches in stadiums to concerts and festivals in open squares. Comparing our results with the best groundtruth data available, our estimates provide a median error of less than 15% of the actual number of attendees.

While a number of existing works deal with the problem of discovering and analyzing events con the basis of cellular network data, the problem of actually estimating the number of attendees is largely unexplored. In particular, to the best of our knowledge, there are not published results of the accuracy of attendance estimation using CDR data.

2. NAIVE APPROACH

A naive approach to estimate the number of people attending an event would be to just count the number of users who generate CDR data in cells covering the event's location during the event time. In particular, we defined the area associated to each event as a circle centered in the event place with a fixed radius of 100m. Then, we record all the CDR produced in the network cells that overlap with the area at the event time. We then counted the number of individual users.

Figure 1 shows correlation results – using the naive approach – for a number of events covered by our dataset. Each point represents an event: the x-coordinate is the CDR estimate for attendance, while the y-coordinate is the groundtruth attendance. It is easy to see that there is almost no correlation ($r^2 = 0.016$) between the two estimates, so the naive approach is highly ineffective.

Figure 1: Correlation result using the naive approach. It is easy to see that there is almost no correlation ($r^2 = 0.016$) among CDR count and groundtruth.

Our goal is to identify a mechanism to create a strong correlation between groundtruth and CDR counts. Once this result is achieved, a simple linear regression can scale up CDR counts to the actual attendees estimate.

3. METHODOLOGY

To overcome the above limitations, we developed a specific methodology to deal with attendance estimation. In particular: **(3.1)** We identify the radius within which are all the cells whose traffic can be associated to the area where the event takes place. **(3.2)** On the basis of the identified cells, we count the number of users who generate CDR data at the event time, but who do not (usually) generate CDR at other times. **(3.3)** On the basis of such data from a number of events, we set up a linear regression to estimate the number of attendees.

3.1 Event Radius Identification

Determining the network cells handling the traffic for an event is a fundamental task. Otherwise it is possible that the cells being considered will include CDR data actually produced elsewhere, or will miss CDR data that were actually produced in the proper area.

We model the event area as a circle with center c - where the event takes place, and with radius r. A

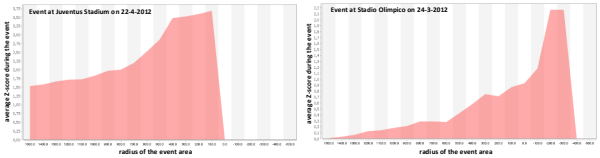


Figure 2: Graph showing for each r_k how much the area had an unusually high number of people during the event.

cell with center b and radius rc is considered relevant for the event if: $dist(c, b) < r + rc$. Where $dist$ is the geographic distance between the points. We could also select $r < 0$ to impose the fact that a cell has to overlap to the center of the event by a certain amount to be considered as relevant.

Our approach starts from the basic consideration that the plot of the number of CDR generated from the event area should have a spike (i.e., an outlier) when the event take place, as the events – we are interested in – will typically attract a large number of people.

Accordingly, we scan different values r_k for the radius of the area of the event and compute a z-score z_k of the number of CDR generated from that area at the event time. The result is a graph showing for each r_k how much the area had an unusually high number of people during the event. Figure 2 shows the result for two events. It is possible to see that once the area is properly identified, the z-score clearly identifies that something unusual is taking place there ($z = 3.7$ with a radius of about 300m for the event on the left, $z = 2.2$ with a radius of about -200m for the event on the right).

We compute the best radius as the average of the r_k values weighted by the associated z-scores.

3.2 Attendance Estimator

Once the event area has been identified, we need a mechanism to count the number of users who attended the event. Since we do not know what the user was doing in the event area, we estimate the probability of the user presence as proportional to the fraction of time in which the user was there during the event $f1$, and inversely proportional to the fraction of time in which the user was there outside of the event time $f2$.

$p = f1 \cdot (1 - f2)$

We then add all such probabilities p together to obtain a raw attendance estimator of the event.

3.3 Piecewise Linear Regression

The above estimator is typically much lower than the actual attendance. This can be naturally explained by the fact that not all the users will use the phone during the match, and by the fact that not all of them adopts the same carrier providing the data for this anal-

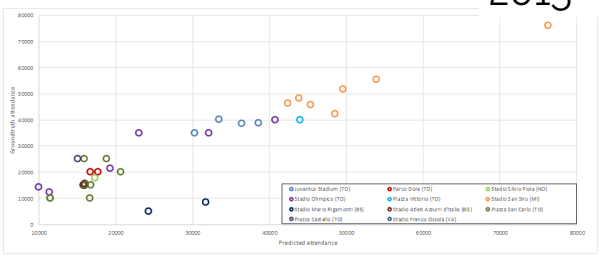


Figure 3: Estimated VS. groundtruth attendance. It is possible to see a good linear correlation between data ($r^2 = 0.78$).

Piecewise Linear R.	
Mean abs. error	6057
Median abs. error	4072
Mean % error	68%
Median % error	14%

Figure 4: Results obtained by using piecewise linear regression.

ysis. In any case, as we will show in the next section, it has a strong positive correlation with groundtruth headcounts. Accordingly, a simple linear regression can scale up the above count to the actual attendees estimate.

We assume the availability of a *training* set of events for which attendance estimate is known to be used to fit the parameters of the linear regression. The resulting coefficients are then used to scale CDR estimates of attendance in a *testing* set of events. specifically, we used piecewise linear regression: for each testing sample, we consider the n closest samples in the training set, create a linear regression on that n points, and use it to scale that predicted testing sample.

4. ANALYSIS AND RESULTS

We applied our approach to estimate the number of attendees to events ranging from football matches in stadiums to concerts and festivals in open squares. The analysis spans large events with ground truth attendance of more than 80000 persons to smaller ones with a ground truth attendance of less than 2000 persons. Overall, we considered a dataset comprising 43 events.

Fig. 3 shows correlation between our estimate and groundtruth values ($r^2 = 0.78$), note the improvement over Fig. 1. Fig. 4 presents mean/median absolute error (i.e., mean/median of the actual difference between estimated attendees and groundtruth). The percentage error divides the actual difference by the groundtruth to understand the relative error to the attendees' size.

It is possible to see that the use of CDR data produces rather good estimates of the number of attendees.

Seasonal Decomposition of Cell Phone Activity Series and Urban Ecology

Blerim Cici, Minas Gjoka, Athina Markopoulou, Carter T. Butts
University of California, Irvine, USA
{bcici, mgjoka, athina, buttsc}@uci.edu

Motivation: It is estimated that within the next forty years, two-thirds of the world's population will be living in expanding urban centers, and the level of urbanization is expected to increase in all major areas of the developing world, with Africa and Asia urbanizing more rapidly than the rest [1]. Given this, there is clearly a need for methods that will cheaply yield up-to-date information on urban environments, without extensive on-the-ground investigation. A promising tool in that regard is the use of aggregate geo-located data on communication activity, a resource that is increasingly available given the near-universal penetration of mobile devices within urban populations. Such data is inexpensive and can be collected and distributed in a privacy-preserving manner.

The question, then, is how aggregate mobile data can be used to provide information on urban ecology — the distribution of population density, social and economic activities, and social interaction. In this abstract, we present an approach to this problem that draws on the notion of seasonal decomposition; by decomposing communication data we create distinct time series that provide information on routine activities and on deviations from those routines.

Key idea: The key to our approach is the spectral decomposition of the original cell phone activity series (see Fig. 1) into seasonal communication series (SCS) and residual communication series (RCS). SCS captures regular patterns of socio-economic activity within an area and can be used to segment a city into distinct clusters. RCS across areas enables the detection of regions that are subject to mutual social influence and of regions that are in direct communication contact. The RCS and SCS thus provide distinct probes into the structure and dynamics of the urban environment, both of which can be obtained from the same underlying data.

Name	Period	Source
Milan Activity	Nov.4-Dec.1 2013	Telecom Italia [2]
Milan Square-to-Square	Nov.4-Dec.1 2013	Telecom Italia [2]
Universities, Businesses, Parks, Population per area, Sport Centers, Bus stops	Jan.1-Dec.31 2013	City of Milan [3]

Table 1: Data Sets

Data: The data that we use in this study (summarized in Tab. 1) consists of time series of aggregate cell phone traffic sent or received by persons within small areal units in the city of Milan, made available for the *Big Data Challenge* competition [2]. In particular, the city of Milan, an area of 550 km^2, was divided into a 100 $\times$ 100 square grid. Each grid square has the same dimensions: a side length of 0.235 km and an area of 0.055 km^2. This is the areal unit we use throughout the paper, and we refer to it as a "square". The temporal unit is the 10-minute interval. In addition, we collected ground truth data from the municipality of Milan and online sources, containing elements such as universities, residential areas, sport centers, parks, etc (last row of Tab. 1). This latter data is used here for external evaluation of our methodology.

Methodology and Findings: We apply our approach to the cell phone activity series, using the decomposed series to characterize

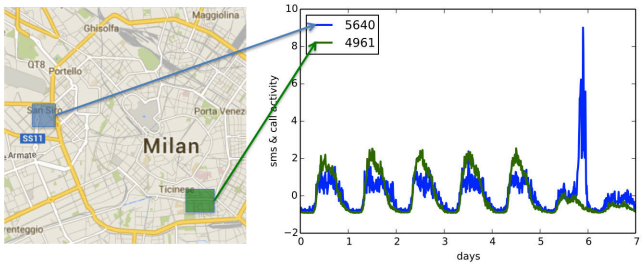


Figure 1: Cell phone activity series (normalized) for two grid squares of Milan, for the week of 11/4/2013–11/10/2013. Square 5640 is located close to the San Siro stadium, while square 4961 is located in a university region. Differences in seasonal patterns (weekday/weekend) reflect stable differences between the university and stadium environments.

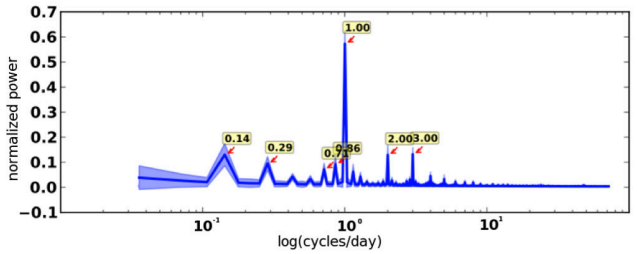


Figure 2: Power spectrum of activity aggregated over all grid squares (blue line indicates mean, shaded area ± 1 std dev); marks indicate high-amplitude frequencies (e.g. daily (1 $cycles/day$) and weekly (0.14 $cycles/day$).

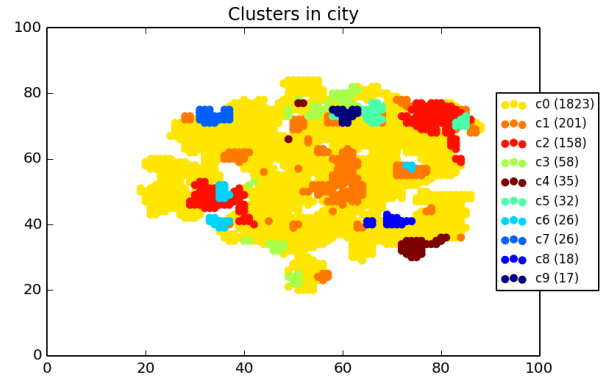
distinct aspects of urban ecology. We proceed as follows:

First, we begin by decomposing the original series, using FFT. For each grid square $i \in S$, we decompose the original time-series $O_i(T)$, into *seasonal* and *residual* through the following steps:

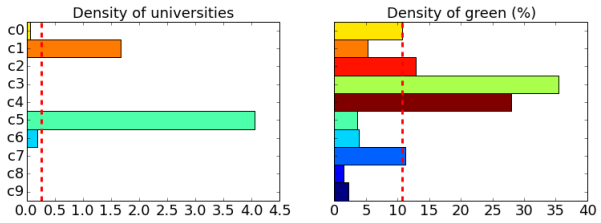
1. We select its k highest-power frequencies[1] (see Fig. 2).
2. We regenerate the seasonal communication series $SCS_i(T)$ using only the top k frequencies of the square, where $SCS_i(T) = \{scs_i(t_1), ..., scs_i(t_m)\}$.
3. We obtain the residual communication series $RCS_i(T)$ by subtracting the basic series from the original series: $RCS_i(T) = \{o_i(t_1) - scs_i(t_1), ..., o_i(t_m) - scs_i(t_m)\}$.

Second, we perform hierarchical clustering of different squares based on the different time series and we validate the results using the ground truth data; we used a correlation based distance function, i.e. $dist(i,j) = 1 - correlation(SCS_i(T), SCS_j(T))$, where $i, j \in S$. Then, we used the cluster-size skewness distribution in order to cut-off the deprogram and get a single clustering; the cut-off is the point where the distribution has the lowest skewness.

[1]We use $k = 30$, since RCS autocorrelation function does not differ significantly from that of a white noise sequence for that value.



(a) Segmentation of Milan



(b) Densities per cluster

Figure 3: Fig. 3(a) shows the city of Milan segmented in different clusters; for clarity we show only the top-10 largest clusters. Fig. 3(b) shows the cluster densities for two of the ground truth elements. The red dashed line indicates the Milan average. SCS clustering separates regions with distinct urban environments (e.g., university vs. green space).

Category	Entropy for hierarchical SCS	Entropy for [4]
Universities	0.96	0.97
Green (%)	0.94	1.27
Businesses	0.82	1.33
Population	0.97	1.34

Table 2: Segmentation performance via Entropy (lower value is better). Hierarchical SCS clustering produces more functionally distinct clusters for all categories.

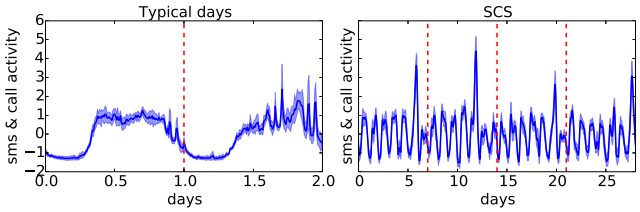


Figure 4: The left figure shows how the typical weekday/weekend approach, of [4], summarizes the cell phone activity series for the San Siro area. Notice that the high peaks are lost when traffic is aggregated. The right figure shows the SCS traffic in the San Siro area for one month (our method).

Finally, we show that our SCS clustering scheme successfully segments areas dominated by distinct types of socio-economic activity, and allows for discovery of regions whose activity patterns differ markedly from the rest of the city (see Fig. 3). Also, the results compare favorably with state-of-the-art approaches such as [4] (see Tab. 2), since SCS is able to flexibly and parsimoniously incorporate regular patterns occurring on any time scale; by contrast, state-of-the-art methods estimate regular patterns using average weekday and weekend days (evaluated using binned averages), and cannot therefore exploit temporal patterns across multiple days (see Fig. 4).

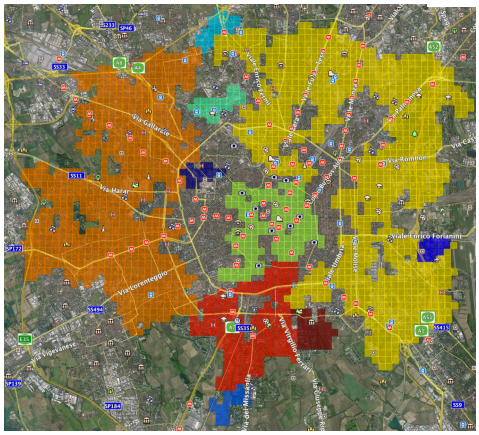


Figure 5: 10 largest strongly connected components for the directed graph of RCS cross-correlations with $lag = 1$.

In addition to using the SCS to identify regular patterns, we show that RCS enables the detection of regions that (i) are subject to mutual social influence or (ii) in direct communication contact; this was not previously possible using mere activity data.

First, we examine how squares affect each other's RCS activity, i.e. for two cross-correlated squares, a perturbation in one square at time t will be associated with a change in the other square at a later point in time. We calculate a cross-correlation digraph for a given time $lag = i$ and compute its strongly connected components to find areas that are subject to mutual social influence. Using this approach in Fig. 5 shows a different, but interesting, segmentation of the city from that obtained by SCS, with a clear structure becoming apparent. The center of the city is a connected component (green color), completely separate from the rest. This means that perturbations occurring inside a square on the center, will most likely propagate to other grid-squares in the center. Another example is the dark blue connected component on the center-right side which corresponds to the Milan Linate airport.

Second, we show that the RCS allows for the detection of areas that are in direct communicative contact. We validate the latter by showing that RCS correlations between areas are significantly related to the volume of inter-area cell phone traffic, and that this relationship is substantially stronger than for SCS: **0.27** for RCS versus **0.05** for SCS; this is done using the Quadratic Assignment Procedure (QAP) [5], which is a technique for testing an matrix correlation against a null hypothesis of no association, while controlling for the underlying structure of the matrices being compared.

Due to space constraints, this abstract does not contain all the details of our methodology. A more detailed version of this work is currently under submission in MobiHoc 2015 [6].

1. REFERENCES

[1] World urbanization prospects: The 2011 revision. United Nations Department of Economic and Social Affairs/Population Division.

[2] Big data challenge. `http://www.telecomitalia.com/tit/en/bigdatachallenge.html`, 2014.

[3] Milan's public data. `http://dati.comune.milano.it/`, '14.

[4] V. Soto and E. Frias-Martinez. Automated Land Use Identification using Cell-Phone Records. In *Proc. of HotPlanet*, 2011.

[5] D. Krackhardt. QAP Partialling as a Test of Spuriousness. *Social Networks*, 9:171–186, 1987.

[6] B. Cici, M. Gjoka, A. Markopoulou, and C. T. Butts. On the decomposition of cell phone activity patterns and their connection with urban ecology.

Social Events in a Time-Varying Mobile Phone Graph

Carlos Sarraute
Grandata Labs
charles@grandata.com

Jorge Brea
Grandata Labs
jorge@grandata.com

Javier Burroni
Grandata Labs
javier.burroni@grandata.com

Klaus Wehmuth
LNCC
klaus@lncc.br

Artur Ziviani
LNCC
ziviani@lncc.br

J.I. Alvarez-Hamelin
UBA–CONICET
ihameli@cnet.fi.uba.ar

Introduction

The large-scale study of human mobility has been significantly enhanced over the last decade by the massive use of mobile phones in urban populations [1, 2]. Studying the activity of mobile phones allows us, not only to infer social networks between individuals, but also to observe the movements of these individuals in space and time [4]. In this work, we investigate how these two related sources of information can be integrated within the context of detecting and analyzing large social events. We show that large social events can be characterized not only by an anomalous increase in activity of the antennas in the neighborhood of the event, but also by an increase in social relationships of the attendants present in the event. Moreover, having detected a large social event via increased antenna activity, we can use the network connections to infer whether an unobserved user was present at the event. More precisely, we address the following three challenges: (i) automatically detecting large social events via increased antenna activity; (ii) characterizing the social cohesion of the detected event; and (iii) analyzing the feasibility of inferring whether unobserved users were in the event.

Data Source and Methodology

Our data source is an anonymized traffic dataset from a mobile phone operator in Argentina, collected mostly in the Buenos Aires metropolitan area, over a period of 5 months. The raw data logs contain about 50 million calls per day. Call Detail Records (CDR) are an attractive source of location information since they are collected for all active cellular users (about 40 million users in Argentina). Further, additional uses of CDR data incur little marginal cost.

Together with the location of the clients of the mobile phone network, the CDRs allow us to reconstruct a social graph derived from the communications among users. We first define a contact graph $\mathcal{G}$ composed of all users participating in the mobile network, where an edge between users exist if they have communicated at any point within the 5 months period. We note that $\mathcal{G}$ includes non client users that have communicated with client users.

Then, we define a time series of subgraphs of $\mathcal{G}$ composed only of users present at a given location at a given time. In this work, these subgraphs are restricted to client nodes (users for which we have mobility data) using the event antenna at a specified day, for a given time window. Note that the existence of a link between such nodes in any subgraph depends on whether there is a link between them in $\mathcal{G}$. These subgraphs allow us to focus on the social dynamics of a set of users present at a particular location at a given time as is the case of client users attending a large-scale event.

We formalize this temporal sequence of subgraphs within the context of a Time-Varying Graph (TVG) based on the representation encountered in [6]. We consider a TVG as an object $H = (V, E, T)$, where V is a vertex set, T is a time instant set, and E is an edge set. In this context, an edge $e \in E$ is primarily a quadruple of the form $\langle u, t_a, v, t_b \rangle$, where $u, v \in V$ are vertices and $t_a, t_b \in T$ are time instants. In this representation, an edge $\langle u, t_a, v, t_b \rangle$ then shows a relation between vertex u at time t_a and vertex v at time t_b, i.e. a directed relation, where the pair $\langle u, t_a \rangle$ is the origin and the pair $\langle v, t_b \rangle$ is the destination. Note that the represented relation is between the composite objects $\mathbf{u} = \langle u, t_a \rangle$ and $\mathbf{v} = \langle v, t_b \rangle$. In this representation, a TVG edge $e \in E \subset V \times T \times V \times T$ can be thus seen as an ordered pair $\langle \mathbf{u}, \mathbf{v} \rangle \in (V \times T) \times (V \times T)$ of composite vertices.

In this work, each edge in a subgraph is represented as a tuple $\langle (x, t_x), (y, t_y) \rangle$, where user x is the caller, user y is the callee, $t_x = t_y$ is the date and time of the call. In addition to the origin and destination composite vertices present on the TVG edge, we also wish to identify the local base station antenna (BSA) used on the call represented by each TVG edge. In order to achieve this, we represent the BSA identity as an edge attribute, so that the TVG edge used in this particular work becomes an ordered 5-tuple of the format $\langle x, t_x, y, t_y, l \rangle$ where in addition to the known edge elements, l represents the BSA identity.

Automatic Event Detection

In this section, we show how we can use the mobile phone data in an endogenous manner to detect past events automatically (as [3, 5] did with social media such as Twitter). The events we are interested in involve a large amount of people, hence they significantly increase the usage demand of antennas nearby, as evidenced by their traffic time series. Our strategy thus is to locate time windows where the antenna activity is significantly higher than some appropriate baseline. Defining an appropriate activity baseline is not a trivial task since this is possibly dependent on the size of the event one is trying to detect and on the regularity of the normal antenna activity patterns.

In this work, the antennas activity is given by the number of calls per hour registered in each antenna, according to the week w_i ($0 \leq i < n$, where n is the number of weeks in the studied dataset), the day of the week d_j ($0 \leq j \leq 6$), and the hour of the day h_k ($0 \leq k \leq 23$). For each time slot (w_i, d_j, h_k), we denote the number of calls as $C_{(i,j,k)}$ which

we then normalise be the average activity over a period of 3 months, defining an *event index* $\mathcal{E}_{(i,j,k)}$ given by

$$\mathcal{E}_{(i,j,k)} = \frac{C_{(i,j,k)}}{\frac{1}{n}\sum_{\ell} C_{(\ell,j,k)}}$$

for the i^{th} week in the 3 month period, j^{th} day of the week, k^{th} hour of the day, and where ℓ sums over all weeks in the whole time period considered ($0 \leq \ell < n$). Figure 1 shows the event index $\mathcal{E}$ for an antenna near the stadium of River Plate, a major soccer team in Buenos Aires, for each time slot (hour) from January 1 to March 31, 2012. We consider a large-scale event to have occurred when the event index is within the highest 99-percentile while considering the whole 3 months period.

During January and February, we observe the event index smoothly oscillating around the mean value except for small sharp increases due to normal peak daily activity. As we approach March, the oscillations' amplitude starts to increase reaching its peak value by mid March. On top of this large oscillations we see very large sharp peaks reaching values of $\mathcal{E} = 9.98$ on March 20. Additional high peaks are also observed on March 7, 9, 10, 12, 14, 15, 17, and 18. Retrospectively, we corroborated that a sequence of 9 Roger Waters concerts took place at this location during March on the same days of the observed peaks. We thus highlight that the event index has clearly spotted the Roger Waters events on the exact days the concerts took place.

An open question is to understand the slight increase in the background oscillation during these peak periods. A plausible explanation is an increase in activity due to gatherings before the event (e.g., fans trying to get tickets, organizers working at the stadium, January and February are vacations month in Argentina, etc.) or it might be a statistical fluctuation one could verify if one had access to a much longer dataset. In any case, further study of this background activity can prove very useful, for instance, allowing for the forecasting of large-scale events. Identifying these early precursors could allow us not only to automatically detect events retrospectively, but also give us the possibility of anticipating a future event.

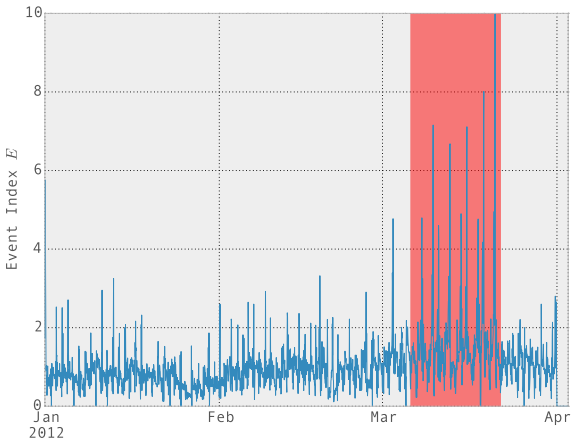


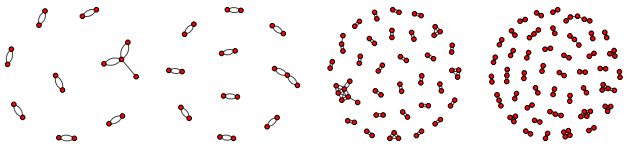
Figure 1: Event index $\mathcal{E}$ for an antenna near the River Plate soccer stadium.

Social Dimension of Events

Having detected a large-scale event, we now take a look at social relationships of users attending these events. We consider all communications from the event antenna for a period of 5 months from November 1, 2011 to March 31, 2012, therefore including the Roger Waters concert series from March 7–20, 2012. For each day, we compute the implicit social subgraph induced by $\mathcal{G}$ restricted to users observed at the event antenna during the time window from 18:00 to 22:00 (time window for the Roger Waters concerts).

Figure 2 shows the induced subgraphs for February 7 and 17, in which no event took place (a, b), along with March 7 and 17 for which two Roger Waters concerts took place (c, d). The total number of nodes on February 7 and 17 were 107 and 130, respectively, whereas on March 7 and 17 we observe 716 and 740 attenders, respectively. Most nodes are singlets and are thus not shown in the plots, leaving only nodes for whom at least one of his/her contacts in $\mathcal{G}$ was also present at the event location, at the same day, and during the defined time window.

Therefore, one can expect that it is somewhat likely that the nodes present in these subgraphs are attending this location together. We call these nodes *social attenders* and all nodes, singlets included, *attenders*. An immediate observation is the larger number of social attenders at the days of the concerts than on a normal day. This is expected given the larger attendance during the event. More interestingly, for the days of the event, especially on March 17, we observe more complex group relationships of up to 7 nodes, possibly indicating larger groups of people attending the concert together.



(a) Feb 7 (b) Feb 17 (c) March 7 (d) March 17

Figure 2: Implicit social graphs: (a), (b) two days with no large event; (c), (d) two days of the Roger Waters concerts.

Probability of Being in an Event

The increased social structure suggested in Fig. 2 motivates the following question: can we estimate the probability that a user attended the event, given that k of his/her contacts in graph $\mathcal{G}$ also attended the event? We compute this probability as follows:

$$p(x \in U | k \text{ contacts} \in U) = \frac{\text{users} \in U \text{ with } k \text{ contacts} \in U}{\text{users with } k \text{ contacts} \in U},$$

where U is the set of all users in the event. The results are shown in Fig. 3, where we also plot a linear regression curve showing how strongly the expectation of finding a user in the event depends on the amount of graph contacts this user has in the event. The scatter plot and the regression curve ($R = 0.83$) show that the probability that a given user attended the event grows linearly with the number of his/her contacts in the event. If several of the user contacts in $\mathcal{G}$ represent real social relationships, such as friends or family members, and a significant fraction of such contacts

are in a social event, such as a music concert, one could expect an increase in the likelihood the user is also attending the event. Nevertheless, it is worth noticing that the linear relation does no saturate even for high values of k.

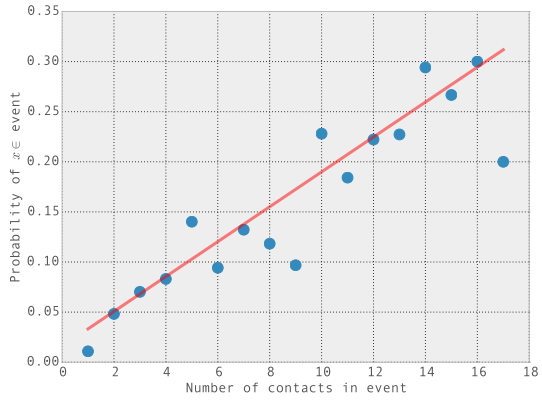


Figure 3: Scatter plot and linear regression curve for the conditional probability of a user being in an event given that k of his/her contacts in graph $\mathcal{G}$ were in the event.

We also computed the ratio between the number of users in the event and the number of users with contacts in the event (such users can be at the event or not), both having at least K contacts in the event U. This ratio is displayed in Fig. 4, which can be interpreted as the conditional probability that a user is in the event given that *at least* K of his/her contacts are in the event, as a function of K. This probability remarks the social character of the event because, as the number of contacts in the event increases, the probability of users with such contacts also increases up to $K = 10$. Then, it starts to decrease due to the less significant proportion of people with more than 10 contacts. Notice that the sample space, where each probability is taken, is defined independently for each K.

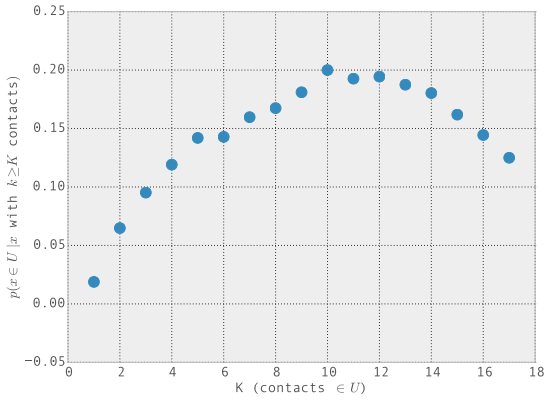


Figure 4: Probability of a user being in the event U, given the user has at least K contacts in the event.

Conclusion and Future Work

In this work, we have shown how we can combine the mobility and social network information present in mobile phone datasets to detect large social events and characterize their social features, such as an increase in local community structures of users present at an event. We have also shown that it is plausible to infer the probability of a user (in the contact graph) attending an event, given the number of his/her contacts who attended the event.

Finding useful inference algorithms of users attendance to a particular location given the mobility of their contacts can be of tremendous use for mobile phone companies. Mobile phone carriers have no mobility information about users in the contact graph $\mathcal{G}$ who are not clients of the carrier. Therefore, having the possibility of making at least weak inferences about the locations and mobility of these users can add a valuable dimension for the carrier to better understand these (non client) users.

We have presented this work within the formalism of time-varying graphs (TVG). Even though this work has not yet explored the potential uses of the TVG to analyse the dynamics of the graph structure, we believe this is a useful formalism to extend this work into further understanding the changes in social dynamic networks as one approaches large-scale events.

Acknowledgements

Authors are grateful to the STIC-AmSud program. Brazilian authors thank CAPES, CNPq, FAPERJ, and FINEP for their support.

1. REFERENCES

[1] R. Becker, R. Cáceres, K. Hanson, S. Isaacman, J. M. Loh, M. Martonosi, J. Rowland, S. Urbanek, A. Varshavsky, and C. Volinsky. Human mobility characterization from cellular network data. *Communications of the ACM*, 56(1):74–82, Jan. 2013.

[2] F. Calabrese, L. Ferrari, and V. D. Blondel. Urban sensing using mobile phone network data: A survey of research. *ACM Computing Surveys*, 47(2):25:1–25:20, Nov. 2014.

[3] F. C. T. Chua and S. Asur. Automatic summarization of events from social media. In *International AAAI Conference on Weblogs and Social Media – ICWSM*, 2013.

[4] N. Ponieman, A. Salles, and C. Sarraute. Human mobility and predictability enriched by social phenomena information. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 1331–1336. ACM, 2013.

[5] A. Ritter, O. Etzioni, S. Clark, et al. Open domain event extraction from twitter. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1104–1112. ACM, 2012.

[6] K. Wehmuth, A. Ziviani, and E. Fleury. A Unifying Model for Representing Time-Varying Graphs. Technical Report RR-8466, INRIA, Feb. 2014.

Real-Time Social Event Analytics

Francesco Calabrese*, Giusy Di Lorenzo*, Gavin McArdle*†, Fabio Pinelli*, Erik Van Lierde‡
*IBM Research - Ireland, Dublin, Ireland
{fcalabre, giusydil, gavinm, fabiopin}@ie.ibm.com
†National Centre for Geocomputation, Maynooth University, Maynooth, Co Kildare, Ireland
gavin.mcardle@nuim.ie
‡Mobistar, Brussels, Belgium
erik.vanlierde@mail.mobistar.be

Managing public safety at large events is important. Crowd control and traffic management are particularly relevant for non-ticketed events in public spaces. In such cases, it can be difficult for organisers to anticipate the number of people who will attend and to validate the event's success [1]. Given the ubiquitous nature of mobile phones, Call Detail Records (CDRs), which are the logs of user transactions with a mobile phone service provider, have been widely used to study urban processes [2], [3], [4]. By building on the work of [5] and [6] our research explores the use of real-time CDR data as a proxy to estimate the density of crowds in different areas of a city while events are taking place. The research has also been extended to estimate the density of vehicles on the main access routes to a city. This has led to the development of an application entitled Social Event Analytics (SEA) which provides both real-time and historic information about crowd and vehicle densities. The application can be used by authorities and event organisers to manage the event and gauge its success. The application was used in January 2015 for monitoring city wide events in Mons, Belgium which marked the launch of Mons as the European City of Culture for 2015. Using SEA local police simultaneously monitored the density of vehicles on the road network and the crowd density in different areas of the city. Below, we briefly describe the new real-time data analysis which we carried out for this specific case in which over 20 million CDRs were analysed each day. The results are useful for authorities but will also help to further our knowledge of human processes in urban environments.

We use data from a Belgium telecommunication operator who provided details about the distribution and azimuth of their cell towers in the city of Mons. Using this information the city is divided into cells using a Voronoi tessellation [7]. In total, 319 distinct cells were created (figure 1). These formed the basic unit on which crowd and vehicle density are analysed. Prior to the opening ceremony, the police and organisers provided details of specific public squares in the city where events were taking place along with a list of important access routes to the city and our analysis focuses on these spatial features.

Anonymous CDR data for individuals connected to cell towers in Mons and the surrounding area are supplied directly by the telecommunication operator. Each row of CDR data consists of an anonymous user ID, a time-stamp, the ID of the cell tower, the home country of the device and the type of record (call, SMS, data). Typically, data connections are an always on service and lead to generation of comprehensive CDRs. In this case the data are recorded at a rate of more than 250 records every second. By analysing the cell IDs, it is possible to determine the number of users connected to each cell tower and determine the Voronoi cell they are in. By combining this data with the known area of each Voronoi cell and the known customer penetration rate of the specific mobile operator, the density of crowds in any given Voronoi cell is estimated. For the public squares which are of particular interest to the police, we also produce a count of the number of individuals in these spaces. Furthermore, an estimation of the number of people in the whole city is calculated by summing the amount people connected to the relevant cell towers.

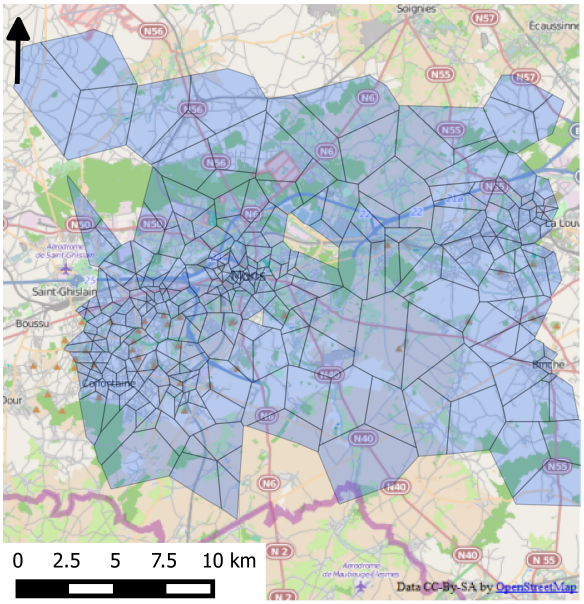


Fig. 1. 319 Voronoi cells were computed from the cell tower distribution and azimuth data provided by the telecommunication operator in Mons.

While the police are interested in public safety and crowd control, organisers are also interested in assessing the success of the event and the marketing campaigns used to attract people. In addition to the numbers attending the event, the organisers also wanted to know where people were travelling from in order to attend the events in the city. Therefore, the number of users by nationality are calculated using the home country of the device which is available directly from each CDR. In order to obtain the home city of Belgian users, further

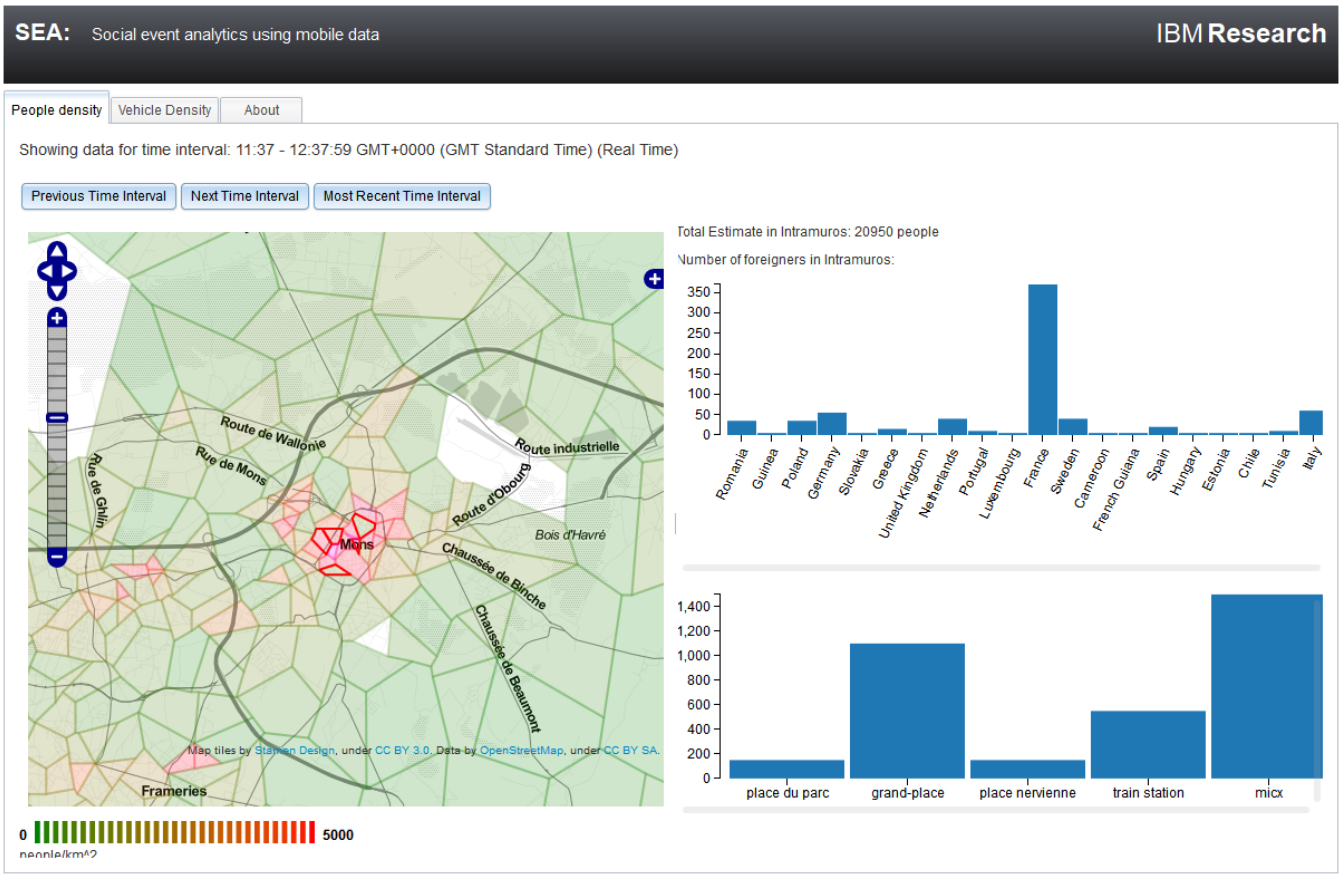


Fig. 3. A screen capture from the SEA application showing the estimated crowd density for the Voronoi cells in Mons. Additional bar charts show the estimated volume of people at several locations in the city and the home country of visitors.

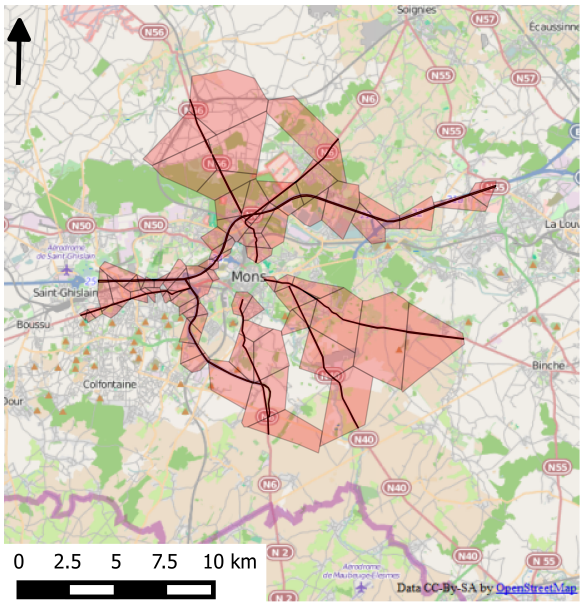


Fig. 2. 86 road segments were computed based on their intersection with the Voronoi cells.

analysis of the complete CDR data for Belgium was carried out off line. Several weeks of CDR data were analysed in order to determine the city where each device generally connected to network during the night time hours between Midnight and 6 AM. This is similar to approaches for calculating the significant places people visit [8], [9].

In addition to crowd estimation, the density of vehicles travelling on the main access roads to the city is estimated using a minimal computational approach. Each road of interest is segmented so each segment is contained within a single Voronoi cell. This produces 86 road segments (figure 2). While the density of crowds in these cells can be estimated using the techniques described above, we are interested in those travelling through the cell. Historic CDR data was analysed to determine users who regularly spend time in that cell. These typically represent people who live or work in the area. These users are removed from the analysis of the density for that cell which allows us to estimate the density of the major roads in the cell. When two or more major roads are covered by a single Voronoi cell, it is necessary to estimate the density of these roads by averaging the known densities of the road segments on either side of the particular cell.

The analysis above was used to produce an application, called SEA, which monitors vehicle and crowd density in near real-time. The data is supplied by the telecommunication operator every 15 minutes and contains the CDR data for the previous 15 minutes. Based on the analysis, dynamic visualisations are produced to present the results to the police and organisers. The main visualisation, as seen in figure 1

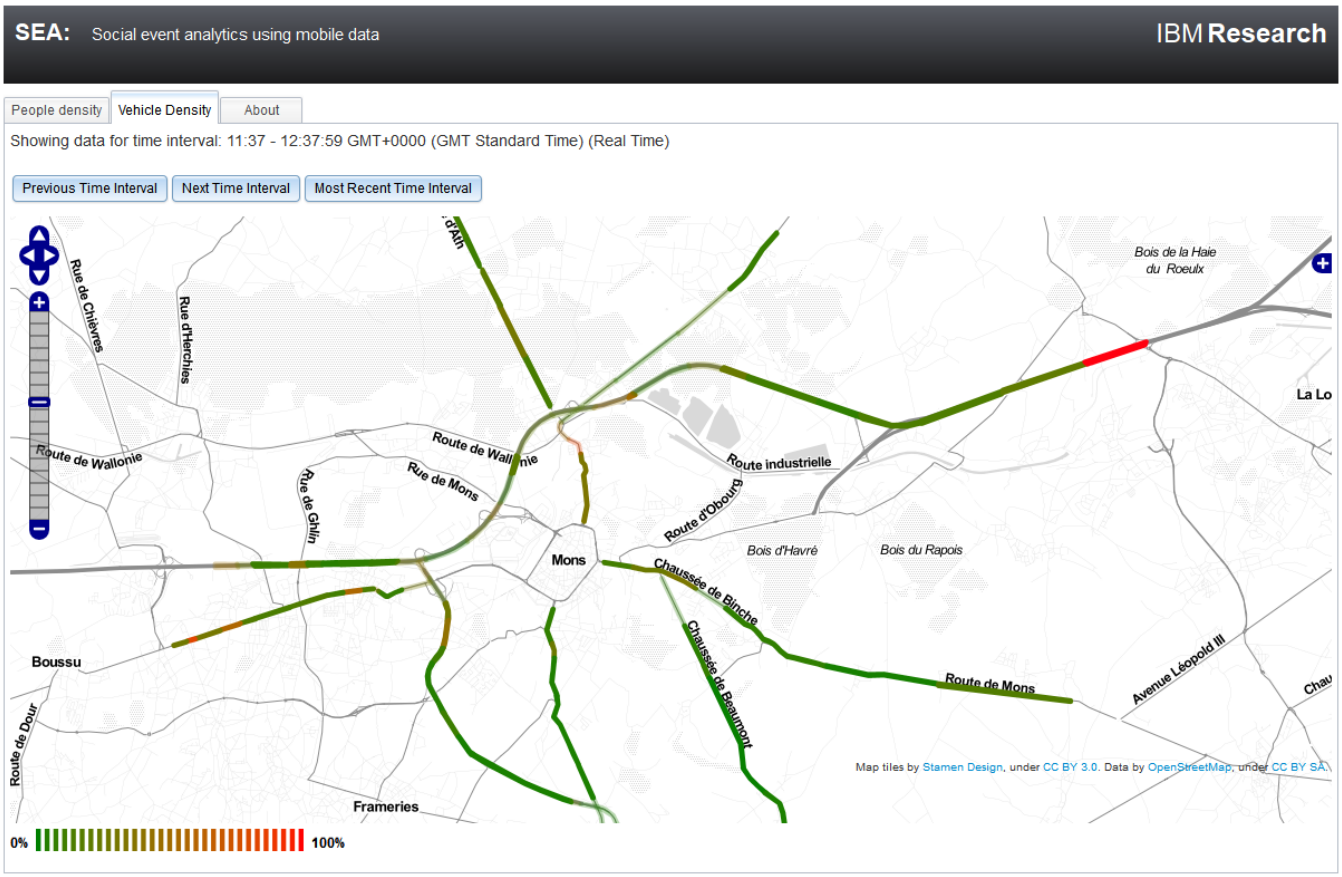


Fig. 4. A screen capture from the SEA application showing the estimated traffic density on several access routes for Mons.

consists of a map with the Voronoi cell structure superimposed. The cells are coloured based on the crowd density. The Voronoi cells containing the squares where events are taking place are further highlighted. A bar chart shows the count of the crowd for each of these squares. An additional bar chart shows the home country of international visitors along with an estimate of the total number of people in the city. A separate interactive map shown in figure 4 presents the roads of interest. Each road segment is coloured based on the estimated traffic density. Roads which share cells are coloured with 50% transparency to signify an average of the surrounding road segments was used to calculate its density.

The real-time application was used by the police and organisers to monitor the events during the opening ceremony of Mons as the European City of Culture 2015. Preliminary feedback has been positive and we are awaiting the results of a more detailed evaluation. We will also use the CDR data to determine the amount of time individuals spend at events and to understand the fine grained mobility occurring in the city. These will serve as indicators regarding the success of events while also furthering our understanding of urban processes.

Acknowledgments

This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant "SFI 13/IF/I2783". The authors would like to thank Mobistar for providing access to the anonymized data.

References

[1] F. Calabrese, F. C. Pereira, G. Di Lorenzo, L. Liu, and C. Ratti, "The geography of taste: analyzing cell-phone mobility and social events," in *Pervasive computing*. Springer, 2010, pp. 22–37.

[2] F. Calabrese, L. Ferrari, and V. D. Blondel, "Urban sensing using mobile phone network data: A survey of research," *ACM Computing Surveys (CSUR)*, vol. 47, no. 2, p. 25, 2014.

[3] 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*, vol. 36, no. 5, pp. 824–836, 2009.

[4] R. Cáceres, J. Rowland, C. Small, and S. Urbanek, "Exploring the use of urban greenspace through cellular network activity," in *Proc. of 2nd Workshop on Pervasive Urban Applications (PURBA)*, 2012.

[5] T. Sohn, A. Varshavsky, A. LaMarca, M. Y. Chen, T. Choudhury, I. Smith, S. Consolvo, J. Hightower, W. G. Griswold, and E. De Lara, "Mobility detection using everyday gsm traces," in *UbiComp 2006: Ubiquitous Computing*. Springer, 2006, pp. 212–224.

[6] C. Ratti, S. Sobolevsky, F. Calabrese, C. Andris, J. Reades, M. Martino, R. Claxton, and S. H. Strogatz, "Redrawing the map of great britain from a network of human interactions," *PloS one*, vol. 5, no. 12, p. e14248, 2010.

[7] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, 2008.

[8] S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, M. Martonosi, J. Rowland, and A. Varshavsky, "Identifying important places in peoples lives from cellular network data," in *Pervasive computing*. Springer, 2011, pp. 133–151.

[9] F. Calabrese, Z. Smoreda, V. D. Blondel, and C. Ratti, "Interplay between telecommunications and face-to-face interactions: A study using mobile phone data," *PloS one*, vol. 6, no. 7, p. e20814, 2011.

Inferring spatio-temporal changes in urban areas from mobile phone data

Sofia Nikitaki
NEC Laboratories Europe, Heidelberg, Germany
Sofia.Nikitaki@neclab.eu

Maurizio Dusi
NEC Laboratories Europe, Heidelberg, Germany
Maurizio.Dusi@neclab.eu

Mobile phone data represents a key source of information for understanding human mobility at a scale and, thank to its ubiquity, has shown that human trajectories exhibit a high degree of temporal and spatial regularity [1]. When we consider the mobility patterns of a set of individuals who share a given area over a common timeframe, we can model the dynamics behind the interactions of people in the area and within the neighbor areas, with several practical applications: public safety, e.g., crowd detection and planning in extreme situations [2]; urban and transportation planning, e.g., forecasting traffic congestions [3]; information diffusion, e.g., spreading of deseases and detecting economic flows [4].

In this paper we focus on the definition of a supervised method that, during a training phase, profiles the communication patterns of users in a given geographical area over time based on antennas' traffic information, in the form of Call Detail Records (CDR). A CDR is produced during a communication transaction and contains attributes such as the caller and callee identification number, date and time, type of service (e.g. voice, text), and duration. We then use such profiles to continuously monitor for future changes in the communication patterns over time within that area. The detection of changes can be correlated with third party sources of information such as news and social media, to identify the reasons of these changes.

The proposed method builds on the Non-Negative Matrix Factorization technique (NNMF) that considers non-negative observed data and explains the observations as a linear combination, represented by the coefficient matrix of a base. NNMF technique has been applied in order to discover latent structuring patterns of sensed data and their changes in time. To this aim, the traffic matrix is defined, that contains the captured information at a specific location over different time intervals. Specifically, NNMF technique solves an optimization problem in order to decompose the traffic matrix into a basis matrix and a coefficient matrix: the bases represent the latent factors, the coefficients the strength of these factors.

Figure 1 outlines the NNMF technique as applied to our case. In particular, we apply the technique to cell phone antenna data[1] collected in Côte d'Ivoire between December 2011 and April 2012 [5]. We split the dataset into ten subsets of two weeks chronologically ordered: we use the first L ($1 \leq L \leq 10$) sets to gather the profiles (*training set* T) and the remaining sets to test the ability in detecting changes in communication patterns by using such profiles (*evaluation set* E).

[1]This Data was made available by France Telecom/Orange Côte d'Ivoire within the framework of the D4D Challenge.

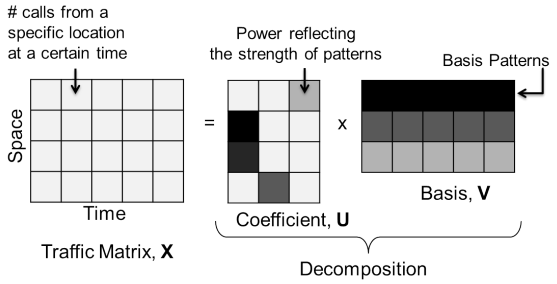


Figure 1: Non-Negative Matrix Factorization technique. The traffic matrix is decomposed into the coefficient matrix and the basis matrix. Each row of the basis matrix represents a communication pattern. Each component of the coefficient matrix represents the strength of each pattern.

In formulating the optimization problem needed to gather the communication patterns during the training phase, our method takes into consideration the spatial and temporal correlations of the calls of the users, to find global communication patterns of an area, which are stable over time. Specifically, we formulate the problem as follows:

$$\min\{\|X - U \cdot V^T\|_F^2 + \alpha(\|U\|_F^2 + \|V\|_F^2) + \\ + \beta(\|S(U \cdot V^T)\|_F^2 + \|(U \cdot V^T)T\|_F^2)\} \quad (1)$$

where $S(\cdot)$ is the spatial constrain, $T(\cdot)$ is the temporal constrains. Specifically, the spatial constraint is captured by creating the weighted communication graph of the matrix X and the matrix S is the Laplacian matrix of that graph. With respect to the temporal constraint, we apply the Toeplitz matrix that captures the temporal smoothness of the collected data. The Frobenius norm of a matrix $X \in \mathbb{R}^{m \times n}$ is defined as $\|X\|_F = (\sum_{i=1}^{m} \sum_{j=1}^{n} |x_{ij}|^2)^{1/2}$.

To cope with the scalability challenge given by the amount of data to consider, we solve the optimization problem by applying a distributed stochastic gradient descent technique, which allows for good convergence guarantees and can be distributed [6].

During the detection phase, we project the traffic matrix X_i of each subset of the *evaluation set* onto the basis matrix, and we estimate the values of the resulting coefficient matrix U_i. We then compute the pairwise differences of two consecutive coefficient matrixes to trigger an alert whenever

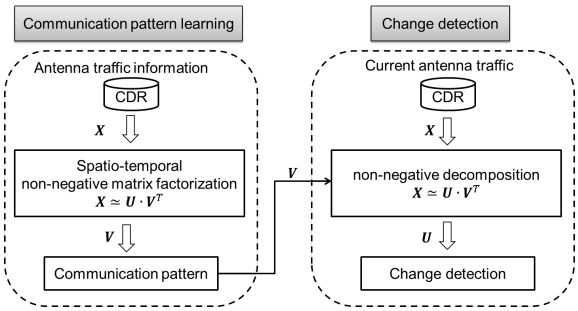


Figure 2: System architecture: overview.

a change is above a pre-defined threshold.

Figure 2 provides an overview of the system.

In the following we show the two main properties of the method: (i) it can gather profiles that are stable over time; and (ii) it can detect changes based on such profiles.

Stability of the bases over time. We start by solving the problem of Equation 1 for each one of the ten subsets that compose the dataset, ending up with ten basis matrix $\{V_i\}_{i=1,\ldots,10}$. We then normalize each row of V_i, so that the $\{V_i\}$ can be compared among each other.

We define as training set T the first L basis matrixes, where $1 \leq L \leq 10$, and as evaluation set E the remaining basis matrixes.

We compute the common basis V_c as the average of the $V_i \in T$:

$$V_c = \frac{\sum_{i=1}^{L} V_i}{L}, \tag{2}$$

and finally, evaluate the normalized reconstruction error defined as the normalized Frobenius norm between V_c and $\{V_i\} \in T$ and between V_c and $\{V_i\} \in E$. In the first case, the reconstruction error gives an indication on how good is the algorithm in finding the common bases; in the latter case, it indicates on how stable are the bases when comparing them to the call patterns of timeframe other than the ones used to gather the profiles.

Figure 3 reports on this, with a comparison between our method, which consider the spatio-temporal constrain among the calls across the antennas, and the basic NNMF.

Detecting spatio-temporal changes. We exploit the above technique to look for changes in call patterns that happen across the areas of Ivory Coast during the four-month period of the dataset. Here we report the changes spotted by our technique when considering as training the first four two-week periods (cell phone antenna data from December 2011 and January 2012), and evaluating the changes on the remaining dataset.

We project each traffic matrix X_i, $5 \leq i \leq 10$, onto the basis matrix V_c, computed from $\{V_i\}_{i=1,\ldots,4}$, thus ending up with the corresponding normalized coefficient matrixes $\{U_i\}_{i=5,\ldots,10}$, which contains the strenghts (proportions) of the communication patterns for each antenna. We then look for the antennas which exhibit changes in the proportion of the patterns.

Interestingly, we found that the biggest differences happen to clusters of neighbor antennas, as if the antennas reveal macro behavior that are common to a wide geographical area. Figure 4 reports the cluster where we register the

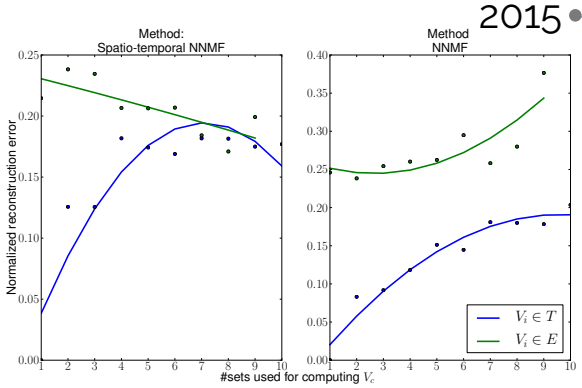


Figure 3: Stability of the common bases V_c. On the left, NNMF with spatio-temporal constrain. On the right, the basic NNMF. By adding the spatio-temporal constrain, the method exhibit a reconstruction error which tends to become similar in case V_i belongs either to the training or to the evaluation sets, in the order of 0.18. Conversely, when using basic NNMF, V_c appears to be more biased towards the V_i used for creating it, and diverges with the $\{V_i\}$ that belong to the evaluation set,$\{V_i\} \in E$.

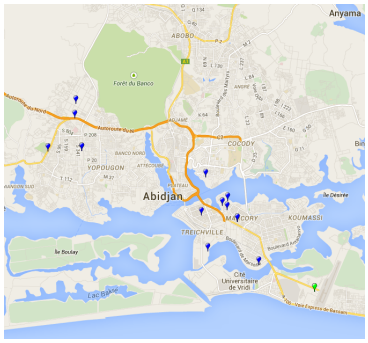


Figure 4: Clusters of antennas which exhibit the biggest differences in communication patterns during the firts half of February as revealed by our technique.

biggest differences during the first half of February, showing two clusters of antennas in the area around the city of Abidjan.

We further inspect the evolution in the communication patterns for the antenna that exhibits the biggest changes (the green antenna in Figure 4). As shown in Figure 5, during the second half of February the area covered by this antenna show a completely different predominant pattern with respect to the previous period: a shift from the diurnal pattern to a morning and nocturnal pattern emerges, and the technique spots it. The following period shows that the communication patterns in the area return to behave as before.

We believe that the proposed technique which allows for continuously monitoring changes in communication patterns presents several advantages. The method is based on a non-negative matrix factorization approach and accounts for the inherent correlation structure of the traffic matrix both in

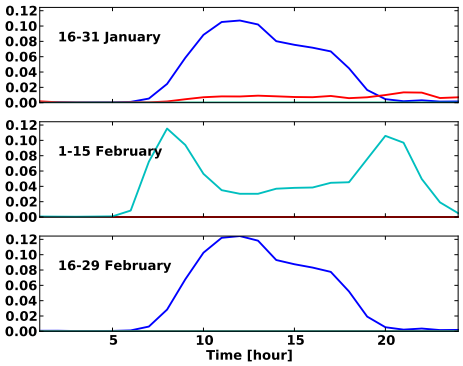


Figure 5: Evaluation of the communication patterns for the antenna that exhibits the biggest changes (green antenna in Figure 4).

time and space in order to find stable global communication patterns, as opposed to the basic formulation of the problem via NNMF. Furthermore, we use such patterns in order to detect for future changes in the communication profiles over time in a given area. Finally, the detection of changes can be associated with third party sources of information such as news and social media, to infer the reasons of these changes. The proposed method is solved via stochastic gradient descent that can be distributed and thus making it appropriate for large-scale learning data.

Acknowledgement

The research leading to these results has received funding from the European Union under the FP7 Grant Agreement n. 318627 (Integrated Project mPlane).

1. REFERENCES

[1] M.C. González, C.A. Hidalgo, and A.-L. Barabási, "Understanding individual human mobility patterns.," *Nature*, vol. 453, pp. 479–482, 2008.

[2] Bagrow JP, Wang D, and Barabsi A-L, "Collective response of human populations to large-scale emergencies.," *PLoS ONE*, vol. 3, pp. e17680, 6 2011.

[3] Chaogui Kang, Stanislav Sobolevsky, Yu Liu, and Carlo Ratti, "Exploring human movements in singapore: A comparative analysis based on mobile phone and taxicab usages," in *Proceedings of the 2Nd ACM SIGKDD International Workshop on Urban Computing*, New York, NY, USA, 2013, UrbComp '13, pp. 1:1–1:8, ACM.

[4] Francesco Calabrese, Massimo Colonna, Piero Lovisolo, Dario Parata, and Carlo Ratti, "Real-time urban monitoring using cell phones: A case study in rome.," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 1, pp. 141–151, 2011.

[5] Vincent D. Blondel, Markus Esch, Connie Chan, Fabrice Clérot, Pierre Deville, Etienne Huens, Frédéric Morlot, Zbigniew Smoreda, and Cezary Ziemlicki, "Data for development: the D4D challenge on mobile phone data," *CoRR*, vol. abs/1210.0137, 2012.

[6] Rainer Gemulla, Erik Nijkamp, Peter J. Haas, and Yannis Sismanis, "Large-scale matrix factorization with distributed stochastic gradient descent," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2011, KDD '11, pp. 69–77, ACM.