NetMob 2015

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

Book of Abstracts :: Posters

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

March 26th version

Organized by

Universidad Carlos III de Madrid

MIT

UCL Université catholique de Louvain

Sponsored by

orange™

RealImpact ANALYTICS

NetMob 2015

Contents

Poster Session 1 :: April 8

Extrapolation in a Big Data Environment: Detect Home Locations in a Continuous Stream of Location Data

Dr. Hendrik Wagenseil, GfK Geomarketing, Nuremberg, Germany
Markus Ziegler, GfK Marketing & Data Sciences, Nuremberg, Germany

Framework and Objective

In recent years, a lot of effort has been put into making use of mobile phone data in the field of Marketing in order to solve business questions. Therefore, GfK, one of the biggest market research companies in the world, has created Location Insights as an independent field of research and, hence, partnered with major carriers in several countries in order to capitalize the spati-temporal content that is inherent in mobile network data.

In this context, GfK receives anonymized tri-laterated location data derived from mobile phone data on individual level, and is fully responsible for the entire data management, data processing and reporting. One of the key methodological questions is how to extrapolate from this data to the full population of a country in order to serve specific business needs, e.g. estimating how many visitors were at a given location. Such information is widely regarded as extremely valuable to a variety of clients, e.g. retailers, OOH media agencies, transportation agencies or urban planning. One of the key requirements for any extrapolation approach is a proper estimation of the place of residence of the devices.

Commercial projects using mobile phone data delivered by a carrier usually face specific challenges. First and foremost, the continuous stream of location data poses a major challenge for data transfer, storage and processing. It is important to install approaches which allow for the processing of millions of device data in a manageable amount of time. Additionally, it is in the best interest of both, operator and commercial partner, to make sure that all privacy legislation is strictly adhered to even if this leads to shorter device lifetimes or a smaller sample to work with. Losing the trust of any device owner or the public can result in major turmoil for future business relationships. Moreover, the value of the results is heavily dependent on the accuracy of the data. Therefore, working with the data always means taking into the account the fact that location data is affected by measurement error. Finally, a proper extrapolation approach is closely linked to overcoming biases in the data. This contains aspects like differing market shares of an operator across country and customer segments or the influence of network technologies on the amount of locations created. All these aspects set the framework for the analytical processes of analyzing the data. Hence, dealing with location data in such a setting requires algorithms which are efficient, easy to implement in a big data environment, and easy to adjust to the boundaries set by the data. The estimation of the home location is the first step of the journey to extrapolating results of device movements. It allows for the linkage of population information and device counts and, as a result, paves the way for the implementation of an extrapolation factor per device.

Methodology

In order to estimate the home location for all devices, GfK has developed a two-step Bayesian estimation process which allows accounting for uncertainty in the data while linking them to information about the universe. In accordance with Isaacman (2011) the estimation of the home location is based on sightings

occurring between 7pm and 7am only.
The first step aims at accounting for measurement error in the data. This is achieved by distributing all relevant locations of a single device to a 100 x 100 meter grid depending on the uncertainty level of every record. As a result, relevant areas on the map can be identified which could serve as potential home areas.
In the second step, coherent clusters are determined with the most likely being chosen for further calculations. Linking the grid information with census information, each grid cell can be assigned a population density value which in turn is used as the prior information in the Bayesian approach. The posterior is the probability of a given cell being the actual home location. Aggregating these cell probabilities allows linking back to the census units and determining the respective home probabilities.

Conclusion

The proposed two step approach allows to link location data of a mobile device with any kind of census information while accounting for uncertainty in the location measures. Our talk will outline the main challenges of doing research with location data derived from mobile network from a commercial point of view, describe the core ingredients of our algorithm and present some validation results based on a large sample of devices demonstrating the feasibility of the approach over time.

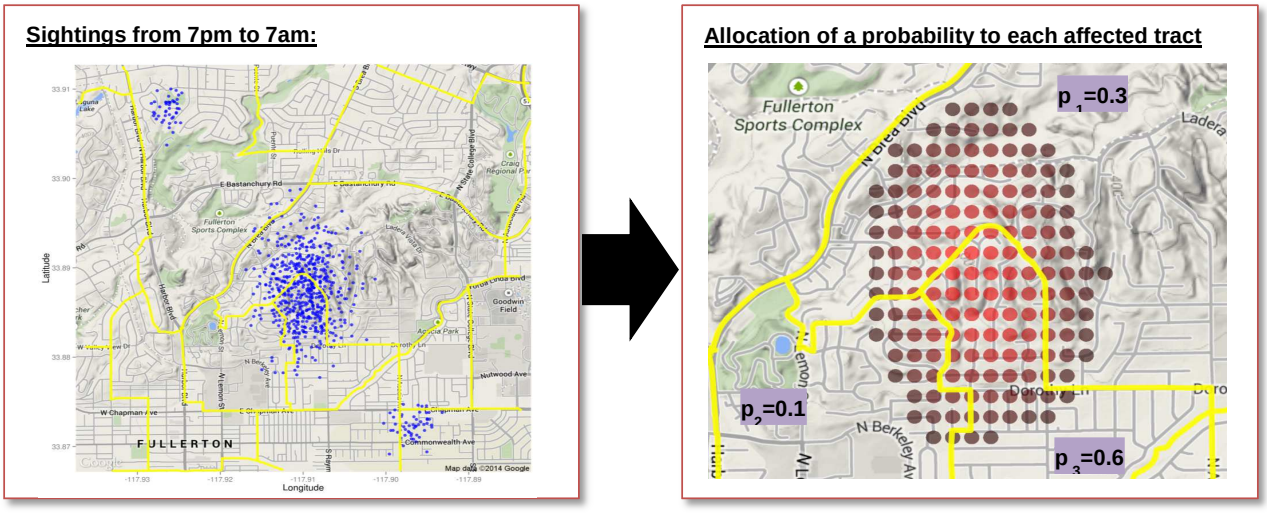


Figure 1: Estimation of probabilities for the home location of one device

References

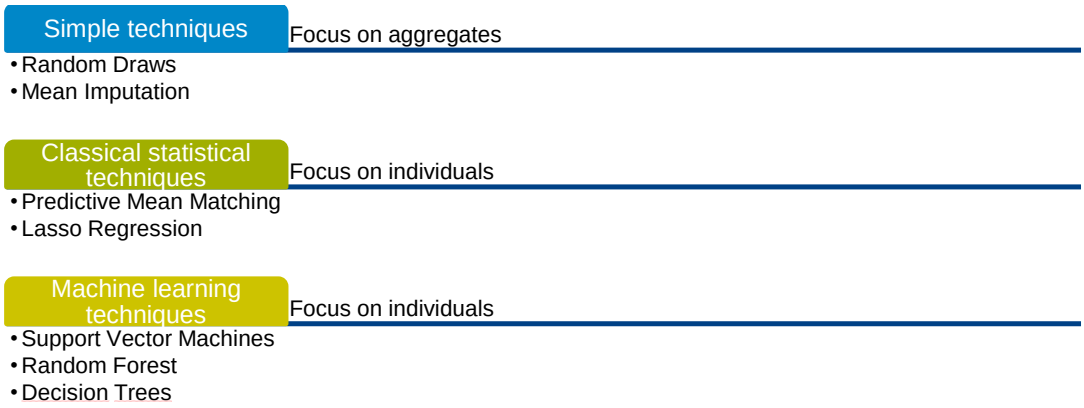
Isaacman, Sibren; Becker, Richard; C´aceres, Ram´on; Kobourov, Stephen; Martonosi, Margaret; Rowland, James; Varshavsky, Alexander (2011): Identifying important places in people's lives from cellular network data. Pervasive Computing, p. 133–151

"Classic meets Machine Learning"

Our world is increasingly online: mobile devices such as tablets or smartphones have become the natural way to be online 24/7, our entry point to a myriad of services, shops, entertainment and information available whenever we need them.

Within different projects, cooperation with mobile carriers allows GfK Marketing and Data Science to analyze data of mobile devices. The challenges in measuring information in mobile context are manifold. Most of these challenges come from technical and / or legal limitations; others are simply due to the sheer size. In some projects mobile data focuses on internet traffic information left in a carrier´s cellular network. The large scale of the data enables to derive app and browsing behavior as well as the long tail and spatial information. The opportunities this presents brands, advertisers and marketers is huge, and to truly maximise them, it's essential to understand how mobile devices are used, and especially who is using them and when. In particular we are interested in characterizing those populations by socio-demographic profiles, by consumer-related attitudes and behaviors, which is within carriers CRM information partly or not available. Being able to enrich those data would add enormous value to the data.

There are different data source and methods, which might be used to enrich those mobile data and predict individual characteristics. More specifically, **an individual enrichment** by using a survey or panel data could be implemented. Hence, we worked on approaches to enrich those mobile data with further information by using different integration methods. A short overview is given in following graphic.



The paper examines these techniques and their prediction performance for specific set of variables. The comparison is done by an internal validation and therefore a GfK recruited opt-in panel of the size of about 3,000 respondents is used. The cooperating carrier provided a month of individual mobile data.

The presentation will demonstrate the process of investigating different approaches by enriching a huge data set and summarizes findings, experiences and gives a recommendation on how to enrich mobile data. The presentation will focus on a limited set of variables showing findings and results on a few selected examples.

Social Networks, Ethnicity, & Political Accountability

Extended Abstract for NetMob 2015

Nicholas Eubank*

January 10, 2015

Network structure impacts political outcomes by shaping the ability of citizens to interact, organize, and share information with one another. This paper uses a novel source of data – six months worth of detailed and geocoded telecommunications data from a cell phone provider in Zambia – to test the relationship between social network properties and the capacity of citizens to engage in social sanctioning, collective action, and to share information about incumbent politicians. This is accomplished by first relating each of these dynamics to different social network structures – transitivity, community fragmentation, and rates of information diffusion – and second, by separately measuring the properties of the social networks of residents of each of Zambia's 150 National Assembly electoral districts and more than 1,000 Local Council electoral districts. In addition, data is used to examine the relationship between co-ethnicity and social network proximity, testing the hypothesis that ethnic fragmentation is correlated with poor development outcomes because ethnically fragmented communities suffer from fragmented social networks.

While it is widely believed in Economics and Political Science that networks affect political outcomes, to date, we have very little empirical evidence on what these inter-citizen networks look like, and how these networks correlate with political outcomes of interest. This paper advances this literature using cell-phone meta-data from more than 9 million Zambians to directly measure the structure of social networks across the entire country and relate these structures to social, economic, and political outcomes of interest.

*PhD Candidate in Political Economy, Stanford Graduate School of Business. nickeubank@stanford.edu. The author is deeply indebted to *Real Impact* and in particular Chief Data Scientist Gautier Krings for extensive support and assistance with this project.

Systematic data on inter-citizen interactions has historically been difficult to come by, especially in developing countries. As a result, existing work has been forced to proxy for inter-citizen networks, either by measuring institutional affiliations [2, 17] or by assuming individuals with similar demographic characteristics (like ethnicity) also belong to common networks [6, 9, 13, 14]. Although reliance on these proxies in the past is understandable, neither proxy captures the complexity of social relations. Institutional affiliations neglect the informal ways in which people might be connected, and ethnicity measures neglect heterogeneity in connectedness within groups. Moreover, these proxies also fail to distinguish between the many distinct properties that fall under the general label of "connectedness" – like the average network distance between individuals, or the degree to which people share common friends. While related, these types of properties are both empirically and theoretically distinct, and may relate differently to different social phenomena.

This paper begins to fill this gap using a novel source of data – namely, six months of detailed telecommunications data on 9 million subscribers from a cell phone provider in Zambia.[1] This data – which is comprised of almost 2 billion telecommunication transactions – is used to test three mechanisms by which networks might shape the ability of citizens to: (1) socially sanction free-riders, (2) organize large groups for political action, and (3) share information about politician activities. In each case, this paper attempts to tie variation in established social phenomena – like the effectiveness of social sanctioning – to observable, measurable properties of social networks – like the degree to which individuals share mutual friends.

The first phenomenon examined is social sanc-

[1] The population of Zambia in 2010 was approximately 12.5 million, suggesting excellent data coverage.

tioning. Social sanctioning is the application of social pressure to induce individuals to participate in pro-social activities like political protests or community projects. Social sanctioning is core to many models of citizen-interactions and and is a key strategy for overcoming the collective action problem [15]. Yet the ability of citizens to social sanction is often simply assumed [13, 3, 4, 1]. This paper presents a novel model of social sanctioning which expands on the work of [10], [6] and [12] to explain both the *ability* of individuals to sanction one another and also the *potential strength* of social sanctioning based on a specific social network property – the degree to which people share mutual friends. This prediction that the degree to which individuals share mutual friends ("transitivity") should be positively associated with measures of political accountability and its consequences is then tested.

The second phenomenon examined is the how network structure impedes or facilities coordinating citizens to engage in collective action. Organizing large groups for political action requires finding a commonly-attractive message, managing logistics, and sharing information among many people. This is difficult in any setting, but this paper argues that some network configurations – those in which people are organized into relatively small, internally well-connected but disparate groups – may make coordination especially challenging. This theoretical formulation gives rise to the prediction that measures of network fragmentation[2] should be associated with lower levels of coordinated mass political activity, like protests.

Finally, this paper examines how network structure affects the diffusion of information within a community. Numerous studies have examined the impact of mass media on political accountability [16, 7, 18], but this study aims to fill an important gap in our understanding of *informal* information sharing, which is likely to be of particular importance in the developing country context where formal media may only be consumed by a limited subset of the population or may not a reliable source of objective information.[3] It does so by simulating information diffusion in the actual networks of Zambians in different electoral districts, and then examines how simulated diffusion rates correlate with levels of political knowledge among voters.

This investigation has the potential to not only improve our understanding of the importance of patterns of inter-citizen interaction, but also directly investigate a proposed explanation for the negative correlation between ethno-linguistic fractionalization (ELF) and development outcomes. In recent years, studies of voting behavior [5, 11], field studies of social sanctioning [13], and even lab studies of ethnic preferences [9, 8] have all pointed to the possibility that (1) co-ethnicity (belonging to the same ethnicity) may actually be a proxy for social network proximity, and (2) that the reason ELF and poor development outcomes are correlated may be that ethnically homogenous communities have integrated social networks in which citizens are better able to monitor and (when necessary) socially sanction one another. This explanation would have the potential to rationalize a diverse set of findings, including the fact that voters in African elections can be better characterized as voting for co-ethnic *individuals* than co-ethnic *parties* [11], and that in lab experiments subjects act in a manner consistent with a fear of social sanctioning, "discriminat[ing] in favor of co-ethnics *if and only if* they can be seen to be doing so." [8, p. 721, emphasis in original]. To date, the link between ethnicity and network structure has not been directly tested empirically. This paper begins to fill this gap by pairing data on network structures with highly disaggregated data on ethnic fractionalization in Zambia to provide one of the first systematic tests of whether ELF is correlated with network structure, as these theories require.

By combining novel geo-referenced social network data with information on social, economic, and political outcomes, this research will ground well-established theories about inter-citizen dynamics in robust empirics. Moreover, this research offers the promise of opening new avenues of research on the effect of social network structure on political outcomes.

[2] This is operationalized as one minus the Herfandahl index of inductively determined modularity-optimizing community structures.

[3] Indeed, Zambia's media has been rated as "Not Free: Not possible to safely criticize government or government officials; government exerts indirect control over media" for most of the last decade [19].

References

[1] George A Akerlof. A Theory of Social Custom, of Which Unemployment May be One

Consequence. *The Quarterly Journal of Economics*, 94(4):749, June 1980.

[2] S. Berman. Civil society and the collapse of the Weimar Republic. *World Politics*, 49(03):401–429, 1997.

[3] T Besley and S Coate. Group lending, repayment incentives and social collateral. *Journal of Development Economics*, 1995.

[4] Timothy Besley, Stephen Coate, and Glenn Loury. The Economics of Rotating Savings and Credit Associations. *The American Economic Review*, 83(4):792–810, September 1993.

[5] Kanchan Chandra. *Why Ethnic Parties Succeed*. Patronage and Ethnic Head Counts in India. Cambridge University Press, 2007.

[6] J.D. Fearon and D.D. Laitin. Explaining interethnic cooperation. *American Political Science Review*, pages 715–735, 1996.

[7] C. Ferraz and F. Finan. Exposing Corrupt Politicians: The effects of Brazil's publicly released audits on electoral outcomes. *The Quarterly Journal of Economics*, 123(2):703, 2008.

[8] J. Habyarimana, M. Humphreys, Daniel N Posner, and J.M. Weinstein. Why does ethnic diversity undermine public goods provision? *American Political Science Review*, 101(4):709, 2007.

[9] James Habyarimana, Macartan Humphreys, Daniel N Posner, and Jeremy M Weinstein. *Coethnicity: Diversity and the Dilemmas of Collective Action*. Russell Sage Foundation, August 2009.

[10] Michihiro Kandori. Social Norms and Community Enforcement. *The Review of Economic Studies*, 59(1):63–80, January 1992.

[11] Philip Keefer. The Ethnicity Distraction: Political Credibility, Clientelism and Partisan Preferences in Africa. *The World Bank*, 2009.

[12] Jennifer M Larson. Cheating Because They Can: Social Networks and Norm Violators. *Workshop on Conflict and*, 2014.

[13] Edward Miguel and Mary Kay Gugerty. Ethnic diversity, social sanctions, and public goods in Kenya. *Journal of Public Economics*, 89(11-12):2325–2368, December 2005.

[14] K Munshi and M Rosenzweig. Networks, Commitment, and Competence: Caste in Indian Local Politics. *Unpublished manuscript*, 2010.

[15] Elinor Ostrom. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press, November 1990.

[16] Rohini Pande. Can Informed Voters Enforce Better Governance? Experiments in Low-Income Democracies. *Annual Review of Economics*, 3(1):215–237, September 2011.

[17] Robert D Putnam, Robert Leonardi, and Raffaella Nanetti. *Making democracy work*. civic traditions in modern Italy. Princeton Univ Pr, May 1994.

[18] Ritva Reinikka and Jakob Svensson. The power of information in public services: Evidence from education in Uganda. *Journal of Public Economics*, 95(7-8):956–966, August 2011.

[19] Jenifer Whitten-Woodring and Douglas A Van Belle. *Historical Guide to World Media Freedom: A Country-by-Country Analysis*. Sage / CQ Press, 2014.

Discovering dependence of tweet inter-arrival times

Balázs Gerencsér* Christophe Cloquet* Vincent Blondel*

January 12, 2015

1 Introduction

The temporal patterns of human communication behavior has been an active field of research. It has been found that the event inter-arrival times follow a scale-free distribution [1], therefore the process of inter-arrival sequences is often modeled as an i.i.d. process of the appropriate scale-free distribution [2].

The goal of this work is to draw attention to the dependence appearing in this process. We claim that we get a much better fit and consequently a more precise understanding if we allow subsequent inter-arrival times to be dependent.

In this pilot project we use Twitter data of Belgium and propose a very simple model of dependence, already showing a considerably better fit to the data.

2 Model description

In this section we describe the model we use for the tweet inter-arrival times W_i of a single user. First of all, we distinguish "short" and "long" waiting times depending on whether W_i exceeds some threshold T_S. This property will be recorded by X_i which can take values S or L.

The key point of the model is the dependence structure we propose. We assume that the distribution of W_i, X_i depends on the type X_{i-1} of the previous waiting time, but is conditionally independent from anything before. Therefore the type X_i is determined using a transition probability matrix:

$$P = \begin{pmatrix} p_{S|S} & p_{S|L} \\ p_{L|S} & p_{L|L} \end{pmatrix},$$

where $p_{S|L}$ stands for the probability of observing a short waiting time after a long one. Similarly, the long waiting times follow the scale free distribution $f_{|S}$ or $f_{|L}$ (parametrized by γ_S, γ_L) depending on whether X_{i-1} is S or L. For short waiting times we disregard the exact value.

We get a reference model of independent inter-arrival times if we assume X_i are i.i.d. and $\gamma_S = \gamma_L$. This means the inter-arrival times W_i follow an i.i.d. process of scale-free variables.

3 Data source

The data we use is gathered using the *Twitter Streaming API* which gives public access and allows to fetch tweets as long as the amount downloaded does not exceed the 1% of the total number of tweets during the actual period.

We focused on geotagged tweets originating from Belgium, therefore the 1% limit was very far from being reached. The tweets were gathered between December 2013 and April 2014. We also applied several filters to

*B. Gerencsér, C. Cloquet and V. Blondel are with ICTEAM Institute, Université Catholique de Louvain, Belgium balazs.gerencser@uclouvain.be, c.cloquet@gmail.com and vincent.blondel@uclouvain.be Their work is supported by the DYSCO Network (Dynamical Systems, Control, and Optimization), funded by the Interuniversity Attraction Poles Programme, initiated by the Belgian Federal Science Policy Office, and by the Concerted Research Action (ARC) of the French Community of Belgium.

remove spammers and other non-human users so that our study really focuses on human tweeting behavior. We select users with at least 1000 tweets during this period so that it is meaningful to discuss inter-arrival times.

4 Results

In order to justify the model we proposed, we statistically compare it with the independent model. We fix the threshold to be $T_S = 60s$.

We perform two tests for comparison. We analyzed which model performed the best, be it (a) for individual users or (b) for all users.

For both tests we use the standard likelihood-ratio test. This means we perform a Maximum Likelihood fitting of both models, we call the resulting likelihoods L_{indep} and L_{new}. Assuming the null hypothesis that the independent model is correct, we have that the difference

$$D = 2 \log \frac{L_{new}}{L_{indep}}$$

follows a chi-squared distribution with degrees of freedom equal to the number of extra parameters in the new model [3]. This is 2 for individual users and $2N$ for the overall comparison, where $N = 1315$, the number of users. We reject the null hypothesis and claim that the new model fits better if D is unreasonably large for the actual chi-square distribution.

For the individual users we accept the new model if D has a p-value at most 0.05. We find that out of the 1315 users, we accept the new model for 997 of them, which is over 75%. For the overall comparison, we get $D_{all} \approx 43239$. Assuming the null hypothesis this should be drawn from a chi-squared distribution with degree of freedom $2N = 2630$, but this has an extremely small p-value (0 up to machine precision). In other words, for an overall view of inter-arrival times it is absolutely clear that the new model should be preferred to the simple, independent one.

5 Conclusions and outlook

In this work we have shown that introducing a simple dependence structure substantially improves the quality of fit to Twitter inter-arrival data.

We understand this result as a first step of better understanding this type of temporal data. The model should be tested for other, similar scenarios, primarily on mobile phone datasets. Moreover, there are several ways to refine the dependence structure currently proposed and these should be explored to get a better insight on the behavior of such processes.

References

[1] J. Candia, M. C. González, P. Wang, T. Schoenharl, G. Madey, and A.-L. Barabási, *Uncovering individual and collective human dynamics from mobile phone records*, Journal of Physics A: Mathematical and Theoretical, 41 (2008), p. 224015.

[2] M. Karsai, M. Kivelä, R. K. Pan, K. Kaski, J. Kertész, A.-L. Barabási, and J. Saramäki, *Small but slow world: How network topology and burstiness slow down spreading*, Physical Review E, 83 (2011), p. 025102.

[3] S. S. Wilks, *The large-sample distribution of the likelihood ratio for testing composite hypotheses*, The Annals of Mathematical Statistics, 9 (1938), pp. 60–62.

A Unified Model for Individual Spatial Temporal Mobility Patterns

Yingxiang Yang, Marta González

Statistical models that can characterize human mobility patterns are of importance in a broad range of research areas from epidemiology, transportation engineering, to urban planning and mobile network communication [1, 2, 3, 4, 5, 6, 7, 8]. Nowadays, even though detailed transportation simulation platforms can already mimic realistic travel behaviors [9, 10, 11], simple statistical models are still needed not only because the detailed information required for the calibration of simulation platforms are usually not available, but also because these statistical models are amenable to mathematical analysis that quantifies the influence of each parameter.

Ubiquitous findings observed in data from human mobility can be expressed by five statistical distributions on the urban population, which are repeatedly found in previous works [12, 13, 14, 15, 16, 17, 18, 19, 20, 21]. These are: the frequency of visits to each location over long term observations, the number of locations visited each day, the duration of the visits, the distance traveled per trip, and the temporal span between visits to the same location. An important research question is to not only include these patterns in a mobility model, but also explain the pervasive mechanisms that produce the observed distributions in diverse scenarios. In this study we model these spatial and temporal human mobility patterns and integrate them into a unified model using stay locations extracted from two million mobile phone users' Call Detailed Records in the Boston metro area.

Spatially, we model the heterogeneously distributed activity locations, i.e, the trip origins(S) and destinations (D) for the entire population, using a bivariate multiplicative cascade model. As is shown by Fig. 1(a), we repetitively divide the entire study area Ω_0 into 2×2, 4×4,… grids. Both the spatial clustering of origins and destinations, and the degree of attraction or repulsion between them at each granularity level i are controlled by bivariate random variables $[W_{D_i}, W_{S_i}]$. The influences of $[W_{D_i}, W_{S_i}]$ at different granularity levels combine multiplicatively. Such heterogeneously distributed activity locations, when combined with a rank based location selection model (Fig. 1(b)), could give rise to heavy-tailed trip length distributions, which is a ubiquitous characteristic found not only in human movement, but also various other systems.

Temporally, stay duration is modeled by a time-inhomogeneous Markovian model that includes both the circadian pattern and burstness of travelling (Fig. 2(a, b)). The circadian pattern is captured by time dependent travel probability at each time step, while the burstness of travelling is captured by distinguishing the steady state, which means being at home, from the active state, which means being at other places. In the proposed temporal model, both the number of daily visited location distribution and the stay duration distribution could be analytically derived using Markov chain embedding technique. They both compare well with the observed distributions from cell phone data, as is shown in Fig. 2(c, d). The long term visitation pattern is incorporated by the exploration and preferential return mechanism, shown in Fig. 2(e).

At last, we show that these components could be integrated into a combined model. In the combined model, each component is controlled by only a few parameters. With these easily tunable parameters, we are ready to evaluate the influence of individual movement patterns on a broad range of research areas including epidemics spreading, wireless network routing, and transportation planning.

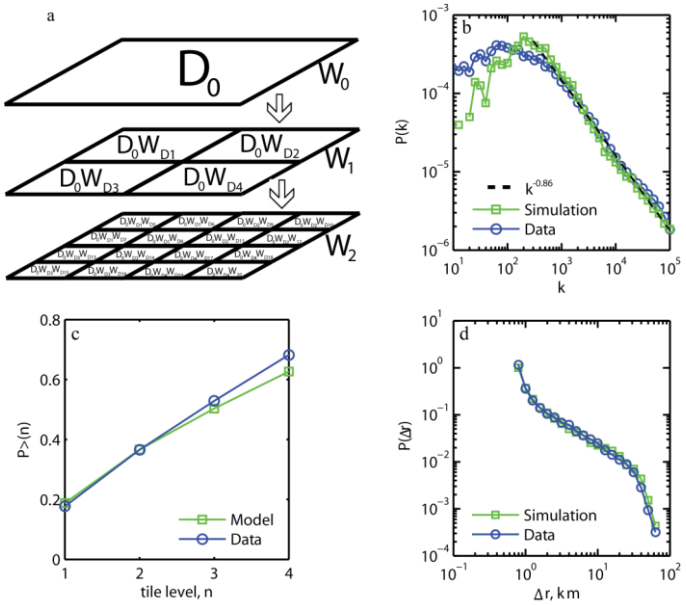


Fig. 1 (a) The scheme of the hierarchical cascade process for generating trip origin density D. Each tile is repetitively divided into 4 smaller tiles. The location density in each tile is controlled by the cascade generator W. (b) The probability to choose the rank k location as the destination is $P(k)\sim k^{-0.86}$. The closest potential destination is rank 1. (c) The probability to travel outside tile level n. The tile size of level 1 is 24 km, level 2 is 12 km. (d) The trip distance distribution $P(\Delta r)$ observed from trips generated using stay locations extracted from cell phone data, and from the simulation of the proposed model.

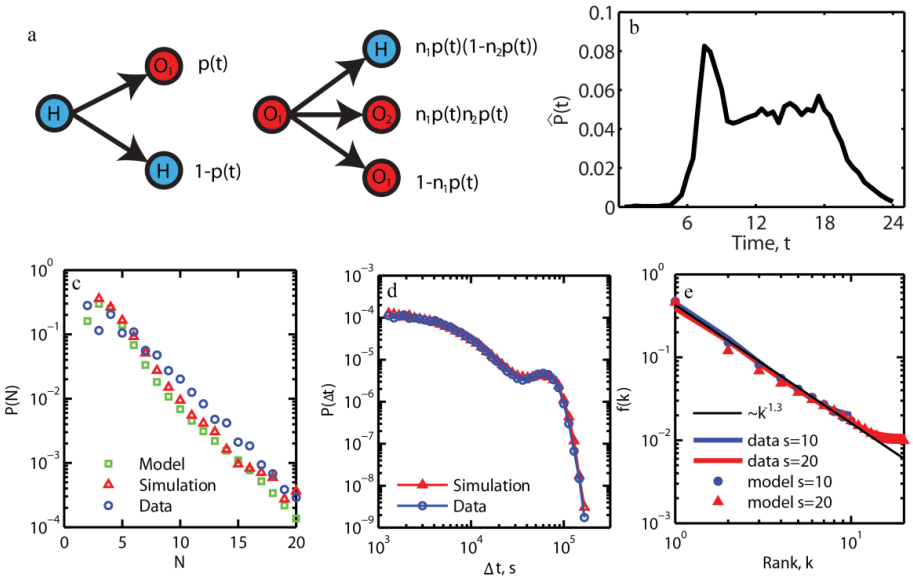


Fig. 2 (a) Illustration of different choices for a non-commuting person when the person is at "home" or "other" location. (b) The time dependent periodic transition probability $\hat{p}(t)$. (c) Number of daily visited location distribution $P(N)$ measured from stays extracted from cell phone data, the model's simulation result, and the model's analytical result. (d) Activity duration distribution $P(\Delta t)$ measured from stays extracted from cell phone data and the model's simulation result. (e) Visiting frequency $f(k)$ to the k^{th} most visited location follows $f(k)\sim k^{-1.3}$ for users visiting different number of locations in the observation period.

Reference

[1] L. Song, D. Kotz, R. Jain, and X. He, "Evaluating next-cell predictors with extensive wi-fi mobility data," *Mobile Computing, IEEE Transactions on*, vol. 5, no. 12, pp. 1633–1649, 2006.

[2] Y. Zheng and X. Zhou, "Computing with spatial trajectories," 2011.

[3] C. F. Choudhury, M. Ben-Akiva, and M. Abou-Zeid, "Dynamic latent plan models," *Journal of Choice Modelling*, vol. 3, no. 2, pp. 50–70, 2010.

[4] M. Ben-Akiva and M. Bierlaire, "Discrete choice methods and their applications to short term travel decisions," in *Handbook of transportation science*. Springer, 1999, pp. 5–33.

[5] P. Waddell, "A behavioral simulation model for metropolitan policy analysis and planning: residential location and housing market components of urbansim," *Environment and Planning B*, vol. 27, no. 2, pp. 247–264, 2000.

[6] M. Batty, *The New Science of Cities*. MIT Press, 2013.

[7] D. Brockmann, L. Hufnagel, and T. Geisel, "The scaling laws of human travel," *Nature*, vol. 439, no. 7075, pp. 462–465, 2006.

[8] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, 2008.

[9] M. Balmer, "Travel demand modeling for multi-agent transport simulations: Algorithms and systems," Ph.D. dissertation, ETH Zurich, 2007.

[10] K. W. Axhausen, A. Zimmermann, S. Schönfelder, G. Rindsfüser, and T. Haupt, "Observing the rhythms of daily life: A six-week travel diary," *Transportation*, vol. 29, no. 2, pp. 95–124, 2002.

[11] N. Schuessler and K. W. Axhausen, "Processing raw data from global positioning systems without additional information," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2105, no. 1, pp. 28–36, 2009.

[12] D. Balcan, V. Colizza, B. Gonçalves, H. Hu, J. J. Ramasco, and A. Vespignani, "Multiscale mobility networks and the spatial spreading of infectious diseases," *Proceedings of the National Academy of Sciences*, vol. 106, no. 51, pp. 21484–21489, 2009.

[13] A. Wesolowski, N. Eagle, A. J. Tatem, D. L. Smith, A. M. Noor, R. W. Snow, and C. O. Buckee, "Quantifying the impact of human mobility on malaria," *Science*, vol. 338, no. 6104, pp. 267–270, 2012.

[14] W.-j. Hsu, T. Spyropoulos, K. Psounis, and A. Helmy, "Modeling time-variant user mobility in wireless mobile networks," in *INFOCOM 2007. 26th IEEE International Conference on Computer Communications. IEEE*. IEEE, 2007, pp. 758–766.

[15] J. Ghosh, S. J. Philip, and C. Qiao, "Sociological orbit aware location approximation and routing (solar) in manet," *Ad Hoc Networks*, vol. 5, no. 2, pp. 189–209, 2007.

[16] S. Petrovskii and A. Morozov, "Dispersal in a statistically structured population: fat tails revisited," *The American Naturalist*, vol. 173, no. 2, pp. 278–289, 2009.

[17] X.-P. Han, Q. Hao, B.-H. Wang, and T. Zhou, "Origin of the scaling law in human mobility: Hierarchy of traffic systems," *Physical Review E*, vol. 83, no. 3, p. 036117, 2011.

[18] K. Lee, S. Hong, S. J. Kim, I. Rhee, and S. Chong, "Demystifying levy walk patterns in human walks," *CS Dept., NCSU, Raleigh, NC, USA, Tech. Rep*, 2008.

[19] R. Hidalgo and A. César, "Conditions for the emergence of scaling in the inter-event time of uncorrelated and seasonal systems," *Physica A: Statistical Mechanics and its Applications*, vol. 369, no. 2, pp. 877–883, 2006.

[20] R. D. Malmgren, D. B. Stouffer, A. E. Motter, and L. A. Amaral, "A poissonian explanation for heavy tails in e-mail communication," *Proceedings of the National Academy of Sciences*, vol. 105, no. 47, pp. 18153–18158, 2008.

[21] A. Noulas, S. Scellato, R. Lambiotte, M. Pontil, and C. Mascolo, "A tale of many cities: universal patterns in human urban mobility," *PloS one*, vol. 7, no. 5, p. e37027, 2012.

Estimating the Wealth of an Individual Based on Individual Patterns of Phone Use

Joshua E. Blumenstock
University of Washington
Seattle, WA
joshblum@uw.edu

ABSTRACT

We provide evidence that mobile phone records can be used to predict the socioeconomic status and other welfare indicators of *individual* mobile phone subscribers. Combining several terabytes of anonymized transactional mobile phone records with data collected through 2,200 phone-based interviews, we test the extent to which it is possible to predict an individual's responses to survey questions based on phone records alone. We observe significant correlations between asset ownership and a rich set of measures derived from the phone data that capture phone use, social network structure, and mobility.

Simple classification methods are able to predict, with varying degrees of accuracy, whether the respondent owns assets such as radios and televisions, as well as fixed household characteristics such as access to plumbing and electricity. More modest results are obtained when attempting to predict a broader set of development indicators such as an individual's response to the question, "Have you had to pay unexpected medical bills in the past 12 months?" While these methods offer a powerful opportunity for policymakers and researchers working in developing countries, we argue that considerable calibration and refinement is needed before such methods can be deployed.

Keywords

Mobile phones, development, big data, call detail records, wealth, mobility, social networks, regression

1. INTRODUCTION AND MOTIVATION

Reliable, quantitative data is a critical input to development policy, social science research, and to the decision-making process of firms and organizations interested in promoting social good. However, the basic measurement of key development outcomes – such as poverty, physical security, and happiness – is notoriously difficult in developing countries, where a lack of physical infrastructure and resources is often compounded by market failures and fragile institutional capacity [10].

Such problems are exacerbated in fragile and conflict-affected regions, where concerns over corruption and the physical security of enumerators and respondents make the regular collection of representative household survey data all but impossible. For example, Angola's last census was in 1970, and covered just 18 districts [11]. As a result, researchers and policymakers typically rely on data from large-scale national surveys (which occur infrequently), or specialized panel survey modules (which are typically administered to small, local populations). Neither traditional source captures fine-grained variation in development outcomes over both space and time.

In this paper, we describe preliminary results from efforts to develop models for predicting an *individual's* socioeconomic status and related development outcomes based upon anonymous, high-frequency data passively registered through use of mobile phone networks. A key innovation of this approach is our ability to link individual survey responses collected in phone interviews with incredibly rich social network and communication data obtained from mobile phone operators. Such an approach can be used to model the relationship between passively collected metrics of mobile phone use and explicitly queried socioeconomic phenomena. For instance, it will be possible to tell whether an individual's communication history can be used to predict whether that individual agrees with a survey-based statement such as, "*I believe the current economic situation will improve in the coming year*," or "*I feel connected to other members of my local community.*"

Here, we focus on results from the analysis of data collected in Rwanda in 2009 and 2010. This work extends a previous workshop paper that used a simple regression model to illustrate the strong relationships between simple metrics of phone use and a composite indicator of socio-economic status [1]. To our knowledge, no other prior work has investigated the relationship between individual communication histories and individual development outcomes. However, a series of recent studies have shown that geographically-aggregated communication records are strong predictors of regional census data [5, 8, 9]. A closely related set of work uses individual phone records to model gender and related (fixed) demographic characteristics [4, 7]. These approaches are strongly complementary, and we expect that over the next several years these methods will significantly advance our ability to measure, model, understand, and improve the lives of historically marginalized populations.

Sample Y_{it} Indicator ("development outcome")	Sample X_{it} Indicator ("feature")
Household owns a motor vehicle	Average number of outgoing phone calls per day
Amount of land owned by individual	Number of unique contacts in social network
Recent illness or other negative economic shock	Number of geographic regions visited in past month
Total expenditures in last month	Total expenditures on mobile phone-based communication
Value of recent agricultural harvest	Eigenvector centrality of respondent
Financial outlook on 7-point Likert scale	Percentage of closed triangles in social network

Table 1: Sample Y_{it} development status indicators to be modeled as a function of sample X_{it} features

	Accuracy	Recall	Precision	F	AUC	% Answered Yes
Panel A: Assets and Housing						
Owns a radio	0.976	1.000	0.976	0.988	0.899	0.973
Owns a bicycle	0.676	0.552	0.678	0.609	0.722	0.456
Household has electricity	0.819	0.533	0.761	0.627	0.828	0.285
Owns a television	0.855	0.497	0.738	0.594	0.814	0.214
Has indoor plumbing	0.887	0.250	0.842	0.386	0.843	0.142
Owns a motorcycle/scooter	0.899	0.011	1.000	0.022	0.772	0.102
Owns a car/truck	0.945	0.213	0.867	0.342	0.849	0.068
Owns a refrigerator	0.954	0.180	1.000	0.305	0.878	0.055
Has landline telephone	0.992	0.125	1.000	0.222	0.562	0.009
Panel B: Social Welfare Indicators						
Hospital bills in last 12 months	0.633	0.890	0.633	0.740	0.653	0.587
Very ill in last 12 months	0.686	0.188	0.550	0.280	0.671	0.325
Death in family in last 12 months	0.665	0.183	0.632	0.284	0.619	0.363
Flood or drought in last 12 months	0.788	0.086	0.607	0.151	0.706	0.219
Fired in last 12 months	0.901	0.022	1.000	0.043	0.731	0.101

Table 2: Model performance at predicting responses from survey respondents based on call records data

2. PRELIMINARY RESULTS

In ongoing work, we are conducting additional phone surveys to collect a broader range of development outcomes such as those listed in column 1 of Table 1. Here, we focus on the simplified task of predicting responses to relatively well-defined questions with concrete answers that were collected in short phone interviews with mobile phone users in 2009 and 2010. Section 2.1 describes results from predicting asset ownership and housing characteristics; section 2.2 describes initial results at predicting more general measures of social welfare; and section 2.3 describes results from predicting a composite index of respondent wealth.

2.1 Predicting asset ownership and housing characteristics

In Panel A of Table 2, we present the results from the use of a logistic regression to predict binary responses to survey questions about fixed assets and housing characteristics such as, *"Does your household own one or more radios?"* or *"Does your household have electricity?"* We fit a version of model (**??**) with regional fixed effects and roughly twenty aggregated measures of phone activity such as those in column 2 of Table 1, including measures of phone use, SMS use, geographic mobility, and social network structure. The model is fit using 10-fold cross-validation on a sample of roughly 900 respondents who answered all survey questions, where the binary classification threshold is determined to maximize accuracy and the other performance metrics are reported at that threshold. Figure 1 shows the ROC curves for three representative questions asked in the survey.

In general, this rudimentary approach to modeling the relationship between phone use and asset ownership shows signs of modest success. For most of the outcome variables we seek to model, we can achieve relatively high accuracy, but these rates are only marginally higher than the naive baseline of predicting the majority class. For instance, the model accuracy of 85% in predicting television ownership is only an 8 percent (6 percentage points) increase over a model that predicts all respondents do not own televisions.

2.2 Predicting welfare indicators

Panel B of Table 2 presents similar results from our attempts to predict more subjective responses to broader development questions such as *"Has your household had to pay significant hospital bills in the past 12 months?"* and *"Have you lost your job in the last 12 months"*. Here, performance is lower than with the asset ownership questions; we find that our models are only able to offer marginal improvements over naive baseline predictions.

2.3 Predicting composite socioeconomic status

Finally, we test the ability of this approach to predict composite index of socioeconomic status. To create this aggregate metric from the survey responses, which we denote by $\widehat{Y_{id}}$, we take the first principal component of the 9 asset and housing characteristics listed in Panel A of Table 2. The first principal component of wealth explains 27.24% of the variance of the 9 asset categories. Similar results ob-

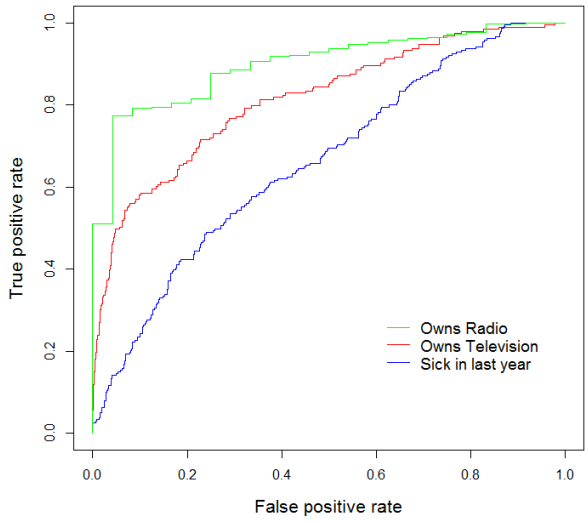


Figure 1: ROC curve for three survey outcomes

tain when creating a composite based on the first principal component of a much larger number of assets and housing characteristcs.[1]

In Table 3, we present the results from fitting an ordinary least squares regression of this first principal wealth component on a representative sample of mobile phone use metrics. While the explanatory power of this regression is rather limited ($R^2 = 0.29$), there are strong relationships between the wealth composite and several of the measures of phone use and network structure. Note that the sign and magnitude of each of the regression coefficients is highly dependent on the set of regressors included; because of the natural dependencies in the phone data, inclusion or exclusion of additional features substantively changes the estimated coefficients (though such tinkering has relatively little effect on the fit of the model).

To further illustrate the strong correlations between phone use and wealth, we perform a second principal component analysis on a large set of different metrics of mobile phone activity. In this case, the first principal component of 97 metrics of phone use explains 34.63% of the variance of the full dataset. In Figure 2, we plot for each of the survey respondents the first principal component of wealth (y-axis) against the first principal component derived from the phone use data (x-axis). The strong positive relationship between these two components is illustrated by the Nadaraya-Watson kernel regression shown in red.

3. DISCUSSION AND CONCLUSION

We have presented preliminary evidence that it is possible to predict a variety of indicators of individual socioeconomic status and welfare using mobile phone call records. If these

[1]In earlier work, we have taken a different approach that develops a composite index of "predicted expenditures" using publicly available household survey (DHS) data to approximate the estimated annualized household expenditures of survey respondent [1]. See [6] for a related approach to developing a composite wealth index from survey data.

Table 3: Regression of first principal component of assets on selected measures of phone use

	Coefficient	(*S.E.*)
Active days	−0.04	(0.03)
Calls per day	2.49	(2.28)
Outgoing calls	0.01	(0.01)
Incoming calls	−0.01†	(0.01)
Degree	0.08**	(0.03)
Int'l outgoing calls	−0.59*	(0.26)
Int'l incoming calls	−1.09	(0.72)
Int'l degree	0.38*	(0.17)
Towers visited	−0.03	(0.21)
Avg. recharge denomination	0.01	(0.03)
Daily recharge	−0.27***	(0.06)
Clustering	−505.80***	(148.09)
Betweenness	137.06*	(56.66)
N	897	
R^2	0.29	

Results show regression of first principal component of the wealth ($\widehat{Y_{id}}$), scaled by 100 to simplify presentation. Standard errors in parentheses. Regression includes district fixed effects but coefficients are omitted from table for clarity. † significant at $p < .10$; $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$

results can be further calibrated and improved upon, this technique could provide policymakers and researchers with a novel quantitative perspective on populations for whom good data has historically been hard to find. Compared to traditional methods for collecting individual and household data, the use of call records represents a considerably cheaper alternative, with dramatically higher spatial and temporal precision. In principle, such fine-grained development indicators could be applied in a variety of settings, from program monitoring and evaluation to social welfare targeting and analysis.

While provocative, we do not want to overstate the accuracy of the methods tested thus far, or imply that such techniques will ever supplant alternative modes of data collection. The predictions presented in this paper are relatively inaccurate, and the methods, models, and data leave considerable room for improvement. In ongoing work, we are working to develop improved statistical and computational models, and are collecting a large amount of survey data that will allow for better calibration and testing.

4. REFERENCES

[1] J. Blumenstock, Y. Shen, and N. Eagle. A method for estimating the relationship between phone use and wealth. *QualMeetsQuant Workshop at the 4th International IEEE/ACM Conference on Information and Communication Technologies and Development*, 2010.

[2] J. E. Blumenstock and N. Eagle. Mobile divides: Gender, socioeconomic status, and mobile phone use in Rwanda. *4th International IEEE/ACM Conference on Information and Communications Technologies and*

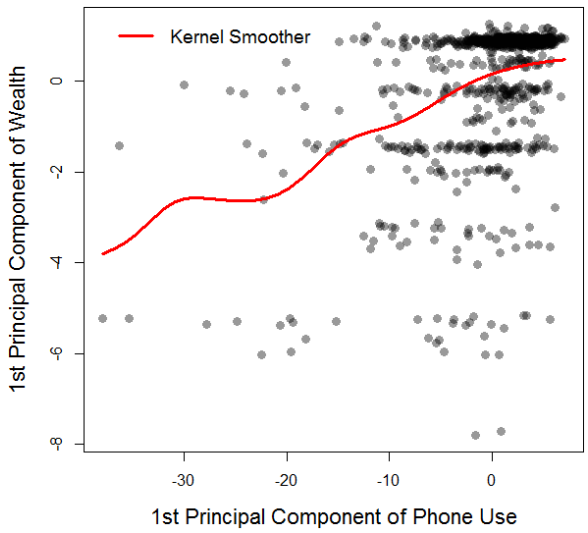


Figure 2: Relationship between 1st principal components of wealth and phone use

Development, Dec. 2010.

[3] J. E. Blumenstock and N. Eagle. Divided we call: Disparities in access and use of mobile phones in Rwanda. *Information Technology and International Development*, 8(2):1–16, 2012.

[4] J. E. Blumenstock, D. Gillick, and N. Eagle. Who's calling? demographics of mobile phone use in rwanda. *AAAI Symposium on Aritificial Intelligence and Development*, 18:116–117, 2010.

[5] N. Eagle, M. Macy, and R. Claxton. Network diversity and economic development. *Science*, 328(5981):1029–1031, May 2010.

[6] D. Filmer and L. H. Pritchett. Estimating wealth effects without expenditure data - or tears: an application to educational enrollments in states of india. *Demography*, 38(1):115 – 132, 2001.

[7] V. Frias-Martinez, E. Frias-Martinez, and N. Oliver. A gender-centric analysis of calling behavior in a developing economy. *AAAI Symposium on Aritificial Intelligence and Development*, Forthcoming, 2010.

[8] V. Frias-Martinez and J. Virseda. 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*, ICTD '12, pages 76–84, New York, NY, USA, 2012. ACM.

[9] T. Gutierrez, G. Krings, and V. D. Blondel. Evaluating socio-economic state of a country analyzing airtime credit and mobile phone datasets. *arXiv preprint arXiv:1309.4496*, 2013.

[10] M. Jerven. *Poor numbers: how we are misled by African development statistics and what to do about it*. Cornell University Press, 2013.

[11] A. J. Tatem and S. Riley. Effect of poor census data on population maps. *Science*, 318(5847):43, Oct. 2007.

Predicting commuting trips based on mobile phone data

Carlos Andre Reis Pinheiro[1], Véronique Van Vlasselaer[1], Alexandre Evsukoff[2,3], Moacyr Silva[2], Bart Baesens[1] and Nelson Ebecken[3]

[1]KU Leuven, Research Center for Management Informatics,
Naamsestraat 69, Leuven, Belgium
{carlos.pinheiro,bart.baeses,veronique.vanvlasselaer}@kuleuven.be

[2]Getúlio Vargas Foundation, School of Applied Mathematics,
Praia de Botafogo 190, Botafogo, Rio de Janeiro, Brazil
alexandre.evsukoff@fgv.br

[3]Federal University of Rio de Janeiro, Department of Civil Engineering,
Centro de Tecnologia, Cidade Universitária, Rio de Janeiro, Brazil
nelson@ntt.ufrj.br

Upon an overall human mobility behavior within the city of Rio de Janeiro, this paper describes a methodology to predict commuting trips based on the mobile phone data. This study is based on the mobile phone data provided by one of the largest mobile carriers in Brazil. Mobile phone data comprises a reasonable variety of information about subscribers' usage, including time and location of call activities throughout urban areas. This information was used to build subscribers' trajectories, describing then the most relevant characteristics of commuting over time. An Origin-Destination (O-D) matrix was built to support the estimation for the number of commuting trips. Traditional approaches inherited from transportation systems, such as gravity and radiation models – commonly employed to predict the number of commuting trips between locations(regularly upon large geographic scales) – are compared to statistical and data mining techniques such as linear regression, decision tree and artificial neural network. A comparison of these models shows that data mining models may perform slightly better than the traditional approaches from transportation systems when historical information are available. In addition to that, data mining models may be more stable for great variances in terms of the number of commuting trips between locations and upon different geographic scales. Gravity and radiation models work very well based on large geographic scales and they hold a great advantage, they are much easier to be implemented. On the other hand, data mining models offer more flexibility in incorporating additional attributes about locations – such as number of job positions, available entertainments, schools and universities posts, among others –and historical information about the trips over time.

Human mobility analysis reveals relevant knowledge about displacements and overall movement behavior, supporting disciplines such as urban and public transportation planning, traffic forecasting, communication networks optimization and insights in the spread of diseases [1]. Human mobility studies based on mobile phone data allow approximating human motion pattern within particular geographic areas such as great metropolitan areas, big cities, entire states and even countries, including travelling behavior and migration trends [2].

This study was conducted by using mobile phone data from one of the largest telecommunications companies in Brazil. The overall analyses were performed based on six months of call detail records, revealing the average behavioral pattern of users travelling throughout the city of Rio de Janeiro over time. In this study, we analyzed approximately 3.1 billion mobile phone records, comprising 2.7 million subscribers, handled by 5,300 antennas, performing 1,700 cell towers.

Mobile phone data contains transactional records about caller and callee information, associating each call or text message to the corresponding cell tower that is spread out through the metropolitan areas. These cell towers process the incoming and outgoing calls, as well as text and multimedia messages sent and received by subscribers, providing relevant information about their geographic locations at particular points in time. This geographical information basically consists of the latitude and longitude, the radius covered by the cell towers and information about the physical addresses such as street, neighborhood, city and state. Even though such locational information only approximates real human mobility, recent studies [5] [19] show that by using the appropriate techniques, mobile phone data may offer the possibility to statistically characterize human trajectories and journeys on an urban area scale.

The use of CDR data as a proxy for human mobility studies must be done with care. Besides of the spatial approximation where the position of the subscriber is assumed to be the position of the tower, subscribers do not make a call every time they move such that the observed trajectory if often not the real trajectory performed by the user. Moreover, the number of call varies strongly among the population such that the sample is highly biased towards people of better socioeconomic levels that trend to perform more calls. It should also be noticed that in Brazil, particularly in the Rio de Janeiro Area, where this study had been carried out, 85% of the subscribers are pre-paid, such that they have very irregular calling pattern

depending on specific commercial offer they get.

Mobile carriers have quite often an unbalanced distribution of prepaid and postpaid mobile phones in their customer database. Particularly in the Brazilian's marketplace, this distribution is very unbalanced, 80% for prepaid and 20% for postpaid. Due to the majority of prepaid mobile phones in the customer database, we lack information about the subscribers' home address. While prepaid subscribers do not need to provide an address, postpaid subscribers do, but they are able to declare any type of address, like their workplace or their home address, a different user's address or the payer's address. As the information about home address – and workplace – has an indispensable value in building the complete sequence of displacements in terms of human mobility behavior, we computed in this study the *presumed domiciles* and *workplaces* for all subscribers, irrespective whether they are prepaid or postpaid.

Both presumed domiciles and workplaces are useful in many human mobility analyses, and could be used to estimate the number of people present in certain urban areas of interest. Information about domiciles and workplaces can also be used to create an Origin-Destination (OD) matrix. The OD matrix is a matrix representation of the network connecting origin with destination locations, where the rows represent the origins, and the columns refer to the destinations. Each cell specifies information about the mobility, e.g. the number of displacements between origin and destination. Using an OD matrix uniquely composed of presumed domiciles and workplaces, we can apply a set of analyses to reveal interesting patterns of human's motion in large metropolitan areas. For example, governmental and industrial decisions about traffic routing and public transportation planning may rely on analysis of home-work and work-home displacements over time. Frequent connections between two areas in the city may receive full attention for instance. Moreover, a shift in traffic can be anticipated by analyzing domiciles and workplaces and the possible routes between the respective locations. Historical analyses of the number of commuting trips between locations over time can also allow organizations to estimate and forecast evolving trends and eventual issues in public transportation, traffic routing and communications network.

According to the presumed domiciles and workplaces identified by the aforementioned method, we are able to clearly describe where people live and work in the city of Rio de Janeiro. Note that domiciliary information was verified by the official city reports obtained through the governmental authorities, and it was therefore included in the overall analysis of human mobility behavior. Home-work and work-home routes are very important in order to understand possible traffic issues in great metropolitan areas. Commuting planning is one of the biggest challenges for big cities, both to establish and arrange properly public transportation resources and to design optimal traffic routes. The analysis of human mobility behavior can definitely help this task and improve its effectiveness.

Gravity and radiation models are good approaches to estimate the number of trips between two distinct locations. In this study, the trip is defined as the movement between the presumed domicile and the presumed workplace, based on a particular geographic scale – cell towers, neighborhood and clusters of cell towers, irrespective of the path performed.

The gravity model tracks its origin from the gravitational law. Two bodies are attracted to one another with a force that is proportional to the product of their masses and inversely proportional to the square of their distance. Gravity models are mapped to human mobility by replacing bodies by locations and masses by importance. Importance can be measured in terms of population, but can also incorporate other attributes like the number of jobs, gross domestic product, public facilities, transportation resources, traffic routes, among others.

The *gravity model* usually incorporates parameters to define constraints to paths and displacements followed by people, such as the cost to travel [10]. The cost to travel consists of several attributes like the distance, the resources to cover the path, the number of people travelling, etc. In particular, it assumes that the commuting activity between two distinct locations is proportional to the product of the population of these two locations and inverse proportional to the distance between these two locations.

Individuals are attracted to other locations as a function of the distance between two different places and the cost of travel between them. The gravity model considers that individuals are more attracted to close locations than to long-distance locations. This last hypothesis is based on the natural limited resources to travel between locations and the higher cost involved in long distances. In problems related to transportation systems, the distance to travel is a crucial factor in users' decision making process when they have to commute between locations. Trips between two locations with the same distance may have different costs, for instance based on possible routes, traffic jam, public transportation resources etc. The population of the locations involved in the trip is also important to predict trips within geographic areas. Large number of people associated to origin and destination locations may imply more trips.

The *radiation model*, on the other hand, tracks its origin based on theories about diffusion dynamics, where particles are emitted at a given location and have a certain probability of being absorbed by surrounding locations.

It uses the spatial distribution for the population as input, not needing any other additional parameters. It basically depends on the populations of the locations involved in the trips and their distances [10].

As the radiation model is parameter-free, the model can be much easier implemented in mobility behavior analyses and trip prediction models, especially when using mobile phone data. Although the radiation model may not seem sufficient to predict human mobility in low geographic scales, this model is successfully applied in reproducing mobile patterns at large spatial scale [16] [17] [18]. As a

result, this type of model can reasonably forecast mobility trends and the number of trips in great metropolitan areas, big cities and even countries, particularly when long-distances travels are involved.

Considering the same geographic scales, we developed linear regression models and applied them on our mobility data set in order to predict the number of commuting trips between two distinct locations. We choose to estimate two types of linear regression models: a Quantile Regression and a Robust Regression. We found that those two models performed slightly better than the models inherited from physics – i.e. the gravity and radiation model. For comparison purposes, the mobility's attributes used to feed these statistical models were the same as used in the gravity and radiation models. These attributes include the distance between the two locations involved in the trip, the number of trips originated from the origin location (regardless the destination), the population of the origin location, the population of the destination location and the radius population (the population in the circle between origin and destination locations).

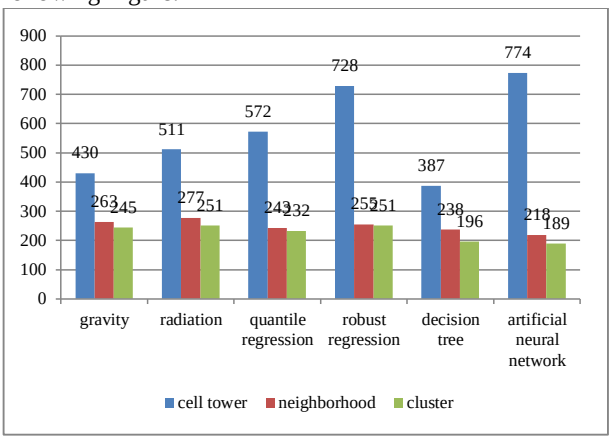
Also considering the same geographic scales, a decision tree model is estimated in order to predict the number of commuting trips between locations. A decision tree is a supervised learning model for classification problems [26]. Each input variable may corresponds to a node in the tree – if it increases the classification rate. Otherwise an input variable may be discarded. Each possible value for the input variable corresponds to edges to split nodes. In the case of continuous values the algorithm estimates cut-off values to properly create the edges. Each leaf node represents a value of the target variable given the values of the input variables, represented by the path from the root of the tree to the leaf. Afterwards, the algorithm prunes away some of the created paths in order to avoid *overfitting*. A tree learns by splitting the input data set into subsets by testing the attributes' value. The model evaluates the data on the target variable by selecting the most promising independent variable to distinguish between the values of the target variable. This process is recursively repeated on each derived subset. This process is called *recursive partitioning*, and it ends when the subset at a node has all the same value of the target variable, or when splitting no longer adds value to the predictions.

The other data mining model that is developed to predict the number of commuting trips between any two locations in the city of Rio de Janeiro is an artificial neural network (ANN). This technique is inspired by the central human nervous systems. In an ANN, neurons are connected together to form a network which simulates a biological neural network. This network consists of sets of adaptive weights associated to each neuron, represented by numerical parameters adjustable by a learning algorithm. These neurons should be able to approximate non-linear functions of their inputs. The weights are activated during training and prediction phases [27].

An ANN is a model formed by the interconnection of basic processing units, called artificial neurons. Every node in a layer is fully connected to every node in the layer above it. In addition to weights associated with a link between nodes, each output node adds a constant weight. The computation feeds forward from input nodes to output nodes without loops. Starting with linear solution of zero hidden units, a net can have a variable number of hidden units, and these determine the complexity of the classifier.

This section describes the models' performance for the transportation systems approach, based on the gravity and the radiation models, and the data mining approach, based on the linear regressions, the decision tree and the artificial neural network models. The models' performance is based on the RMSE – Root Mean Square Error – which is a good measure to describe the differences between observed and predicted values. This measure represents the standard deviation of the differences between these values, called residuals (based on the data sample used for the prediction).

The comparison of all the models deployed in this study in terms of the root mean square error for the number of trips, considering the observed and predicted trips, based on the different geographic scales is presented in the following Figure.



[1] González, M., Hidalgo, C., Barabási, A. 2008. "Understanding individual human mobility patterns." Nature, Vol. 453: 779-782.

[2] Simini, F., González, M., Maritan, A., Barabási, A-L. 2012. "A universal model for mobility and migration patterns." Nature, Vol. 484: 96–100.

[3] Rubio, A., Sanchez, A., Martinez, E. 2013. "Adaptive non-parametric identification of dense areas using cell phone records for urban analysis." Engineering Applications of Artificial Intelligence, Vol. 26: 551-563.

[4] Liu, F., Janssens, D., Wets, G., Cools, M. 2013. "Annotating mobile phone location data with activity purposes using machine learning algorithms." Expert Systems with Applications, Vol. 40, Issue 8: 3299-3311.

[5] Candia, J., González, M., Wang, P., Schoenharl, T., Madey, G., Barabasi, A-L.2008. "Uncovering individual and collective human dynamics from mobile phone records." Journal of Physics A: Mathematical and Theoretical, Vol. 41 N. 224015.

[6] Schneider, C., Belik, V., Couronné, T., Smoreda, Z.,

González, M. 2013. "Unraveling daily human mobility motifs." Journal of The Royal Society Interface, vol. 10 no. 84 20130246.

[7] Yan, X-Y., Zhao, C., Fan, Y., Di, Z., Wang, W-X. 2013. "Universal predictability of mobility patterns in cities." Physics and Society, arXiv:1307.7502.

[8] Park, J., Lee, D., González, M. 2010. "The eigenmode analysis of human motion." Journal of Statistical Mechanics: Theory and Experiment Vol. 2010.

[9] Jiang, S., Fiore, G., Yang, Y., Ferreira, J., Frazzoli,E., González, M. 2013. "A review of urban computing for mobile phone traces: current methods, challenges and opportunities." Proceedings of the ACM SIGKDD International Workshop on Urban Computing.

[10] Masucci, A., Serras, J., Johanson, A., Batty, M. 2012. "Gravity vs radiation model: on the importance of scale and heterogeneity in commuting flows." arXiv:1206.5735.

[11] Lee, A., Chen, Y-A., Ip, W-C. 2009. "Mining frequent trajectories patterns in spatial-temporal databases." Information Sciences, Vol. 179: 2218-2231.

[12] Järv, O., Ahas, R., Witlox, F. 2014. "Understanding monthly variability in human activity spaces: A twelve-month study using mobile phone call detail records." Transportation Research Part C, Vol. 38, 122-135.

[13] Sun, J.B., Yuan, J., Wang, Y., Si, H.B., Shan, X.M. 2011. "Exploring space-time structure human mobility in urban space." Physica A, Vol. 390, 929-942.

[14] Zong, E., Tan, B., Mo, K., Yang, Q. 2013. "User demographics prediction based on mobile data." Pervasive Mobile Computing, Vol. 9, Issue 6, 823-837.

[15] Makse, H. A., Havlin, S., Stanley, H. E. 1995. "Modelling urban growth patterns." Nature, Vol. 377, 608–612.

[16] Bettencourt, L. M. A., Lobo, J., Helbing, D., Kühnert, C., West, G. B. 2007. "Growth, innovation, scaling, and the pace of life in cities." Proceedings of the National Academy of Sciences of the United States of America, Vol. 104, 7301–7306.

[17] Batty, M. 2008. "The size, scale, and shape of cities." Science, Vol. 319, 769–771.

[18] Becker, R., Cáceres, R., Hanson, K., Isaacman, S., Loth, J. M., Martonosi, M., Rowland, J., Urbanek, S., Varshavsky, A., Volinsky, C. 2013. "Human mobility characterization from cellular network data." Communications of the ACM, Vol. 56, No. 1, 74-82.

[19] Balcan, D., Colliza, V., Bruno, G., Hu, H., Ramasco, J. J., Vespignani, A. 2009. "Multiscale mobility networks and the spatial spreading of infectious diseases." Proceedings of the National Academy of Sciences of the United States of America, Vol. 106, No. 51, 21484-21489.

[20] Wang, L., Hu, K., Ku, T, Yan, X. 2013. "Mining frequent trajectory pattern based on vague space partition." Knowledge-Based Systems, Vol. 50: 100-111.

[21] Bayir, M.-A., Demirbas, M., Eagle, N.. 2010. "Mobility profiler: A framework for discovering mobility profiles of cell phone users." Pervasive and Mobile Computing, Vol. 6, Issue 4, 435-454.

[22] Lin, M., Hsu, W.-J. 2013. "Mining GPS data for mobility patterns: A survey." Pervasive and Mobile Computing, Available online 8 July 2013.

[23] Koenker, R. 2005. "Quantile Regression. Cambridge University Press."

[24] Andersen, R. 2008. "Modern Methods for Robust Regression." Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-152.

[25] Howard, R.-A. 1968. "The Foundations of Decision Analysis." IEEE Transactions on System Science and Cybernetics, Vol. SSC–4, No. 3, 211–219.

[26] Bishop, C.-M. 1995. "Neural Networks for Pattern Recognition." Oxford University Press.

[27] Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., Vanthienen, J. 2003. "Benchmarking state-of-art classification algorithms for credit scoring." Journal of the Operational Research Society, Vol. 54, No. 6, 627-635.

[28] Sarle, Warren S.. 1983. "Cubic Clustering Criterion." SAS Technical Report, Vol. 108.

[29] Ward, J. H.. 1963. "Hierarchical grouping to optimize an objective function." Journal of the American Statistical Association, Vol. 58, 236-244.

[30] Aurenhammer, F. 1991. "Voronoi diagrams – a survey of a fundamental geometric data structure." Journal ACM Computing Surveys, Vol. 23, Issue 3: 345-405.

Real-time Event Detection Method using Location Information from Mobile Phone Handsets

Satoshi Nishiyama, Kento Kimura, Jun Nakajima, and Shinji Kashima
Technology Strategy Department, R&D Strategy Division, KDDI, Chiyoda-ku, Tokyo, Japan
{sa-nishiyama, kn-kimura, ju-nakajima, s-kashim}@kddi.com

Our contributions in this paper are new real time event detection method using location information from mobile phone handsets and wide-area evaluation using large telecom data set in Japan. Our method is combination of real time version of LPE algorithm and the automatic generation of training set for LPE. Evaluation result shows that our method is better than statistical approach.

I. INTRODUCTION

Human mobility models are useful for many applications, such as city infrastructure planning, area marketing and disaster management. Recently, many researches on human mobility model from location information of mobile phone users are reported [1][2]. However these researches only represent regular human activities, i.e. commuting between home and work place, but not for irregular activities such as attending concerts or sport games (here we call them "events" in this paper). If we use these models to estimate the population distribution when a popular concert is given at a stadium, we may underestimate it. In this paper, we propose a novel real-time event detection method using the estimated number of mobile phone users in the target area to compensate the underestimation of population distribution by the existing human mobility models.

II. BACKGROUND

A. *Event detection problem*

We consider "events" as sports games, concerts, conventions, conferences, festivals, troubles of public transportations or even bad weather days, which show different patterns of human mobility (especially positive increase cases) to the regular days. We do not consider the regularly crowded people in the Tokyo Disney Land (TDL) as the event. If an unusual number of people gather at TDL for the Final Count Down on New Year's Eve, it is an event. In the evaluation, we selected relatively larger events (i.e. event with more than 10,000 people). But our approach is applicable to relatively smaller events, such as surprise concerts or small troubles of public transportation systems.

Problem of finding an event from the number of users in a region is considered as a kind of outlier detection problems [3][4]. In the case where no training set is available, typical approaches are based on statistical model (e.g. average, standard deviation) and various kinds of models using proximity (e.g. distance based, density based and clustering). However, as the statistical characteristic differs for each mesh, simple statistical model does not work well. Due to the computational complexity, proximity based models are not suitable for real-time methods.

Neumann et al. proposed a proximity based method called LPE (Localized P-value Estimator) [5] using CDRs. LPE uses the number of CDRs for each hour of whole of day (i.e. 24 data per day) and calculates a distance value (e.g. Euclid distance) between the training data (days without events). Then LPE set the distance to the *k*-nearest neighbors (k-NNs) of each training day as the score of the day. A day is an event day if the score of the day is larger than those of α percent of the training set. They showed that LPE has better performance to one-class SVM using CDRs in Barcelona city for event detection. However, methods which require training sets are not suitable for event detection method to wider region since it is not practical to provide training sets manually (Issue 1). Also as LPE uses whole of the day for the feature vector, it is not suitable for real-time detection of events (Issue 2).

B. *CDR data used for this research*

We used the CDR data from opt-in users of KDDI's mobile phone service in this research. The approximate number of opt-in users is about one percent of the population in Japan. We use the estimated location from base station ID and the delay information from the base station for the location of the user. Firstly we anonymize the CDRs by hashing the user-related information in them and generate the estimated locations of the anonymized users in chronological order. Next, we count the estimated locations of the anonymized users in every five minutes for 500m square meshes in Kanto region. Kanto region is an eastern part of Japan including the Greater Tokyo area and contains approximately one third of the total population of Japan. The numbers of 500m meshes are 307,200.

III. REAL-TIME EVENT DETECTION METHOD

The computational cost of LPE classifier and the training cost of it are $O(n)$ and $O(n^2)$, where n is the size of the training set, respectively. If the size of the training set is not so large, LPE could be used for real time event detection. Our approach is that 1) generate the training set for LPE automatically, and 2) use LPE with shorter time window size (e.g. 4 hours) and limited size of training set (e.g. 30 days) for real time event detection, slide the time window along time axis and calculate the LPE classifier by differential computation from the previous time window, as shown in Fig. 1. To realize our approach, two issues mentioned above are discussed as follows.

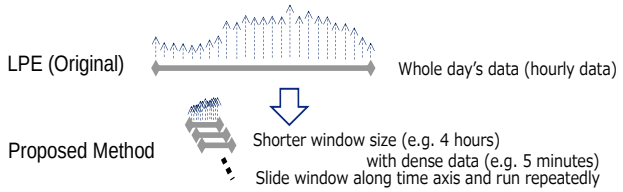


Fig. 1 Proposed method for real time event detection

A. *Issue 1: Automatic generation of training set*

Currently there are no general methods for classifying regular days (days without events) from days with events for various places (or meshes) automatically with high accuracy even if we use longer time window (e.g. whole day) for feature vector. But if we allow some error to the training set, there are many classification methods. Fortunately LPE uses k-NN distance for the feature vector metric, LPE allows some errors in its training set. Here we cluster the days in the training period into 2 clusters by k-means and use the larger cluster as the training data set, assuming that 'event' is not usual. Practically we separate the training period into weekdays and weekends + holidays before constructing the training sets.

B. *Issue 2: Shortening window period*

With shorter window period, LPE tends to generate false reports. To improve the recall rate, we set additional thresholds, a) the Euclid distance of the target > 2 times of average of those in training set, b) number of estimated users > 20 (which corresponds to minimum of 2,000 people in the mesh) and c) number of estimated users > average number of users in the training set for the LPE classifier.

IV. Evaluation

We implemented the proposed method on a Hadoop cluster and simulated the real time event detection for each 500m mesh in the Kanto region. We used the number of users in the mesh as the feature, used 4 months from June 2014 to the September of this year for generating training set and tested our method during 2 months (July and August of 2014).

For evaluation, we selected 50 events, which are considered initially as large (events with approximately more than 10,000 people), including sports, concerts and festivals and calculated the F-score for the 33 meshes where the 50 events took place, compared with conventional statistical approach (using average as the threshold parameter). Here we use the criteria of 'true' as two consecutive positives from the method before 30 minute later of the start time of an event. Table 1 shows the overall result. Our method report 261 two consecutive positives while we found manually 209 events from the Internet. Thus the precision is 209/261=0.80. For the pre-selected 50 events, the method reported 31. The recall is 31/50=0.62. The recall value is low because 15 of the pre-selected events are smaller than we had expected and from the number of estimated users they are not distinguishable to the no-event days. If we remove such events from the pre-selected set, the recall value will increase up to 31/35=0.89 and F-score will be 0.85, respectively.

Fig. 2 shows an example of automatic generation of training set at the mesh with a large arena (Saitama Super Arena) and Fig. 3 shows the detected events using the training set. Fig. 2 shows that most of the events are clustered into non-training sets, except a few events for weekends and holidays. Fig.3 shows that all events reported in this mesh are corresponded to actual events including the concert on July 5th.

TABLE I. Evaluation Results

	Precision	*Recall*	*F-score*
Statistical Approach[a)]	192/234=0.82	25/50=0.50	0.62
Our Proposal	209/261=0.80	31/50=0.62	0.70

[a.] *At the maximum of F-score (1.8 times average value as a threshold).*

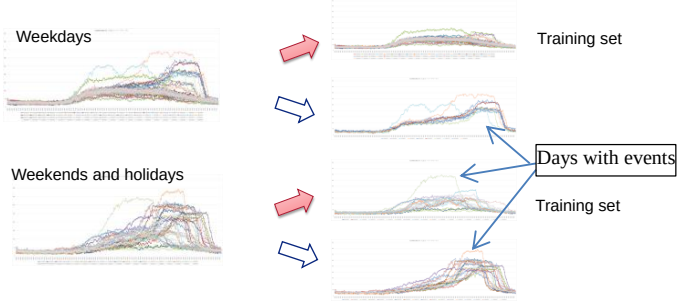


Fig. 2 Example of generated training set (Saitama Super Arena)

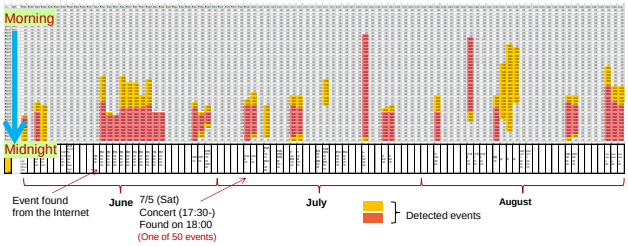


Fig. 3 Events detected (Saitama Super Arena)

V. Conlusion

In this paper, we presented a real time event detection method from CDRs based on LPE algorithm. Through empirical evaluation of large area in Japan with actual telecom dataset, we showed that our method is better than the conventional approach. This research is supported by Ministry of Internal Affairs and Communications on "Research and Development of Technologies for the Utilization of Real-time Information on Geospatial Platforms".

References

[1] Gautier Krings et al., "Urban gravity: a model for inter-city telecommunication flows." J. Stat. Mech., 2009.

[2] Sibren Isaacman, et. al., "Human Mobility Modeling at Metropolitan Scales," in Proceedings of Mobisys 2012, 2012.

[3] Varun Chandola, Arindam Banerjee and Vipin Kumar, "Anomaly Detection : A Survey", ACM Computing Surveys, Sept 2009.

[4] Hans-Peter Kriegel, Peer Kroger and Arthur Zimek, "Outlier Detection Techniques", tutorial notes, the 2010 SIAM International Conference on Data Mining, 2010.

[5] Neumann, J., Zao, M., Karatzoglou, A. and Oliver, N., Event Detection in Communication and Transportation Data, in Proceedings of IbPRIA 2013, 828-838, 2013.

Measuring global and regional influence of cites using geolocated tweets

José J. Ramasco,[1] Maxime Lenormand,[1] Bruno Gonçalves,[2,3] and Antònia Tugores[1]

[1] *Instituto de Física Interdisciplinar y Sistemas Complejos IFISC (CSIC-UIB), 07122 Palma de Mallorca, Spain*
[2] *Aix-Marseille Université, CNRS, CPT, UMR 7332, 13288 Marseille, France*
[3] *Université de Toulon, CNRS, CPT, UMR 7332, 83957 La Garde, France*

The study of competition and interactions between cities has a long history in fields such as Geography, Spatial Economics and Urbanism [1–3]. This research has taken traditionally as basis information on finance exchanges, sharing of firm headquarters, number of passengers transported by air or tons of cargo dispatched from one city to another. One can define a network relying on each of these indicators and identify the so-called World Cities, those with a higher level of centrality as the global economic or logistic centers. In most of these analysis, London and New York rank as the most central cities in terms of economic influence and transport network centrality.

In this work, we take a radically different approach to measure quantitatively the influence of a city in the world. Nowadays, geolocalized devices generate a large quantity of real time and geolocated data allowing us to understand how people move throughout time and space. It is now possible to explore human mobility in detail using geolocalized data generated by mobile phones, credit cards, transport cards or GPS. In this study we use a Twitter database containing 20 millions of geo-located tweets worldwide recorded during a period of 1000 days to measure and compare the influence of 58 cities. The aim here is to search for an answer to the questions raised by the following thought experiment: an announcement with a particular message is displayed in the most centric place of a city. People seeing it, whether tourist or locals, will later travel throughout the world. Which would be the city most efficient as source for these travels? Understating as efficient that able to spread further or more extensively the message by personal user movements. Is there a difference between the classification obtained with locals or visitors?

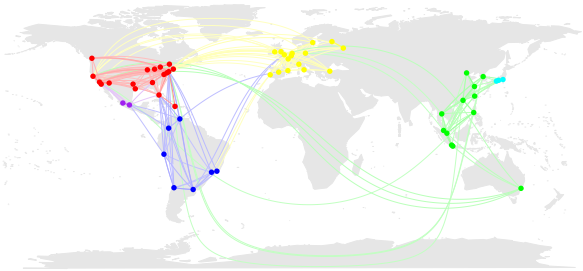


FIG. 1: Local Twitter users mobility network between the 58 cities. Only the flows representing the top 95% of the total flow have been plotted. The flows are drawn from the least to the greatest.

We use the Twitter data in order to tackle these questions. First , we consider the displacements of users visiting each city. The urban areas are ranked according to the area covered and the radius traveled by these users as a function of time. These metrics are inspired by the framework developed for random walks and Levy flights, which allows us to characterize the evolution of the system with well defined mathematical tools and with a clear reference baseline in mind. The ranks change dramatically when one has into account these measures, showing as two top rankers cities such as Rome and Hong Kong that either are large centers of global touristic attraction or are gates to the extensive hinterland of countries such as China. When the users are separated by visitors and locals, we find that the main contribution to the general ranking comes from visitors and that if only locals are taking into account the raking is modified even though not dramatically.

For getting these previous results all the Twitter population is considered, regardless of the fact that the users mobility takes place in a city or in rural areas. The only condition is that the user passed at certain moment through the city under consideration but the analysis is based on all his/her posterior movements without discerning travels to rural and urban areas However, many previous studies were focused mostly on cities and interchanges between them. To be able to compare, we restrict the analysis to users residing in an urban area and to their movements toward other cities. In this way, we obtain a weighted directed network between cities, whose links weights represent the (normalized) fluxes of users traveling from one city to another. This network provides the basis for a more traditional centrality analysis, in which we recover London and New York as the most central cities at a global scale. Still, the network also allows us to run clustering techniques and divide the world city network in communities or areas of influence. This division of the network with some of the most intense connections is shown in Figure 1. When the centrality is studied only within each community, we obtain a regional perspective that induces a ranking of cities. The comparison between the global and the regional ranking provides also with important insights in the hierarchy of world urban areas.

[1] Christaller, W. Die Zentralen Orte in Süddeutschland: eine Ökonomisch-Geographische Untersuchung Über die Gesetz Massigkeit der Verbreitung und Entwicklung der Siedlungen mit Städtischen Funktionen, Fischer Verlag, Jena (1933). (English translation: Christaller, W. & Baskin, C.W. Central places in Southern Germany, Prentice Hall, Englewood Cliffs, N.J. (1966))
[2] Knox, P.L. & Taylor, P.J. (eds.). World cities in a world-system. Cambridge University Press, 1995.
[3] Batty, M. The New Science of Cities. The MIT Press (2013).

Empirical Evaluation of Disaggregated Trip Data Collected from Cellular Networks

Peter Widhalm
Austrian Institute of Technology
Giefinggasse 2, A-1210 Vienna
Email: peter.widhalm@ait.ac.at

Michael Ulm
Austrian Institute of Technology
Giefinggasse 2, A-1210 Vienna
Email: michael.ulm@ait.ac.at

I. Introduction

Over the last decade a number of methods have been proposed in the literature to use trip data collected from cellular networks for estimating travel demand, i.e. Origin-Destination matrices, and analyzing human mobility patterns. But how reliable are the trips extracted from cellular traces? In previous work validation was conducted by comparing the spatial and temporal distribution of trips to traditional mobility surveys [1]. The authors in [2] compare different interpolation methods for mobile phone traces and evaluate their spatial errors by subsampling smartphone user position trajectories with different subsampling ratios. An evaluation of trajectory filtering techniques to reduce spatial errors in cell phone tracks is presented in [3]. However, there is still very little empirical knowledge about achievable trip detection rates and the spatio-temporal precision of the extracted trip data at a disaggregated level. With this study we hope to fill this gap.

We report the results of an empirical evaluation conducted with 241 individuals who used a smartphone to record all their trips over a period of one week. A specialized logging software collected their GPS traces along with the cells the mobile was connected to and the points in time when the device communicated with the mobile network (phone usage and Location Area Update events). In addition, we asked the participants to annotate their tracks with trip start and end points. We generate cellular trajectories by mapping the recorded cell IDs to estimated geographic locations and sub-sampling the trajectory according to the empirical frequency of mobile communications (phone usage) and logged Location Area transitions. From the resulting cellular trajectory we try to recover the original trips and use the annotated GPS positions and time stamps of the recorded trips as "ground truth" for validation. We apply a simple and frequently used trip extraction method described, e.g., in [4], [5], [6] and also test three extensions to the algorithm (see Sect. II). We analyze trip detection and error rates measured as sensitivity (true positive rate) and precision (positive predictive value) of the trip extraction method and plot it against trip length, stay duration at the trip destination and time of day. We evaluate the temporal errors of the extracted trips and the spatial uncertainty of trip origins and destinations. In addition, we demonstrate the impact of different parameter settings and evaluate how the results are affected by the inclusion of Location Area Update records, which are independent from phone usage and are issued whenever the mobile device enters a new Location Area.

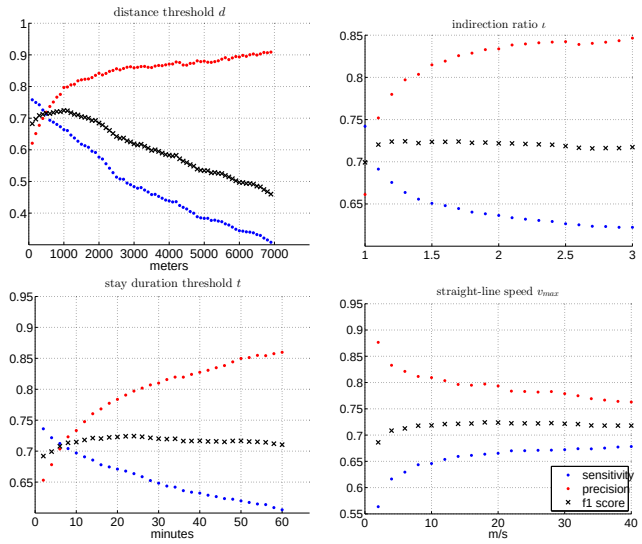


Fig. 1: Sensitivity analysis of the parameters used for trip extraction.

	basic algorithm	interpol.+ geometry	+low-pass filter	without LA-Updates
sensitivity:	47.0%	68.7%	66.4%	49.5%
precision:	91.6%	74.2%	79.7%	86.5%
F1-score:	62.1%	71.3%	72.4%	63.0%

TABLE I: Sensitivity, precision and F1-score of the basic trip extraction algorithm and the proposed extensions.

II. Trip Extraction

The basic trip extraction procedure can be summarized as follows: consecutive cell locations where all pairwise distances are below a threshold d are fused together and their coordinates are averaged to compute a centroid position. The timestamps of the first and the last record are used to approximate the time of arrival and departure. Trip origins and destinations are distinguished from "passing-by points", i.e. points along a traveled route, by introducing a minimum stay duration t.

In practice there are several kinds of errors and situations when this trip extraction approach may fail:

1) trips shorter than d are likely not to be detected, unless spatial errors add to the actual trip length;

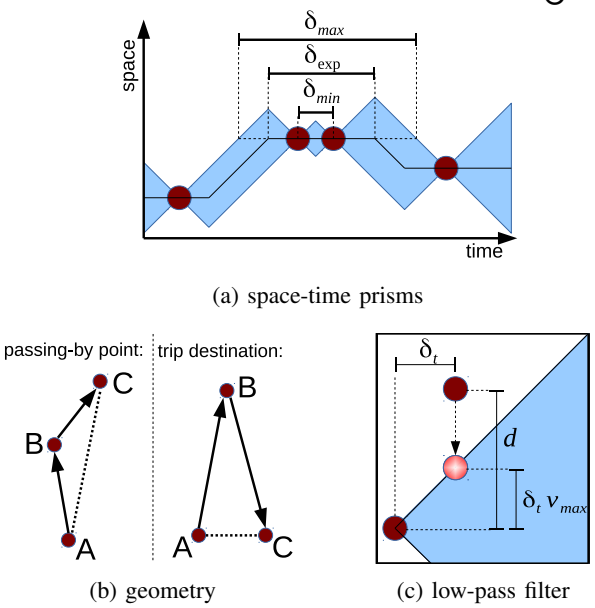
on the other hand, if d is too small, pure signal movement will generate fictitious trips;

2) if the stay duration at the trip destination is shorter than t the trip is likely not to be detected, unless the stay duration is overestimated due to the spatial uncertainty and the fusion of cell locations; on the other hand, if t is too small, many points along a traveled route will be wrongly detected as trip destinations (e.g. waiting time at transport hubs or due to road congestion);
3) the sparse and irregular temporal sampling of cellular trajectories introduces interpolation errors and uncertainties in the estimated arrival and departure times; the trip detection rate therefore depends on phone usage and actual stay duration;
4) Location Area (LA) Update events are triggered by motion rather than phone usage and can therefore improve trip detection; on the other hand these records can cause large interpolation errors: it can be assumed that devices issuing an LA Update event are currently moving and that the actual trip destination is somewhere else. As a consequence, LA Update records can introduce large spatial and temporal errors to the extracted trip data or generate fictitious trip destinations.

In addition to the basic trip extraction algorithm we evaluate the following extensions to reduce these kinds of errors:

1) interpolating the trajectory based on space-time-prisms [7] and estimating upper and lower bounds as well as an expected value of the stay duration in order to improve the arrival, departure and stay duration estimates (Fig. 2a);
2) including the trajectory's geometry in the trip extraction procedure to better distinguish trip destinations from "passing-by points" along the traveled route: for triplets (A, B, C) of successive locations we define a threshold ι of the "indirection" ratio $(\overline{AB}+\overline{BC})/\overline{AC}$ (Fig. 2b).
3) low-pass filtering based on an assumed straight-line travel speed v_{max} to reduce noise and fictitious trips (Fig. 2c);

These extensions are described in detail in [8].



(a) space-time prisms

(b) geometry

(c) low-pass filter

Fig. 2: Extensions to the basic trip extraction procedure.

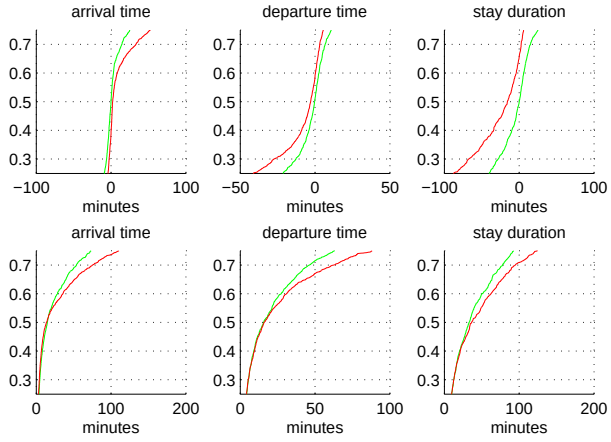


Fig. 3: Errors (top row) and error magnitudes (bottom row) of arrival time, departure time and stay duration estimates resulting from linear interpolation (red) and interpolation based on space-time prisms (green).

III. Results

The results of the sensitivity analysis of the parameters used in the trip extraction algorithm are shown in Figure 1. The optimal value of distance parameter d that maximizes the F1-score (i.e. the harmonic mean of sensitivity and precision) was $d = 1000$m, but all settings with $500 < d < 1500$m resulted in similar scores. For large ranges of duration parameter t, geometry parameter ι, and speed parameter v_{max} there was almost a one-to-one trade-off between sensitivity and precision: fixing the distance parameter to it's optimal value the F1-score varied only little for all parameter settings with $\iota > 1.1$, $v_{max} > 8$m/s, and $6 < t < 60$min. The highest F1-score was obtained for $\iota = 1.2$, $v_{max} = 18$m/s and $t = 24$min. Choosing parameter values maximizing the F1-score resulted in approximately 66% sensitivity and 80% precision.

The results achieved with the basic trip extraction algorithm and it's extensions are detailed in Table I. Combining the space-time prism approach (extension 1) with analysis of the geometry of the cellular trajectory (extension 2) significantly improved the trip detection rate as compared to the basic algorithm. The reason is that linear interpolation underestimates the stay durations at visited locations. As shown in Figure 3, the interpolation scheme based on space-time prisms improved the estimates of trip start times, end times, and stay durations. However, in some cases it overestimated the stay durations, and as a consequence the number of fictitious trips was increased, which is reflected by a lower precision percentage. Applying a low-pass filter (extension 3) reduced the number of fictitious trips due to noise. This improved precision by 5.5%, but on the other hand it also reduced sensitivity by 2.3%. Excluding

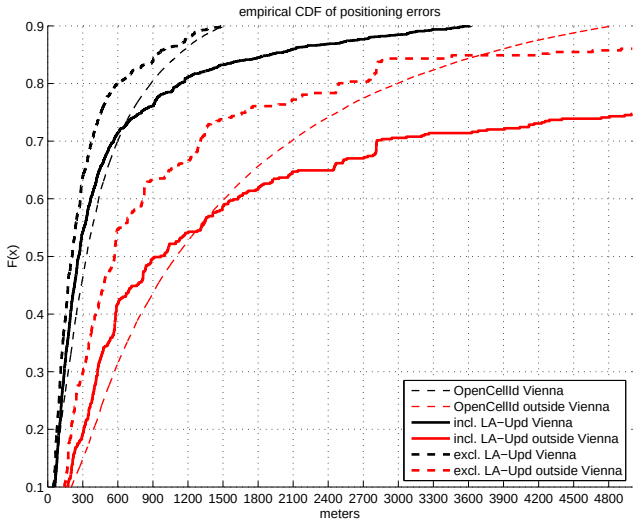


Fig. 4: Spatial error.

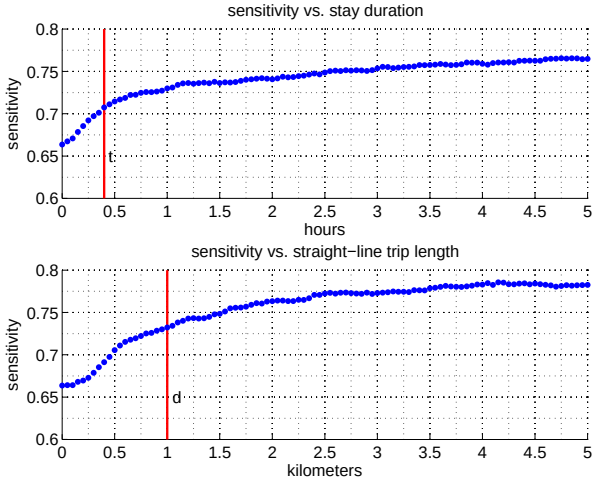


Fig. 5: Sensitivity vs. stay duration and straight-line trip length.

Location Area Update events from the cellular traces reduced the trip detection rate by 16.9% but improved precision by 6.8%.

As expected, the detection rate (sensitivity) increases with the straight-line trip length and the stay duration at the trip destination, which is shown in Figure 5. For example, for trips with straight-line lengths greater than 2km and stay durations greater than 1 hour, the detection rate rose to 82%.

The spatial error distribution of the trip origins and destinations is given in Figure 4. Including Location Area Update events in the cellular traces increased the spatial error significantly. There was also a significant difference in the positioning errors between locations within and outside the metropolitan area of Vienna, due to different network density and antenna ranges. The positioning errors could be roughly approximated by analyzing the distances of GPS measurements in the OpenCellID database [9] to the estimated cell centroid. This could be useful to estimate spatial errors for other countries and regions without re-conducting empirical investigations.

In summary, these results show that the evaluated extensions of the basic trip extraction algorithm improved the results as measured by the achieved F1-score. However, there are strong trade-offs between detection sensitivity and precision scores, and spatio-temporal errors, which means that the optimal trip extraction algorithm and parameter settings depend largely on the intended application and it's requirements. As expected, the trip detection rate is correlated to stay duration at the trip destination and straight-line trip length. Including Location Area Update records increases sensitivity but at the same time reduces precision and worsens the spatio-temporal errors. Since Location Area Update events are issued by devices moving between Location Areas and generally do not indicate trip origins and destinations, these records require a different interpolation approach than event records due to phone usage.

Future work includes analysis of selection biases in cellular data, i.e. distortions by over- or under-representation of groups of people depending on their age, sex, income, social role or other socioeconomic factors, or biases in the detection of trips depending on their purpose.

Acknowledgment

This work was partially funded the Austrian Federal Ministry for Transport, Innovation and Technology (BMVIT) within the strategic program FIT-IT under grant 835946 (SEMAPHORE).

References

[1] S. Colak, L. P. Alexander, B. G. Alvim, S. R. Mehndiretta, and M. Gonzalez, "Analyzing cell phone location data for urban travel: Current methods, limitations and opportunities," in *Transportation Research Board Annual Meeting, 94th, 2015, Washington, DC, USA*, 2015.

[2] S. Hoteit, S. Secci, S. Sobolevsky, C. Ratti, and G. Pujolle, "Estimating human trajectories and hotspots through mobile phone data," *Computer Networks*, vol. 64, pp. 296–307, 2014.

[3] C. Horn, S. Klampfl, M. Cik, and T. Reiter, "Detecting outliers in cell phone data: Correcting trajectories to improve traffic modeling," in *Transportation Research Board 93rd Annual Meeting*, no. 14-3690, 2014.

[4] F. Calabrese, G. D. Lorenzo, L. Liu, and C. Ratti, "Estimating origin-destination flows using mobile phone location data," *IEEE Pervasive Computing*, vol. 10, no. 4, pp. 36–44, 2011.

[5] H. Wang, F. Calabrese, G. Di Lorenzo, and C. Ratti, "Transportation mode inference from anonymized and aggregated mobile phone call detail records," in *Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on*. IEEE, 2010, pp. 318–323.

[6] T. Horanont, "CSIS discussion paper no. 115 - a study on urban mobility and dynamic population estimation by using aggregate mobile phone sources," 2012.

[7] T. Hägerstraand, "What about people in regional science?" *Papers in regional science*, vol. 24, no. 1, pp. 7–24, 1970.

[8] P. Widhalm, Y. Yang, M. Ulm, S. Athavale, and M. C. González, "Discovering urban activity patterns in cell phone data," *Transportation - Special Issue on emerging, passively generated data for travel behavior and policy analysis*, 2015, accepted.

[9] Opencellid database. [Online]. Available: http://opencellid.org

"I don't have a photograph, but you can have my footprints."

Chris Riederer, Sebastian Zimmeck, Coralie Phanord, Augustin Chaintreau, Steven M. Bellovin
Computer Science Department, Columbia University, New York, NY
{mani,sebastian,augustin,smb}@cs.columbia.edu, Coralie.S.Phanord.16@dartmouth.edu

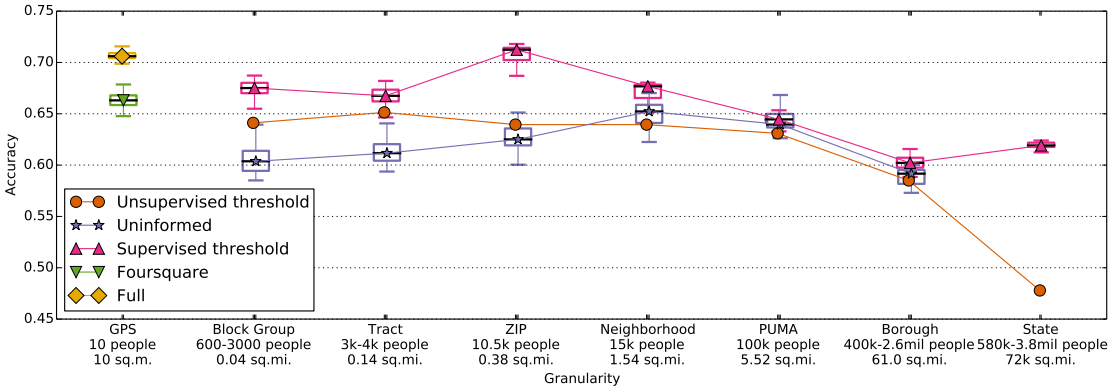


Figure 1: Accuracy of predicting ethnicity for five algorithm classes and various granularity ranges of location data.

Location data collected through GPS are routinely available to a plethora of websites, apps, and third party services. They are increasingly used to learn user characteristics particularly for purposes of behavioral and contextual advertising. While the *identification* risk of location information has been widely reported, its *discriminative* risk has received much less attention so far. Documenting this characteristic of location data requires an understanding of mobility at a demographic level; it raises methodological challenges of reproducibility, extensibility, and accuracy, all of which are hard to solve from a technical and ethical standpoint.

In our work, we fill this gap for the first time demonstrating which demographic traits can be inferred from users' geographical footprints. We leverage the growing body of public evidence in photo-sharing services. Using a corpus of geo-tagged photos from Instagram and Foursquare, we show that discriminative demographic information can be inferred from visited locations. In particular, we show that it is possible to predict a user's ethnicity with reasonable accuracy (Figure 1). Furthermore, we show that human location trends observed in census data can be reliably reproduced from the publicly available Instagram data (Figure 2).

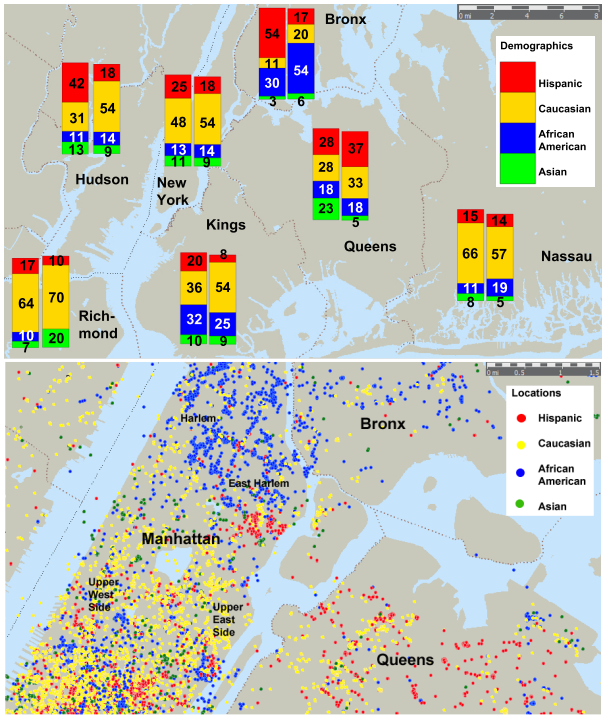


Figure 2: Top: Comparison of ethnicity according to census (left column) and Instagram location data points (right column) for various counties in New York State. Bottom: Instagram location data points colored by ethnicity in New York City.

Longitudinal Human Mobility and Real-time Access to a National Density Surface of Retail Outlets

Thomas R. Kirchner[1,2] Hong Gao[2], Andrew Anesetti-Rothermel[2], Heather Carlos[3], Brian House[4]

[1] New York University, NY, USA
[2] The Schroeder Institute at Legacy, Washington, DC, USA
[3] Norris Cotton Cancer Center, Dartmouth College, Hanover, NH, USA
[4] Brown University, Providence, RI, USA

ABSTRACT

Health-related behaviors occur as part of a broad socio-ecological context that unfolds dynamically over time. It is important to improve our understanding of the way health-related features within cities affect the health of citizens traveling through their streets. In this paper we present the development of a national density surface of convenience, grocery and gas outlets, and explore the way real-time access to these outlets varies as a function of both their density and the mobility patterns of residents.

1. INTRODUCTION

The link between cities and the health of their citizens represents a rapidly evolving area of scientific inquiry[1]. It is increasingly clear that there are large individual differences in mobility patterns that drive a dynamic interaction between individuals and their surroundings over time[2-4]. Conceptions of the urban environmental impact on health as static fail to account for the mobility and preferences of citizens actively engaging with their real-time context. Evidence from the growing literature on human mobility[2, 3] supports the notion that mobility patterns determine environmental exposure levels more than static factors like a person's place of residence. Yet methods for continuous quantification of accumulating levels of exposure to features in the built environment are only now being developed.

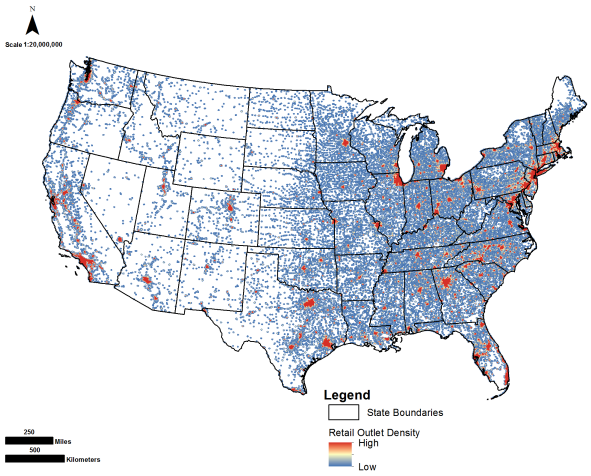
In this paper we present the development of a framework for the study of human mobility and real-time access to the landscape of point-of-sale products available across the US, a factor with indisputable implications for a range of health related behaviors[5-7]. This begins with a national probability density surface representing the continuous landscape of convenience, grocery and gas outlets in the US. We then overlay longitudinal human mobility data collected from a sample of 550 people who have voluntarily recorded their real-time geographic location via their cellular phone every 10-minutes over an average of 3 months, resulting in an average of 10,635 observations per user. Analyses examine the extent to which the dynamic nature of participants' mobility patterns interacts with their real-time surroundings and thus determines their day-to-day access to products sold in retail convenience stores. Variations across urban areas and over time provide insight and identify targets of intervention for both urban planners and public health practitioners.

2. METHOD

2.1 Retail Density Surface

A nationwide density surface of convenience and related retail outlet locations was generated using kernel density estimation (KDE). The empirical basis for this probability density surface was a national dataset of retail outlets, identified by North American Industry Classification Systems (NAICS) codes[8]. The final dataset included N = 269,781 retail outlets (Figure 1). To quantify individuals' real-time access to retail outlets, a static bandwidth KDE approach was carried out with the spatial analyst density toolset in ArcGIS v.10.1 software. The resulting density surface had a fixed 5-mile bandwidth and a cell size of 250 meters. Zonal statistics of the final density surface was calculated for zip code tabulation areas (ZCTAs)[9] in the United States. The average retail outlet density per ZCTA was then linked to each real-time mobility coordinate contributed by the participants in our geo-location tracking sample.

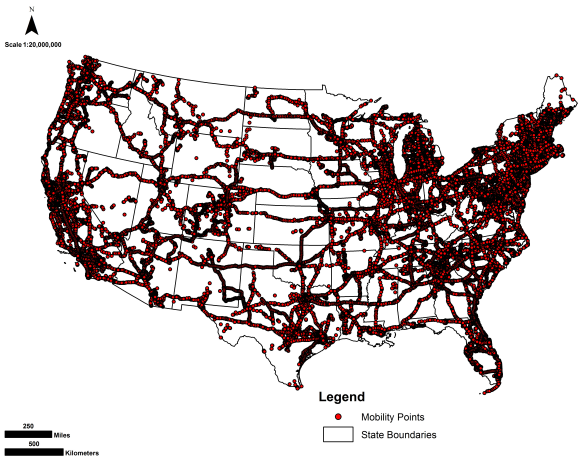
Figure 1. National Density Surface: US Convenience Retail



2.2 Longitudinal Mobility Data

Geolocation tracking made it possible to physically link each person's real-time location to the probabilistically continuous landscape of convenience stores across the US. Mobility data comes from OpenPaths<https://openpaths.cc>, launched by the New York Times Company Research and Development Lab in May of 2011[10]. OpenPaths collects GPS location information through iOS and Android location tracking applications. These applications utilized multiple approaches to geographic location capture, including both direct satellite GPS coordinates (when available) and wireless network-based "assisted" geo-location estimation via trilateration among cellular towers and wireless data access points.

Figure 2. Human Mobility Cohort



The longitudinal mobility dataset contains 3,440,821 observations collected from 859 individuals worldwide from 03/01/2012 to 12/31/2013, which was then clipped to United States using state outline polygons published by United States Census Bureau, yielding a US cohort of 550 individuals with 2,013,042 observations recording during the present observation period. The total number of locations contributed by each individual ranges from 1 to 79,602 with a mean at 10,635.39 and standard deviation of 15,189.44. The number of days falls between 1 and 670 with a mean of 202.52 (SD=181.79 days).

2.2.1 States of residence

Noteworthy, 19.33% of the data (N= 389,058) falls in California, and 14.45% (N= 290,854) in New York. All other states have 66.22% (N= 1,333,130). Given nice contrast in terms of geo-locations, climate, built-in environment, etc., between New York and California, the two states are entered as covariates in all models.

2.3 Radius of Gyration

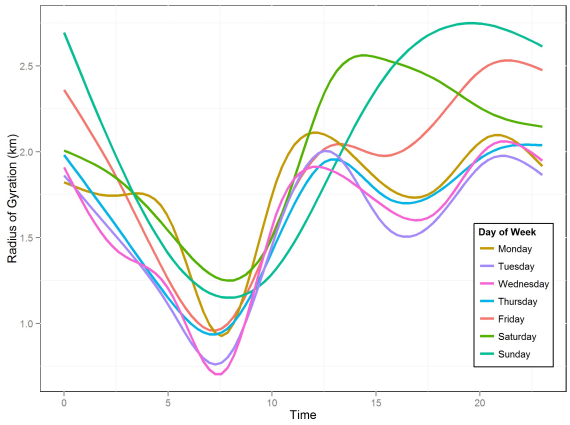
Radius of gyration measures the distance a person travels within a certain time period[3]. It defines by the standard deviation between locations and their center of mass:

$$r_g^2 = \frac{1}{N}\sum_{k=1}^{N}(r_k - r_{mean})^2$$

In this paper, hourly radius of gyration was obtained for each individual, resulting in 747,347 total observations. Validity of the data is supported by observation of expected patterns of mobility, such as that related to weekdays and weekends. Figure 3 illustrates the daily drop in mobility across the early morning hours, followed by a steep rise across the middle of the day, and then divergence on Friday, Saturday and Sunday, with late Sunday revealed as the window of greatest mobility, as travelers who departed on either Friday or Saturday return home.

Overall, mean radius of gyration is 1.93 kilometers per hour with a 4.41 kilometers standard deviation. Minimum and maximum are 0 and 119.51 kilometers. Within each day, r_g was the lowest early and higher across the remainder of the day. We see a spike in r_g mid-day on weekdays – and generally more variation on weekdays.

Figure 3. r_g over 24-hr clock as function of day of week.



2.4 Real-time Retail Access

Real-time access to the retail density surface was defined as the product of each participant's radius of gyration within each hour under observation and their average retail outlet density value for each mobility coordinate recorded within the same hour. Conceptually, this "Access" variable accumulated the number of retail options participants had as they moved, and as expected from a count variable of this kind, the observed distribution was heavily skewed right, and thus not a reasonable fit for the assumptions of the general linear model (see Section 3.1). Access was therefore stratified into deciles (plus an additional level for values of zero), effectively transforming it into a categorical variable with 11 levels from 0 to 10 corresponding to growing access to retail products. Non-parametric categorical data analysis methods were then employed as describe in Section 3.1.

3. STATISTICAL ANALYSES

3.1 Best-in-class Model Selection

Table 1. Step-down contrasts of "best-in-class" models

Model	(Model Terms)	Deviance	df	p	G2	Δ df	χ2 (Δdf; 0.001)
1	(N, W, T, S, A)	1.11E+08	698	0.00	110496519.6	337	135.807
2	**(NTA, WTA, NWT, NTS, TSA, NSA, WSA)**	**258972.5**	**361**	**1.00**	**62639.4**	**22**	**67.459**
3	(NTA, WTA, NWT, NWA, NAS, NTS, TSA, WSA, WTS, NWS)	196333	339	1.00	8189.4	75	100.425
4	(NTSA, NWTA)	188143.5	264	1.00	164905.4	174	135.807
5	(NWTA, NWST, NTSA, WTSA, NWSA)	23238.13	90	1.00	23404.8	90	124.116
6	(NWTSA)	-166.6045	0	1.00	-	-	-

N: NY versus CA
W: weekend
T: time of day
S: season
A: access to retail

Hierarchically nested model comparison techniques were used to iteratively identify the most parsimonious combination of factors required to explain the observed data. Table 1 presents an overview of the best-in-class model selection process that sought to identify the most parsimonious model form, defined as the minimal set of parameters required to provide and adequate fit to the observed data. The initial basis for comparison is Model 6, which is the saturated model that corresponds perfectly to the raw data, having degrees of freedom (df=0) equivalent to the total number of cells minus all interactive combinations of the 5 factors under study here: retail Access (the conceptual DV), State (i.e., NY as reference), Weekend, TOD, Season. Stepping-down from the saturated model, Model 2 in Table 1 fits the data well while only including 342 of the total 704 cells under study. This is thus the most parsimonious model, effectively isolating an informative pattern in the data that then becomes the basis for inference and conclusions.

Figure 4. Access by Time of Day in California and New York

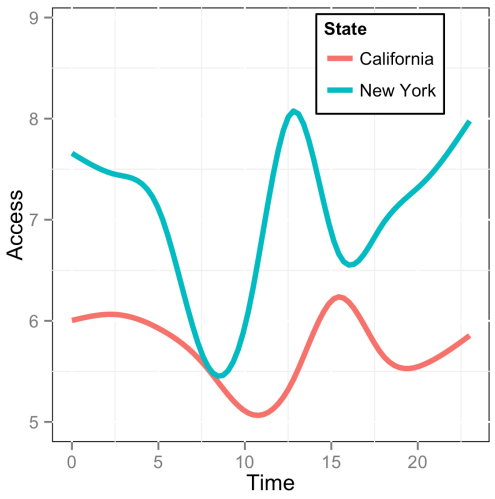


Figure 4 highlights the pattern of interaction NTA (State, Time of Day, and Access) in Model 2, Table 1. It reveals the different Access level across time of day between New York and California. Elevated Access spikes in the middle of the day in New York and again in late night, while in California, Access peaks in late afternoon.

Figure 5. Access by Time of Day between Weekday and Weekend

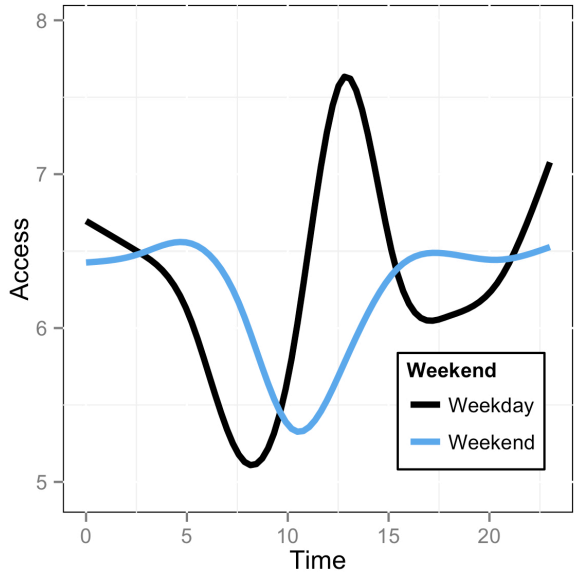


Figure 5 presents the different Access patterns in time of day between weekday and weekend. On weekdays, Access spikes in the middle of the day, probably due to high movement and high exposure to retail outlets during lunchtime. In contrast, on weekends, Access elevates in early evening and late night, which could probably be explained by going out and returning home for weekend activities.

4. CONCLUSIONS

Results of this paper shed additional light on the nature of real-time retail Access, especially as it compares to the static aggregated density of outlets alone. The contrast between New York and California is particularly useful on this point. First of all, outlet density peaks much higher in New York than in California, while California and its urban sprawl is characterized by more widely distributed yet moderate to low levels of density. Second, radius of gyration is consistently higher in California than in New York, and exhibits an entirely different cyclic pattern. This pattern of results demonstrates the kind of insight that can be gained by considering both POIs and dynamic mobility patterns.

REFERENCES

1. Cromley E.K., Mclafferty S.L., *GIS and Public Health*. 2002: The Guilford Press.

2. Yuan J., Zheng Y., Xie X., *Discovering regions of different functions in a city using human mobility and POIs*, in *In proceedings of the ACM KDD*. 2012, ACM Press: Beijing, China.

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

4. Kirchner T.R., Cantrell J., Anesetti-Rothermel A., Pearson J., Cha S., Kreslake J., Ganz O., Tacelosky M., Abrams D., Vollone D., *Individual mobility patterns and real-time geo-spatial exposure to point-of-sale tobacco marketing*, in *In Proceedings of the ACM Wireless Health*. 2012, ACM Press: San Diego, CA.

5. U.S. Department of Health and Human Services., *Preventing tobacco use among youth and young adults: a report of the surgeon general*. 2012, U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health: Atlanta, GA.

6. Institute of Medicine., *Food marketing to children and youth: threat or opportunity?* 2006.

7. Cairns, G., et al., *Systematic reviews of the evidence on the nature, extent and effects of food marketing to children. A retrospective summary.* Appetite, 2013. **62**(0): p. 209-215.

8. U.S. Census Bureau. *Introduction to NAICS*. 2014 [cited 2015 January 8th]; Available from: http://www.census.gov/eos/www/naics/.

9. U.S. Census Bureau. *ZIP Code Tabulation Areas (ZCTAs)*. 2014 [cited 2015 January 8]; Available from: https://www.census.gov/geo/reference/zctas.html.

10. House B., *OpenPaths: empowering personal geographic data*, in *In proceeding of the ISEA*. 2013: Sydney, Australia.

Bandicoot: a Python toolbox to extract behavioral indicators from mobile phone metadata

Yves-Alexandre de Montjoye[1], Luc Rocher, Alex 'Sandy' Pentland
*Corresponding author (yvesalexandre@demontjoye.com)

Mobile phone data, also called call detail record (CDR), have been used extensively by researchers in computational social science [4]. Researchers have compared the recent availability of large-scale behavioral data sets to the invention of the microscope [3]. Mobile phone metadata have for example been used to detect communities inside countries [1], analyze the impact of mobility on malaria [7], and to predict the personality [2] or gender [6] of users. There exist, however, a growing need for standardized and privacy-preserving methods to analyze mobile phone metadata datasets.

Bandicoot is an open-source Python toolbox to process mobile phone metadata. Bandicoot loads individuals metadata and attribute text files for a given user and all its contacts as well as a tower file. It then computes, on a weekly basis, 30 behavioral indicator for the user and return their mean, median, and standard deviation. The behavioral indicators falls into 3 categories: individual level (number of call, text response rate...), spatial patterns (radius of gyration, entropy of places...), and at the social network level (clustering coefficient, assortativity...). Bandicoot then provides easy to use functions to export in a CSV file the behavioral indicators of a set of users as well as potential missing data.

Overall, Bandicoot provides a complete easy-to-use environment for researchers using mobile phone metadata. It allows them to easily load their data, perform analysis, and export their results with a few line of codes. Its standardized metrics and tested code helps researcher compare results across studies, see figure 1 for a visual overview. The toolbox is easy to extended and contains an extensive documentation with guides and examples.

Bandicoot has already been used in combination with machine learning algorithms to label datasets and to run large-scale experiments [5].

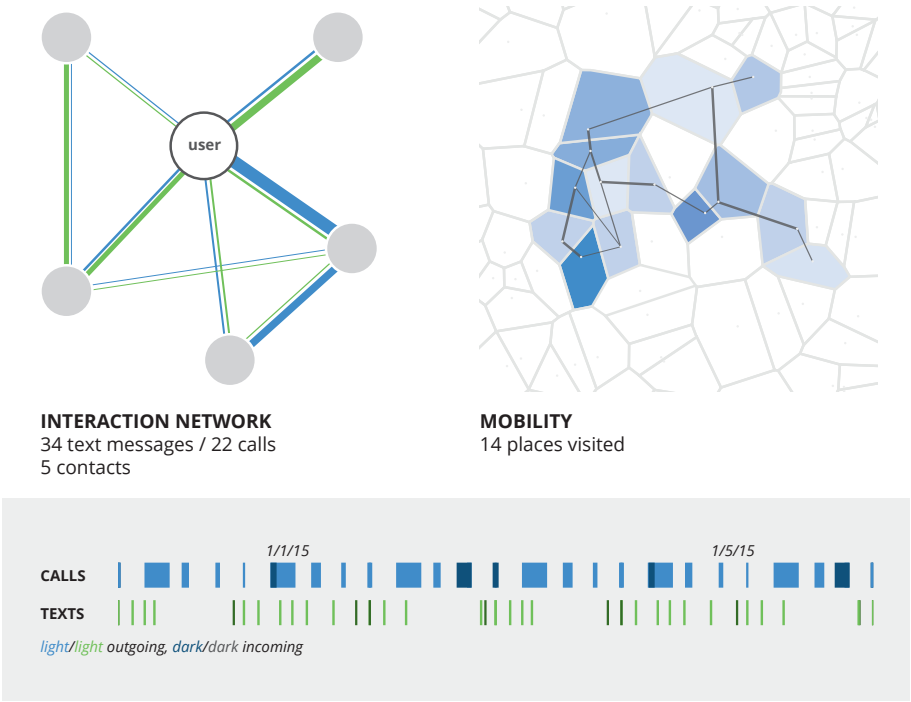


Figure 1: Visual features of a user's records

References

[1] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[2] Yves-Alexandre de Montjoye, Jordi Quoidbach, Florent Robic, and Alex Sandy Pentland. Predicting personality using novel mobile phone-based metrics. In *Social Computing, Behavioral-Cultural Modeling and Prediction*, pages 48–55. Springer, 2013.

[3] S. Higginbotham. "For science, big data is the microscope of the 21st century", 2011. URL `http://gigaom.com/2011/11/08/for-science-big-data-is-the-microscope-of-the-21st-century/`.

[4] David Lazer, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. Life in the network: the coming age of computational social science. *Science (New York, NY)*, 323(5915):721, 2009.

[5] Pål Sundsøy, Johannes Bjelland, Asif M Iqbal, Yves-Alexandre de Montjoye, et al. Big data-driven marketing: How machine learning outperforms marketers' gut-feeling. In *Social Computing, Behavioral-Cultural Modeling and Prediction*, pages 367–374. Springer, 2014.

[6] Pål Sundsøy, Johannes Bjelland, Asif M Iqbal, Alex Pentland, Yves-Alexandre de Montjoye, et al. A cross country study of gender prediction using mobile phone metadata. In *Netsci-X*. 2015.

[7] Amy Wesolowski, Nathan Eagle, Andrew J Tatem, David L Smith, Abdisalan M Noor, Robert W Snow, and Caroline O Buckee. Quantifying the impact of human mobility on malaria. *Science*, 338(6104):267–270, 2012.

Detecting Train Commuters using CDRs and GIS information

Hiroki Ishizuka
KDDI R&D Laboratories., Inc.
hk-ishizuka@kddilabs.jp

Nao Kobayashi
KDDI R&D Laboratories., Inc.
no-kobayashi@kddilabs.jp

Shigeki Muramatsu
KDDI R&D Laboratories., Inc.
mura@kddilabs.jp

Chihiro Ono
KDDI R&D Laboratories., Inc.
ono@kddilabs.jp

ABSTRACT

When a large-scale disaster occurs in downtown Tokyo, Japanese government has to provide a sufficient number of temporary camp sites or shelters for stranded commuters who cannot go back their home except that they have to keep walking several hours. Hence, It is necessary for the government to estimate exactly a number of potential candidate of stranded commuters in daily life. As commute types, train has been 76% of motorized transports (Train, Bus, Car) in Tokyo. Therefore, we suppose train commuters to be almost candidates of stranded commuters. For disaster measure for government, we proposed detecting method of the train commuters from CDRs and GIS information of train routes. As a result of our experiment, the average of the accuracy for detecting a train commuter is 75%.

INTRODUCTION

There are a lot of train commuters in Downtown Tokyo, since Tokyo has a huge train network including underground lines has been grown compared with other city in the world. As transport types for commute, train has actually been 76% of motorized transports (Train, Bus, Car) in Tokyo. In other words, train has been a main transportation for commutes in metropolitan Tokyo.

When a large-scale disaster occurs in a metropolitan area, a government has to provide sufficient number of temporary camp sites or shelters for stranded commuters who cannot go back their home except that they have to keep walking for many hours. At the Great East Japan Earthquake, practically there were 3.5 million people as stranded commuters in city of Tokyo. For securing temporary evacuations for such stranded commuters, we should estimate a number of potential candidate of stranded commuters in daily life.

In case of Tokyo, we suppose train commuters to be almost potential candidates of stranded commuters. Therefore, we proposed estimating method of train commuters in daily life. To detect train commuters, our method utilizes algorithms of supervised machine learnings with CDRs and GIS information of train routes. And also we evaluate the accuracy of our proposed method using CDRs data of 1,000 volunteers during 1 month (March, 2014). As a result of our experiment, the average of the accuracy for detecting a train commuter is 75%. Figure 1 is shown an example result of our proposed method. Black dots, Blue dots and Red dots mean locations from CDRs which a user is on a train, all locations from CDRs and train stations, respectively.

RELATED WORKS

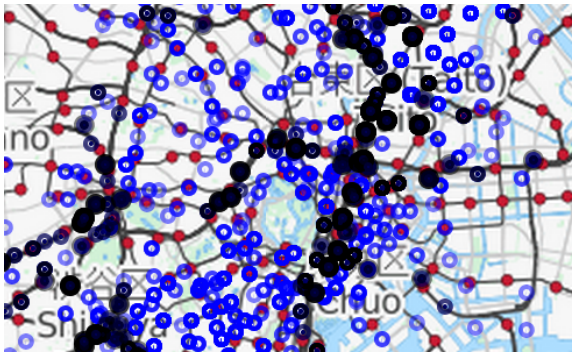


Figure 1. An Example result of our proposed method

There have been a number of conventional approaches[1-8] to determine the transportation mode including train commute. Hemminki et al.[3] uses only accelerometers for distinguishing different modalities while testing the system across a few cities. They obtain impressive results around 80%. However, they needs to keep sensing accelerometer during detecting a mode of transit without considering energy consumption. And also the method needs to install special sensing application to a user's smartphone.

Zhou et al.[4] utilize accelerometer, audio sensor and cell tower sequences of a smart phone to identify whether the user is on public transit or car, which mostly works for bus detection. The proposed method is energy efficient because they use cell tower sequences rather than GPS. However, a training of a cell tower sequence is required for all routes. it takes a long time to add new routes.

Rahul et al.[5] uses several sensors on smart phone and GIS information. Their contributions are that they reduce learning time of sensor data using General Transit Feed Specification (GTFS) which describes schedule for public transits. An accuracy of the proposed system is around 85%. However, they might not deploy new countries since opened GTFS are not so sufficient in the world.

Our proposed system utilizes location information from CDRs which generate when a user just uses his phone in daily life. Therefore, the system does not impose a user to consume energy of his phone and install a specific app for detecting his mode of transit.

PROPOSED SYSTEM

We show two phases of operation as the system architecture in Figure 2. Our proposed system estimates train commuters

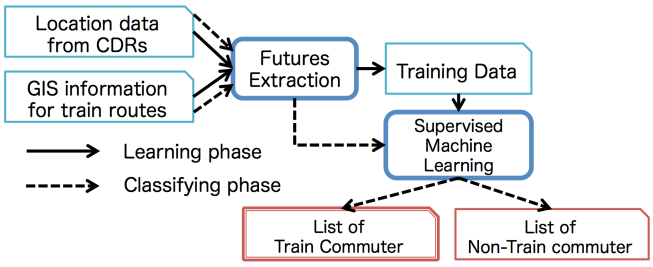


Figure 2. A System architecture of a proposed method

using supervised machine learnings. In the learning phase, the system gets the location data (learning data) from CDRs of some train commuters who exactly utilize a train as transport types for commute. Next, it extracts features of such data and generates a classifier model. In the classifying phase, the system classifies whether a train commuter or not using the generated classifier model. In next section, we describe details of a method for feature extraction.

Feature extraction

Before feature extraction, as data cleaning process, we exclude location data which moving velocity is 0 km/h or over 150 km/h between previous location and it from a data set for feature extraction. Table 1 is shown all features of our method.

Similarity to a train line

We calculate a minimum distance d_i between a shape of a train line and location data $(l_1, l_2, ...l_i)$ from CDRs. GIS information of train routes is provided from Geographical Survey Institute of Japan. A similarity to a train line is expressed as a rate of a number of d_i not exceeding a threshold d_{rail} to a number of all d_i. We set $100m$, $300m$, and $500m$ as threshold d_{rail} in our system.

Moving velocity

A moving velocity v_i is expressed as a velocity between 2 locations in a sequence of time series. We adopt an average, a median, and a maximum as value of features.

EXPERIMENT AND EVALUATION

We evaluate an accuracy of estimation of train commuters using a generated classifier . To collect learning data for train commuters, we interview to applicants (commuters) to get a type of transport for a daily commute. At the same time, we obtain an individual permission for getting CDRs in advance from the applicants.

Table 1. Features for Estimation of Train Commuters

Features
1. Distance between train line and location data Threshold : $d_{rail} = 100, 300, 500(m)$
2. Moving Velocity over v_{rail} and Dist between train line and location Threshold : $v_{rail} = 20, 40(km/h)$
3. The Average of Moving Velocity
4. The Median of Moving Velocity
5. The Maximum of Moving Velocity
6. The Average Moving Distane in a day
7. Distance between home and work

Table 2. A comparison of accuracies among machine learning algorithms

Metric	k-NN	LR	SVM	RF
Accuracy	0.734	0.721	0.742	0.751
Precision	0.734	0.721	0.735	0.751
Recall	0.733	0.731	0.747	0.751
F-measure	0.733	0.726	0.741	0.751

Experiment of a public interview

We held a public interview throughout the internet for gathering a type of main transport for a daily commute and for getting an individual permission to analyze CDRs. The target of this interview is from 18 old to 60 old. The period for getting CDRs is 1 month, March 2014. At the end of the experiment, we got results of the interview and CDRs of $1,000$ people in this experiment.

Evaluation

We created learning data for a machine learning from 1,000 people results of the experiment. In the results of the interview, there were 460 people who utilize trains for daily commute and 540 people who does not. We calculated features from CDRs of 1000 people in 1 month. Finally, we generated a classifier using the results of the interview and the features. As an assumption of this work, we do not care about subway lines, since location from CDRs does not support a localization in underground. Therefore, above 460 people did not use subway lines.

We use k-Nearest Neighbor(k-NN), Logistic regression(LR), Support Vector Machine(SVM) and Random Forest(RF) for classifier algorithms of this evaluation. we evaluate our classifier model using k-fold cross-validation($k = 3$). Table2 is shown accuracy, precision, recall, F-measure for each algorithm. There is not much different in meaning among each algorithms. We seem that Random Forest is the better since the average of each metric is 0.75.

CONCLUSION AND FUTURE WORK

In this paper, we proposed detecting method of the train commuters from CDRs and GIS information of train routes. As contributions of this work, our system utilizes only location information from CDRs which generate when a user just uses his phone in daily life. Therefore, the system does not impose a user to consume energy of his phone and install a specific app for detecting his mode of transit. From the result of the evaluation, the estimation accuracy of mode of transportation in this work is less than other conventional works that use sensors on mobile phones. However, our method still needs to optimize the features from CDRs as a future work.

ACKNOWLEDGMENTS

This work was supported by Research and Development of Technologies for the Utilization of Real-time Information on Geospatial Platforms of Ministry of Internal Affairs and Communications of Japan.

REFERENCES

1. Stenneth L.,Wolfson O.,Yu P.S. and Xu B. Transportation mode detection using mobile phones and GIS information. In Proc. SIGSPATIAL International Conference on Advances in Geographic Information Systems, 2011.
2. Zheng Y., Chen Y., Li Q., Xie X. and Ma W. Understanding transportation modes based on GPS data for web applications. In Journal ACM Transactions on the Web, vol. 4, issue 1, Jan 2010.
3. Hemminki S., Nurmi P. and Tarkoma S. Accelerometer based transportation mode detection on smartphones. In ACM SenSys 2013.
4. Zhou P., Zheng Y. and Li M. How long to wait?: Predicting bus arrival time with mobile phone based participatory sensing. In Proc. MobiSys 2012.
5. Rahul C. Shah, Chieh-yih Wan, Hong Lu, Lama Nachman. Classifying the Mode of Transportation on Mobile Phones using GIS Information. In Proc. The 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing(UBICOMP 2014), 2014.
6. Kjaergaard M. B., Bhattacharya S., Blunck H. and Nurmi P. Energy-efficient tracking for mobile devices. In ACM MobiSys 2011.
7. H. Lu, W. Pan, N. D. Lane, T. Choudhury, and A. T. Campbell. Soundsense: scalable sound sensing for people-centric applications on mobile phones. In Proceedings of ACM MobiSys, pages 165178, 2009.
8. S. Reddy, M. Mun, J. Burke, D. Estrin, M. Hansen, and M. Srivastava. Using mobile phones to determine transportation modes. ACM Transactions on Sensor Networks, vol. 6(issue 2): pages 127, March 2010.

Agent-Based Dynamic Pricing for Wireless Services

Dina Elreedy[1,*], Amir Atiya[1], Hatem Fayed[2]

dinaelreedy@eng.cu.edu.eg,amir@alumni.caltech.edu,h_fayed@eng.cu.edu.eg
[1]Department of Computer Engineering, Cairo University, Egypt
[2]Department of Engineering Mathematics and Physics, Cairo University, Egypt

The tremendous surge in demand for wireless broadband data is testing the limits of the resources of wireless networks. Wireless operators have to keep a satisfactory quality of service (QoS) by guaranteeing certain levels of blocking/dropping probabilities. However, bandwidth-hungry applications frequently result in congestion. To cope with QoS guarantees, operators have the option to either expand their wireless capacity or manage the traffic using price as a control tool. Typically, the former is a costly approach, while the latter is the worthwhile approach. In this paper, we present a new dynamic pricing model using the so-called agent based approach for two purposes: to promote the agent-based approach as an effective and flexible model for pricing any type of service, and to develop the agent-based approach for the wireless dynamic pricing problem (with focus on data services), as a detailed case study. Agent-based systems have been introduced in financial markets [1], and recently in economics [2]. The main feature of the agent-based approach in wireless pricing is that it treats each subscriber separately and models how pricing will affect his demand. Moreover, it accommodates the existence of several rate plans from the same operator as well as rate plans from different operators. This way it manages to account for the heterogeneity that exists in real situations. It will also exhibit how this heterogeneity manifests itself globally in aggregate behavior.

We propose here a dynamic pricing methodology that is based on considering several candidate rate plans and optimize their parameters in order to maximize the operator's revenue. Each rate plan has aspects of dynamic pricing. This means that we have a variable pricing that is based on traffic, capacity, time of the day and other factors. *The goal here is not to propose specific rate plans per se, but to present a whole framework for designing and testing rate plans.* In other words we propose a complete and large scale simulation testbed that can be used to accurately and in flexible manner simulate the users' actions, their response to rate plans, and to price changes, and the effect on QoS, and aggregate traffic.

The dynamic pricing is framed as premiums and discounts over the prevailing rate, in order to make it more palatable to the operators, as it allows them to relate the proposed prices to their own set reference price. Embedded in this proposed system is an agent-based model that simulates the behavior of the users, including the initiation of data service usage, the dynamics of the traffic generated for each session, the effect of pricing on the traffic generated by the user and the long term effect of pricing in selecting the rate plan. This simulation yields an estimate of the revenue and the load and its output will therefore directly influence the pricing parameters and function.

To test the proposed model, we consider a number of rate plans as examples of static and dynamically priced plans, and optimize their parameters in an attempt to maximize the revenue subject to QoS constraints. We limit our study to data services, and do not consider voice, as data services is the most pressing issue concerning bandwidth. These rate plans are:

Rate plan 1: Dynamic Pricing for a Fixed Internet Usage

Assume that the mobile operator provides a standard monthly flat rate P_{ref} for a certain bundle. We propose an alternative adjusted dynamic rate plan as follows. The price ranges from $P_{ref} - w$ to $P_{ref} + w$ where $w$is a fixed predetermined value that limits the minimum and maximum price values. So the customer is guaranteed not to pay more than $P_{ref} + w$ and will not pay less than$P_{ref} - w$. The exact paid price will be determined dynamically according to peak/off-peak hours and cell load (measured by both the number of online users in the cell, and the total bandwidth they consume).

Rate plan 2: Fixed Price for Dynamic Internet Usage

This rate plan is simply the dual of the above rate plan. As some customers prefer to pay a fixed amount monthly, we propose a rate plan with a fixed price P_{ref} while the usage amount (in Mbytes or GBytes) U_{ref} ranges from $U_{ref} - \Delta$ to $U_{ref} + \Delta$ where Δ is a fixed predetermined value that limits the minimum and maximum usage amount.

Rate plan 3: Floating Price

In this rate plan, the price is dynamically determined according to peak/off-peak hours and the cell load. However, unlike the first rate plan there are no minimum or maximum bounds on the paid price. So, if a user did not use the internet, he/she will pay nothing (and vice versa, extensive use leads to a large bill). On the other hand, the price itself is not limitless. There are built-in pricing bounds that guarantee that it does not diverge much from the reference price.

System Overview

We collect the data from 21 users using an Android mobile application, that we have developed, that stores and retrieves internet data usage (3G). The average collected data length is 72 days per user.

In order to be able to simulate the users' actions, we have a first stage, where we learn the dynamics of the users' data services usage. We assume that the learned dynamics of the 21 users will provide a diverse set of usage patterns hopefully similar to what is observed in a typical cell. There are two main aspects about the usage that we have to learn from the gathered users' data. For each period (for example a five minute or a ten minute period) we need to predict whether the user is online or not (i.e. whether the user is using some data service or not). The other aspect is to predict how much traffic is he using in the period (typically in Mbytes). To obtain these dynamics we develop a classifier for the former aspect, and a time series forecasting model for the latter aspect. Once these two models are estimated for each user, we can generate more data. Moreover, we can generate additional hypothetical users that follow dynamics similar to the existing users. Also, one can anticipate changes in the dynamics in response to changes in the pricing. Limiting the use to the original 21 users' real data would be too rigid and therefore not flexible enough to achieve all these goals.

We developed one classifier for each user that estimates his/her probability of being online or in a session using Support Vector Machines (SVM) classifier [3].We then developed one time series forecasting model for each user that forecasts the usage amount in Megabytes (MB) for future time periods using Autoregressive (AR) [4] model with a number of lags determined using Bayesian Information Criterion (BIC).

$$y_t = \sum_{k=1}^{L} \varphi_k y_{t-k} + \varepsilon_t$$

where y_t is the forecasted demand usage, L is number of lags and ε_t is a white noise term.

How the price varies dynamically is designed as follows. We consider that the price should be a function of the remaining cell capacity, the time of day and the immediate preceding rate of usage. This will make the price as an effective control tool for reigning in congestion and for shifting demand towards low usage periods. We consider certain *price multipliers* that vary around *1* and provide a varying discount/premium over the reference price (which is typically taken as the prevailing rate). Each multiplier is a function (typically linear or piecewise linear) of one of the aforementioned factors. These multipliers will be multiplied by the reference price to obtain the final price. They vary around the value of *1*, where a value that is lower than *1* corresponds to a discount with respect to the reference price (for example, 0.9 means that the price is 10% lower) while a value that is higher than *1* represents a premium over the reference

The multipliers considered are: *time of the day multiplier* which has higher values for peak hours, *cell capacity multiplier* which is directly proportional to cell load, *number of active users in cell multiplier* which is directly proportional to the number of users and *average usage per session multiplier* which rewards a user for uniform usage over long period and penalizes him/her for higher usage in a concentrated period (to avoid network jamming).

An meta-heuristic optimization algorithm called Covariance Matrix Adaption Evolution Strategy (CMA-ES) [5,6] is applied to optimize the parameters defining the functional form of the multipliers such that the operator's revenue is maximized. Any price modification will influence the demand via the price elasticity. An x % change in demand is translated into an x % change in generated traffic, which in turn will influence the revenue. The optimization algorithm will keep probing the space intelligently until reaching the optimal solution. The optimization is subject to strict a constraint on the QoS in the form of caps on the blocking probability. Thus, any solution leading to a violation of the blocking probability will be discarded. A training period is used to determine customers' choices of the competing rate plans. We define a utility function for the rate plan evaluation, in the form of a proposed function of the average monthly payment, and the monthly payment variability experienced for each customer during the training period. A *softmax* function is applied to the different rate plans' utility to determine the probability of a customer to select a particular rate plan. Figure 1 shows a block diagram of the proposed system.

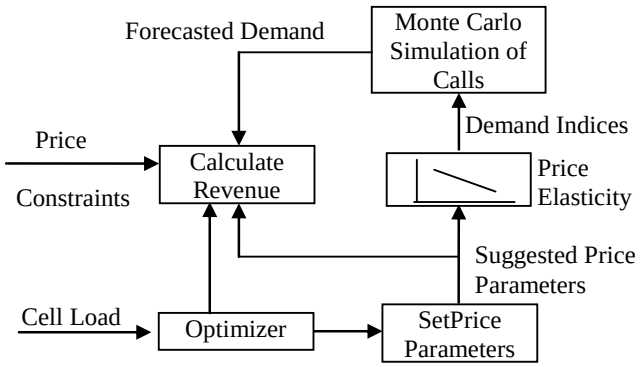


Figure 1: The block diagram of the dynamic pricing model.

Results

We applied two simulation runs for 50 generated users (whose dynamics are extracted from the 21 available real users' data). In the first run, only a static plan (fixed-price but with a cap on total usage) of one of the Egyptian operators was offered to the users. This is considered the baseline run, over which we would like to improve by developing new rate plans, and tuning their parameters. In the second run, both the static plan and our proposed dynamic plans were offered. The simulations were done for 3-months period, after the training period at which the user recognizes his average payment and subsequently makes a rate choice. (In reality the customer does not necessarily do that, because he knows his own usage, and can effectively predict his payment for each possible rate plan.) Our proposed model yields an increase of the total revenue by about 66% compared to that of the static plan. So the proposed plans in conjunction with existing rate plans achieved both acceptable QoS, and improved utility to the users and higher operator's revenue.

References

[1] LeBaron, B. (2001) Empirical regularities from interacting long and short memory investors in an agent based stock market, IEEE Trans. Evol.Comput.5.5: 442–455.

[2] Tesfatsion, L., Judd, K.L. (2006), Handbook of Computational Economics, Volume 2: Agent-Based Computational Economics, Elsevier, Amsterdam.

[3] Cortes, Corinna, and Vladimir Vapnik. "Support-vector networks." Machine learning 20.3 (1995): 273-297.

[4] Box, G., Jenkins, G. M., Reinsel, G. C. (1994). Time Series Analysis: Forecasting and Control 3rd ed. Prentice-Hall.

[5] Auger, A., Hansen, N. (2005) A restart CMA evolution strategy with increasing population size, IEEE Congress on Evolutionary Computation 2, 1769-1776.

[6] Hansen, N. (2009) Benchmarking a BI-Population CMA-ES on the BBOB-2009 function testbed. In: GECCO Genetic and Evolutionary Computation Conference, 2389-2396.

Measuring the predictability of interstate travel times using real time traffic monitoring data for Boston

Morgan R. Frank[1,*], Serdar Çolak[1], Jameson Toole[1], Lauren P. Alexander[1], and Marta C. González[1,2]

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

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

January 15, 2015

Abstract

The recent availability of high-definition traffic flow data from sources, such as TomTom, allows for empirical investigation into traffic prediction. Here, we examine interstates in the Boston area to evaluate the predictability of travel times over a route throughout the day. The resulting distributions exemplify the fluctuations in travel times as a result of temporally dynamic traffic conditions. Examining the shape of these distributions over the course of the day shows us the difficulty of predicting travel times accurately with other means. We compare these empirical averages to travel times resulting from iterated traffic assignment using mobile phone data. Our results expose natural fluctuations in travel time statistics, while examining the use of mobile phone data for aggregate traffic assignment in conjunction with traffic feeds to move beyond only solving an optimization problem, such as iterated traffic assignment. Our results suggest that more realistic statistical models for urban traffic networks might be generated by including empirical traffic data into the traffic assignment process; this would in turn allow for improved mobility solutions and traffic applications.

Flow dynamics on networks are a thoroughly investigated topic as a theoretical problem [1,2] and in several applications, such as airway transportation [3–5] and social systems [6–8]. In particular, the effects of traffic congestion across a road network presents interesting properties with an increasing number of methods for modeling such phenomenon [9–11]. The discrepancies among these models may be attributable to the sparsity of empirical data for calibration [12].

A common goal for the transportation community is to provide methods that produce efficient routes for commuters. Currently, this task requires approximating system-wide road conditions by assigning road demand from aggregate commuter origin-destination data [13–15]. Some methods for simulating population level movement based on spatially localized information include gravity models [16–18] and the radiation model [11, 19]. However, the trend of using increasingly available mobile phone data to approximate population levels by reception area is an appealing alternative because of the temporal resolution and spatial coverage [12, 20]. We can estimate road demand through methods, such as iterated traffic assignment, using data to approximate population level transitions between regions, such as commuting from residential to commercial areas.

High-definition traffic flow and traffic incident data made available by TomTom [21] is a promising source of empirical traffic information. By enhancing official government traffic reports with data from 80 million anonymous mobile phone users and 1.6 million TomTom device users, the TomTom traffic feed provides accurate glimpses of congestion on the road network. Note that the TomTom feed does not allow us to track individual trajectories, but is a suitable tool for calibrating traffic assignment methods as it allows us to verify flow distributions across the road network.

Previous studies have revealed the non-linear behaviors inherent in traffic flow as the percolation of shock waves in continuous models [22–24] and in agent-based investigations [25, 26]. This non-linearity may cause difficulty in predicting travel times for individual routes. However, empirical investigation into the deviations of route travel times as a function of the time of the day are lacking due to the sparsity of relevant data. This issue is in part responsible for the current interest in the traffic assignment problem.

While empirical data feeds are sufficient to validate traffic assignment, these feeds are often too sparse or unavailable for particular U.S. cities to replace current traffic assignment methods [27]. Our method uses census tract commuting data from the U.S. Department of Transportation Federal Highway Administration [28] and demand estimations through real-time mobile phone data [12, 20]. We implement an iterated traffic assignment strategy fol-

*corresponding author: mrfrank@mit.edu

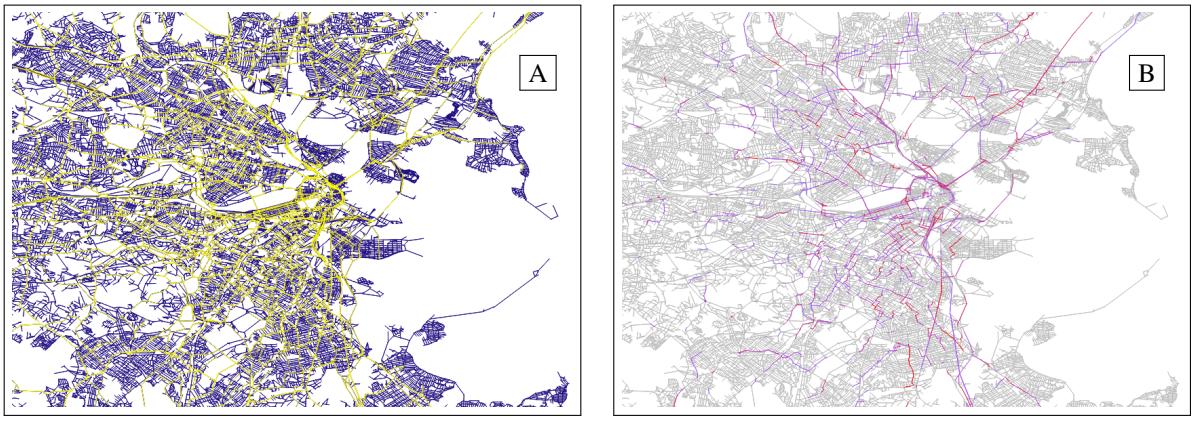


Figure 1: A snapshot of traffic flow on the Boston road network. **(A)** We plot locations of Boston roads in the OpenLR dataset in blue. Yellow road segments indicate areas where traffic flow may be observed with the TomTom high-definition traffic flow data. **(B)** Projected volume-over-capacity (*voc*) using iterated traffic assignment from mobile phone data during a morning commute with purple indicating $voc \approx 2.0$ and red indicating $voc \approx 5.0$.

lowing the methods explained in [20] using the OpenLR road network [29]. This method allows for realistic individual routing by distributing aggregate traffic across a road network. The output of this algorithm includes an estimation of traffic flow on each road in the road network, which can be validated against the TomTom flow data. Figure 1 shows example predicted areas of congestion using our iterated traffic assignment method. Furthermore, integrating data from traffic feeds may allow researchers to move beyond relying on only an optimization problem for providing accurate traffic assignment.

In conclusion, we show the natural variation in route travel times over the course of a week day using empirical TomTom data for the Boston area collected over two months. This result exposes the innate difficulties in producing accurate travel time predictions. Despite this, we implement a traffic assignment strategy using mobile phone data to successfully produce realistic travel predictions for interstates in the Boston area when compared to the empirical travel times. This promising result highlights the predictive power of traffic assignment methods using aggregate commuter data through mobile phones in the absence of sparsely available traffic feeds. Furthermore, our results suggest that more realistic statistical models of the urban traffic network are obtainable by integrating empirical traffic data into the traffic assignment process. Such models could be used to inform better mobility solutions and traffic applications.

References

[1] Vanderbei, R. J. "Network Flow Problems". Linear Programming. Springer US, 2014. 199-224.

[2] Angulo, G., Ahmed, S., and Dey, S. S. "Semi-continuous network flow problems." Mathematical Programming 145.1-2 (2014): 565-599.

[3] Guan, X., Zhang, X., Zhu, Y., Sun, D., & Lei, J. (2014). "An Airway Network Flow Assignment Approach Based on an Efficient Multiobjective Optimization Framework". The Scientific World Journal.

[4] Cook, A., Blom, H. A., Lillo, F., Mantegna, R. N., Miccich, S., Rivas, D., ... & Zanin, M. (2014). Applying complexity science to air traffic management. Journal of Air Transport Management.

[5] Xu, D., Zhang, C. W., Miao, Z., & Cheung, R. K. (2014). "A flow allocation strategy for routing over multiple flow classes with an application to air cargo terminals". Computers & Operations Research, 51, 1-10.

[6] Bagrow, J. P., Desu, S., Frank, M. R., Manukyan, N., Mitchell, L., Reagan, A., Bloedorn, E. E., Booker, L. B., Branting, L. K., Smith, M. J., Tivnan, B. F., Danforth, C. M., Dodds, P. S., & Bongard, J. C. "Shadow networks: Discovering hidden nodes with models of information flow" (2013). preprint: http://arxiv.org/abs/1312.6122

[7] Coviello, L., Sohn, Y., Kramer, A. D., Marlow, C., Franceschetti, M., Christakis, N. A., & Fowler, J. H. (2014). "Detecting Emotional Contagion in Massive Social Networks". PloS one, 9(3), e90315.

[8] Bliss, C. A., Frank, M. R., Danforth, C. M., & Dodds, P. S. (2014). "An evolutionary algorithm approach to link prediction in dynamic social networks". Journal of Computational Science.

[9] Li, D., Fu, B., Wang, Y., Lu, G., Berezin, Y., Stanley, E., & Havlin, S. "Percolation transition in dynamical traffic network with evolving critical bottlenecks" (2014), Proceedings of the National Academy of Sciences.

[10] Çolak, S., Schneider, C. M., Wang, P., & González, M. C. "On the role of spatial dynamics and topology on network flows" (2013), New Journal of Physics, 15(11).

[11] Ren, Y., Ercsey-Ravasz, M., Wang, P., González, M. C., & Toroczkai, Z. "Predicting commuter flows in spatial networks using a radiation model based on temporal ranges", Nature Communications, in press, 2014.

[12] Jiang, S., Via-Arias, L., Zegras, C., Ferreira, J., & González, M. (2011). "Calling for validation: demonstrating the use of mobile phone data to validate integrated land use transportation models". In International Conference Virtual City and Territory (7: 2011: Lisboa)

[13] Gupta, A., Xu, W., Perrine, K., Bell, D., & Ruiz-Juri, N. (2014, October). "On scaling time dependent shortest path computations for Dynamic Traffic Assignment". In Big Data (Big Data), 2014 IEEE International Conference on (pp. 796-801). IEEE.

[14] Shafiei, M., Nazemi, M., & Seyedabrishami, S. (2014). "Estimating time-dependent origin-destination demand from traffic counts: extended gradient method". Transportation Letters: The International Journal of Transportation Research.

[15] Wismans, L., de Romph, E., Friso, K., & Zantema, K. (2014). "Real Time Traffic Models, Decision Support for Traffic Management". Procedia Environmental Sciences, 22, 220-235.

[16] Barthélemy, M. "Spatial networks". Phys. Rep. 499, 1?101 (2010)

[17] Jung, W. S., Wang, F. & Stanley, H. E. "Gravity model in the Korean highway". EPL 81, 48005 (2008)

[18] Thiemann, C., Theis, F., Grady, D., Brune, R. & Brockmann, D. "The structure of borders in a small world". PLoS ONE 5, e15422 (2010).

[19] Simini, F., González, M. C., Maritan, A., & Barabási A.-L. "A unviersal model for mobility and migration patterns" (2012). Nature, 96(484) doi:10.1038/nature10856

[20] Colak, S., Alexander, L. P., Alvim, B. G., Mehndiretta, S. R., & González, M. C. "Analyzing Cell Phone Location Data for Urban Travel: Current Methods, Limitations and Opportunities", in press, Transportation Research Records, ("Practice Ready" for Presentation in the annual TRB meeting 2014)

[21] www.tomtom.com

[22] Simini, F., Maritan, A., & Néda, Z. "Human Mobility in a Continuum Approach" (2013). PLoS ONE, 8(3)

[23] Flynn, M. R., Kasimov, A. R., Nave, J.-C., Rosales, R. R., & Selbold, B. "Self-sustained nonlinear waves in traffic flow". Phys. Rev. E 79, 056113. Published 26 May 2009.

[24] Aw, A., & Rascle, M. "Resurection of the 'Second Order' Models of Traffic Flow" (200), SIAM Journal of Applied Mathematics, 60(3), 916-938, doi:10.1137/S0036139997332099

[25] Gazis, D. C., Herman, R., & Rothery, R. W. "Nonlinear Follow-the-Leader Models of Traffic Flow" (1961). Operations Research, 9(4), 545-567, doi:10.1287/opre.9.4.545

[26] Khalesian, M., & Delavar, M. R. "A multi-agent based traffic network micro-simulation using spatio-temporal GIS" (2008). Intl. Archives Photogrammetry, Remote Sensing & Spatial Info. Sci, 12(7), 31-35.

[27] Yang, Y., Herrera, C., Eagle, N., & González, M. C. "Limits of Predictability in Commuting Flows in the Absence of Data for Calibration" (2014). Nature Scientific Reports, 4(5662), do:10.1038/srep05662

[28] U.S. Department of Transportation Federal Highway Administration. CTPP 2006-2010 Census Tract Flows. http://www.fhwa.dot.gov/planning/census issues/ctpp/dataproducts/2006-2010 tract flows/index.cfm.

[29] http://www.openlr.org/

Analysis and Modeling of Mobile Data Traffic in Mexico City

Eduardo Mucelli Rezende Oliveira[◇,×], Aline Carneiro Viana[×], K. P. Naveen[×], and Carlos Sarraute[⋆]

[◇] École Polytechnique, France [×] INRIA, France [⋆] Grandata Labs, Argentina

I. Introduction

The 3G cellular networks are struggling with the recent boost of mobile data consumption led by the pervasive era. The steady growth of smartphones, the very rapid evolution of services and their usage is accentuated in metropolitan scenarios due to the high urbanization and concentration of mobile users. In this context, *understanding mobile data traffic demands per user* is crucial for the evaluation of data offloading solutions designed to alleviate cellular networks [1].

The pervasive era also brought new facilities: currently smartphones provide the best means of gathering users information about content consumption behavior on a large scale. In this context, the literature is rich in work studying and modeling users mobility, but little is publicly known about users content consumption patterns.

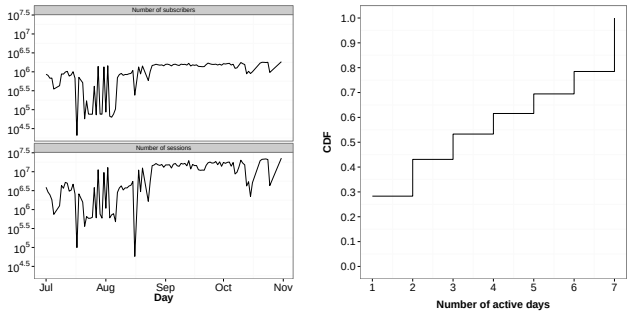
Our contributions in this work are twofold: First, our analyses provide a *precise characterization of individual subscribers traffic behavior clustered by their usage pattern*, instead of a network-wide data traffic view [2]. Second, we provide *a traffic generator that synthetically, still consistently, reproduces real traffic demands*. A synthetic traffic generator has positive implications on network resource allocation planning and testing, or hotspot deployment. Moreover, the synthetic traffic carries no direct personal information from the original users, thus greatly reducing privacy issues.

II. Dataset Analysis

Our study is performed on an anonymized dataset collected at the core of a major 3G network of Mexico's capital, consisting of data traffic associated with 6.8 million subscribers. The data includes information about subscribers' *sessions* that took place from 1st July to 31st October, 2013. The studied dataset contains more than 1.05 billion sessions and each of them has the following fields: (1) amount of upload and download volumes (in kilobytes) during the session; (2) session duration in seconds; and (3) timestamp indicating when the session starts.

Due to the routinary behavior of people [1] and the large scale of the dataset, it suffices to study a subset of the whole dataset in order to capture the daily behavior of subscribers. We have selected one week to more deeply assess the subscribers' behavior, which spans from 25th August to 31st August 2013 and contains information of about 2.8 million smartphone devices and activity that totalizes 104 million sessions.

In the following, we start our analysis by studying the behavior of mobile subscribers in terms of traffic they generate and their activity on the temporal scale.



(a) (b)

Fig. 1. (a) Number of subscribers and sessions on the whole dataset. (b) CDF of number of days in which subscribers generate traffic.

A. Traffic Dynamics

Fig. 1(a) shows the total number of subscribers and the total number of sessions from the whole dataset. Both parameters are highly correlated (Spearman's correlation is 98%). On average, the number of active subscribers is higher during the weekdays than during the weekend (also observed in [3]).

Fig. 1(b) shows the CDF of the number of active days of the subscribers within the week. Additionally, we observed that most subscribers generate traffic on few hours during the day.

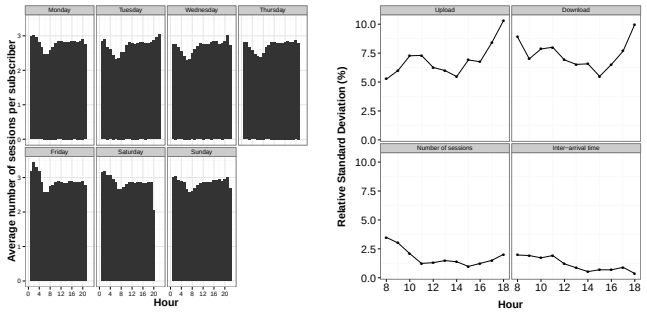
Furthermore, our analysis shows that uploaded and downloaded session volumes are similar and correlated (Spearmans's correlation is 88%). Owing to this high correlation, in our evaluation and traffic modeling, we take into consideration the total volume per session.

B. Temporal Dynamics

Some hours are more active than others when it comes to subscribers daily activities. Two features are important to highlight: (1) there is a repetitive behavior during different days at the same hours; (2) there are peak and non-peak hours in traffic demands.

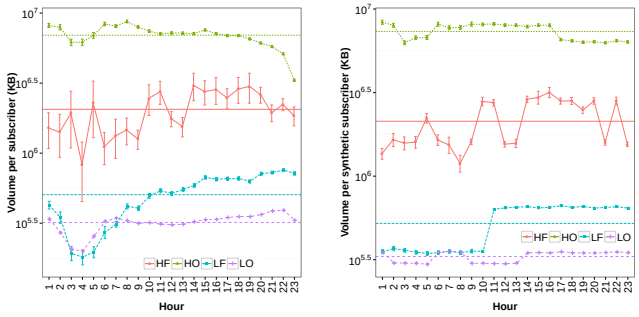
Indeed, Fig. 2(a) shows the hourly dynamics of the number of sessions during the week (for a complete evaluation of other parameters, refer to [4]). We can see a clear gap on the average number of sessions from 4am to 8am.

Our analysis of all parameters shows a high similarity on number of sessions, volume of traffic, and inter-arrival time, when compared day-wise (even comparing weekdays and weekends). Fig. 2(b) shows the per-parameter average Relative Standard Deviation (RSD), which considers the hour-wise variation from all 7 days. We have also calculated the maximum RSD of the parameters when compared day-wise. The results show that *the parameters from the same hours*



(a) (b)

Fig. 2. (a) Average number of sessions per user during the week. (b) Relative Standard Deviation per parameter.



(a) (b)

Fig. 3. Volume of traffic per class per hour (a) for real and (b) for synthetic subscribers.

on different days present less variability than the parameters within the same day on different hours.

The similarity of the temporal activity patterns among different days of the week allows us to select one day to perform our extensive per-hour analysis and distinguish different profiles of users.

III. Subscriber Profiling Methodology

To describe the behavior of subscribers, we group them into a limited number of profiles, generated according to two traffic parameters: traffic demands (i.e., volume of traffic) and activity behavior (i.e., number of sessions). We detail below our profile definition procedure.

A. Similarity Computation

Our development can hold in general for any time interval D chosen from the week. Let $\mathbb{S}$ be the set of all subscribers that generate some traffic during D, and $\mathbb{S}' \subseteq \mathbb{S}$ be a randomly selected sample of subscribers. Our objective is to partition the subscribers in $\mathbb{S}'$ into a set of *clusters* $\mathbb{P}$, such that subscribers belonging to the same cluster are "similar" in terms of traffic demands. We use the Euclidean distance to measure the *similarity* between two subscribers [5].

Each subscriber $i \in \mathbb{S}$ can be effectively represented as a sequence of sessions generated by i. Let t_k^i denote the time instant at which his k-th session begins, and let v_k^i be the volume of traffic (both upload and download) generated.

To reduce the memory and processing time required, we divide D into time slots of length T (we use $T = 1$ hour). Let τ_t^i denote the set of all sessions starting within time slot t. Now, the volume of traffic generated by subscriber i, in time slot t, is given by $V_t^i = \sum_{k \in \tau_t^i} v_k^i$. Similarly, we define the number of sessions N_t^i.

Using the above expressions, the total volume and the total number of sessions generated by subscriber i during D are $\vartheta^i = \sum_{t \in D} V_t^i$ and $\eta^i = \sum_{t \in D} N_t^i$. Finally, we define the *traffic volume similarity* between two subscribers i and j as the difference $w_{ij}^{\vartheta} = \|\vartheta^i - \vartheta^j\|$, and the *number of sessions similarity* as $w_{ij}^{\eta} = \|\eta^i - \eta^j\|$.

Using the subscribers in $\mathbb{S}'$ as the vertices, and using either $w_{i,j}^{\vartheta}$ or $w_{i,j}^{\eta}$ as the edge weights, we obtain a complete graph $G(\mathbb{S}', \mathbb{E})$, which is given as input to our clustering algorithm.

B. Subscriber Clustering and Classification

Instead of a-priori fixing a value for the number of profiles (i.e., clusters) $|\mathbb{P}|$, our goal is to obtain from the data the number of profiles which best represent the subscribers' traffic activities. For this purpose, we use the UPGMA hierarchical clustering algorithm [6], that iteratively aggregates vertices from the similarity graph $G(\mathbb{S}', \mathbb{E})$. To find the best number of clusters, we have implemented and used 23 stopping rules (see [4] for a complete list).

Profiling occurs then in four stages: (1) building a similarity graph with $\mathbb{S}'$ subscribers, (2) hierarchically clustering it using a similarity metric, (3) determining the best number of clusters $|\mathbb{P}|$, and (4) classifying the remaining subscribers in $\mathbb{S} - \mathbb{S}'$.

These four stages are performed in two rounds. In the first round, the graph $G(\mathbb{S}', \mathbb{E})$ weighted according to the traffic volume similarity w_{ij}^{ϑ} is used. According to stopping criteria results, $|\mathbb{P}| = 2$ weighted subgraphs G_1 and G_2 are created. In the second round, G_1 and G_2 are weighted according to the number of sessions similarity w_{ij}^{η}. Two new clusters are found for G_1 and G_2, totalizing four subscribers profiles.

C. Subscriber Profiles

To obtain the profiles for our dataset, we set D as 27th of August 2013, which contains information of about 1.5 million smartphone devices, and randomly sampled $|\mathbb{S}'| = 10000$ subscribers. Our profiling methodology resulted in *four profiles*, named as follows: Light Occasional (LO), Light Frequent (LF), Heavy Occasional (HO) and Heavy Frequent (HF). Table I shows the characteristics of each of the profiles.

TABLE I
Characteristics of the resulting profiles

	Light		Heavy	
Volume	29 KB to 20 GB		21 GB to 625 GB	
Subscribers	1489242		27659	
	Occasional	Frequent	Occasional	Frequent
Sessions	1 to 278	279 to 8737	1 to 278	279 to 8737
Subscribers	1486496	2746	27593	66

In Fig. 3(a), we show the dynamics of the volume of traffic per subscribers class per hour, calculated using V_t^i. The error bars correspond to a 95% confidence interval. We can see that our methodology well separates the profiles. For each curve in this plot, we also show the respective mean values (horizontal

lines). We classify, for each profile of subscribers and for each parameter, the hours above the mean as *peak hours*, and hours below the mean as *non-peak hours*.

IV. Measurement-Driven Traffic Modeling

The goal of the traffic model is to generate synthetic subscribers, whose usage pattern is consistent with the observations made about the real subscribers in the previous section.

A. Fitting Empirical Distributions

Using the original subscribers' data, we first study for each profile in peak and non-peak hour, the empirical distribution functions (i.e., CDF) of the traffic parameters. Detailed CDFs analyses are reported in [4].

Once the CDFs are obtained, we estimate the set of distributions that best fit them. More specifically, when considering the volume of traffic and the inter-arrival time (consisting of continuous values), the Kolmogorov-Smirnov statistic test is used. The test estimates the parameters for a set of continuous distributions (namely, Log-normal, Gamma, Weibull, Logis, and Exponential) that best fit the corresponding empirical distribution. Similarly, when considering the number of sessions (consisting of discrete values), the Chi-squared statistic test is used to estimate the best fitting parameters for a set of discrete distributions (Negative binomial, Geometric, and Poisson). In both cases, with the results of the fitting tests, we select the distribution function that best fits each corresponding CDF.

B. Synthetic Subscriber Generation

Generating a synthetic subscriber will first require us to generate a profile type (HO, HF, LO or LF) for the subscriber. Profile types are assigned randomly, based on the distribution of profiles population observed in the real data (see Table I).

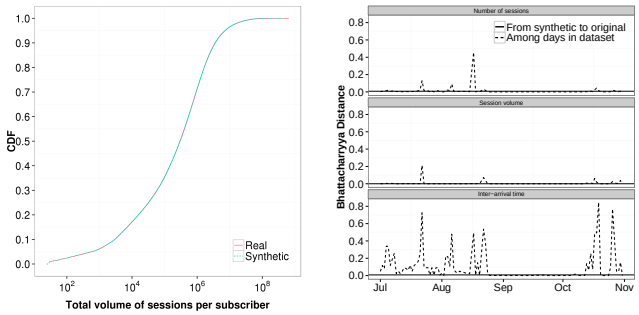
After obtaining the profile type, for a given hour t, we randomly sample values for each traffic parameter according to the corresponding fitted distribution functions. That is, for each subscriber i and time slot t, we sample a number of sessions N_t^i, mean inter-arrival time IAT_t^i, and average session volume V_t^i from the appropriate (fitted) distributions. The volume per session v_k^i (for $k \in \tau_t^i$) and initial timestamp of each session are computed accordingly. By varying t over the 24 hours in a day, we obtain a synthetic subscriber traffic for one day.

C. Synthetic Traffic Model Evaluation

In order to evaluate our traffic modeling, we generate a synthetic dataset and compare it with the original dataset. Towards this goal, we generate a set $\mathbb{R}$ of synthetic subscribers, with $|\mathbb{R}| = |\mathbb{S}|$, and one day of traffic, denoted as D'.

Let p_E^ϑ denote the PDF (Probability Distribution Function) of the total volume generated per subscriber in day E. Fig. 4(a) shows the CDFs corresponding to p_D^ϑ and $p_{D'}^\vartheta$. We can observe an *almost complete overlap of the two CDFs due to a high similarity between the real trace and the synthetic trace*.

Furthermore, to measure the similarity between datasets, we use the Bhattacharyya distance d (see [3], [7]). Let $\mathbb{D}^*$ denote the set of days contained in the dataset, excluding the original day D. Fig. 4(b) shows the distances $d(p_D, p_E)$ between D and the remaining days $E \in \mathbb{D}^*$, and as horizontal line the distance $d(p_D, p_{D'})$, per parameter. We verified that the



(a) (b)

Fig. 4. (a) CDF of the total volume per session for real and synthetic subscribers (b) Per-parameter Bhattacharyya distances between original and synthetic trace, and between original trace in D and other days from the original trace.

distance $d(p_D, p_{D'})$ is within the 95% confidence interval of the distances $d(p_D, p_E)$ for $E \in \mathbb{D}^*$, for all three parameters.

Finally, we applied the profiling methodology on the synthetic users. Fig. 3(b) depicts the per-class behavior for the volume of traffic per session for the classified synthetic users. *This result is coherent with the behavior from the original trace presented in Fig. 3(a).*

V. Conclusions and Next Steps

In this paper we have first presented a characterization of a 4-month dataset that contains more than 1.05 billion session connections from about 6.8 million smartphone users. Moreover, we propose a framework that automatically classifies those users by their traffic demands into a limited number of profiles. Our approach takes advantage of repetitive user behavior due to their daily routines. Furthermore, we have calculated distributions that describe their traffic demands into peak and non-peak hours. Finally, from these distributions we create a traffic generator and evaluate the synthetic trace it generates. Our results show that the synthetic trace presents a consistent behavior when compared to original dataset.

As future work, we aim to investigate models to describe sessions' transfer rate and duration. Additionally, we intend to apply and evaluate our traffic generator on different problems such as network planning.

References

[1] E. M. R. Oliveira and A. C. Viana, "From routine to network deployment for data offloading in metropolitan areas," in *Proc. of IEEE SECON*, Jun. 2014.

[2] D. Naboulsi, R. Stanica, and M. Fiore, "Classifying call profiles in large-scale mobile traffic datasets," in *Proc. of IEEE Infocom*, Apr. 2014.

[3] U. Paul, A. Subramanian, M. Buddhikot, and S. Das, "Understanding traffic dynamics in cellular data networks," in *Proc. of IEEE Infocom*, Apr. 2011.

[4] E. M. R. Oliveira, A. C. Viana, K. P. Naveen, and C. Sarraute, "Measurement-driven mobile data traffic modelling in a large metropolitan area," INRIA, Tech. Rep., 2014.

[5] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "On clustering validation techniques," *Journal of Intelligent Information Systems*, vol. 17, no. 2-3, pp. 107–145, Dec. 2001.

[6] R. R. Sokal and C. D. Michener, "A statistical method for evaluating systematic relationships," *University of Kansas Scientific Bulletin*, vol. 28, pp. 1409–1438, 1958.

[7] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by their probability distributions," *Bulletin of the Calcutta Mathematical Society*, vol. 35, pp. 99–109, 1943.

Inference of Users Demographic Attributes based on Homophily in Communication Networks

Jorge Brea
Grandata Labs
Buenos Aires, Argentina
jorge@grandata.com

Javier Burroni
Grandata Labs
Buenos Aires, Argentina
javier.burroni@grandata.com

Carlos Sarraute
Grandata Labs
Buenos Aires, Argentina
charles@grandata.com

1. INTRODUCTION

Over the past decade, mobile phones have become prevalent in all parts of the world, across all demographic backgrounds. Mobile phones are used by men and women across a wide age range in both developed and developing countries. Consequently, they have become one of the most important mechanisms for social interaction within a population, making them an increasingly important source of information to understand human demographics and human behaviour.

In this work we combine two sources of information: communication logs from a major mobile operator in Mexico, and information on the demographics of a subset of the users population. This allows us to perform an observational study of mobile phone usage, differentiated by age groups categories [1, 2]. This study is interesting in its own right, since it provides knowledge on the structure and demographics of the mobile phone market in Mexico.

We then tackle the problem of inferring the age group for all users in the network. We present here an exclusively graph based inference method relying solely on the topological structure of the mobile network, together with a topological analysis of the performance of the algorithm. The equations for our algorithm can be described as a diffusion process with two added properties: (i) memory of its initial state, and (ii) the information is propagated as a probability vector for each node attribute (instead of the value of the attribute itself). Our algorithm can successfully infer different age groups within the network population given known values for a subset of nodes (seed nodes). Most interestingly, we show that by carefully analysing the topological relationships between correctly predicted nodes and the seed nodes, we can characterize particular subsets of nodes for which our inference method has significantly higher accuracy.

2. DATA SET

The dataset used in this work consists of cell phone calls and SMS (short message service) records (CDRs) collected over a three month period. We aggregate this information into an edge list $(x, y, w_{x,y})$ where $w_{x,y}$ is a boolean value indicating whether users x and y have communicated at least once within the three month period. This edge list represent our mobile phone graph $\mathcal{G} = \langle \mathcal{N}, \mathcal{E} \rangle$ where $\mathcal{N}$ denotes the set of nodes (users) and $\mathcal{E}$ the set of communication links. Our graph has about 70 million nodes and 250 million edges. We are also given the age of a subset of 500,000 nodes, which we use as a ground truth (denoted $\mathcal{N}_{GT}$).

3. AGE HOMOPHILY IN THE COMMUNICATION NETWORK

Graph based methods like the one we present in this work rely strongly on the ability of the graph topology to capture correlations between the node attributes we are aiming to predict. A most fundamental structure is that of correlations between a node's attribute and that of its neighbors. Figure 1(a) shows the correlation matrix C where $C_{i,j}$ is the number of links between users of age i and age j for the nodes in the ground truth $\mathcal{N}_{GT}$. Though we can observe some smaller off diagonal peaks, we can see that it is most strongly peaked along the diagonal, showing that users are much more likely to communicate with users of their same age.

To account for a population density bias, we compute a surrogate correlation matrix R as the expected number of edges between ages i and j under the null hypothesis of independence. This matrix represents a graph with the same nodes as the original, but with random edges (while maintaining the same number of edges as the original). Both C and R are represented with a logarithmic color scale. If we subtract the logarithm of R to the logarithm of the original matrix C, we can isolate the "social effect" (i.e. homophily) from the pure random connections, as can be seen in Fig. 1(b).

4. REACTION-DIFFUSION ALGORITHM

For each node x in $\mathcal{G}$ we define an initial state probability vector $g_{x,0} \in \mathbb{R}^4$ representing the initial probability of the nodes age belonging to one of $d = 4$ age categories: below 24, 25 to 34, 35 to 50, and over 50 years old. More precisely, each component of $g_{x,0}$ is given by

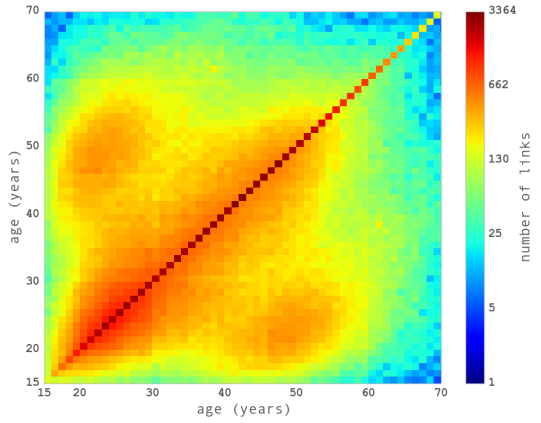
$$(g_{x,0})_i = \begin{cases} \delta_{i,a(x)} & \text{if } x \in \mathcal{N}_S \\ 1/d & \text{if } x \notin \mathcal{N}_S \end{cases} \quad (1)$$

where $\delta_{i,a(x)}$ is the Kronecker delta function, $a(x)$ the age category assigned to each seed node x, and $\mathcal{N}_S \subset \mathcal{N}_{GT}$ are the seed nodes, whose known age attribute is diffused across the network. For non seed nodes, equal probabilities are assigned to each category. These probability vectors are the set of initial conditions for the algorithm.

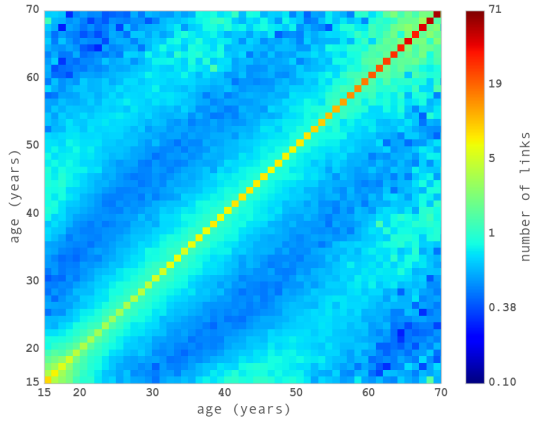
The evolution equations for the probability vectors g are then as follows:

$$g_{x,t} = (1 - \lambda)\, g_{x,0} + \lambda\, \frac{\sum_{x \sim y} g_{y,t-1}}{|\{y : x \sim y\}|} \quad (2)$$

where $x \sim y$ is the set of x's neighbours and $\lambda \in [0, 1]$ defines



(a) Communications matrix C.



(b) Difference between C and R.

Figure 1: Age homophily plots showing (a) the communications matrix C and (b) the difference between C and the surrogate random links matrix R.

the relative importance of each of the two terms. It is not hard to show [1] that the above equation can be rewritten as a discrete reaction diffusion equation on the mobile phone graph given by

$$g_t^a - g_{t-1}^a = (1-\lambda)(g_0^a - g_{t-1}^a) - \lambda \mathcal{L} g_{t-1}^a \qquad (3)$$

where $\mathcal{L}$ is a normalized graph Laplacian. At each iteration, each $g_{x,t}$ in (2) updates its state as a result of its initial state and the mean field resulting from the probability vector of its neighbors.

Preserving Age Demographics

A salient feature of our algorithm is that the demographic information being spread is not the age group itself, but a probability vector for each age group. In each iteration, the algorithm does not collapse the information in each node to a preferred value; instead it allows the system to evolve as a probability state over the network, which allows us to impose further external constraints on the algorithms results. For instance, after the last iteration, we can select the age category of each node based on the final probability state of each node's category, but constrained so that the age group distribution for the whole network be that of the ground truth set $\mathcal{N}_{GT}$ as was described in [1, 2].

5. RESULTS

In this section we first present the results for the predictive power of the reaction-diffusion algorithm over the whole network $\mathcal{G}$. The overall performance for the entire validation set (20% of $\mathcal{N}_{GT}$) was 46.6%, and we note that a performance based on random guessing without prior information would result in an expected performance of 25%, or an expected performance of $\sim 36\%$ if we set all nodes age group to the most probable category $(35-50)$. We now take a closer look and see how the performance can increase or decrease as we select particular subsets of nodes.

Topological Metrics and Performance

We first look at how our algorithm performs for different subsets of $\mathcal{N}$ selected according to three topological metrics: (i) number of seeds in the node's neighborhood, (ii) distance of node to $\mathcal{N}_S$ and (iii) node's degree.

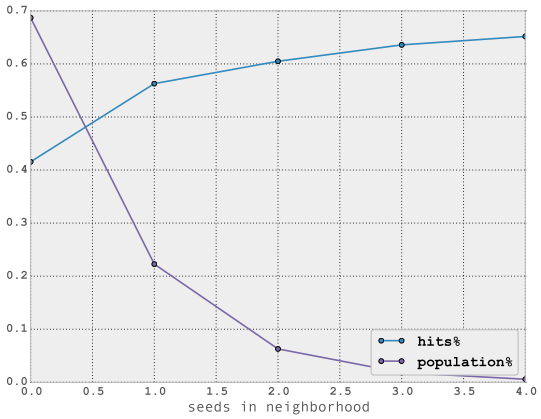


Figure 2: Performance and population as function of *SIN* (seeds in neighborhood).

Figure 2 plots the performance (*hits*) of the algorithm as a function of the number of seeds in the node's neighborhood (*SIN*). The algorithm performs worse for nodes with no seeds in the immediate neighborhood with *hits* = 41.5%, steadily rising as the amount of seeds increase with a performance of *hits* = 66% for nodes with 4 seeds in their neighbourhood. We also see that the population of nodes decreases exponentially with the amount of seeds in their neighbourhood.

Next we examine how the algorithm performs for nodes in $\mathcal{G}$ that are at a given distance to the seed set (*DTS*). In Fig. 3 we plot the population size of nodes as a function of their *DTS*. The most frequent distance to the nearest seed is 2, and almost all nodes are at distance less than 4. This implies that after four iterations of the algorithm, the seeds information have spread to most of the nodes in $\mathcal{G}$. This figure also shows that the performance of the algorithm decreases as the distance of a node to $\mathcal{N}_S$ increases.

In Fig. 4 we see that the performance of the algorithm is lowest for nodes with small degree and gradually increases

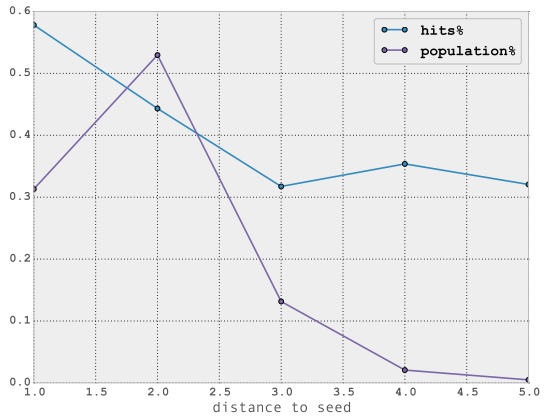


Figure 3: Performance and population as function of *DTS* (distance to seeds set).

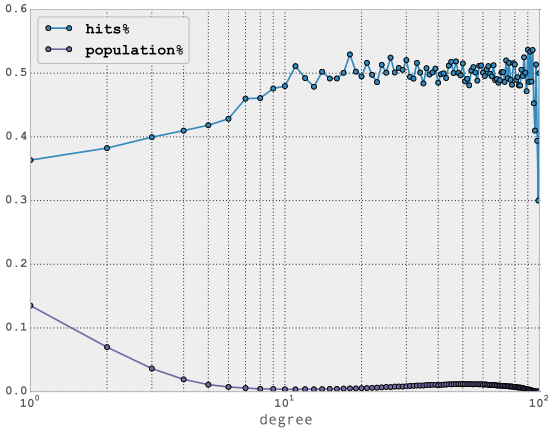


Figure 4: Performance and node population as function of the nodes degree.

as the degree increases reaching a plateau for nodes with $d(x) > 10$.

Probability Vector Information

An orthogonal approach to find an optimal subset of nodes (where our algorithm works best) is to exploit the information in the probability vector for the age group on each node. Namely, we examine the performance of the reaction-diffusion algorithm as we restrict our analysis to nodes whose selected category satisfies a minimal threshold value τ in its probability vector.

In Fig. 5 we observe a monotonic increase in the performance as the threshold is increased but, as expected there is a monotonic decrease of the validation set. We note that for $\tau = 0.5$ the performance increased to 72% with 3,492 out of the 143,240 (2.4%) of the validation nodes remaining. The performance of the algorithm increases to $\sim$ 81% for $\tau = 0.55$ but the validation set remaining sharply decreased to only 201 nodes (0.1%).

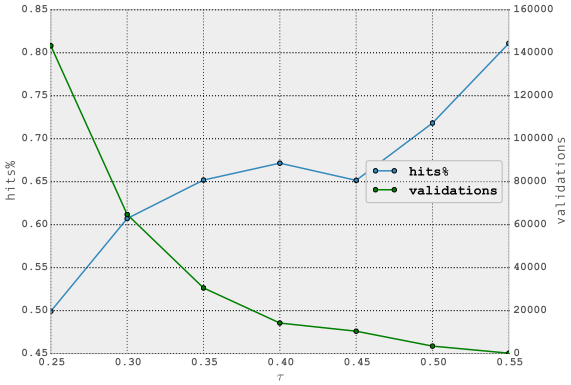


Figure 5: Performance as function of τ.

6. CONCLUSIONS

In this work we have presented a novel algorithm that can harness the bare bones topology of mobile phone networks to infer with significant accuracy the age group of the network's users. We show that an important reason for the success of the algorithm is the strong age homophily among neighbours in the network as evidenced by our observational study of the ground truth sample $\mathcal{G}_{GT}$.

We have shown the importance of understanding nodes topological properties, in particular their relation to the seed nodes, in order to fine grain our expectation of correctly classifying the nodes. Though we have carried out this analysis for a specific network, we believe this approach can be useful to study generic networks where node attribute correlations are present.

As future work, one direction that we are investigating is how to improve the graph based inference approach presented here by appropriately combining it with classical machine learning techniques based on node features [2]. We are also interested in applying our methodology to predict variables related to the users' spending behavior. In [3] the authors show correlations between social features and spending behavior for a small population. We are currently tackling the problem of predicting spending behavior characteristics on the scale of millions of individuals.

7. REFERENCES

[1] J. Brea, J. Burroni, M. Minnoni, and C. Sarraute. Harnessing mobile phone social network topology to infer users demographic attributes. In *Proceedings of the 8th Workshop on Social Network Mining and Analysis (SNA KDD)*. ACM, 2014.

[2] C. Sarraute, P. Blanc, and J. Burroni. A study of age and gender seen through mobile phone usage patterns in Mexico. In *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ACM, 2014.

[3] V. K. Singh, L. Freeman, B. Lepri, and A. S. Pentland. Predicting spending behavior using socio-mobile features. In *Social Computing (SocialCom), 2013 International Conference on*, pages 174–179. IEEE, 2013.

The City Pulse of Buenos Aires

Carlos Sarraute[1], Carolina Lang[1], Nicolas B. Ponieman[1], and Sebastian Anapolsky[2]

[1]Grandata Labs, Argentina
[2]Mobility and transport specialist

I. Introduction

Cell phone technology generates massive amounts of data. Although this data has been gathered for billing and logging purposes, today it has a much higher value, because its volume makes it very useful for big data analyses. In this project, we analyse the viability of using cell phone records to lower the cost of urban and transportation planning, in particular, to find out how people travel in a specific city (in this case, Buenos Aires, in Argentina). We use cell phones data to estimate the distribution of the population in the city using different periods of time. We compare those results with traditional methods (urban polling) using data from Buenos Aires origin-destination surveys. Traditional polling methods have a much smaller sample, in the order of tens of thousands (or even less for smaller cities), to maintain reasonable costs. Furthermore, these studies are performed at most once per decade, in the best cases, in Argentina and many other countries. Our objective is to prove that new methods based on cell phone data are reliable, and can be used indirectly to keep a real-time track of the flow of people among different parts of a city. We also go further to explore new possibilities opened by these methods.

II. Mobile Data Source

We applied our methodology to Buenos Aires city, the capital of Argentina, which has 2,890,151 inhabitants [1] and is the main political, financial and cultural center of the country. Buenos Aires city is formally divided in 48 neighborhoods, which are grouped for political and administrative purposes in 15 communes.

We have a dataset of geolocalized CDR (call detail records), from which we examine the mobility patterns of mobile phone users. The high penetration of cell phone technology in the city allows us to estimate the mobility patterns of all the inhabitants from this data.

Our dataset has about 4.95 million mobile phone users (1,000 times the number of people in the Buenos Aires survey [2]); it also contains more than 200 million call records generated by these users during a period of five months (from November 1st, 2011 to March 30th, 2012). Each record contains the origin (caller), destination (callee), timestamp, duration of the call and antenna used to connect. In addition, we have the geolocalization of the antennas. We used that information to map the antennas to a certain commune, and we used the map [call→antenna] as dataset of geolocated calls.

III. Methodology

In this section we explain the methodology we used to adapt the Call Detail Records (CDRs) to our objective.

The first step of our method generates, for each particular user, a *Location Distribution Matrix* (LDM) that shows the probability of the user being in a commune c at a given time t of the week. The second step defines a criteria to consider only users whose LDMs give us enough information about their mobility patterns. The last step scales our sample using the population values from the census data.

A. Generation of Location Distribution Matrices

We separated a typical week into four day groups and four hour groups, as shown in Table I.

TABLE I
Day groups and hour groups used in our analysis

Day groups	**Hour groups**	
Monday to Thursday	Morning	5am - 11am
Friday	Noon	11am - 3pm
Saturday	Afternoon	3pm - 8pm
Sunday	Night	8pm - 5am (of next day)

This selection is based on the fact that Monday to Thursday are typical working days, Fridays show different mobility patterns (specially at night), and weekends present a completely different pattern.

The hour group selection corresponds to an analysis realized with the data of [2], from which we determined the peaks and valleys of mobility, for a typical working day in the city.

Let $\mathcal{C}$ be the set of communes and $R_{u,d,h,c}$ the number of calls made by user u on day group d, hour group h, in commune c. The proportion of calls (i.e., the cell values of the LDM) that a certain user u made in commune c during a combination of day group d and hour group h is

$$P_{u,d,h,c} = \frac{R_{u,d,h,c}}{\sum_{c' \in \mathcal{C}} R_{u,d,h,c'}}$$

or 0 if the denominator is zero. The matrix P_u is the *Location Distribution Matrix* of user u.

B. Criteria for Filtering Users

We filter the users that don't provide enough information on their location; more precisely we only take into account the users that have enough calls in every one of the 16 day/hour groups. That is, the user u is kept if

$$\sum_{c' \in \mathcal{C}} R_{u,d,h,c'} \geq \tau$$

for any combination of d and h, given a threshold τ (in our study $\tau = 1$). After filtering, we obtain a set of 73,000 users which we denote $\mathcal{U}$.

C. Scaling up to Census Population

First, we determine the home commune H_u for every user $u \in \mathcal{U}$. We consider that a user is at home on weekdays, at night:

$$H_u = \arg\max_{c \in \mathcal{C}} R_{u,\text{weekday},\text{night},c}$$

In case of a tie, we decide randomly. We registered only 395 ties among the set of valid users $\mathcal{U}$ (0.56% of the cases).

With that information, we extend our predictions using the census data [1]. The scaling factor F_c for commune c is:

$$F_c = \frac{\text{pop}_c}{\#\{u \in \mathcal{U} | H_u = c\}}$$

where pop_c is the population of commune c according to the census. The range of scaling factors goes from 17.26 in commune 2 to 93.29 in commune 8.

We now define the expected quantity of people in a commune c, during a combination of day group d and hour group h, in terms of the proportion of calls of each user in c and the scaling factor of their home commune:

$$EP_{d,h}[c] = \sum_{u \in \mathcal{U}} (P_{u,d,h,c} \cdot F_{H_u})$$

Additionally, the expected quantity of people found in commune c, during a day group d and hour group h, and that live in commune c' is given by:

$$EP_{d,h}[c][c'] = \sum_{u \in \mathcal{U} | H_u = c'} (P_{u,d,h,c} \cdot F_{c'}) .$$

Note that $EP_{d,h}[c] = \sum_{c' \in \mathcal{C}} EP_{d,h}[c][c']$. Having presented the methodology, we now describe the results obtained.

IV. Results

A. Validation Against the Survey

We first validated the proposed methodology, by comparing it with the most traditional method used among the urban mobility studies in Buenos Aires: the origin-destination survey [2], [3].

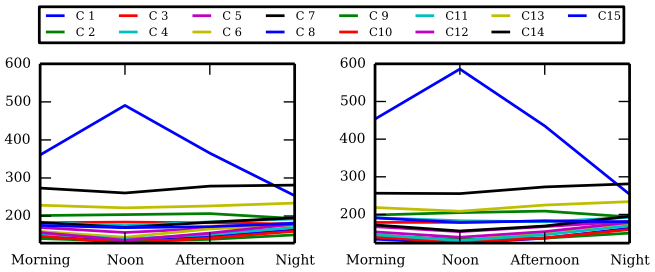


Fig. 1. Comparison between the ENMODO survey (left) and our analysis (right), for a typical working day, and for all the communes in Buenos Aires. The numbers in the legend correspond to the commune numbers. The y axis shows the estimations in thousands of people.

In Fig. 1, we see that the results obtained are similar (both plots show the same growth patterns for each commune). A more detailed analysis of the differences between the two data sources shows that the average difference is 5%. The highest variation appears in Commune 1 (20% in the morning hour group) and the second highest in Commune 6 (11% in the noon hour group). For a more detailed analysis, we refer the reader to [4].

B. Extension to Weekends

Given that we have successfully validated our proxy methodology with the origin-destination survey, we can now use it to extend the analysis to other time periods. We examine here the mobility during the weekends. The mobility survey [2] does not include this information; we are thus presenting here new results on the mobility of the citizens of Buenos Aires.

The patterns for weekends (Fig. 2) are very different: Commune 1, the central business district of the city, is not a major pole of attraction (as it is during weekdays), whereas other communes (mainly Commune 14) are more attractive for citizens on weekends. Commune 14 is well known for its bars, restaurants and night clubs, so this pattern coincides with our insight on the social life in this commune.

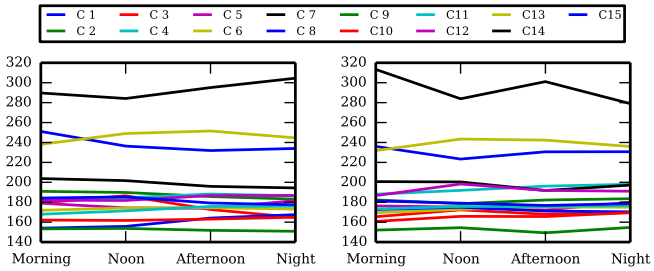


Fig. 2. Predictions for a typical Saturday (left) and Sunday (right) according to our methodology, for all the communes in Buenos Aires. The numbers in the legend correspond to the commune numbers and the y axis shows the estimations in thousands of people.

C. Analysis of Commune 14

We analyse in more detail Commune 14 (Palermo), which has very particular characteristics (see Fig. 3). First of all, we remark it has a typical residential commune pattern for weekdays (with a lower concentration of people during working hours, and a higher concentration at night). During weekends,

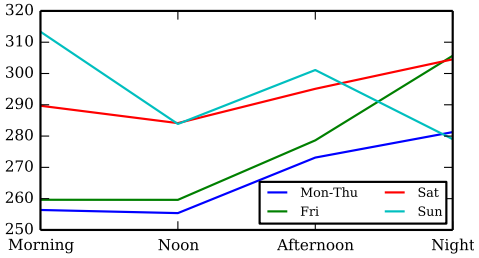


Fig. 3. Predictions according to our method, for the different day types, for Commune 14 (Palermo). The y axis shows the estimations in thousands of people.

however, Commune 14 shows a special behavior due to its role as social and nightlife hub. During Fridays, we notice an increase of people during the night when compared to other weekdays, which we attribute to people going out. Saturdays show an increase in population across all time groups, with a peak at night that is similar to the one on Friday, and Sunday night has the same quantity of people than a regular working day at night, probably because people will have to go to work on the following day. Moreover, we notice a similar number of people on Friday night compared with Saturday morning, and on Saturday night compared with Sunday morning. This fact may be explained considering nightlife in Buenos Aires extends into the morning (even until 8am). All these observations are coherent with our knowledge of the city.

D. The City Pulse Matrix

The urban mobility information can be used to generate what we call the *City Pulse Matrix* (CPM), a 2-dimensional matrix such that, for any day group d and hour group h,

$$CPM[i][j] = EP_{d,h}[i][j].$$

Fig. 4 shows our visualization of the matrix generated by our predictions, on a typical weekday noon (which is the time period that varies the most with respect to weekday nights).

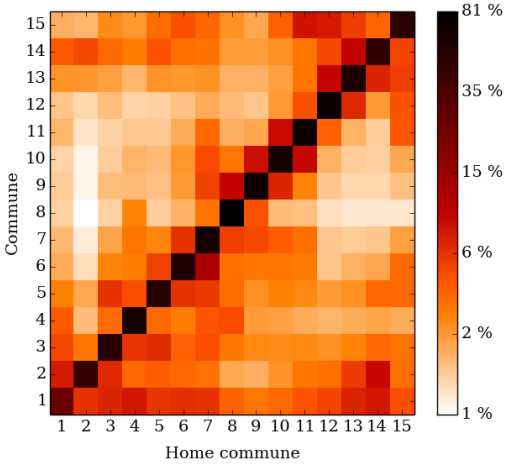


Fig. 4. Visualization of the *City Pulse Matrix* generated with our methodology, for a weekday (Monday to Thursday) noon, with values normalized by row.

We can see in Fig. 4 that there is a darker diagonal, meaning that in all the communes, most of the people that spend their weekday noon in a given commune also live there. The lightest element in the diagonal corresponds to Commune 1 (with 24%, followed by Commune 2 with 43%), because of the flow of people from the rest of the city that work there.

E. Visualizing the City Pulse

Finally, Fig. 5 presents a visualization of the *city pulse*. We plotted a map showing for Commune 1 and Commune 6 the number of people present there on a typical working day (Monday to Thursday) at noon, according to their home communes. Commune 1 is the central business district so many people work there during the day, coming from very diverse locations. Commune 6, on the other hand, is one of the most populated and dense communes of the city (and represents its geographical center), but is mainly residential. The difference in the number of people and variety of provenance between a central business district as Commune 1 and a more residential district as Commune 6 can be seen clearly in Fig. 5. We have also done a more complete analysis including other communes and day and hour groups (as shown in Table I) achieving similar results.

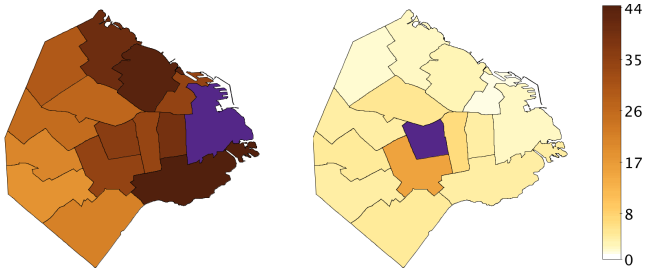


Fig. 5. Visualization of the number of people present on Monday to Thursday noon period in (a) Commune 1 and (b) Commune 6 (colored in violet) that live in each of the other communes. The scale shows the number of people (in thousands) each color represents.

V. Conclusions and Future Work

We presented a methodology to estimate the flow of people between different parts of the city using mobile phone records. According to our validation, the method is reliable, presenting an average difference of 5% with the origin-destination survey [2].

We extended the analysis to weekends using the proposed methodology, and found many interesting patterns which are coherent with our knowledge of the city. For instance, we showed how Commune 1, the central business district, yields during the weekends its role as major pole of attraction to Commune 14, which is a social and nightlife hub. We also presented a visualization where a business and a residential district can be clearly differentiated. A more detailed analysis of this methodology was published in [4].

We finally introduce ideas for future work: (i) achieve a finer spatial granularity with a richer dataset; (ii) consider the metropolitan region (suburbs) of the city in the analysis, as many people travel between the capital and its suburbs every day; (iii) analyze the mobility of citizens during particular situations or events (for example, an evacuation or a holiday).

References

[1] Instituto Nacional de Estadística y Censos (INDEC). *Censo Nacional de Población, Hogares y Viviendas 2010*, volume 1. INDEC, October 2010.

[2] Secretaría de Transporte. Ministerio del Interior y Transporte. ENMODO (2009-2010). Resultados de la encuesta origen destino. Movilidad en el area metropolitana de Buenos Aires, 2010.

[3] Sebastian Anapolsky. Los flujos de movilidad territorial: un análisis de la población y la movilidad en el área metropolitana de Buenos Aires. *Revista Digital Café de las Ciudades*, 133-134, 2013.

[4] Sebastián Anapolsky, Carolina Lang, Nicolás Ponieman, and Carlos Sarraute. Exploración y análisis de datos de telefonía celular para estudiar comportamientos de movilidad en la Ciudad de Buenos Aires. In *XVIII CLATPU, Congreso Nacional de Transporte Público y Urbano*, 2014.

Detecting and understanding big events in big cities

Barbara Furletti, Lorenzo Gabrielli, Roberto Trasarti
KDDLAB - ISTI CNR, Pisa, Italy
name.surname@isti.cnr.it

Zbigniew Smoreda, Maarten Vanhoof, Cezary Ziemlicki
Sociology and Economics of Networks and Services dept., Orange Labs, Paris, France
name.surname@orange.com

Recent studies have shown the great potential of big data such as mobile phone location data to model human behavior. Big data allow to analyze people presence in a territory in a fast and effective way with respect to the classical surveys (diaries or questionnaires). One of the drawbacks of these collection systems is incompleteness of the users' traces; people are localized only when they are using their phones. In this work we define a data mining method for identifying people presence and understanding the impact of big events in big cities. We exploit the ability of the Sociometer for classifying mobile phone users in mobility categories through their presence profile. The experiment in cooperation with Orange Telecom has been conduced in Paris during the event *Fête de la Musique* using a privacy preserving protocol.

The objective of this study is to investigate the impact of big events in big cities on the contemporary composition of the population. The method foresees the application of a data mining tool, called Sociometer [5, 4] on a mobile phone dataset collected in Paris in the month in which the *Fête de la Musique* occurs. This event, also known as World Music Day, is an annual music festival taking place on June 21, the first day of summer in cities around the world. The Sociometer, by analyzing aggregated presence profiles of mobile phone users, is able to classify a population in mobility categories hereby differentiating between residents, dynamic residents, commuters, and visitors. The presence profiles are represented by an aggregated presence matrix on weekly basis (weekdays and weekends) and are obtained by counting the cell phone registrations of individuals in the areas of interest. By means of a data mining strategy, the Sociometer classifies each profile. Starting from this partition, we design a strategy for identifying how the event impacts on the composition of the population, exploring both a temporal and a spatial dimension. The spatial dimension is characterized by the partitioning of Paris in three areas (identified as P1, P2, and P3 so that P2 includes P1, and P3 includes P2 - Fig. 1) based on the grouping of several administrative borders. The temporal dimension is a window of one month of mobile phone observations analyzed with weekly and daily granularity.

Fig. 2 shows the variation of the categories population categories over the three areas during the whole period of observation. It is evident that the number of visitors decreases from the city center (the more touristic area) to the larger peripheral areas. Of course, this confirms the impact

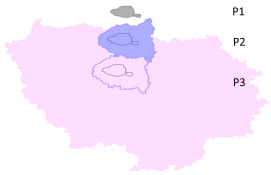


Figure 1: Administrative partitioning of Paris area: P1, P2, and P3.

of visitors in the city center of Paris, but it also implicates a sort of interplay between the city center and it's wider that could be interesting to define the complex usage of the city center by its surrounding inhabitants.

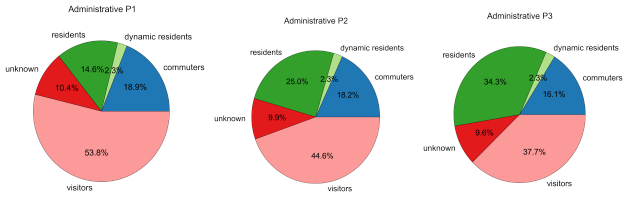


Figure 2: Variation of the composition of the population in the three Administrative areas of Paris.

To deeper investigate this result, we perform a sort of multi-classification analysis starting from P1 and seeing how the classification of each category of people may change enlarging the observation area (Fig. 3). Let us consider for example the set of visitors in P1 (clearly are 100% in P1), this set, in P2 and P3 become progressively 70% and 55%. This means that the 45% of them, originally classified as visitors in P1, indeed belong to a different category if we look at a bigger area. In other words they are not "foreign" in a strict sense, in fact almost the 20% of them, are actually resident in P3. Moreover, Some of the visitors in P1 become commuters in P2 and resident in P3. This may happen for the users that work in P2, live in P3, and that visit the city center only once in a while. This multi-classification analysis adds a new dimension to the classification allowing the analyst to refine and extend the categories with a new class "Tourist" for users which remain visitor in all three zones, and "Occasional visits of Resident" for users which are visitors in P1 but residents in P2 or P3.

In general, an event in the city can be detected through the study of the distributions of the presence of people categories, and in particular of the visitors. As shown in Fig. 4,

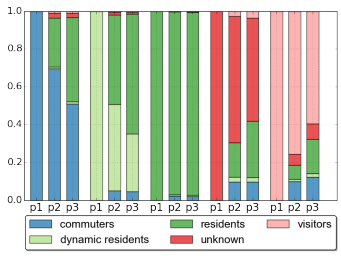


Figure 3: Multi-Classification of each category of users over the 3 areas.

for Paris a peak of presences is not so evident in the week of the event (the weekdays labeled with "25 weekdays") when we consider the weekly distributions. This can be justified with the fact that the event is held for only one day, as well as by the fact that June forms the unofficial start of the tourist season which explains the increasing trend of presences in the whole month. The event is thus hidden by the normal activities and daily dynamics of the city.

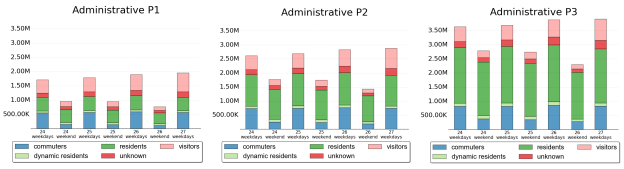


Figure 4: Weekly distribution of the presences in the three Administrative areas of Paris.

For detecting big (but short in time) events in very big cities (like Paris), we come to the conclusion that it is necessary to lower the temporal granularity (in this case from weekends/weekday to days) when we study the presence distribution of people already classified. It is important to notice that the first step of the analysis, i.e. the classification with the Sociometer, uses profiles aggregated on weekly basis. As shown in Fig. 5, the daily distributions on daily basis of the presences and the calls actually highlight a peak on June 21st.

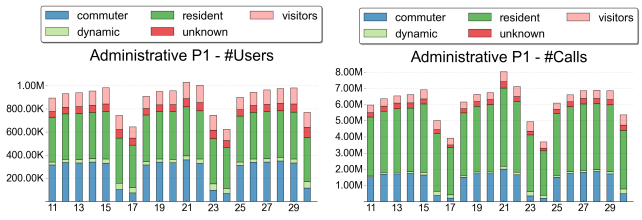


Figure 5: Daily distribution of presences and calls in P1 in the month of June.

Computing the multi-classification during the only day of the event, we find that the event is mostly a big attractor for people around Paris rather than the classical tourists coming from outside. In fact the 41% and 58% of the visitors in P1 does not remain visitors in P2 and P3, respectively. This observation gives rise to the interpretation that the *Fête de la musique* is a festival for the Parisians themselves rather than for people coming from a long distance. Such an interpretation is not surprising at all as the fe one day (a Thursday even) and imbeds within a nationwide event in which all French cities have festivities.

Due the sensitive nature of the data, we have taken into account the privacy issues during the entire process of analysis customizing and applying the privacy risk analysis method presented in [3] and already tested in the work presented at CPDP in 2013 [6]. This methodology implements and satisfies the constraints issued by the European Union for data protection in [2] and follows the principle given in [1]. The risk analysis follows the idea that, given a dataset and a specific application, it is possible to define the set of attacks w.r.t. different levels of knowledge in order to evaluate the risk of linkability and re-identification. After a risk is detected, a technique for anonymizing the data is chosen, realizing a good trade-off between privacy guarantee and quality of service.

Conclusions

The analytical process we described shows how to use the Sociometer to classify people in categories during an event, and it allows to reason about the event attractiveness. It also points out how the concept of city may change depending on the spatial granularity. The study of an event with reference to the different categories of population instead of an undefined group of people brings out how differently an event impacts (attracts) people at urban level. Through the weekly analysis of the call behaviors we are able to identify a general increasing of presences across the month, while the event *Fête de la musique* emerges by computing distributions on daily basis. In the case of Paris, the fact that it is a very important city from the touristic point of view and that attracts many visitors especially in the period of analysis, contributes to hide the event behind the daily dynamics of the city. The *Fête de la musique*, as reported by the domain experts, is actually a very important event that attracts tourists and Parisian, and that involves all the city, nevertheless, it does not affect the presences on weekly basis. With this analysis we confirm that this event has a big effect on local residents more than external visitors. In summary, with this work we meet the following objectives: (1) Verify how our proposed data mining methodology performs in the discovering of big events in big cities; (2) Identify the presence of visitors during the Festival by means of the Sociometer; (3) See how the composition of the population changes along the period of observation; (4) See how the classification of the population changes considering different spatial resolution of Paris.

Acknowledgments. This work has been partially funded by EIT ICT Labs - Project City Data Fusion for Event Management (activity n. 14189).

References

[1] E. U. for data protection. Article 6.1(b) and (c) of directive 95/46/ec and article 4.1(b) and (c) of regulation ec (no) 45/2001, 2001.

[2] E. U. for data protection. Opinion 05/2014 on anonymisation techniques, 2014.

[3] B. Furletti, L. Gabrielli, F. Giannotti, A. Monreale, M. Nanni, D. Pedreschi, F. Pratesi, and S. Rinzivillo. Assessing the privacy risk in the process of building call

habit models that underlie the sociometer, 2014.

[4] B. Furletti, L. Gabrielli, C. Renso, and S. Rinzivillo. Analysis of gsm calls data for understanding user mobility behavior. In *Proceedings of the BigData Conference*, pages 550–555. IEEE, 2013.

[5] B. Furletti, L. Gabrielli, C. Renso, and S. Rinzivillo. Tourism fluxes observatory: deriving mobility indicators from gsm calls habits. In *In the Book of Abstracts of NetMob.*, 2013.

[6] S. Mascetti, A. Monreale, A. Ricci, and A. Gerino. Anonymity: A comparison between the legal and computer science perspectives. In S. Gutwirth, R. Leenes, P. D. Hert, and Y. Poullet, editors, *European Data Protection*, pages 85–115. Springer, 2013.

Vehicle-Relative Positioning System

Márton Hunyady
hunyadym@hunyadym.hu

Gergely Lukács
lukacs@itk.ppke.hu

András Oláh
olah.andras@itk.ppke.hu

Faculty of Information Technology and Bionics, Pázmány Péter Catholic University
Práter u. 50/A, 1083 Budapest, Hungary

1. INTRODUCTION

Personal mobility data is useful both for helping the journey of the data collector and, when aggregated, for improving travel infrastructure. In some application areas, such as local or long-range public transport, taxi or car sharing, absolute location and trajectories based on them are not sufficient; the more important question is whether a particular person is actually in a particular vehicle.

The goal of this paper is vehicle-relative positioning with smartphones, i.e. deciding whether a person is in a particular vehicle or just near it, possibly in another vehicle. Simply matching the spatio-temporal trajectories is not sufficient for this, because of the inherent quality limits of the data. Our approach combines spatio-temporal data with data from other sensors, which serves as the basis for creating a comprehensive system.

2. RELATED WORK

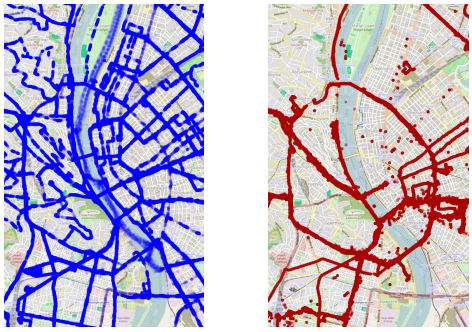
The concept of *trajectory flock* is used to describe a set of objects moving close to each other for some period of time [1], though the question of a shared single vehicle is not analysed. Activity recognition, typically using accelerometer data, [2] is used to recognize physical activities, e.g., walking, sitting, running. Activity recognition is also becoming part of the mobile operating systems, and special sensors are also appearing in the devices (e.g. step detector). Vehicle and passenger accelerator data were analyzed in [3] and [4]. Indoor positioning approaches typically use WiFi or Bluetooth scanning or magnetometer data [5]. Similar approaches are used to detect face-to-face meetings [6].

3. APPROACH

Our approach is based on large-scale data collection, analysis of trajectory data and sensor data mining.

3.1 Data collection

The area of study was Budapest (2 million inhabitants). Dynamic public transport vehicle data is available in a Web-application and was collected by our software robot. Vehicle positions are refreshed ca. every 30 seconds. The daily number of vehicle positions is ca. 3-4 millions (Fig. 1a). At the time of the analysis data for about one month was available. We developed a smartphone application to assess location and sensor data about the personal movements. Data was collected by 15 persons, covering altogether 140 hours of travel time, ca. 340 000 locations (ca. 1.5 seconds temporal granularity; Fig. 1b), and sensor data (accelerometer, gravitational, magnetometer and gyroscope, appr. 200 Hz sampling frequency).



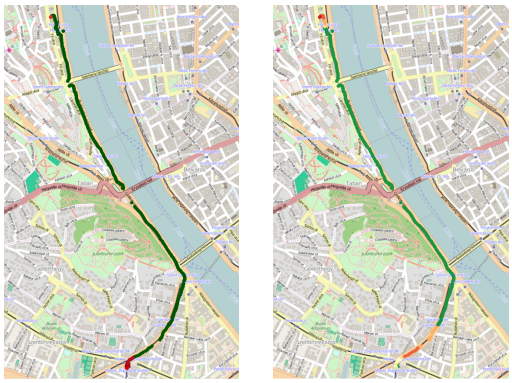
(a) Vehicle locations for one day (b) User locations

Figure 1: The collected vehicle and user location data.

3.2 Trajectory matching

We matched personal trajectories to vehicle trajectories in the following steps:

1. Efficient spatial and temporal prefiltering of distant vehicles;
2. Linear interpolation to compensate for the unsynchronized timing of vehicle and personal position data;
3. Sophisticated weighed multiple point-pair matching, providing also a mean squared error based confidence measure. This is necessary because the matching of



(a) Correctness of detection (green when correct) (b) Calculated confidence (green is better)

Figure 2: The results of the trajectory-based vehicle detection

a single personal location to the nearest vehicle position performs poorly. First, position data has limited accuracy. Second, in long vehicles a passenger can be closer to the tracking device of another vehicle than that of their own vehicle. Our approach eliminates both problems.

The performance of the trajectory matching algorithm is high in the middle of the trajectories and low when boarding or disembarking from a vehicle, as correctly reflected by the confidence measure also delivered by the algorithm (Fig. 2). Thus, additional measures are required.

3.3 Sensor data based matching

To create a model for sensor based activity and vehicle type recognition, various features of the time series data (maximum, integral, root mean square, spectrum values, all on a window) were first calculated. Data was labelled by activity and vehicle type partly by hand, partly using trajectory matching with past data. (Fig. 3).

Data mining algorithms were trained and tested on this data, where test data was selected across different vehicles and persons. The binary prediction (vehicle/on foot) performed best with the C4.5 algorithm, achieving 97.5% accuracy. The multiclass prediction of the vehicle type (bus/tram) had a performance of 86% using Bayesian network. The learning algorithms selected mainly accelerometer and, to a smaller extent, magnetometer and gyroscope data as relevant. After training, prediction with these algorithms is efficient and can be run on smartphones.

The combination of trajectory matching and sensor based matching achieves better overall results. The uncertainty of trajectory matching at the beginning and at the end of the journey is compensated for by sensor data. Past data can also be used to improve the quality of sensor data based matching, as it can be automatically labelled by trajectory matching.

4. POTENTIAL APPLICATIONS

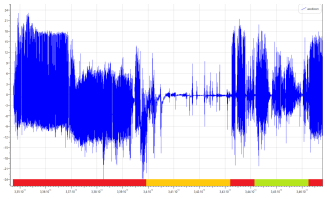
One possible application area of vehicle-relative positioning is public transportation. Here individually customized transportation information for users as well as improved statistics and control information for public transportation companies can be obtained from the data. Beside public transportation, dynamic car-pool organization and the detection of face-to-face meetings inside vehicles can be supported by our approach.

5. SUMMARY AND OUTLOOK

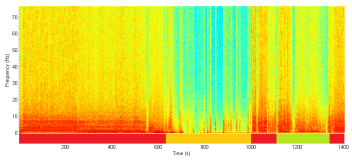
We presented an approach for vehicle-relative positioning based on trajectory matching and extended by sensor data based matching. The approach was applied for a reasonable amount of data and yielded promising results.

Future work is planned on improving the algorithm, testing its performance and streamlining data collection and processing so that the system can be scaled up for a large number users.

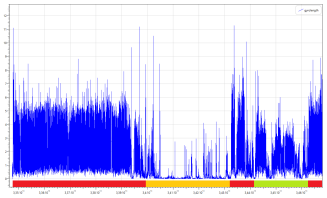
The research was partially financed by the Rectors's Office of Pázmány Péter Catholic University, project number KAP-1.2-14/007.



(a) Vertical acceleration



(b) Spectrum of vertical acceleration



(c) Magnitude of gyroscope values

Figure 3: Some of the calculated features of the sensor values. Below the graphs the colors show the actual activity: red is walking, yellow is tram, green is cycling.

6. REFERENCES

[1] Chiara Renso, Stefano Spaccapietra, and Esteban Zimányi. *Mobility Data*. Cambridge University Press, 2013.

[2] Thomas Bernecker, Franz Graf, Hans-Peter Kriegel, Christian Moennig, Dieter Dill, and Christoph Tuermer. Activity recognition on 3d accelerometer data (technical report). 2012.

[3] Ina Partzsch. Positioning in real-time public transport navigation: Comparison of vehicle-based and smartphone-generated acceleration data to determine motion states of passengers. In *Networks for Mobility 2012. 6th International Symposium. Proceedings*, page 10, Stuttgart, September 2012. Universität Stuttgart.

[4] Samuli Hemminki, Petteri Nurmi, and Sasu Tarkoma. Accelerometer-based transportation mode detection on smartphones. SenSys '13, pages 13:1–13:14, New York, NY, USA, November 2013. ACM.

[5] Jaewoo Chung, Matt Donahoe, Chris Schmandt, Ig-Jae Kim, Pedram Razavai, and Micaela Wiseman. Indoor location sensing using geo-magnetism. In *Proceedings of the 9th International Conference on Mobile Systems, Applications, and Services*, MobiSys '11, pages 141–154, New York, NY, USA, June 2011. ACM.

[6] Piotr Sapiezynski, Arkadiusz Stopczynski, and Sune Lehmann. Detecting face-to-face meetings using smartphone sensors. In *Analysis of Mobile Phone Datasets (NetMob), 2013 3rd International Conference on. MIT*, pages 52–53, Cambridge, MA, May 2013.

Change Detection in Human Mobility Patterns from Successive OD Matrices

Julio C. Chaves[1,2], Moacyr A. H. B. da Silva[1], Alexandre G. Evsukoff[2]

[1]Getúlio Vargas Foundation, School of Applied Mathematics,
Praia de Botafogo 190, Botafogo, Rio de Janeiro, Brazil

[2]Coppe/Federal University of Rio de Janeiro,
Cidade Universitária, Rio de Janeiro, Brazil

Origin-Destination (OD) matrices have been widely used for transportation planning and traditionally is computed from expensive surveys, made once each several years. The availability of mobile phone data allow to compute OD matrices easily from Call Detail Records (CDR) data. Recent studies have shown that mobile phone data are a reliable source of information to compute OD matrix [1,4,5] and the mobility of phone users can be extrapolated to the whole population.

This work present a study to detect changes in human mobility patterns in the Rio de Janeiro metropolitan area. The study was carried out from a dataset collected during 2014, comprising the activity of an average of 2,7 millions of users and 1,700 cell towers. The dataset was provided by one mobile carrier and only outgoing calls were available. The Voronoi regions of cells are aggregated to represent district of the Rio de Janeiro city or suburb cities in the metropolitan area.

The dataset were cleansed before used to compute OD matrices. As only outgoing calls were available, the effect of pre-paid users was expressive. In Rio de Janeiro, about 80% to 85% of the users correspond to pre-paid contracts and those users are mostly low-income population, in contrast, post-paid users correspond to medium to high-income population.

These asymmetries were corrected from census data, in order to gather user mobility as an approximation of the population mobility. The place of residence is estimated by the most frequent call location during the night and weekends (provided some constraints). The place of residence is used to compute the user/population ratio for each neighborhood and adjust the number of users to the number of inhabitants.

In this work, daily OD matrices are computed and only users with a recorded displacement within a day are considered. If an user has visited more than two different places within a day each pair of visited locations are used to compute the OD matrix in that day. The OD matrix computed in the number of trips is divided by the total number of trips. All the process is being implemented in a high performance database system, such that a daily OD matrix can be computed in seconds.

The procedure is employed to compute successive OD matrices, one for each day, during the whole year of 2014. The OD matrix is computed with respect to the neighborhoods and correcting for the user/population ratio for each neighborhood. For each day t, the element $OD_{ij}(t)$ represent the mobility from neighborhood i to neighborhood j on that day. As result, there is one time serie $OD_{ij}(t)$, $t = 1,...,N$ for each pair of neighborhoods. The analysis of these

time series will allow to identify differences of mobility in workdays and weekends as so as changes due to seasonality or events. The change detection is computed by standard statistical methods like CUSUM [2,3] as reference, but other methods will also be investigated.

The year of 2014 is particularly interesting for this study as it was the year of the FIFA 2014 World Cup, which caused a great impact in the city mobility, specially in the days of the national team other games and games in the city. Moreover, the city is under work with the introduction of new transportation systems a weel as a complete refurbishing of the harbor region with the dismantlement of a former important expressway. Therefore we expect to verify the impact of these changes in the mobility of the city.

[1] J. Park, D.-S. Lee, and M. C. González, “The eigenmode analysis of human motion,” Journal of Statistical Mechanics: Theory and Experiment, vol. 2010, no. 11, p. P11021, Nov. 2010.

[2] M. Bassevile and I. Nikiforov, “Detection of abrupt changes:Theory and Application”. Prentice Hall, 1993.

[3] OAKLAND, J. S., "Statistical Process Control Paperback - September 26, 2007". Routledge, 2008. ISBN: 978-0-7506-6962-7.

[4] CANDIA, J., GONZÁLEZ, M. C., WANG, P., et al. "Uncovering individual and collective human dynamics from mobile phone records", Journal of Physics A: Mathematical and Theoretical, v. 41, n. 22, pp. 224015, jun. 2008.

[5] SLINGSBY, A., BEECHAM, R., WOOD, J. "Visual analysis of social networks in space and time using smartphone logs", Pervasive and Mobile Computing, v. 9, n. 6, pp. 848?864, dez. 2013.

Inferring social status and rich club effects in enterprise communication networks[1]

Yuxiao Dong†, Jie Tang‡, Nitesh V. Chawla†,*, Tiancheng Lou♯, Yang Yang†, Bai Wang¶
† Interdisciplinary Center for Network Science and Applications, Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN, U.S.A.
‡ Department of Computer Science and Technology, Tsinghua University, Beijing, China
♯ Google Inc., Mountain View, CA, U.S.A.
¶ Department of Computer Science and Technology, Beijing University of Posts and Telecommunications, Beijing, China
∗ **Email: nchawla@nd.edu**

Abstract

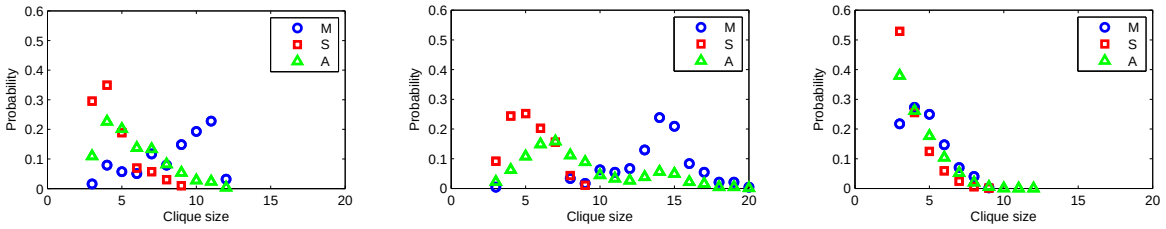
Social status, defined as the relative rank or position that an individual holds in a social hierarchy [3], is known to be among the most important motivating forces in social behaviors [5]. In this paper, we consider the notion of status from the perspective of a position or title held by a person in an enterprise. We study whether enterprise communication logs can help reveal how social interactions and individual status manifest themselves in social networks. To that end, we use two enterprise datasets with three communication channels — voice call (CALL), short message (SMS), and email (EMAIL) — to demonstrate the social-behavioral differences among individuals with different status. Specifically, we investigate the interplay of social status and several well-known social theories, including structural hole, social balance, homophily and social clique.

Structural Hole. The principle, that individuals can benefit from serving as intermediaries between others who are not directly connected, forms the underpinning for the theory of structural holes [1]. Our analysis clearly shows that managers are more likely (70% in CALL, 55% in SMS, and 43% in EMAIL) to be spanned as structural holes across the three networks. In other words, the structural holes extracted from enterprise communication network structure reveal the social status of staff in their company. This can be explained by the fact that managers usually need to operate the responsibility of correspondents and organizers within the company, especially for the experience for connecting different departments or groups to cooperate.

Link Homophily. Homophily is the tendency of individuals to associate and bond with similar others [4]. Lazarsfeld and Merton [4] argued that individuals with similar social status are more likely to associate with each other, which is called status homophily. Our analysis shows that two individuals who share more common neighbors will have a tendency to have similar social status in the company. Managers' ability of creating and maintaining social connections in enterprise networks is more prominent that subordinates'. This could have the potential to further promote their status in companies, which further highlights the rich club effect.

Social Balance. Triad is one of the simplest grouping of individuals that can be studied and is mostly investigated by microsociology [3]. We find that the managers' overall balance ratios are larger than the subordinates' across all the three channels. Moreover, the individuals in organizations have the tendency to maintain balanced relationships with people of the same status; this phenomenon coincides with the link homophily observed above. We conjecture that

[1]This work is accepted at PLOS ONE, 2015 and the abstract is also submitted to the International Conference on Computational Social Science (IC^2S^2).



(a) CALL (b) SMS (c) EMAIL

Figure 1. Social Clique vs. Social Status. Distribution of social clique in enterprise communication networks. M: Managers; S: Subordinates; A: All employees;

the relatively high status empowers the managers to connect with more people and maintain the relationships within the enterprise, enhancing the chance to promote their status.

Social Clique. Clique is a concept in both social sciences and graph theory. In social sciences, clique is used to describe a group of persons who interact with each other more regularly and intensely than others in the same setting. Here we aim to examine how managers and subordinates form cliques and to which extent they are connected. We build two sub-networks that only contain mangers or subordinates respectively for each type of a network derived from each of the communication channels. Figure 1 shows the distributions of clique size, conditioned on the status of individuals (employees in the enterprise). For reference, we also plot the overall clique distribution in each full network. It is obvious that the distributions of managers and subordinates are quite different and the maximal cliques for managers are much larger than these for subordinates.

The correlations between social status and several social theories provide the evidence of "rich club" [2] maintained by high-status individuals. Inspired by the observations around the social structure and characteristics, and their potential to infer social status in a network, we also developed a computational model—Factor Graph Model—to predict social status using the aforementioned characteristics as features. We demonstrate that the social status of more than 85% – 93% of individuals can be inferred from their communication interactions among their colleagues. This prediction results further confirm our observations on communication behaviors and social theories are general across different companies, even with different communication channels (CALL vs. SMS vs. EMAIL).

References

1. R. S. Burt. *Structural Holes : The Social Structure of Competition.* Cambridge, Mass.: Harvard University Press, 1995.
2. V. Colizza, A. Flammini, M. A. Serrano, and A. Vespignani. Detecting rich-club ordering in complex networks. *Nat Phys*, 2(2):110–115, Feb. 2006.
3. D. Easley and J. Kleinberg. *Networks, Crowds, and Markets: Reasoning about a Highly Connected World.* Cambridge University Press, 2010.
4. P. F. Lazarsfeld and R. K. Merton. Friendship as a social process: A substantive and methodological analysis. In M. Berger, T. Abel, and C. Page, editors, *Freedom and Control in Modern Society*, pages 18–66. Van Nostrand, New York, 1954.
5. L. Milroy and J. Milroy. Social network and social class: Toward an integrated sociolinguistic model. *Language in Society*, 21:1–26, 3 1992.

Home and Work Estimation from Mobile Phone Data: Improving Accuracy and Privacy through Spatial Aggregation

Bradley Sturt[1], Jameson L. Toole[1], Serdar Çolak[2], Marta C. González[2]
[1] *Engineering Systems Division, MIT, Cambridge, MA, 02139*
[2] *Department of Civil and Environmental Engineering, MIT, Cambridge, MA, 02139*
bsturt@mit.edu

1 ABSTRACT

Analysis of mobile phone data is becoming increasingly complex. Many algorithms and applications still use simple methods of detecting and assigning home and work locations to users from mobile phone metadata. There are clearly errors in these simplistic approaches, but it is unclear how much they affect other results. There are also issues with privacy. We make progress on both fronts through demonstrating a new approach for measuring home and work. Through applying our method to the Boston metropolitan area, we show that the proposed approach can improve the accuracy of home and work commuting flow estimations while simultaneously improving and preserving the privacy and anonymity of users.

2 INTRODUCTION

Detecting the home and work locations of residents of cities is vital to numerous domains, including transportation, urban planning, and sociology. The emergence of mobile phone metadata offers an opportunity to detect home and work locations of users [1]. Through leveraging techniques to find aggregate home and work locations at the census tract-level, mobile phone metadata can be used to dynamically estimate population densities [2], commuting flows [3], and congestion [4].

Despite the ubiquitous use of home and work locations, the simplicity of the techniques to detect them from mobile phone data at census tract aggregation have problems.

First and foremost, there are quality problems with any Cell Detail Record (CDR) dataset, ranging from the noise inherent to the spatial component of calls, to irregularity and sparsity in CDR entries for individual users. The accurate estimation of home and work locations from a dataset hinges on the dataset's quality, and in the presence of these common problems, the accuracy will undoubtedly be degraded.

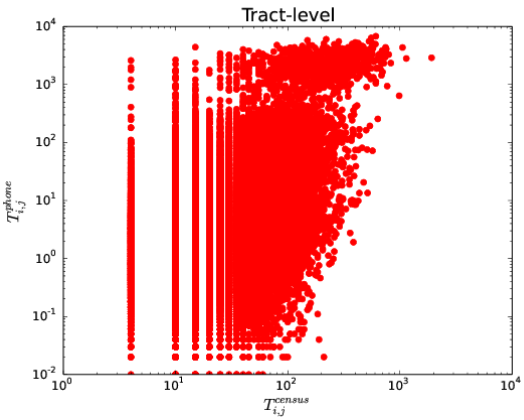


Figure 1: The correlation between home and work trips from the census (x-axis) and mobile phone data (y-axis) for the Boston metropolitan area. The Pearson Correlation Coefficient equals 0.5312

However, the extent to which the accuracy of home and work locations is affected is difficult to quantify, due to the absence of a high-accuracy, ground truth dataset. The most common candidate to compare against is census data. At the census tract level, the United States government publishes the Census Transportation Planning Product (CTPP), which provides estimations of true tract-to-tract home and work commuting flows (i.e. the number of residents of one tract who commute to work on another). Figure 1 shows the correlation between CDRs and CTPP at the tract level.

Due to the often low commuting flows between tracts and the limited sample size used to conduct the survey, there is high uncertainty with many of the commuting flow estimations in the CTPP. Thus, while it is tempting to interpret Figure 1 as the accuracy of the home/work estimations from a CDR, its correlation is of limited insight given the inaccuracies in both datasets.

Second, there are privacy issues from finding home

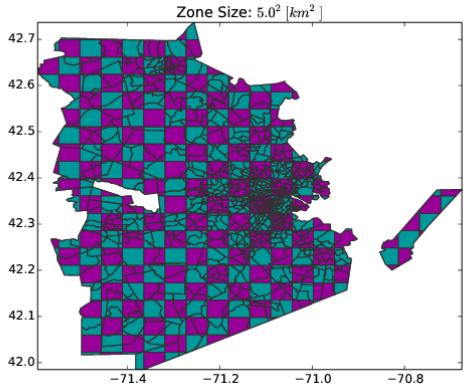


Figure 2: The Boston metropolitan census tracts, aggregated to zones of size $5.0^2[km^2]$

and work commuting flows at the census tract level, stemming from the low commuting flow counts between many tracts. The privacy and anonymity of phone users can be compromised with relatively small data samples [5], and in the census data, many census tract pairs have commuting flows with fewer than 10 individuals. While there are certainly errors with the CTPP, due to the prevalence of tract pairs with low commuting flows, it is safe to assume the trend is likely to reflect reality. Thus, calculating the home and work locations of users has potential to infringe upon the anonymity of users, for users that commute between two tracts with low commuting flows.

3 METHODS

Lenormand et al.[6] analyzed the similarities and differences of the aspects of a city that were inferred using different metadata sources (Twitter, cell phone records, and the census) Because the data came at different spatial regions and different levels of granularity, the authors rasterized the spatial area of a city into square-regions of 1 and 2 kilometers.

Adopting the rasterizing approach, we demonstrate that quality and anonymity of the tract-level commuting data can be ameliorated through spatially-aggregating the data through rasterization at various granularities.

The intuition behind spatial aggregation is that flows between square regions that encompass multiple census tracts will be greater and less noisy than flows between individual census tracts. The rasterizing of the metropolitan area of Boston at a cell size of $5.0^2[km^2]$ is visualized in Figure 2. For the rest of this paper, we will refer to the square cells as *zones*, and we will denote any spatial area as a *region*.

3.1 COMPUTING AGGREGATED FLOWS

First, we aggregate the tract-to-tract commuting flows from the census data. Let T^t_{ij} and T^z_{ab} denote the commuting flows between tracts i and j or zones a and b, respectively. Furthermore, let P^t_i and P^z_a denote the polygons corresponding to tract i and zone a, respectively, and let $|P|$ denote the area of a polygon P.

Our goal is to compute the commuting flows T^z_{ab} between each pair of zones a and b. Therefore, for each pair of zones a and b, we compute

$$T^z_{ab} = \sum_{\forall i,j} \left(\frac{|P^t_i \cap P^z_a|}{|P^t_i|} \right) \left(\frac{|P^t_j \cap P^z_b|}{|P^t_j|} \right) T^t_{ij} \quad (1)$$

where $\frac{|P^t_i \cap P^z_a|}{|P^t_i|}$ is the portion of tract i that overlaps with zone a.

Second, we aggregate the home and work locations from the CDR data. For each user, we compute their home and work coordinates [3]. Then, for each pair of zones, we compute the number of users in the CDR dataset with home and work locations in the two zones.

4 RESULTS

We compute the home and work commuting flows and calculate the correlations between the census and CDRs at various zone sizes. Figure 3 shows the correlations between the datasets at zone sizes of $1^2[km^2]$,$2^2[km^2]$,$5^2[km^2]$, and $10^2[km^2]$. The correlation between the datasets increases drastically as the zone size increases. The correlation between the datasets at the $1^2[km^2]$ zone size is 0.5041, which is not any improvement over the tract level. At greater zone sizes, however, the correlations improve to 0.9360 and 0.9598 at zone sizes of $5^2[km^2]$ and $10^2[km^2]$, respectively.

First, we interpret the correlation as follows: there are underlying quality problems in both datasets, regardless of aggregation. However, if we assume that the errors between the datasets are uncorrelated, we can conclude that the high correlation at larger zone sizes implies that the computed flows are likely to be close to reality. Furthermore, since the results of home and work estimations are highly correlated to the census data, the results suggest that mobile phone

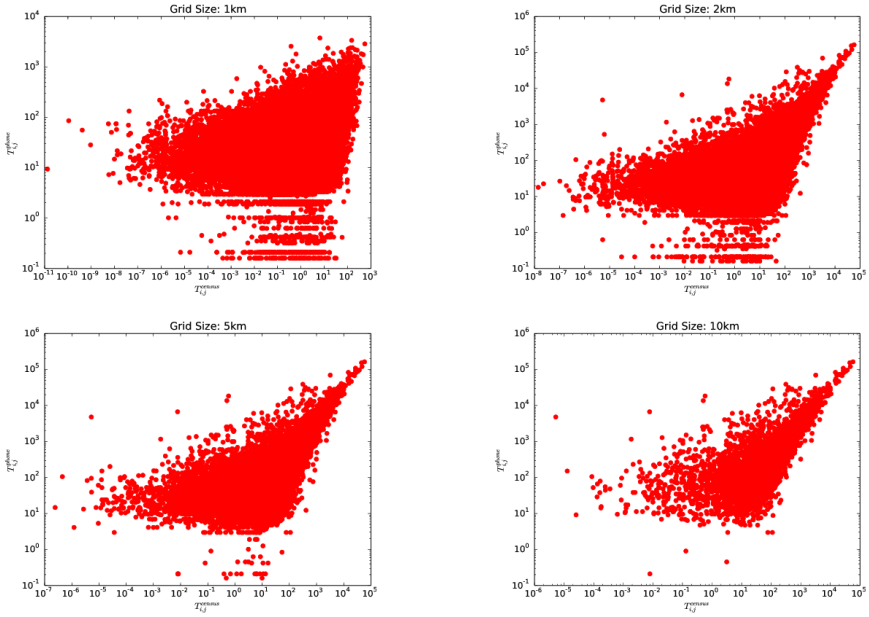


Figure 3: Correlations between aggregated census and mobile phone Home and Work commuting flows. The correlation coefficients between the two datasets at grid sizes of $1^2[km^2]$,$2^2[km^2]$,$5^2[km^2]$, and $10^2[km^2]$ are 0.5041, 0.8459, 0.9360, and 0.9598, respectively.

data is as good as census surveys in estimating home and work locations for residents at higher levels of aggregation.

Second, we observe that the commuting flows between zones, compared to commuting flows between tracts, are generally much larger. This relationship increases as the zone size increases. Thus, as we increase the zone size, we can also improve privacy of users by increasing the difficulty to deanonymize individuals who have home and work locations in tracts with low commuting flows.

Spatial aggregation is a necessary step for researchers analyzing and integrating geographic and mobile phone data. Through our results, we have demonstrated that our method for spatial aggregation with zone size as a tunable parameter allows us to understand the tradeoffs between accuracy, anonymity, and resolution. In contrast to aggregating to fixed regions like census tracts, our method enables researchers to quantitatively tune the size of those regions to best match the accuracy and privacy needs of the application.

REFERENCES

[1] Sibren Isaacman, Richard Becker, Ramón Cáceres, Stephen Kobourov, Margaret Martonosi, James Rowland, and Alexander Varshavsky. Identifying important places in people's lives from cellular network data. *Pervasive Computing*, pages 133–151, 2011.

[2] Pierre Deville, Catherine Linard, Samuel Martin, Marius Gilbert, Forrest R Stevens, Andrea E Gaughan, Vincent D Blondel, and Andrew J Tatem. Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences*, 111(45): 15888–15893, 2014.

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

[4] Pu Wang, Timothy Hunter, Alexandre M. Bayen, Katja Schechtner, and Marta C. Gonzalez. Understanding road usage patterns in urban areas. *Sci. Rep.*, 2, 12 2012.

[5] Yves-Alexandre de Montjoye, César A Hidalgo, Michel Verleysen, and Vincent D Blondel. Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, 3, 2013.

[6] Maxime Lenormand, Miguel Picornell, Oliva G Cantu-Ros, Antonia Tugores, Thomas Louail, Ricardo Herranz, Marc Barthelemy, Enrique Frias-Martinez, and Jose J Ramasco. Cross-checking different sources of mobility information. *arXiv preprint arXiv:1404.0333*, 2014.

Studying Human Behavior through the Lens of Mobile Phones during Floods

A. J. Morales[1], D. Pastor-Escuredo[1], Y. Torres[1], V. Frías-Martínez[2], E. Frías-Martínez[3], N. Oliver[3], A. Rutherford[4], T. Logar[4], R. Clausen-Nielsen[4], O. De Backer[4], M. A. Luengo-Oroz[4]
[1]Universidad Politécnica de Madrid, [2]University of Maryland
[3]Telefónica Research,[4]United Nations Global Pulse

Natural disasters affect hundreds of millions of people worldwide every year. Emergency response efforts depend upon the availability of timely information, such as information concerning the movements of affected populations. The analysis of Call Detail Records (CDR) captured from the mobile phone infrastructure provides new possibilities to characterize human behavior during critical events. In this study, we combine remotely sensed data and CDRs to understand how people communicated during severe floodings in the Mexican state of Tabasco in 2009. This research demonstrates that CDR data has the potential to provide useful information on human behavior for improved emergency management and humanitarian response. Our results could also serve as a potential proxy indicator for flood impact and risk awareness.

The lack of timely, accurate information about movements and communications of affected populations during natural disasters can limit the effectiveness of humanitarian response. However, the growing ubiquity of mobile phones has revealed new opportunities for accessing such information. Mobile phone data can provide valuable insights, in order to tackle issues related to economic and humanitarian development [2], such as understanding the behavior of affected populations during a natural disaster [1]. For example, recent research has demonstrated the potential of mobile phone data to help study population movements after an earthquake in Haiti [3] or to model malaria outbreaks in Kenya [4]. During these critical events, the patterns of collective human behavior are disrupted, as the population faces the ongoing emergency. Such effect is closely related to the emergence of large information cascades, since people tend to communicate with others, triggering chain reactions in the social network [5]. The geographical distribution of the activity can also be used to characterize the catastrophe. For instance, Twitter activity allows to locate an earthquake's epicenter with extraordinary accuracy by geographically measuring the volume of related tweets [6].

The goal of this study is to develop and apply methods to assess the suitability of CDR data to characterize the impact of floods on the population. Our vision is to build CDR-based decision-support tools to help the public sector better respond to floods and other natural disasters. We investigate the viability of using CDR data combined with other sources of information to characterize the floods that occurred in Tabasco, Mexico in 2009. In particular, we analyzed CDRs of the geographical area affected by the floods during a period of nine months (July 2009 to March 2010). The main technical contribution of this work is the development of a multimodal data integration framework that facilitates the combination of CDR data with data from other sources, in order to characterize changes in the communication patterns during the floods. We also contrast our results with external ground truth information. For a longer description of the research presented in this short paper please refer to [1].

The methodological framework proposed in this study is composed of the following steps: First, we evaluate the representativeness of the data by using the 2010 census [10] of Tabasco as the ground truth. For this purpose, we compare the census data with a data-driven social baseline that we built based on the location of the home antenna tower (HAT) for each phone, meaning the antenna tower most used at night during the baseline (BL) period [7]. Second, we integrate additional and diverse data sources to further understand the phenomena. We use remote sensing through medium resolution (15 to 60 meters) ETM+ Landsat7 [8] satellite images to detect and geographically confine the submerged land. Moreover we analyze the Tropical Rainfall Measuring Mission [9] data, in order to build a temporal series of precipitations and to understand the relationship between the natural phenomenon and mobile phone activity.

In order to detect abnormalities in the activity, we examine mobile phone activity data before, during and after the disaster. We propose the variation metric that relies on the comparison of the number of phones placing or receiving calls per antenna $x(t)$, against their characteristic variation obtained during the baseline period (BL). Mathematically, the variation metric is defined as the z-score from $x(t)$ –*i.e.* the normal distribution characterizing the baseline– and defined as $x_{norm} = (x(t) - \mu_{BL})/\sigma_{BL}$, where the pair (μ_{BL}, σ_{BL}) statistically characterizes the activity during the baseline period. By means of analyzing the normalized series, we establish a baseline understanding of emergency behavior which enables us to measure the rate of disaster recovery and to show how affected populations behave in response to flooding. The variations in the number of active phones connected to each cell tower reveal abnormal activity patterns in the most affected locations during and after the floods that could be used as signatures of the floods – both in terms of infrastructure impact assessment and population information awareness.

This research demonstrates that mobile phone data has the potential to provide timely information about human

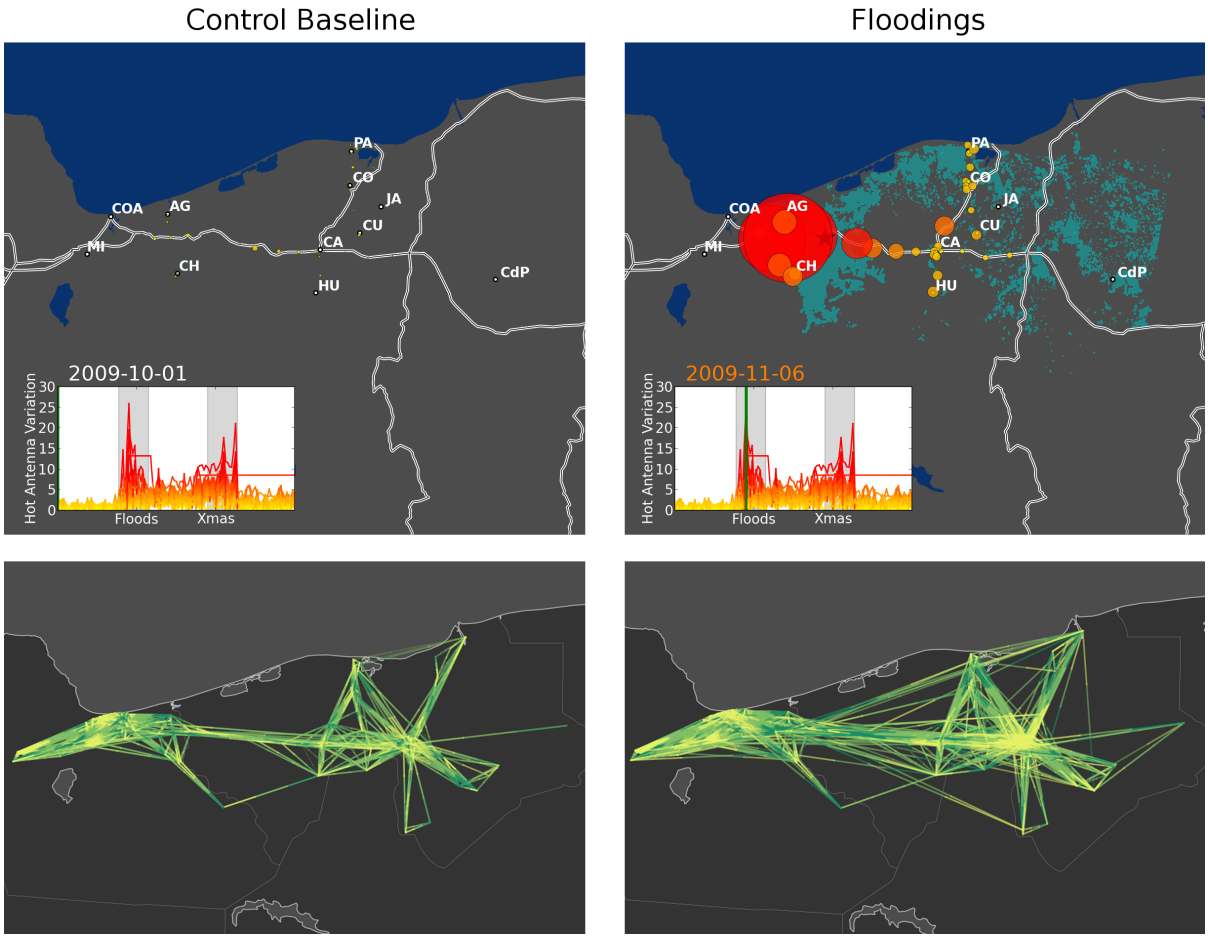


FIG. 1. Tabasco impact maps. The top panels show the absolute value of the antenna variation metric during an arbitrary day before floods (top left) and another one during floods (top right). In these panels, each antenna is represented by a circle with color and size proportional to the daily variation. The segmented flooded area has been colored in light blue. The insets display the temporal series of the antenna variation metric, and the green line indicates the day of observation. In the top right panel, antennas near the flooding area dramatically increased their variation during the floods. In the bottom panels, we show two visualizations of human displacement networks among cell towers. Towers are connected if a person makes two consecutive calls. We show an average week of the network before the floods (bottom left), and an aggregated network across the floods (bottom right). The edge color means the direction of the displacement, from green to yellow. It can be noticed that during floods (bottom right) the graph is denser, and more connections are established between the towers.

behavior for improved emergency management and humanitarian response. Insights gained from CDR analysis could also serve as a potential proxy indicator for flood impact and risk awareness. On one hand, mobile phone data can be highly representative of the population's behavior. A comparison between CDR data and census data yields a strong linear relation between official population statistics and population estimates computed from CDR data. Furthermore, civil protection warnings are not necessarily an effective way to raise awareness. In fact, a civil protection warning was issued on the day of highest rainfall in Tabasco in 2009. However, big spikes in phone activity were only observed in two cell phone towers along the most affected road when floods already showed initial impacts, meaning that the civil protection warning did not generate similar levels of awareness. This finding reveals important behavioral insights for emergency responders on how and when affected populations are made aware of a disaster. Finally, mobile activity can provide signals of flooding impact. When analyzed against the baseline activity, cell phone towers with the highest variation metric in the number of calls made during the floods were located in the most affected locations (see top panels in Fig. 1). Note that mobility patterns also changed significantly near the main cities and the capital as the ground transport system and the physical size of the cities constrain how the people could move during the floods (see bottom panels in Fig. 1).

In summary, aggregated and anonymized mobile phone data can be used to assess risk awareness, understand the

effect of public communications such as disaster alerts and measure the direct impact of floods on the population. The research findings described in this paper show that CDR data could be a beneficial source of information for both emergency management and resilience assessment. Analyzing mobile activity during floods could be used to potentially locate damaged areas, efficiently assess needs and allocate resources (for example, sending supplies to affected areas). Identifying cell phone towers in the most affected areas of flooding might also serve to improve and target public communications and safety alerts, as well as help measure the effectiveness of such early warning announcements.

[1] Pastor-Escuredo, D., Morales, A. J. et al., *Flooding through the Lens of Mobile Phone Activity.* IEEE Global Humanitarian Technology Conference, GHTC 2014.

[2] Decuyper, A. et al., *Estimating Food Consumption and Poverty Indices with Mobile Phone Data*, arXiv:1412.2595, (2014)

[3] Bengtsson, L. et al., PLoS Med **8** (2011), no. 8, e1001083. *Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: A post-earthquake geospatial study in haiti*,

[4] Wesolowski, A. et al., Noor, Robert W. Snow, and Caroline O. Buckee, *Quantifying the Impact of Human Mobility on Malaria*, Science **338** (2012), no. 6104, 267–270.

[5] Bagrow, J. P. et al., *Collective Response of Human Populations to Large-Scale Emergencies*, PLOS ONE **6** (2011), no. 3, e17680.

[6] Sakaki, T. et al. *Earthquake shakes twitter users: real-time event detection by social sensors*, Proceedings of the 19th international conference on World wide web (New York, NY, USA), WWW '10, ACM, 2010, pp. 851–860.

[7] Becker, R. et al., *Human mobility characterization from cellular network data*, Commun. ACM **56** (2013), no. 1, 74–82.

[8] http://earthexplorer.usgs.gov/

[9] http:// http://trmm.gsfc.nasa.gov/

[10] http://www.censo2010.org.mx/

Earthquakes, Hurricanes and Mobile Communication Patterns in the New York Metro Area: Collective Behavior during Extreme Events

Christopher Small[1] Richard Becker[2] Ramón Cáceres[3] Simon Urbanek[2]

[1] *Lamont-Doherty Earth Observatory, Columbia University, Palisades, NY, USA*
[2] *AT&T Labs – Research, Bedminster, NJ, USA*
[3] *Google, New York, NY, USA*

We use a spatially and temporally extensive collection of voice call and SMS text message volumes to quantify spatiotemporal communication patterns in the New York Metro area before, during and after the Virginia earthquake and the passage of Hurricane Irene in the same week of 2011. We compare and contrast spatial and temporal disruptions to normal patterns of voice and text communication in response to each of these extreme events in a diverse range of environments within the New York Metro area. We have been careful to preserve privacy by using only anonymous and aggregate data. Results show both similarities and differences in call and text responses to both the earthquake and the hurricane. The earthquake produces an instantaneous and pervasive response followed by a ~90 minute temporal disruption to both call and text volume patterns. The hurricane produces a two day, spatially varying disruption to normal call and text volume patterns. Comparison of call and text response to these events yields less intuitive results. Both call and text volumes increase abruptly following the earthquake but call volume anomalies are much larger than text volume anomalies. The magnitude of the call volume anomaly diminishes with distance from the epicenter with multiple spatial localizations of high response sectors in Manhattan. On the day preceding the arrival of the hurricane, coastal evacuation zones show varying response both in location and in call versus text volume. These spatial patterns suggest partial, but not full, compliance with evacuation orders for most low lying areas in NYC and surroundings. In most coastal areas call volumes dropped anomalously in the afternoon before the hurricane's arrival, but text volumes showed a much less consistent pattern and often did not decrease in parallel with calls. In terms of total daily volumes, most low lying coastal areas show a decrease in both calls and texts relative to the previous week.

Understanding dynamics of collective human behavior during extreme events has obvious relevance to both preparedness and response. Most current knowledge of human behavior during extreme events comes from relatively small numbers of retrospective observations – often qualitative with unknown accuracy or degree of representation of the impacted population. In contrast, mobile communication data can provide pervasive, quantitative observations of human communication patterns before, during and after extreme events [1]. In this study we use a spatially and temporally extensive collection of voice call and SMS text message volumes to quantify spatiotemporal communication patterns in the New York Metro area before, during and after the magnitude 5.8 Virginia earthquake (2011-08-23) and the passage of Hurricane Irene (2011-08-28), both of which occurred in the same week of 2011.

Numbers of voice calls (calls) and SMS text messages (texts) were measured at 1 minute intervals over the course of a year (2/1/2011 through 1/31/2012) at the spatial resolution of

azimuthal sectors of almost 11,000 mobile network antennas within a 50 mile (80.5 km) radius of Times Square in NYC. We aggregate these call and text volumes at different spatial and temporal resolutions and use spatial correlation matrices to quantify normal spatial and temporal patterns and their disruption before, during and after both events. We have been careful to preserve privacy throughout this work. In particular, this study uses only anonymous and aggregate data.

Spatial and temporal correlation matrices can be used to quantify both regularities and disruptions of spatiotemporal patterns in call and text volumes. The spatial correlation matrices of daily call and text volumes clearly resolves differences between weekday and weekend spatial patterns, as well as decorrelations associated with holidays when the normal weekly spatial patterns are disrupted. The largest decorrelation observed in 2011 is associated with Hurricane Irene. Spatial correlation matrices of hourly call and text volumes in the weeks before, during and after the earthquake and hurricane show the evolution of the spatial disruptions in greater detail (Fig. 1).

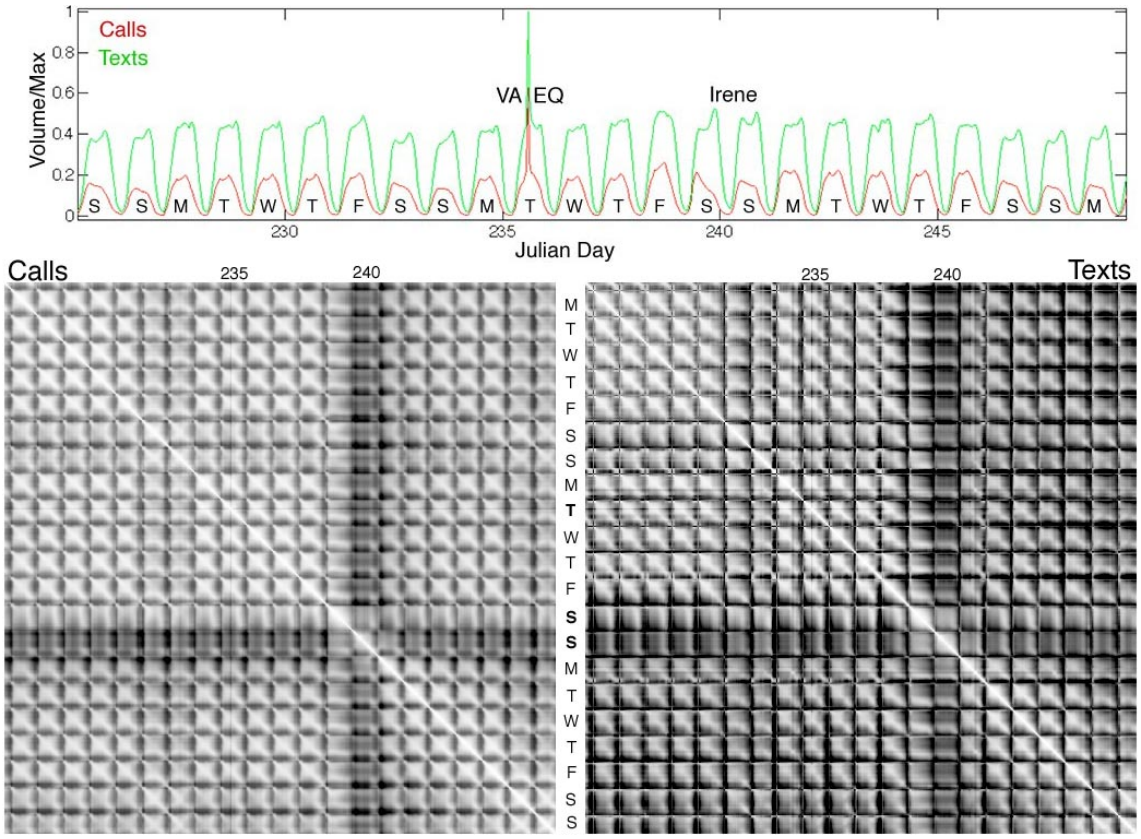


Figure 1 Spatial sum call/text volume time series and spatial correlation matrices for the weeks spanning the earthquake (jd 235) and hurricane (jd 240). Over the region calls decreased and texts increased the afternoon before the hurricane but both quickly returned to normal after the hurricane passed. Hourly correlation matrices clearly show the weekly cyclicity of spatial structure and the disruptions caused by the earthquake and hurricane. Spatial correlation typically peaks at ~0.96 on workdays but drops briefly to ~0.93 immediately after the earthquake. Correlations remains high (>0.9) the day before the hurricane indicating a stable spatial pattern that is also strongly correlated with early evening patterns on normal weekdays. The spatial pattern the day of the hurricane shows the weakest overall correlation (< 0.7) to other days but with strongest correlation to weekday evenings.

[1] Bagrow, J.P., Wang, D., & Barbási, A.-L. (2011). Collective response of human populations to large-scale emergencies. *PLoS ONE, 6*, 1-8

Recent Results/Work in Progress

Application of Floating Phone Data (FPD) in Germany

January 2012 – March 2015

Dipl.-Ing. Moritz von Mörner, vonmoerner@verkehr.tu-darmstadt.de
Technische Universität Darmstadt, Germany
Department of Civil and Environmental Engineering
Chair of Transport Planning and Traffic Engineering

Objective:

Determine the state of science and technology for the use of Floating Phone Data (FPD) in traffic and transport research. Feasibility study for the application of FPD as input value for e.g. transport models and transport planning.

Focus:

- Long distance travel in Germany
    1. rail
    2. motorway
- Regional public transport with busses in rural areas.

This project is conducted for : Deutsche Bahn AG
FPD provider: Motionlogic GmbH / T-Systems International GmbH

Floating Phone Data restrictions at the moment:

- Raw data cannot be obtained, only aggregated and extrapolated data is available,
- Aggregated data is available in 90 minutes blocks,
- Max. duration for tracking: 90 minutes,
- Min. five phones registered per time unit and cross section/ the same origin destination connection are needed to obtain data,
- The FPD is aggregated and extrapolated from signals of one carrier to the overall amount of people in the area, assuming one SIM card per person.

Other Data:

- Cross section counts along the A8 between Ulm (Baden-Württemberg) and Munich (Bavaria) – all cross sections that were available in the study period are included,
- 9 cross sections in both directions,
- Sum of all vehicles and distinction of 8 vehicle types,
- Extrapolation to number of people crossing the cross section (with average occupancy rates),
- Train occupancy for one long distance train between Munich and Ulm in passenger counts.

Preliminary Results:

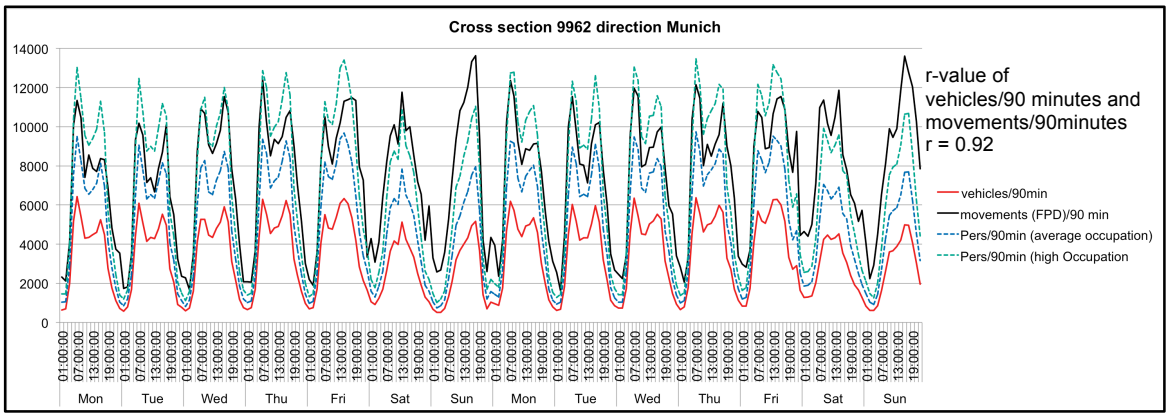
1. Stationary vehicle counts were compared to extrapolated number of people from FPD (see Picture 1).
The correlation between extrapolated data and vehicle counts is quite good, especially during workdays with only a few r-values below 0.9. At two cross sections some anomalies were detected, which might be due to interference in the measurements through other roads close-by the surveyed sites. Converting these numbers to passenger per vehicle showed a very high fluctuation over the day, where occupancy rates were especially high during the night. Trying to understand the fluctuation, with an average for vehicle occupancy the overall number of people was calculated to be compared to the extrapolated data from FPD. This comparison showed that some of the patterns did match and others did not. However, we were not yet able to identify a clear pattern.

2. Passenger counts on one long distance ICE-train from Munich to Stuttgart were conducted and compared to the corresponding extrapolated FPD counts (see Picture 2). The comparison shows that during workdays and weekend there are different offsets. In general extrapolated FPD counts seem to be higher than passenger counts on weekdays, whereas the FPD counts on weekends are lower than the corresponding passenger counts. However, due to the limited number of data samples no direct relation can be determined.

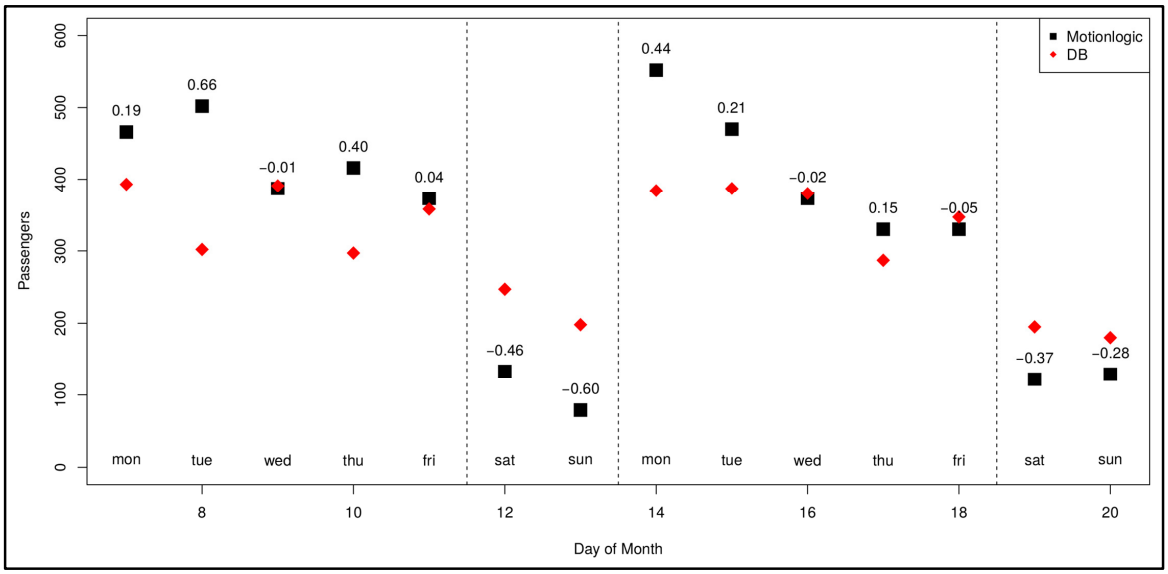
Now we're in the midst of identifying possible flaws in the extrapolation and comparison. Furthermore, in this project a comparison of passenger counts in busses in a rural area in Germany is planned. Passenger counts are done manually on busses and these numbers will be compared to extrapolated occupancy from FPD.

Conclusions:

To obtain detailed information more insight into the extrapolation process is needed as well as more data samples to obtain viable conclusions from the data. Towards the feasibility on using FPD as input data for e.g. transport models, regarding the mentioned restrictions, no comment can be made at this stage of the project.



Picture 1: Comparison of movement counts and number of vehicles (with theoretical occupation)



Picture 2: Comparison of long distance train passenger counts and extrapolated FPD (Source: Deutsche Bahn)

Evolving classification based on CDR-derived behavior patterns

Extended abstract

Michal Mucha, Dominik Filipiak, Agata Filipowska
Department of Information Systems
Poznan University of Economics
Poznan, Poland
firstname.lastname@kie.ue.poznan.pl

The ubiquity of mobile phones, and the immense amount of data that is generated with their use, open up vast possibilities for research and real life applications. Use of mobile phones by all parts of the population enables research on both macro and micro scale. Mobile phone datasets describe user telecommunication activity, tagged with approximate geographical location & time; on top of that, social relations can be modeled with the use of graphs [1], [2]. Such a wide spectrum of information about a great number of individuals could prove to be a valuable resource for research in the fields of psychology and sociology. In fact, this potential is already being successfully explored by [3]–[7]. The authors of this paper wish to contribute towards further efforts of this kind by providing a useful tool.

We introduce an approach, stemming from *Behavior Informatics* (BI) [8], to prepare behavior vector sequences using *Call Detail Records* (CDR). A behavior vector is prepared from each telecommunications activity represented by a CDR entry. Spatial and social information contained in CDR datasets can be interpreted with relevant methods and used in the composition of behavior vectors; i.e. the importance of certain locations to individual users may be estimated [9], and then used to label users' actions. The resulting behavior vector sequence may then be used for creation of econometric models, finding particularly interesting behaviors, etc. An example, anonymized dataset of 1 month, 8000 unique locations, and 3.35 million active users, served for the creation of such a behavior vector sequence. Elements of the vector, derived from CDR, are: *user ID, location category, relation strength category.*

Spatial description of each CDR entry is given as the transceiver station, which was used to handle the service. One or more of such stations are placed together with the aim to cover a certain area [10], [11]. Thus a CDR entry means that a certain user was at a certain area. To give an example of incorporating location into the behavior vector sequence, we propose the use of information on identified places of importance to individual people, as proven to be possible by [9] and [12]. In our experimental setting, we use a simplified approach to estimate home and workplace locations for the users in the dataset available to us. The personal categories assigned to locations are then used in the composition of behavior vectors. This demonstration shows the possibility of inclusion of research such as [13], and other kinds of methods, into social and psychological studies based on mobile-phone data.

In this way results of other methods may be used to compose the behavior vector. To illustrate how research of relation strength estimation, such as [14]–[17], can be included the BI framework, we propose a simplified approach towards categorization of relations between users in the dataset available to us. We use a simple categorization of relations into strong and weak. Top 5 most contacted people, who also satisfy the arbitrary threshold of 15 connections per month, are deemed *strong* relations; all others - *weak*. The result of this simplified categorization is included in the behavior vectors, and can easily be replaced by more elaborate and accurate methods, i.e. based on community structure [18], [19].

Similarly, other methods may be used to complement or replace the proposed modules with which the behavior vector sequence is constructed (i.e. analysis of physical meetings [20]).

As the main contribution, we propose the use of an *Evolving Agent-based Classifier (EVABCD)* (originally proposed in [21]) with corrections and modifications that allow temporary (instead of the default permanent) memory of user behavior. The original version of the classifier is designed by its authors to allow efficient processing of data streams with the use of recursion to reduce memory and computational requirements of the method. Moreover, the evolving nature of the classifier means that during its functioning new classes appear when new pattern of behavior emerge in the population, and classes representing behavior patterns that ceased to be expressed by users are removed. These features are retained, with the additional benefit of a limited memory for past behavior patters.

Such classification of behavior patterns is suggested by authors to serve e.g. as a foundation for scientific comparison of results of psychological tests (as in [4]) with mobile dataset-derived behavior observations, studies of sociological phenomena, as well as any other scientific work that would employ the possibility of the macroscopic scale of mobile phone datasets.

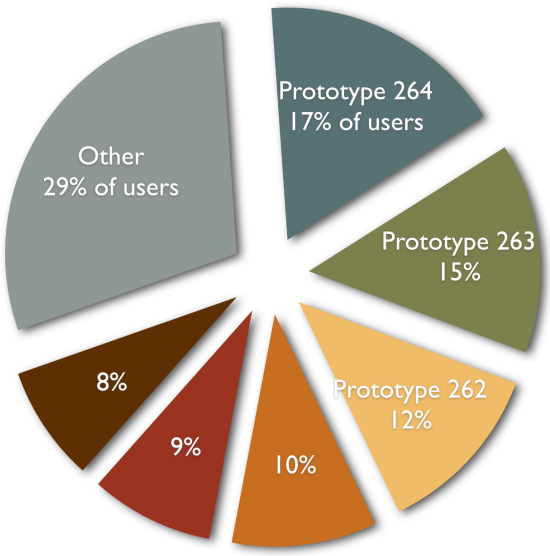


Fig. 2. Distribution of the population into behavior-based classes

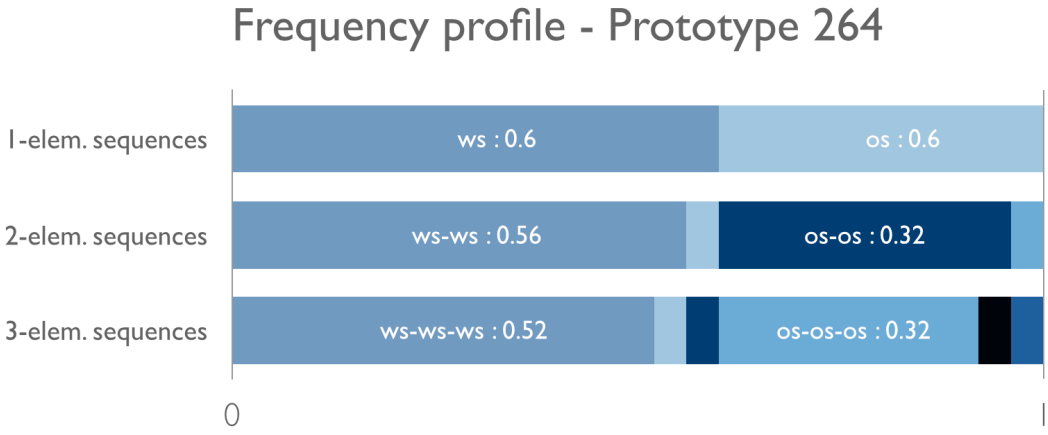
The classification was carried out on the modified, recursive EVABCD. In a single-machine experimental setting we analyzed 550 thousand users who performed 8 million actions (one day of activity for the selected population). During the life of the classifier (designed to process data streams), we took snapshots of its state. Overall, 267 prototypes appeared, describing various clusters of the population. Appearance of a new prototype usually results in the removal of one or more of the existing prototypes, when the new one describes them well enough. Thus the number of prototypes reported in the snapshots varied between 5 and 20, and in the final there were 19. Fig. 1 presents the distribution of the population into classes, and fig. 2 contains the probability profile of the most popular prototype. Notably, in our experiment the classifier has proven capable of identifying small clusters with distinct behavior patterns, i.e. users calling only from home and only their close contacts.

In our experiments, the samples were not labeled. Different research endeavors that are enabled with this foundation can employ the Evolving Classifier for supervised learning – analysis of behavior prototypes together with other data.

The contribution thus consists of:

- A way to represent CDR and layers of information derived from it in the form of behavior vector sequences, in accordance with the *Behavior Informatics* framework
- An example of a simplified method of estimating strength of relations, as well as inclusion of relation strength categorization into the BI framework
- An example of the application of personalized location-tagging in the composition of the behavior vector sequence
- A classifier that allows for dynamic discernment of behavior patterns within the population, also keeping track of the dynamics of behavior of individuals

All of which may prove useful for research of sociological and psychological phenomena with the application of mobile phone datasets.



1-elem. sequences		2-elem. sequences				3-elem. sequences					
os	ws	os-os	ws-ws	ws-os	os-ws	os-os-os	os-os-ws	os-ws-ws	ws-ws-ws	ws-os-os	ws-ws-os
0,4	0,6	0,36	0,56	0,04	0,04	0,32	0,04	0,04	0,52	0,04	0,04

Fig. 1. A profile of behavior vector occurrence frequency that is the closest representation of the profiles of 17% of the population. *WS* means that the call was made from *work* to a *strong relation* contact. *OS* means that the call was made from *other location* (other than home or work), to a *strong relation* contact. 2- and 3-element sequences show the frequencies of patterns of consecutive actions. This particular prototype represents users who call only a stable group of contacts, don't call from home, and do slightly more calls from work than from other locations. Only non-zero frequencies are displayed in this figure.

REFERENCES

[1] R. Xiang, J. Neville, and M. Rogati, "Modeling Relationship Strength in Online Social Networks," in *Proceedings of the 19th International Conference on World Wide Web*, 2010, pp. 981–990.

[2] M. Saravanan, G. Prasad, S. Karishma, and D. Suganthi, "Analyzing and labeling telecom communities using structural properties," *Social Network Analysis and Mining*, vol. 1. pp. 271–286, 2011.

[3] A. Rubio, V. Frias-Martinez, E. Frias-Martinez, and N. Oliver, "Human mobility in advanced and developing economies: A comparative analysis," in *AAAI Spring Symposium - Technical Report*, 2010, vol. SS-10–01, pp. 79–84.

[4] R. de Oliveira, A. Karatzoglou, P. Concejero Cerezo, A. de Vicuña, and N. Oliver, "Towards a Psychographic User Model from Mobile Phone Usage," in *CHI '11 Extended Abstracts on Human Factors in Computing Systems*, 2011, pp. 2191–2196.

[5] V. Frias-Martinez and J. Virseda, "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 - ICTD '12*, 2012, p. 76.

[6] E. Frias-Martinez, "An agent-based model of epidemic spread using human mobility and social network information," in *... conference on social ...*, 2011, pp. 57–64.

[7] Y. A. De Montjoye, J. Quoidbach, F. Robic, and A. Pentland, "Predicting personality using novel mobile phone-based metrics," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2013, vol. 7812 LNCS, pp. 48–55.

[8] L. Cao, "In-depth Behavior Understanding and Use: The Behavior Informatics Approach," *Inf. Sci.*, vol. 180, no. 17, pp. 3067–3085, Sep. 2010.

[9] S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, M. Martonosi, J. Rowland, and A. Varshavsky, "Identifying important places in people's lives from cellular network data," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2011, vol. 6696 LNCS, pp. 133–151.

[10] Y.-A. de Montjoye, C. a Hidalgo, M. Verleysen, and V. D. Blondel, "Unique in the Crowd: The privacy bounds of human mobility.," *Sci. Rep.*, vol. 3, p. 1376, 2013.

[11] M. R. Vieira, V. Frías-Martínez, N. Oliver, and E. Frías-Martínez, "Characterizing dense urban areas from mobile phone-call data: Discovery and social dynamics," in *Proceedings - SocialCom 2010: 2nd IEEE International Conference on Social Computing, PASSAT 2010: 2nd IEEE International Conference on Privacy, Security, Risk and Trust*, 2010, pp. 241–248.

[12] Y. Qu and J. Zhang, "Trade Area Analysis Using User Generated Mobile Location Data," in *Proceedings of the 22Nd International Conference on World Wide Web*, 2013, pp. 1053–1064.

[13] R. Trasarti, F. Pinelli, M. Nanni, and F. Giannotti, "Mining Mobility User Profiles for Car Pooling," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2011, pp. 1190–1198.

[14] A. Filipowska, M. Mucha, B. Perkowski, E. Szczekocka, J. Gromada, and A. Konarski, "Towards Social Telco Applications Based on the User Behaviour and Relations Between Users," in *ICIN 2015, in press*, 2015.

[15] N. Eagle, A. S. Pentland, and D. Lazer, "Inferring friendship network structure by using mobile phone data.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 106, pp. 15274–15278, 2009.

[16] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A.-L. Barabási, "Structure and tie strengths in mobile communication networks.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 104, pp. 7332–7336, 2007.

[17] K. Janakiraman and S. Motahari, "How are you related? predicting the type of a social relationship using call graph data," in *Proceedings - 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust and 2012 ASE/IEEE International Conference on Social Computing, SocialCom/PASSAT 2012*, 2012, pp. 111–116.

[18] V. D. Blondel, J. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of community hierarchies in large networks," *Networks*, pp. 1–6, 2008.

[19] M. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Phys. Rev. E*, 2004.

[20] C. Brown, N. Lathia, C. Mascolo, A. Noulas, and V. Blondel, "Group Colocation Behavior in Technological Social Networks," *PLoS One*, vol. 9, p. e105816, 2014.

[21] J. A. Iglesias, P. Angelov, A. Ledezma, and A. Sanchis, "Creating Evolving User Behavior Profiles Automatically," *Knowl. Data Eng. IEEE Trans.*, vol. 24, no. 5, pp. 854–867, May 2012.

Sms transmission using phone users density in big cities

Floran Berthaud[1] , Yannick Léo[1] , Carlos Sarraute[2] , Anthony Busson[1] , and Eric Fleury[1]

[1] *Université de lyon, UMR 5668 CNRS - ENS Lyon - INRIA - UCB Lyon 1, IXXI*
[2] *Grandata Labs, Bartolome Cruz 1818 Vicente Lopez. Buenos Aires, Argentina*

This work answers the question : is it possible to transmit a sms using phones as relay in a big city such as Mexico City? We defined a simple transport protocol to transmit sms from a source to a destination. This protocol does not need routing, it is based on locality of sms, the density of phones in Mexico City and mobility of phone users. We studied a mobile dataset including 8 millions users living in Mexico city. This gave use a precise estimation of the average transmission time and the global performances of our approach. After 30 minutes, half of the sms were delivered successfully to destination.

The need of communicating in a dense city is always increasing. Every day, millions of sms are sent in a big city like Mexico City. Phone operators have to adapt their infrastructures to provide an efficient service. At present times, sms are not only routed with base stations. The way to communicate and exchange sms between each other has become diversified these last years. We can now send messages with applications like WhatsApp [1], Tango, Skype and Viber while connected to a wireless spot [4]. During rush hours, the capacity of the operator service are almost saturated. It is becoming a great challenge to increase the capacity of the service with the same number of relays.

In this study, we propose a new way to transmit sms and more generally data from a source to a destination. Instead of using classical routing, we use relays close to the source and phone users that are connected to those local relays to reach the destination. A big advantage is that we do not perform a routing algorithm as we do not need to know where the destination is. Moreover, as we only use local relays that are close to the source, the bandwidth cost of a sms is smaller. On top of that, the density of phones and the mobility of users are even higher when the capacity of classic relay network is challenged during rush hours.

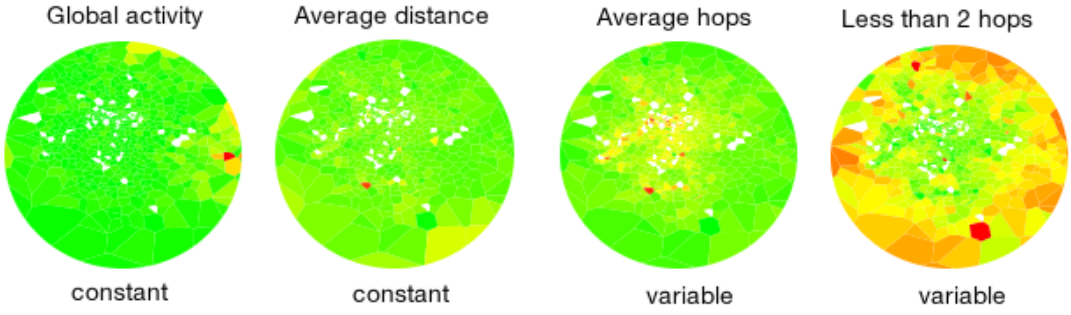


Figure 1: Geographical heat maps of static network parameters : average number of base station hops (left), average distance (middle) and global activity (right) around Mexico City. The Voronoi cell for a base station represents the area in which users are connected to this base. The green colour represents low values and the red one high values.

We used a communication dataset [3] containing the mobile phone interactions of 8 millions of people in Mexico City covered by 775 base stations that are part of the classic network. This anonymised dataset contains sms and calls with some location information defined by the base station of the source and destination. Over three months, we managed to extract around 10 millions fully located sms for our study. Most of the sms had Mexico City as source and destination.

Protocol and results

We analyzed our dataset through time and space for each base station. We showed the variability of the global activity according to time. We noticed a constant activity for every base station, the distance between two stations depends on the local activity. On figure 1, for each sms of the trace, we computed the distance and the number of relay hops from the source to the destination base station. The average distance of sms is constant whereas the number of relay hops is higher in the center where base stations are closer.

This study provides an empirical proof of the close proximity of messages. Many sms are very local with a very small number of relay hops. We applied a neighboor protocol that consists in delivering the sms to the phone users that are attached to the same base station when the source sends the sms. Then we let these neighboors moving with the sms. If any of the neighboors reach the destination, the sms is delivered. If after half an hour the message has not reached the destination, then the message is dropped. In our network, one over three messages were delivered after 10 minutes and one over two after thirty minutes. As some locations are missing in our dataset, in reality, the results are likely to be even better.

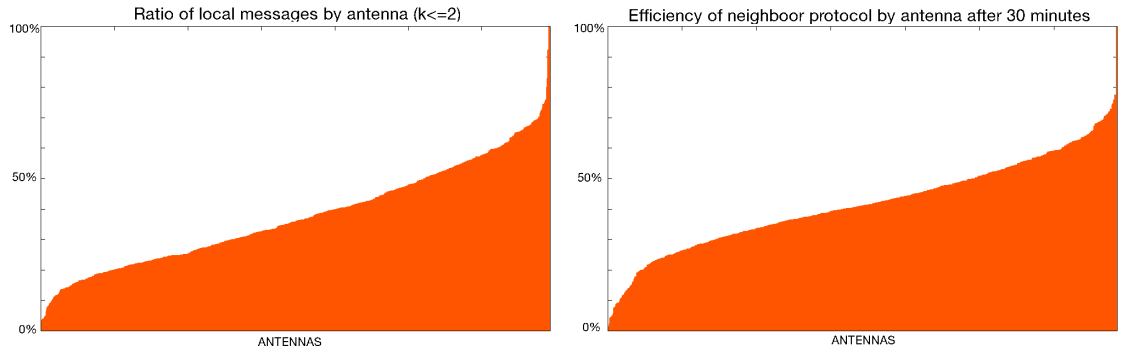


Figure 2: For each base station, we performed the ratio of local sms that had less than two relay hops (left) and the ratio of sms that had successfully reached the destination by the neighboor protocol (right).

References

[1] Karen Church and Rodrigo de Oliveira. What's up with whatsapp?: Comparing mobile instant messaging behaviors with traditional sms. In *Proceedings of the 15th International Conference on Human-computer Interaction with Mobile Devices and Services*, MobileHCI '13, pages 352–361, New York, NY, USA, 2013. ACM.

[2] Yves-Alexandre de Montjoye, César A Hidalgo, Michel Verleysen, and Vincent D Blondel. Unique in the crowd: The privacy bounds of human mobility. *Nature srep.*, 3, 2013.

[3] C. Sarraute, P. Blanc, and J. Burroni. A study of age and gender seen through mobile phone usage patterns in mexico. In *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*, pages 836–843, Aug 2014.

[4] Chai K Toh. *Ad hoc mobile wireless networks: protocols and systems*. Pearson Education, 2001.

Topological Properties and Temporal Dynamics of Place Networks in Urban Environments

Anastasios Noulas
Computer Laboratory
University of Cambridge

Blake Shaw
Foursquare
568 Broadway, New York

Renaud Lambiotte
Department of Mathematics
University of Namur

Cecilia Mascolo
Computer Laboratory
University of Cambridge

1. INTRODUCTION

Mobile user trajectories are known to exhibit structural and temporal regularities associated with the daily and weekly cycles of human activity. The spatial network formed by user movement, and its topological characteristics in particular, have been explored in recent research including the detection of urban neighborhoods [4], place recommendation to mobile users [9], touristic route identification [7] and a broad range of applications in epidemiology [1].

However, the majority of models of human mobility focus exclusively on its spatial characteristics [2, 5, 11], and neglect both network topology and temporal dynamics. More advanced computational methods proposed recently realize some of these aspects, for instance by incorporating information about the users' social network [10] and their spatiotemporal dynamics [3]. The applicability of these approaches is limited though, as they rely on complete knowledge of a user's historic whereabouts and social connections as input, which might not be readily available in most domains.

The goal of this paper is to bridge the gap between universal mobility models and complex computational methods in mobility modeling. As opposed to tracking the whereabouts of individual users, our key idea is to use the aggregate trajectories of users between real-world places to define a network of venues in the city. Using a longitudinal dataset from the location-based service Foursquare we empirically analyze *place networks* in one hundred metropolitan areas across the globe.

Figure 1 presents a visualization of the place network shaped by the movement of Foursquare users in New York City. One can spot hubs being formed at multiple areas across the urban plane, with local transitions connecting them to nearby places and occasional long jumps connecting places located further apart from each other, for example when users move between Manhattan and Brooklyn.

Exploiting a set of insights on the growth patterns, temporal dynamics and topological properties of these place networks, we then build a new human mobility model that accurately predicts the future interactions between places in urban environments with minimal parameterization and computational costs. Our work is articulated into three parts:

Place network growth and temporal pattern analysis. We first consider the temporal properties of place networks, and focus on their growth over time in terms of edge and node addition processes. In accordance with previous observations in online social networks [6], we observe a densification pattern, as *the number of edges grows superlinearly to the number of nodes in the system*. A saturating effect for node growth is reached quickly nonetheless, when the large majority of Foursquare venues is added to the network. *It takes almost 10 weeks for mobile users to crowdsource a large fraction (more than 95%) of public places in a city*. Subsequently, we compare instances of place networks across consecutive time windows of observation; we find that *a significant number of new links are generated over time as users form new spatial trajectories when they navigate between places*. The set of places that generate those edges, however, remains remarkably stable over long periods of time. These results reveal the importance of viewing connections as fleeting entities that emerge dynamically in the network.

Topological properties of place networks. We then empirically analyze the topological properties of place networks. We make two key observations: first, we note that *place networks exhibit the well-known characteristics of social networks* such as heavily skewed degree distributions, scale-free properties, small-world behavior and high clustering coefficients. We trace this relationship to the inherent inter-dependence between mobility and social link formation in geographic space [10, 3]. In contrast, we also find a striking difference compared to social networks: *they show a resemblance to the web graph presenting a balanced assortative mixing pattern with hub nodes connecting to each other but also to low degree nodes.* This non-social property arises from the different roles played by places in the network, and in particular the existence of travel spots, such as train stations or airports, acting as intermediate hubs between nearby places, e.g., food places, the most frequent place type in the network, typically characterized by low degrees. These characteristics are consistent across one hundred cities.

A new gravity model for link prediction in place networks. Finally, the turnover of links in the network over time motivates the following prediction task: given past observations about the connectivity of public venues in Foursquare, we would like to predict the pairs of places that are likely to connect at a future time. Candidate prediction models need to rank highly the pairs of venues

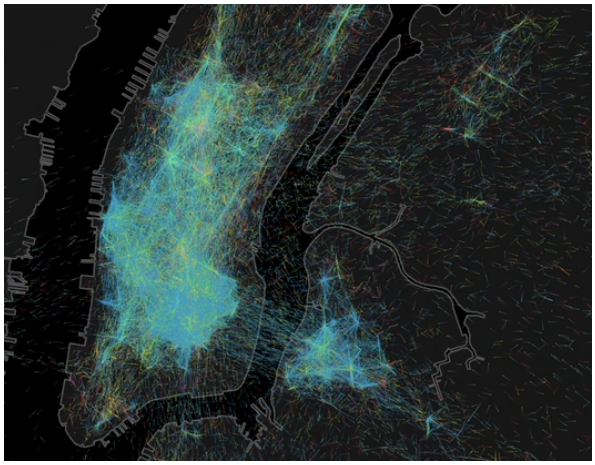


Figure 1: A visualization of the place network for New York City at 11pm. Each dot represents a user traveling between venues, and is color-coded by the category of the destination with blue being nightlife and green being food. We clearly see the edges of the network formed by people moving between places.

that are most likely to interact, a task complicated by a number of challenges. In particular, the highly volatile, time dependent, link generation process and sparse data setting may hinder the use of complex prediction algorithms that can be prone to overfitting. The inherently spatial embedding of the network suggests the need for models which integrate appropriately geographic distance as a factor. We therefore develop a generalization of *gravity models* [2, 5, 8], popular in the mobility and transport literature, where we incorporate the temporal aspects of the system: the model combines information on venue synchronization in terms of user activity, in and out-bound movement towards places and geographic distance. In practice, it captures the observation that nodes may act as sources or sinks of users in the course of time, depending on their cycle of activity. Finally, it incorporates information about the interaction of places on the network level, a valuable aspect of attraction that has been ignored by past mobility modeling approaches. The ranking strategy put forward by the model outperforms by at least two points in the Area Under the Curve (AUC) score even popular supervised learning algorithms and by a large margin the model adhering to the standard formulation of gravity in the literature (AUC score 0.905 versus 0.811). This is achieved with minimal requirements for training and optimization, making it ideal in practical application scenarios where expensive computations can pose a trade-off against the real time demands of many mobile applications.

2. NETWORK GROWTH AND DYNAMICS

Network densification is a fundamental phenomenon in network dynamics and relates to the different rhythms with which nodes and edges are added to the network. Previous work by Leskovec et al. [6] characterizes empirically the densification process in online social and technological networks showing that the number of edges grows superlinearly with the number of nodes in the network. Specifically, given the number of nodes $n(t)$ observed at a point in time t, one is interested in the number of edges $e(t)$ and the way this relationship forms as t grows. Formally we have:

$$e(t) \propto n(t)^{\alpha} \tag{1}$$

Different values of the exponent α, imply differences in the expected number of edges over time. A graph with $\alpha = 1$ maintains

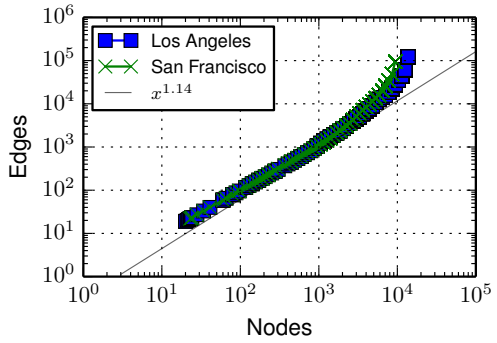


Figure 2: Number of edges versus number of nodes in Los Angeles and San Francisco as the cities become crowdsourced by Foursquare users.

a stable average degree over time, whereas $\alpha > 1$ corresponds to an increase in the average degree. The findings reported in [6] suggest that the latter is the case in many real world networks and here we investigate whether it holds also in urban place networks. We pick a random point in time t_0 where we begin monitoring the evolution of a place network and then measure the number of new nodes and edges added by users sequentially. Figure 2 shows the number of edges versus the number of nodes, in log-log scale, in the cities of Los Angeles and San Francisco, as venue information in these cities becomes crowdsourced over time by Foursquare users. Initially the number of links grows superlinearly with the number of nodes. We have measured using the least squares optimization method an exponent $\alpha = 1.14$ with a standard deviation ± 0.06 across a set of one hundred cities. However, at a specific city size-dependent threshold, this scaling breaks, as the number of nodes ceases to increase whereas new links continue to appear. At that point, a majority of places have been discovered by the users, and finite-size effects induce a slowing down of new place discovery.

3. REFERENCES

[1] P. Bajardi, C. Poletto, J. Ramasco, M. Tizzoni, V. Colizza, and A. Vespignani. Human mobility networks, travel restrictions, and the global spread of 2009 h1n1 pandemic. *PloS ONE*, 6(1):e16591, 2011.

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

[3] E. Cho, S.A. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In *KDD'11*.

[4] J. Cranshaw, R. Schwartz, J. Hong, and N. Sadeh. The livehoods project: Utilizing social media to understand the dynamics of a city. In *ICWSM'12*.

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

[6] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *KDD'05*.

[7] C. Lucchese, R. Perego, F. Silvestri, H. Vahabi, and R. Venturini. How random walks can help tourism. In *Advances in Information Retrieval*, pages 195–206. 2012.

[8] J. H. Niedercorn and B. Bechdolt. An economic derivation of the gravity law of spatial interaction. *Journal of Regional Science*, 9(2):273–282, 1969.

[9] A. Noulas, S. Scellato, N. Lathia, and C. Mascolo. Mining user mobility features for next place prediction in location-based services. In *ICDM'12*.

[10] A. Sadilek, H. Kautz, and J. Bigham. Finding your friends and following them to where you are. In *WSDM'12*.

[11] F. Simini, M. González, A. Maritan, and A.-L. Barabási. A universal model for mobility and migration patterns. *Nature*, 484(7392):96–100, 2012.

The Effect of Geographical Proximity on Mobile Communication

Hyungtae Kim
Columbia University

Tony Jebara
Columbia University

1 INTRODUCTION

Geographical proximity can significantly influence communication patterns. At small scales, proximity faciliates face to face interaction and relationships are naturally formed between individuals that coincide in similar public spaces as well as work places. A user's location pattern has been shown to be highly correlated with the pattern of his neighbors in a social network [4] .Call detail records (CDRs) have been used to study a variety of user behaviors such as the statistics of location patterns [2] and the probability of communication given a distance or proximity between pairs of users [3] . The correlation between user proximity and social connectivity have also been studied using social networks endowed with IP-based geolocation [1]. This article analyses a large data-set of location-augmented call detail records (LACDRs) to more precisely characterize the influence of geographical proximity on communication frequency and the length of average communication time.

2 DATA

The data-set under consideration is a large LACDR data-set that represents the majority of users in an undisclosed city-scale region. The data is sampled each time a communication is performed by one of the users which includes phone calls and SMS messages. Each such event generates a record containing the event-type, a start time, the end time as well as the GPS location. The data set spans approximately 2.3 million devices over a period of one month and contains approximately 300 million distinct communication and location events. Devices are given a unique ID which is derived from anonymizing the phone number associated with the device. This renders the devices uniquely identifiable (up to telephone number changes which we assume occur with limited frequency in the data-set). The GPS coordinates have also been transformed using a translation and a rotation such that distances between pairs of geographical points are preserved.

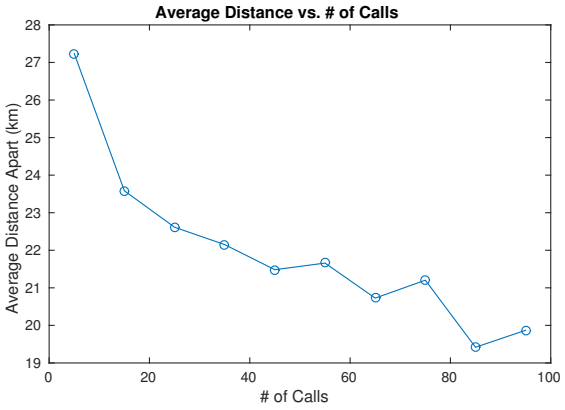


Figure 1: Average Distance vs. Number of Calls. The number of calls were grouped by buckets of 10 and the distance was averaged for each bucket.

3 METHODOLOGY

With the LACDR data spanning n devices, a sparse $n \times n$ matrix X is constructed to represent the weighted communication graph of the devices. Let $X(i, j)$ be the number of communication events between device i and j. Both SMS messages and phone calls are considered communication events. Through X, we have a surrogate measure of the degree of social interaction between pairs of individuals.

To analyze the average calling time, a sparse $n \times n$ matrix A is estimated from data. Therein, $A(i, j)$ is the average time spanned by each communication event between device i and device j. The scalar $A(i, j)$ is calculated by dividing the total (undirected) communication time between i and j by the total number communication events between device i and device j. Fo A, only phone call events are considered since our SMS records do not convey duration or extent of communication.

The location history of the devices consists of the GPS coordinate along with a timestamp of when the coordinate was recorded. For device i, let $Y_i = \{Y_{1,i}, Y_{2,i}, \ldots, Y_{T_i,i}\}$ where each $Y_{T,i} \in \mathbb{R}^3$ is a three-dimensional vector containing the latitude, longitude

and event of the t'th event generated by device i. Every device is considered stationary at is previously reported location until the a new location is observed. This interpolation allows us to efficiently compute a measure of average proximity and the time spent at that proximity for a pair of devices. All static proximity computations were performed using the Haversine forumla between the GPS coordinates of two devices. The average distance between two devices is then calculated as a weighted average of the Haversine distances each weighted by the amount of time spent at that distance. The average distance between two users (or devices) i and j (or their average proximity) is stored as element $D(i, j)$ in a sparse matrix D. It is important to maintain a sparse representation of the $n \times n$ D matrix since n is so large. We therefore explicitly computed the distances precisely for pairs of users which communicated at least once (i.e. we apply the sparsity pattern in X to the D matrix). For the non-communicating pairs, we estimate the distances through random sampling rather than exhaustively enumerating all possibilities.

4 RESULTS

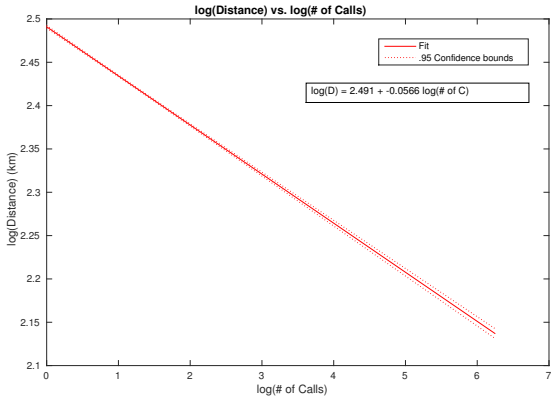


Figure 2: Relationship between Distance and Number of Calls.

The regression analysis of C in Figure 2 shows that the average distance between two devices decreases as the number of communication events increases. A linear relationship between log(distance) and log(calling frequency) provides an especially good fit for this phenomenon.

Analysis of A in Figure 3 provides insight into another relationship. The average call time increases as the average distance between devices increases. This emperical result hasn't been previously studied to our knowledge.

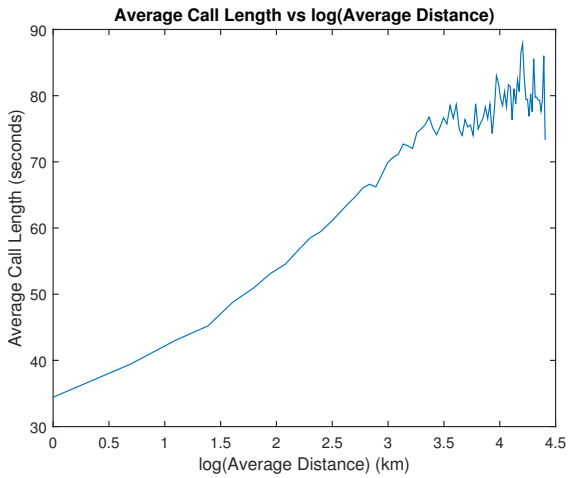


Figure 3: Average Call Length vs. log(Average Distance). The distances were rounded to the nearest kilometer before the log function.

5 CONCLUSION

In this article, we present emperical evidence of two social phenomenons in the LACDR data set. Our analysis shows that individuals communicate with those in their close proximity with greater frequency than those further apart and the average communication time increases as the distance between two devices increase.

References

[1] M. Burke, R. Kraut, and C. Marlow. Social capital on facebook: Differentiating uses and users. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, page 571–580. ACM, ACM, 2011.

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

[3] L. G. Moyano, O. R. M. Thomae, and E. Fras-Martnez. Uncovering the spatio-temporal structure of social networks using cell phone records. In J. Vreeken, C. Ling, M. J. Zaki, A. Siebes, J. X. Yu, B. Goethals, G. I. Webb, and X. Wu, editors, *ICDM Workshops*, pages 242–249. IEEE Computer Society, 2012.

[4] D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A.-L. Barabasi. Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 1100–1108, New York, NY, USA, 2011. ACM.

How mobile positioning data can contribute to urban geography: measuring ethnic segregation in daily activity spaces in Estonia

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

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

[2] Positium LBS, Estonia

The aim of this presentation is to introduce the theoretical and methodological aspects of using passive mobile positioning data in studying ethnic segregation. Passive mobile positioning data is secondary data, which is recorded in various phone use logs, for example, the Erlang data of mobile phone data (1) or Call Detail Record (CDR) data (2) is used. The presentation is based on two academic articles on the ethnic segregation analyses in Estonia.

Ethnic segregation as spatial separation of some population groups from others is one of the most important population processes in urban areas (3). Segregation of minorities is usually deemed to be negative because the isolation is associated with problems in education, employment, poverty, safety, and health care (4). A traditional analysis of segregation on the basis of a study of activity places (residence, place of work, leisure) may not show the complete picture of the population processes because the activities may take place across many different places or activities (5).

Thus, researches have increasingly highlighted the need to study segregation in the whole extent of the places of activity of people and in the whole extent of the 24-hour activity cycle (6,7). Such studies are becoming possible and interesting in connection with taking into use of various modern tracking data, which enable to observe people ubiquitously in time and space. The use of such data is limited by many restrictions related to the right of use of the data and privacy. There are also huge methodological and theoretical challenges. How to enrich quantitative tracking data to a level interesting to modern social sciences? How to make space-time tracking data statistically processable? How to interpolate them spatially and temporally?

For the first case, we will be introducing the use of mobile data in studying the temporal differences in the use of urban space.

Original paper was published in Social Science Research in 2014 (8). The objective of this paper was to determine the temporal dimension of ethnic segregation in Tallinn. We used the CDR data over the course of three years to study temporal segregation levels in urban space through the day, the week, and the year.

The methodological challenge of the analysis was aggregation of the mobile network operator's (MNO) CDR data into units of suitable temporal and spatial accuracy. It is necessary to select units of time of the most suitable length and units of the space of the most suitable size. This also depends on the number of respondents in the CDR database and the number of calls made by them.

As result of methodological development of data enrichment algorithms we measured segregation for three-hour periods in city district level using traditional segregation indixes – the index of dissimilarity (ID) and the location quotient (LQ) method – and compared the results with residence-based indices based on 2000 census data.

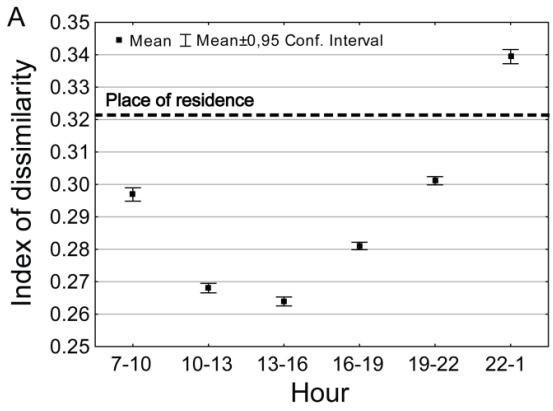


Figure 1A. The index of dissimilarity during the day compared to the places of residence by census data.

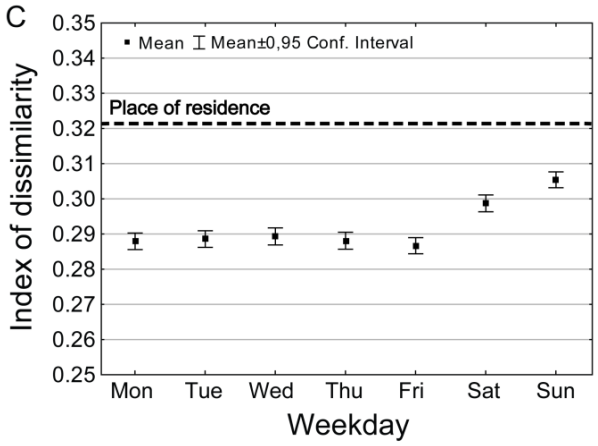


Figure 1B. The index of dissimilarity during the week compared to the places of residence by census data.

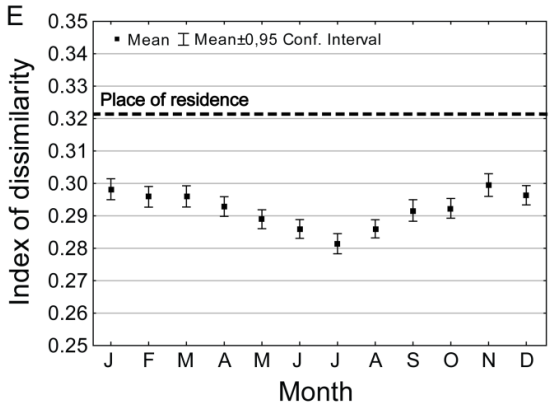


Figure 1C. The index of dissimilarity during the annual cycle compared to the places of residence by census data.

The results indicated that the locations of people are more segregated at night, with considerably less segregation during the daytime (Figure 1A). The segregation is significantly lower on workdays compared to weekends (Figure 1B). Segregation is also lower during summer holidays compared to the winter working period (Figure 1C). The results show that although places of residence are segregated, different ethnic groups use the city together during the day, which increases the potential for interethnic contacts. The results demonstrate also that temporal segregation based on mobile-phone use are considerably lower than segregation of places of residence that are derived from the census.

For the second case, we will be introducing the use of mobile data in studying ethnic variations occurring in the space usage. The original paper was published in the Annals of Association of American Geographers in 2014 (9). The aim of the article was to find the differences in the leisure time space usage of ethnic groups. From the theoretical perspective, this is an interesting challenge, because people spend the majority of their time at home and at work and thus researches have studied these anchor points most. However, many researchers claim that ethnic and social differences gain their true meaning outside of the routine, e.g. outside of the daily places of activity and in the course of leisure time activities. Then people get to choose with whom and where they go.

From the methodological perspective, the challenge of this study is to develop an algorithm which would enable to distinguish home, the place of work, and the places of leisure time space usage from the CDR data. For the theoretical basis, we used the "out-of-home non-employment activities" (10) concept of travel behaviour researchers, and developed a special algorithm to find it from the CDR data. This is based on the anchor points model with the help of which we can find regularly visited places, home, work and other anchor points. Development of the anchor point model was based on a survey (11). Using the different spatial units of GIS databases, we developed an algorithm with the help of the elimination method, which extracts the places of activity that can be defined as out-of-home non-employment activity places from the CDR data. Based on the analysis of these places of activity, we conducted a statistical analysis to find ethnic differences.

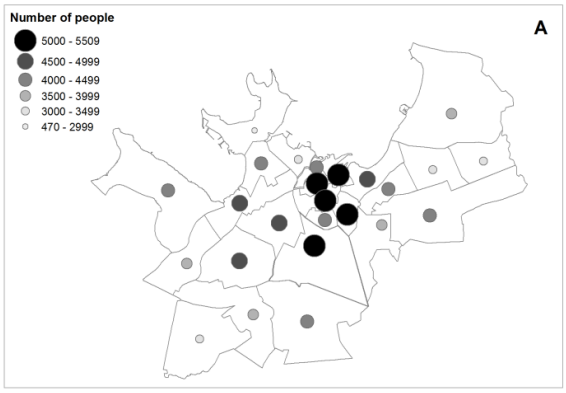


Figure 2A. Differences in districts in Tallinn visited by Estonians.

Figure 2 B. Differences in districts in Tallinn visited by the Russian-speaking minority (B).

The results show that ethnicity has a significant influence on the leisure mobility of individuals. The biggest differences between the two population groups occur in Estonia outside the

respondents' home city of Tallinn. The Russian minority were found to visit 45 percent fewer districts in Estonia (excluding Tallinn) than Estonians. Moreover, they exhibit a preference for districts in Tallinn and Estonia generally that are more heavily populated by the Russian-speaking population. With respect to international travel, the Russian-speaking minority visit fewer countries and have a 3.6 times higher odds of visiting former Soviet Union countries than Estonians. The space usage in out-of-home non-employment activities have fewer differences between the two groups in Tallinn. Overall, our results show that ethnic differences has less effect on everyday space usage and a greater influence on the choices made regarding long-distance travel (Figure 2A and 2B).

Having introduced the two approaches to using mobile data in examining a “classic” social science problem, the generalising part of this presentation will discuss the challenges involved in using mobile data. The results showed that by using digital behavioural data, it is possible to discover new aspects of segregation and further develop the theoretical approach to segregation. Segregation at the place of residence can be related to inertia and the residential property market, people may have many interethnic contacts on the daily basis. Everything also depends on the scale of the city and the functionality of the urban space. A certain warning sign of “ghettoization”, however, is the encapsulation of all of the people’s trajectories over 24 hours in their neighbourhood or places connected to the same ethnic group. This dimension must not necessarily be ethnic, social isolation and financial stratification have the same effect.

From the methodological perspective, we will be discussing the advantages arising from using the data and the limitations arising from the peculiarity of CDR data in conducting such urban geographical researches. A reasonable amount of theoretical task setting and methodological courage enable to develop data processing algorithms, which make the CDR data of little promise at first sight interesting even for a more demanding researcher.

References

(1) 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.

(2) 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.

(3) Ellis, M., Wright, R., & Parks, V. 2004. Work Together, Live Apart? Geographies of Racial and Ethnic Segregation at Home and at Work. Annals of the Association of American Geographers, 94 (3), 620–637.

(4) Massey, D. S., & Denton, N. A. 1988. The dimensions of residential segregation. Social Forces, 67, 281–315.

(5) Strömgren, M.; Tammaru, T.; Danzer, A.; van Ham, M.; Marcinczak, S.; Stjernström, P.; Lindgren, U. (2014). Factors shaping workplace segregation between natives and immigrants. Demography, 51(2), 645 - 671.

(6) Schnell, I., & Benjamini, Y. (2001). The sociospatial isolation of agents in everyday life spaces as an aspect of segregation. Annals of the Association of American Geographers, 91 (4), 622–636.

(7) Järv, O., Müürisepp, K., Ahas, R., Derudder, B., Witlox, F. 2015. Ethnic differences in activity spaces as a characteristic of segregation: A study based on mobile phone usage in Tallinn, Estonia Urban Studies 0042098014550459, first published on September 22, 2014 as doi:10.1177/0042098014550459

(8) Silm, S. & Ahas, R. 2014.The temporal variation of ethnic segregation in a city: evidence from a mobile phone use dataset, Social Science Research 47: 30-43

(9) Silm, S. & Ahas, R. 2014. Ethnic differences activity spaces: The study of out-of-home non-employment activities with mobile phone data, Annals of Association of American Geographers 104(5): 542-559.

(10) Kwan, M.-P. 1999. Gender, the home-work link, and space-time patterns of non-employment activities. Economic Geography 75 (4): 370-94.

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

Protecting the Privacy of Location Data using the openPDS/SafeAnswers Framework

Fatima Makki[a], Wassim El-Hajj[a], Abbas El-Hakim[b], Yves-Alexandre de Montjoye[c]

a Computer Science Dept., American University of Beirut, Lebanon, [fmm26,we07]@mail.aub.edu
b Department of Mathematics, American University of Beirut, Lebanon, aa145@aub.edu.lb
c Media Lab, Massachusetts Institute of Technology, Cambridge, MA, USA, yva@mit.edu

***Abstract*—Billions of applications have already been downloaded by smartphone users. To successfully download an application, the application asks the user to accept a set of permissions allowing it to access sensitive information on the phone. It is mostly unclear when such data is being accessed, how it is being used, and for what purposes. A framework called openPDS, was recently suggested to limit such privacy invasion. openPDS gives the user full control over her data and only allows access to the data/metadata via safe answers. Although openPDS protects the user's raw data, sensitive personal information, such as location trace, can still be inferred by the service provider by analyzing the accumulated answers. In this work, we focus on location privacy and append openPDS with a module that prevents the service provider from reconstructing the trace of users. The module might provide a correct answer, a wrong one, or might not answer at all. However, the module guarantees quality of service (QoS) by abiding with the QoS requirements necessitated by the provider, which we define as the tolerance of the application to inaccurate answers. We tested our approach on 10 users whose locations traces were recorded for 10 months. The results show that no user trace was successfully reconstructed even when high QoS levels were required. Moreover, the adversary knowledge gained through the recurrent months was not maintained.**

Keywords-component; openPDS; location privacy; smartphones; mobile apps

I. Introduction and Background

Mobile phone, these little devices we carry around pretty much 24/7, are probably the most ubiquitous behavioral and locational sensor currently available. A recent study by CNIL [1], the French data protection authority, showed that 22% of Android and 31% of iOS applications are getting permission to access our location at any point in time.

While many of these applications access location data for legitimate purposes, the proliferation of raw and large-scale location databases is a source of concern; 62% of Americans consider data about their exact location to be moderately or extremely private [2], and research has shown that it is easy to re-identify individuals in simply anonymized mobility dataset.

Numerous solutions have consequently been developed to protect the privacy of individuals on smartphones [3-8]. Aquifer [3] protects the user from unwanted information disclosure that is permitted to applications by controlling the data access after each interaction with the server provider. This is done by designing applications that present to users a clear flow of their information and adds restrictions to data accesses that appear to it as unlawful. Authors in [4] suggest market-aware privacy protection framework that includes a feedback control loop that adjusts the mobile privacy settings based on the monetary revenue generated by advertisements. MockDroid [5] and AppFence [6] exchange sensitive data with fake ones, for example, by submitting fake GPS coordinates for an application requesting them; however, it is not location specific and works for other data as well. TaintDroid [7] monitors the flow of privacy-sensitive data and identifies potential misbehavior by third-party applications. Other solutions such as the ones in [8-12] protect the user from specific attacks and target specific data. However, these solutions are so extreme to the extent that they do not permit service providers from accessing historical data, or incorporating temporal factors [13], making their approach less practical especially for applications that target data-science statistics and findings. Yet, other approaches achieve privacy by protecting the user identity within a group (mix-zone) using approaches similar to k-anonymity [14,15] (or its variants), or by obfuscating spatial or temporal location related data [16].

openPDS/SafeAnswers [17], on which our solution is based, takes a different approach. In short, location information about where a user was and is, is collected and stored under his control on his PDS. The user can then allow applications to ask questions, in the form of code, to his PDS. For example, an application might want to know whether a user is currently on AUB campus. In this case, the code sent by the application would be run in a sandbox and a yes/no answer will be generated depending on whether the user is on AUB campus or not. The fundamental difference between openPDS and other solutions is that openPDS allows users to use and answer questions using their full mobility trace without sharing the raw data.

Although openPDS never shares raw mobility data, a malicious application might try to infer more information through a specific sequence of questions-answers. For instance, openPDS does protect the user from having the application provider know her exact location at a certain time instance. In its current form, however, it does not have a module to prevent the application provider from inferring the mobility trace of the user after a certain number of questions-answers.

We here propose such a module for openPDS/SafeAnswers. The module processes location related questions, whether a user is or is not within a given region, by telling the truth, lying, or by not providing an

answer. Our module prevents application providers from inferring a user's trace while ensuring a good quality of service. We here define quality of service (QoS) as the percentage of wrong/no answers the application can tolerate. This percentage can be anywhere between 10% and 100%. We also define privacy to be the percentage of knowledge gained about the user's mobility. This privacy metric will be quantified in section II.

The module works as follow: the module records questions that have been previously asked by the application and creates spatiotemporal profiles of both, what the application knows and that of the actual user. Both spatiotemporal profiles are created using first-order Markov mobility chain. Using these profiles, the module can decide on providing or not a correct answer. Whether the module provides an answer is a trade-off between the QoS and the maximal difference between the adversary's profile and the actual profile, which actually translates to privacy level. If the question is answered, the spatiotemporal profile of the adversary is updated. This process is repeated for every question. It is to be noted that the location privacy problem we are tackling here is different than the traditional location privacy problems mentioned before where the privacy approach relies on achieving anonymity within a group by using k-anonymity [14,15] or its variants, obfuscating temporal and spatial information related to locations, tainting private data, or protecting against specific attacks. Our approach however is user-centric that balances between QoS and privacy and poses no restrictions on data needed by the application.

We tested our proposed location privacy method on the mobility traces of 10 students in the American University of Beirut over the course of 10 months. Results show a good privacy level with high QoS. Our module successfully prevents an adversary from accumulating knowledge about a user's location across time.

II. Location Privacy Module

As a first step, we create spatiotemporal profiles (mobility profiles) of the user (P_{real}) and of the application as it is learning from the recurrent questions-answers across time (we call this profile $P_{adversary}$). Both spatiotemporal profiles are created using first-order Markov mobility chain. Based on the notion that users act similarly in the same time period of the day, we integrate the time domain in the mobility profile as done in [16]. The week is divided into four time periods, three for weekdays (morning – 7 to 11 am, noon – 12 to 6 pm, and night – 7:00 pm to 6:00 am) and one representing weekends. The spatial resolution considered in the mobility profile is as coarse as city level; for example, we consider the area Hamra (AUB campus included) as visited location instead of AUB campus. It is to be noted though that our approach allows the application to ask at anytime using any spatial resolution. Each state s of the mobility chain is thus the spatiotemporal event of the transition, which is the combination of the visited location with the particular time period. An example of a state s is the probability of a user being in Beirut (location) in the morning (time period). Since the profile is regular and aperiodic, based on the Ergodic theorem, the profile (the first-order Markov mobility chain) will converge to a unique steady state distribution π. Each entry π(s) in π is the presence probability of the user in state s i.e. each entry represents the percentage of a user being in location *x* at time period *t*. Our objective is to answer the adversary in such a way to keep the distance between the stationary distribution of the adversary model $\pi_{adversary}$ and that of the real profile π_{real} the maximum possible, knowing that this will lead to distinct profiles and hence different traces, which is confirmed in the experimental results.

Since the adversary gains his knowledge from the questions he poses, the next step is to update the adversary's stationary distribution based on the answer provided by the location privacy module. Three kinds of answers can be provided by the privacy module: true, false, and no answer. We consider an answer to be true if it is correct, and false if it is a noisy one. When not answering, we are sure that the attacker is learning nothing, while when submitting an answer, the attacker is not sure if the submitted answer is correct or noisy. Hence, noisy answers will help in misleading the adversary. Going forward, we propose having three privacy zones: Favorable, Steady, and Danger. We categorize the user to be in a Steady zone if the states in $\pi_{adversary}$ are uniformly distributed with a small tolerance, which means that the attacker cannot infer the user's location. If both $\pi_{adversary}$ and π_{real} include a peak at location *i* in time period *t*, and the adversary's question in time period *t* is whether the user is in location *i* or not, and the user is actually in *i,* the zone is considered Danger, because the adversary is getting closer to the actual most visited location by the user in time period *t*. Otherwise, the zone is considered Favorable. If the user is in a Favorable zone *f*, the privacy module tends to answer truly more often than false, for example T_f=85%, $F_f = 10\%$ and N_f=5% where T_f, F_f, and N_f are the percentages of answering truly, falsely, and not answering. If the user is in a Danger zone *d*, the privacy module tends to answer falsely more often than truly, for example T_d=45%, $F_d = 50\%$ and N_d=5%. If the user is in a Steady zone *a*, the privacy module tends to answer in a way to satisfy the required QoS, for example T_a=70%, $F_a = 25\%$ and N_a=5% when the QoS is 70%, which means that the location privacy module will answer truly 70% of the time. For the QoS to be respected, the following inequality should be enforced

$$dT_d + fT_f + aT_a \geq Q$$

where d, f, and a are the normalized number of times the user enters into Danger, Favorable, and Steady zones respectively, and Q is the percentage of true answers generated by the privacy module.

Instead of fixing the percentages of answering true, false, or no answer in every zone, we resort to calculate these percentages based on an optimization problem where we aim at maximizing the distance between $\pi_{adversary}$ and π_{real}, as follows:

$$F_z \left\| \pi_{real} - \pi^{F}_{adversary} \right\|_2^2 + T_z \left\| \pi_{real} - \pi^{T}_{adversary} \right\|_2^2$$

where F_z and T_z are the optimization variables representing the percentages of answering falsely or truly given zone z, and $\pi^{F}_{adversary}$ and $\pi^{T}_{adversary}$ are the stationary distributions of the adversary knowledge after answering falsely or truly (discussed next). The optimization problem is subject to $dT_d + fT_f + aT_a \geq Q$, where Q is a given percentage (QoS constraint).

After obtaining the optimal percentage T_z and F_z in every zone z, we need to update the adversary's knowledge based on the answer provided by the privacy module, as follows:

$$\pi_{answer}(i) = \begin{cases} T_z, & q = i \text{ and } Answer = true \\ F_z, & q = i \text{ and } Answer = false \\ \dfrac{T_z\,\pi_{adversary}(i)}{1-\pi_{adversary}(q)}, & q \neq i \text{ and } Answer = true \\ \dfrac{F_z\,\pi_{adversary}(i)}{1-\pi_{adversary}(q)}, & q \neq i \text{ and } Answer = false \end{cases}$$

where, π_{answer} is what the adversary is going to learn, i is the location the adversary is learning about, and q is the location the adversary is asking about (hence, i and q are indices in π).

Having determined the knowledge gained by the adversary after the submitted answer, we update the adversary profile ($P_{adversary}$) by updating all the possible transitions (pt) in $P_{adversary}$ that could have been made to reach the current period. The new adversary profile is thus calculated as follows:

$$P_{adversary} = \sum_{pt}^{PT} (P^{old}_{advesary}(pt) * \sum_{i} \pi_{answer}(i) * P^{plus}_{advesary}(pt, i))$$

where $P^{plus}_{advesary}(pt, i)$ is the adversary profile with only the element at (pt, i) incremented by one transition. A new stationary distribution $\pi_{adversary}$ is then calculated from the updated $P_{adversary}$. The same steps are repeated every time a new answer is given. After every answer, we quantify the privacy level of the user. In a given time period, the user's top visited location is identified, meaning that the most visited location in a specific time period in $\pi_{adversary}$ is the same as the most visited location in the same time period in π_{real}. We consider a user's most likely trace to be exposed or breached (privacy level) if the adversary was able to identify the top visited location in each period of the available four periods. The more locations identified, the higher the privacy breach.

III. Experimental Results

Ten student volunteers from AUB were provided with an Android application to record their traces for a period of 10 months. 6 months were used to build their real profiles (P_{real}). The other 4 months were used to test the effectiveness of the privacy module. Since there are no clear statistics to indicate the number of questions a normal app might ask and at what rate, we assumed that the application would ask, on average, 12 questions/day. Every question asks about a location i.e. are you in location X? X can be any location that the user visited in the past, in addition to other locations that the user might visit. In our experiments we set the number of locations to 10. Hence, the maximum number of locations visited by our volunteers was 6. The aim of the location privacy module is to provide answers to the app within the corresponding QoS limitations while achieving a good privacy level (few locations identified as mentioned in the end of section II).

We considered two types of adversaries: Regular and Malicious. A Malicious adversary targets in his questions the top visited locations by the user in each time period based on his findings. For example, if an attacker noticed that 80% of a user's presence in morning is in location x, the adversary will target this location. Questions about such locations will be increased in number and rate. Consequently, he can identify the user trace faster. On the other hand, a Regular adversary does not bias his questions.

We tested our privacy module using three QoS levels: 70%, 85%, and 90%. In total 6 experiments were conducted by switching between the regular and malicious adversary and the various QoS levels.

After each month of testing, we calculate the average number of locations identified for a user by the adversary. Since the maximum breach is to identify the four most visited locations by the user, we calculate the average as follows:

$$\frac{\sum_{n=0}^{4} n * number\ of\ users\ with\ n\ locations\ identified\ in\ a\ given\ month}{total\ number\ of\ users}.$$

To know the maximum breach level for a user profile reached by the adversary throughout the whole testing period, we report the highest breach level reached across the months by both malicious and regular adversary. Figure 1 shows the maximum an adversary (Malicious or Regular) can know (highest breach level reached) vs. the 3 QoS levels used. As shown, our privacy module achieved 60% privacy level with 90% QoS level (40% breached). We can notice that as the QoS increases, the privacy level decreases. Moreover, results show that profile knowledge acquired by the adversary in one month is not maintained in the following months. This is in part due to the lying strategy adopted by the location privacy module when the user is categorized to be in a danger zone.

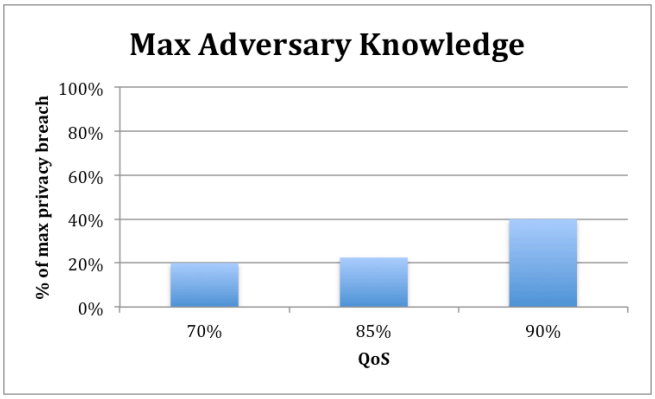


Figure 1. Max privacy breach level vs. QoS level

In conclusion, regardless of the adversary type and the QoS level used, the adversary was only able to know maximum 40% of a user profile on average, after 4 months. We can deduce that the adversary cannot maintain the knowledge achieved across time, and hence the locations that were identified were identified based on pure luck and this will remain the case throughout the months to come.

REFERENCES

[1] http://www.cnil.fr/fileadmin/documents/La_CNIL/publications/DEIP/Lettre_IP_N-8-Mobilitics.pdf

[2] https://www.bcgperspectives.com/content/articles/information_technology_strategy_consumer_products_trust_advantage_win_big_data/?chapter=2

[3] Nadkarni, Adwait, and William Enck. "Preventing accidental data disclosure in modern operating systems." Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security. ACM, 2013.

[4] Leontiadis, Ilias, et al. "Don't kill my ads!: balancing privacy in an ad-supported mobile application market." Proceedings of the Twelfth Workshop on Mobile Computing Systems & Applications. ACM, 2012.

[5] Beresford, Alastair R., et al. "MockDroid: trading privacy for application functionality on smartphones." Proceedings of the 12th Workshop on Mobile Computing Systems and Applications. ACM, 2011.

[6] Hornyack, Peter, et al. "These aren't the droids you're looking for: retrofitting android to protect data from imperious applications." Proceedings of the 18th ACM conference on Computer and communications security. ACM, 2011.

[7] Enck, William, et al. "TaintDroid: an information flow tracking system for real-time privacy monitoring on smartphones." Communications of the ACM 57.3 (2014): 99-106.

[8] Beresford A, Stajano F (2003) Location privacy in pervasive computing. Pervasive Computing, IEEE 2: 46-55.

[9] Gedik B, Liu L (2005) Location privacy in mobile systems: A personalized anonymization model. In: Distributed Computing Systems, 2005. ICDCS 2005. Proceedings. 25th IEEE International Conference on. Ieee, pp. 620-629.

[10] Zhong G, Goldberg I, Hengartner U (2007) Louis, lester and pierre: Three protocols for location privacy. In: Proceedings of the 7th international conference on Privacy enhancing technologies. Springer-Verlag, pp. 62-76.

[11] Mascetti S, Freni D, Bettini C, Wang X, Jajodia S (2011) Privacy in geo-social networks: proximity notification with untrusted service providers and curious buddies. The VLDB JournalThe International Journal on Very Large Data Bases 20: 541-566.

[12] Reades J (2010) Finite state machines: preserving privacy when data-mining cellular phone networks. Journal of Urban Technology 17: 29-40.

[13] Monreale A, Andrienko G, Andrienko N, Giannotti F, Pedreschi D, et al. (2010) Movement data anonymity through generalization. Transactions on Data Privacy 3: 91-121.

[14] Sweeney, Latanya. "k-anonymity: A model for protecting privacy." International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10.05 (2002): 557-570.

[15] Gedik, Bugra, and Ling Liu. "Location privacy in mobile systems: A personalized anonymization model." Distributed Computing Systems, 2005. ICDCS 2005. Proceedings. 25th IEEE International Conference on. IEEE, 2005.

[16] Shokri, Reza, et al. "Quantifying location privacy." Security and Privacy (SP), 2011 IEEE Symposium on. IEEE, 2011.

[17] de Montjoye, Yves-Alexandre, et al. "openpds: Protecting the privacy of metadata through safeanswers." PloS one 9.7 (2014): e98790.

Characterizing preferences and returns in human mobility

Hugo Serrano*[1], Fernando Buarque de Lima-Neto†[2], Alexandre Evsukoff‡[3], and Ronaldo Menezes§[1]

[1]BioComplex Lab, Computer Science, Florida Institute of Technology, USA
[2]Computational Intelligence Research Group, Polytechnic School of Pernambuco, Brazil
[3]COPPE, Federal University of Rio de Janeiro, Brazil

In the past few years, our understanding of the fundamental laws of human mobility has improved considerably thanks to the increasing availability of time-resolved human mobility data, such as Call Detail Records (CDRs) [3, 7], credit card transactions [6] and location-based services [4, 2] data.

One fundamental characteristic of human mobility, called *preferential return* (PR), was proposed by Song et al. [11] and it models the strong tendency of humans to return to previously visited locations. More precisely, it defines the probability Π_i for returning to a location i as $\Pi_i = f_i$, where f_i is the visitation frequency of the location i. It implies that the more visits a location receives, the more visits it is going to receive in the future, which in different fields goes by the names of *Mathew effect* [8], *cumulative advantage* [9], or *preferential attachment* [1]. Although the focus of the PR mechanism (as part of the Individual Mobility–IM model) was to reproduce some of the scaling properties of human mobility, its general principles are grounded on plausible assumptions from the human behavior point of view.

However, in the long-term, the PR assumption as a property of human motion leads to two discrepancies. First, the earlier a location is discovered, the more visits it is going to receive. Or, in other words, the first visited location will always be the most visited one. Second, if the cumulative advantage indeed holds true for human movements, people would never change their favorite restaurant, change jobs or move to a different city, which is clearly not true.

In this work, we explore the visitation return patterns under a temporal perspective. We analyzed different ranking approaches and tested their respective correlations with the return probabilities. Our approach is based on the premise that the longer the time since the last visit to a location, the lower is the probability of observing a user at this location [11, 3], as depicted in Fig. 1A. The proposed approach overcomes the those

*hbarbosafilh2011@my.fit.edu
†fbln@ecomp.poli.br
‡alexandre.evsukoff@coc.ufrj.br
§rmenzes@cs.fit.edu

two limitations of the PR mechanism by giving a higher importance to recently visited locations.

Our findings are based on the analysis of 6 months of anonymized mobile phone data from Brazil,[1] The fundamental question our approach answers is: *from the visitation frequencies point of view, what is the most likely destination of a person? One of the most visited places but whose last visit was a long ago, or a recently discovered place?*

To answer these questions we compared two different visitation ranks, one based on the visitation frequencies (k_f) and the other based on the recency (k_s), measured as a function of the elapsed time (i.e., number of steps) since the last visit to a location. Both ranks were measured in a rolling basis from the accumulated sub-trajectories. The most visited location will have $k_f = 1$ whereas the most recently visited location will have a rank $k_s = 1$. For each return, we collected both location ranks. For instance, a return to the $10th$ most visited place right after visiting it is accounted as the pair ($k_f = 10$, $k_s = 1$). Conversely, a return to the most visited place (e.g., home) after 10 steps is represented by ($k_f = 1$, $k_s = 10$).

When we look at the distribution of the two variables, we can see that most of the returns are concentrated to recently and frequently visited locations (Fig. 1B). As we can see, both rank distributions can be better approximated by similar truncated power laws. The lower limits of the heavy-tails ($k_s min = 4$ and $k_f min = 2$) suggest that the visitation patterns to the most visited location and the last 3 steps in human trajectories are different and deserve further investigation.

Inspecting the returns heat map (Fig. 1C) we can see that even though the two variables are positively correlated, for the most visited sites (first column of the heat map) the return data points are heavily clustered around low k_s values, indicating that sub-trajectories starting and finishing at these locations are much shorter than of the other points. It is consistent with human travel patterns: on a daily basis we tend to return back to home [5, 10]. On the other hand, when it comes to the recently visited locations, the data points corresponding to $k_s \leq 3$ (bottom-most rows in the plot) span over a large range of k_f values, supporting our hypothesis that the most recent visits have an important role in determining our future steps. Not surprisingly, the particular range of values of k_s corresponds to recurrent visits to low-ranked locations which is the expected outcome of long-term changes in users' visitation patterns, such as due to a relocation or new visitation preferences.

To measure to what extent users tend to return to recently-visited places, we defined $\Pi_s(r) = p(k_f = r \mid k_s = 1)$ as *the probability of returning to the r^{th} most-visited location right after visiting it ($k_s = 1$)* and $\Pi_f(r) = p(k_s = r \mid k_f = 1)$ as *the probability of returning to the most visited location ($k_f = 1$) after r steps since the last visit to it.*

The next step is to compare these two probabilities and how they vary with r. For such we defined $\rho(r)$ simply as

$$\rho(r) = \frac{\Pi_s(r)}{\Pi_f(r)} = \frac{p(k_f = r \mid k_s = 1)}{p(k_s = r \mid k_f = 1)}. \quad (1)$$

A $\rho(r) > 1$ suggests a preference for recently visited locations. On the other hand, $p(r) < 1$ implies that $\Pi_f(r) > \Pi_s(r)$ and hence a preference for highly visited locations.

[1] Approximately 8.9 million records of 30,000 randomly sampled users from one of the largest phone carriers in Brazil.

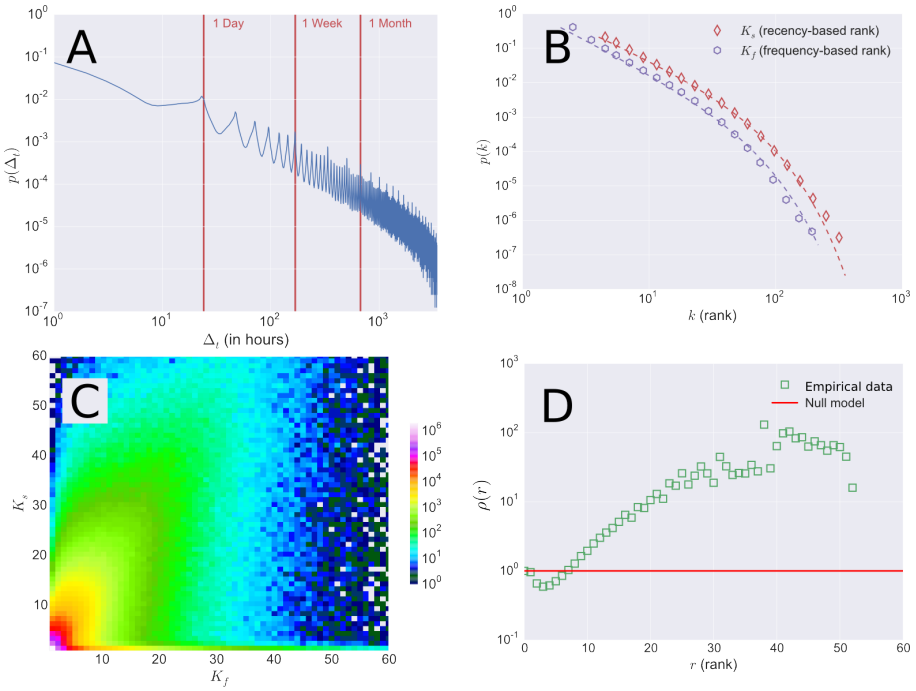


Figure 1: **Return characterization A.** Return probabilities as a function of the elapsed time Δ_t since the last visit. Peaks are observed at 24h intervals, capturing the temporal regularity of which humans return to previously visited locations. Also, it is possible to see that the return probability decays very quickly as the time increases, which corroborates with our main hypothesis. **B. Return ranks (probability density function)**. Both distributions can be better approximated by truncated power laws (dashed lines). The recency-based rank has exponents $\alpha_{K_s} = 1.64$ and exponential cut-off $\kappa_{K_s} = 40.94$, whereas the frequency-based rank distribution has a better fit for $\alpha_{K_f} = 1.86$ with $\kappa_{K_f} = 36.88$. **C. Return ranks**. Each point represents a return step, whereas the color encodes the density of points. This plot represents the probability of returning to a location of rank $k_f = K_f$ and $k_s = K_s$. **D. Recency over frequency**. From the empirical data, we can see that a recent visit to place increases significantly the probability of returning to it. That is true especially for locations with $k_f \geq 4$.

The case where $\rho(r) = 1$ means that both ranks have the same influence on the visitation probabilities (null hypothesis).

From the empirical data, our analyses on the behavior of $\rho(r)$ suggest that the preference for recently visited locations can be more than 50 times higher than the visitation frequency which corroborates with our thesis that visitation preferences are biased in favor of the recent past trajectory. A further investigation on the $\rho(r)$ value–including other datasets–has shown that the range of $k_f < 4$ in which the visitation frequency seems to have a stronger importance than the recency, is due to the visits to the most visited location. When we analyze the same data after removing the most visited location (most likely the users' homes), the influence of the recency becomes even more evident with $\rho(r) > 1$ for all values of r.

In this work we explored the fundamental mechanisms of human preferences and returns. Our results has shown that a recent visit to a location increases dramatically the probability of returning to it in a near future, supporting the assumption of a preference for recently visited locations, regardless their frequency-based ranks. The results can be used to explain not only long-term changes in visitation preferences but also short-term transient trajectory changes such as temporary relocation and seasonal trends.

References

[1] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, page 11, Oct. 1999. doi: 10.1126/science.286.5439.509.

[2] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '11*, page 1082, New York, New York, USA, 2011. ACM Press. ISBN 9781450308137. doi: 10.1145/2020408.2020579.

[3] M. C. González, C. A. Hidalgo, and A.-L. Barabási. Understanding individual human mobility patterns. *Nature*, 453(7196):479–482, June 2008. ISSN 1476-4687. doi: 10.1038/nature06958.

[4] P. Grabowicz, J. Ramasco, B. Gonçalves, and V. Eguíluz. Entangling mobility and interactions in social media. *PloS one*, pages 1–16, 2014.

[5] S. Hasan, C. M. Schneider, S. V. Ukkusuri, and M. C. González. Spatiotemporal Patterns of Urban Human Mobility. *Journal of Statistical Physics*, 151:304–318, 2012. ISSN 0022-4715. doi: 10.1007/s10955-012-0645-0.

[6] C. Krumme, A. Llorente, M. Cebrian, A. S. Pentland, and E. Moro. The predictability of consumer visitation patterns. *Scientific reports*, 3:1645, Jan. 2013. ISSN 2045-2322. doi: 10.1038/srep01645.

[7] X. Lu, E. Wetter, N. Bharti, A. A. J. Tatem, and L. Bengtsson. Approaching the limit of predictability in human mobility. *Scientific reports*, 3:2923, Jan. 2013. ISSN 2045-2322. doi: 10.1038/srep02923.

[8] R. K. Merton. The Matthew Effect in Science. *Science (New York, N.Y.)*, 159: 56–63, 1968. ISSN 0036-8075. doi: 10.1126/science.159.3810.56.

[9] D. Price. A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information ...*, 1976.

[10] C. M. Schneider, V. Belik, T. Couronné, Z. Smoreda, M. C. González, and T. Couronne. Unravelling daily human mobility motifs. *Journal of the Royal Society, Interface / the Royal Society*, 10:20130246, 2013. ISSN 1742-5662. doi: 10.1098/rsif.2013.0246.

[11] C. Song, T. Koren, P. Wang, and A.-l. Barabási. Modelling the scaling properties of human mobility. *Nature Physics*, 6(10):818–823, Sept. 2010. ISSN 1745-2473. doi: 10.1038/nphys1760.

Utilizing Origin Destination Information obtained from Mobile Phones in Equilibriated Path Choice

Serdar Çolak[1,*] and Marta C. González[1,2]
[1] *Civil and Environmental Engineering, MIT*
[2] *Engineering Systems Division, MIT*
(Dated: January 30, 2015)

Congestion is an inherent problem in networks on which agents compete for limited resources and incurred costs are a function of how many people choose to use a supply element. In the context of mobility on road network, we can elaborate this problem as follows: Any driver with a specific destination makes a choice about the route they will take, most often depending on their utility which by definition tries to reflect their preferences, hence utility is inversely proportional to the travel time of the chosen route. When a driver makes this choice, the total cost incurred to the whole system is not only the travel time experienced by that driver, but also the marginal cost that one driver creates on all other users of that road segment. In this work, we begin by mining billions of mobile phone call records to obtain a typical morning commute and validate our findings against surveys. Then we make use of the existing literature to implement a framework to solve the problem of static path selection and how it influences overall congestion. We compare our findings with previously used, unequilibriated traffic assignment methods.

I. INTRODUCTION

In this work, our goal is to utilize an equilibriated traffic assignment procedure to analyze the effects of individualistic path choice on congestion. In addition, we aim to build on the routing framework to further analyze urban mobility from a sociodemographic perspective to try and answer questions that relate commuting times, path choices, and roads used to income, population density and other similar qualities.

In the US the first works using mobile phone data for transportation applications refer to traffic monitoring. Departments of Transportation in different states carry out these studies in collaboration with private data providers. Less is known about how the massive amount of information hidden in several months of anonymized mobile phone bills; also known as call detailed records (CDRs), can support the models of travel behaviors. Here we implement a method of using CDRs that can be used to extract ODs by purpose and time of the day. The presented results are validated against surveys and existing ODs available from the local DOT. We show the robustness of the method comparing the results over two types of phone data sets and in two cities.

In road networks, travel time is known to be a function of the flow on that road segment. Though many expressions of this relationship exist, we will use the most common choice in the literature: The BPR function relating the observed travel time t_e on an edge e with capacity k_e, free travel time t_{e0}, and x_e vehicles traveling on it is,

$$t_e(x_e) = t_{e0}\left(1 + \alpha\left(\frac{x_e}{k_e}\right)^{\beta}\right). \tag{1}$$

where generally $\alpha = 0.15$ and $\beta = 4$.

With the assumption that the observed cost for all users is equivalent to the observed travel time, we can define the total cost incurred by all users in a network as,

$$\text{total cost} = C = \sum_{e \in E} x_e t_e(x_e) \tag{2}$$

The flow configuration obtained by trying to minimize C to decrease the average time an average user will spend for their trip is referred to as the *socially optimal* flows.

However, it is known that the actual flow configuration is far from optimal. As drivers make selfish choices, they push the system away from optimality. The set of flows that occur when every driver individually and selfishly minimizes their own cost is referred to as the *user equilibrium* flows. This problem is actually a routing game, where players are the drivers and their strategy spaces are available paths between their origin and destination. In this context, a Nash equilibrium refers to a set of strategy choices of every user that results in a final outcome where no user can unilaterally deviate from their strategy to get a better payoff, which in the context of road networks is less travel time. These principles have also been summarized under Wardrop's principles [1] in transportation literature.

In solving for equilibrium flows, potential functions are made use of. By definition, a potential game is a game that can be represented by a potential function that when optimized yields Nash equilibria of the problem at hand. A function $\phi_e(x_e)$ such that $\phi_e'(x_e) = t_e(x_e)$ can be found by $\phi_e(x_e) = \int_0^{x_e} t_e(x)dx$. One can therefore write a potential function for the user equilibrium problem as follows:

* serdarc@mit.edu

$$\Phi(f) = \sum_{e \in E} \int_0^{x_e} t_e(x)dx$$
$$= \sum_{e \in E} t_{e0} x_e \left(1 + 0.03 \left(\frac{x_e}{k_e}\right)^4\right) \quad (3)$$

Minimization of this function results in equilibrium flows.

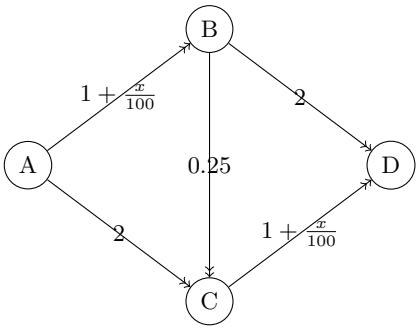


FIG. 1. Example network and an analysis of its equilibrium and optimal flows. For a demand of 100 drivers going from node A to node D, user equilibrium would allocate the flows between paths ABD, ACD and ABCD such that the travel times are equal: 25 people each choose paths ABD and ACD, and 50 people choose path ABCD; resulting in an average travel time of 3.75 for all drivers. Optimal flows would minimize total travel time, yielding in 50 people each on paths ABD and ACD, and zero people on path ABCD; resulting in an average travel time of 3.5 for all drivers. The fact that edge BC remains unused captures what's known as the Braess' Paradox [2]; when additional capacity does not decrease social cost. The price of anarchy under these conditions is $^{3.75}/_{3.5} = 1.07$.

II. METHODOLOGY

Current methods to solve the traffic assignment problem vary in their approaches: Typical gradient descent type algorithms, although fast to solve the problems, do not help with the generation of the possible paths but only generate aggregate link flows that path flows create. On the other hand as network sizes increase, solving explicitly for paths becomes infeasible very quickly. Therefore algorithms that can efficiently compute desired equilibrium path and link flows are preferred. The literature spans many such algorithms for this purpose, with most being derivatives of three main sets of methods: Link based, path based, origin based. Due to it's advantages in generating paths, fast convergence and efficiency, we utilize an origin-based approach in this work. We will follow Algorithm B, proposed in [3] along with modifications and improvements outlined in [4], an origin based algorithm that focuses on the equilibriation of a graph structure referred to as a bush, a directed acyclic graph (DAG) emanating from every origin node. These structures are used with the assumption that in the equilibrium flows, no directed cycles should exist. In fact the computational efficiency of this algorithm stems from this property, as DAGs can be traversed in linear time.

III. RESULTS

We began by comparing our findings of demand from the mobile phone data in Fig. 2. Our findings are in good agreement with the results obtained from traditional surveys. Comparison of the commuting trips for each origin-destination pair in the morning peak shows strong correlations for both inter-town and intra-town trips, reaching $\rho = 0.84$ for Rio de Janeiro and $\rho = 0.99$ in Boston. Fig. 2 also illustrates spatially the flow distribution of the model and the CDR ODs for both cities by mapping color-coded and width adjusted lines between OD pairs whose flow values exceed 0.10% of the total study area trips. By visual inspection, it can be said that CDR data manages to capture the flow distribution of that of the model ODs, with majority of the flows concentrated towards downtown in Boston and downtown and across the bay in Rio de Janeiro.

Theoretically, for a user-equilibrium solution, the total cost obtained by the travel time between an origin and destination multiplied by the demand for that pair should be equal to the flow on every link multiplied by the travel time on that link; as both measure the total travel time. To assess convergence, we use the following measure:

$$\text{relative gap} = 1 - \frac{\sum_{o,d} traveltime_{od} * demand_{od}}{\sum_{i,j} traveltime_{ij} * flow_{ij}} \quad (4)$$

Fig. 3 depicts the convergence of Algorithm B compared to ITA, the incremental assignment approach, as well as a more reasonable and accurate VoC (volume over capacity) for Boston. These findings are preliminary, but do have implications in further usage of user equilibrium models for assignment procedure. Consequently, it is possible to carry out further analysis about path selection, and the effects of urban sociodemographics on congestion.

[1] John Glen Wardrop. Road paper. some theoretical aspects of road traffic research. In *ICE Proceedings: Engineering Divisions*, volume 1, pages 325–362. Thomas Telford, 1952.

[2] Dietrich Braess, Anna Nagurney, and Tina Wakolbinger. On a paradox of traffic planning. *Transportation science*, 39(4):446–450, 2005.

[3] Robert B Dial. A path-based user-equilibrium traffic assignment algorithm that obviates path storage and enumeration. *Transportation Research Part B: Methodological*, 40(10):917–936, 2006.

[4] Yu Marco Nie. A class of bush-based algorithms for the traffic assignment problem. *Transportation Research Part B: Methodological*, 44(1):73–89, 2010.

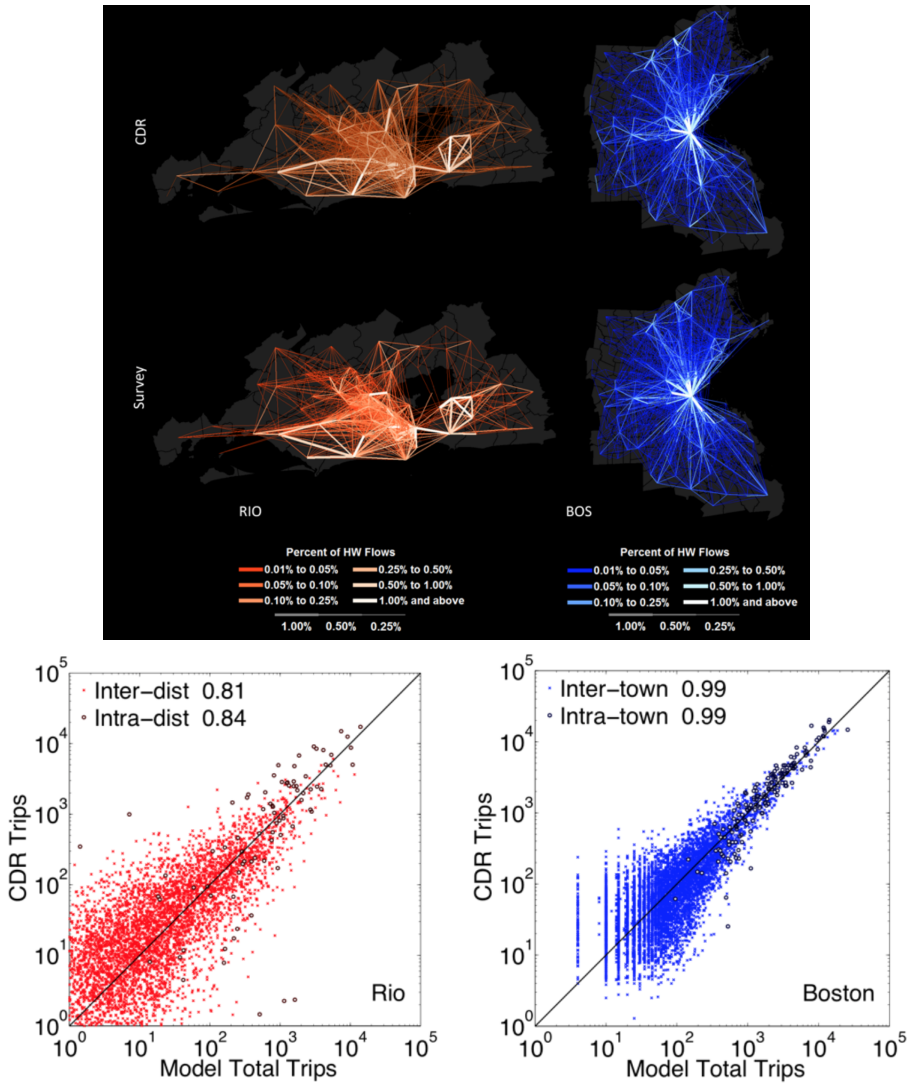


FIG. 2. The flow distributions in Rio and Boston as a comparison between those obtained from the mobile phone data and those from the survey. The correlations for the subdistrict level and the town level are depicted in the legend, for Rio and Boston, respectively.

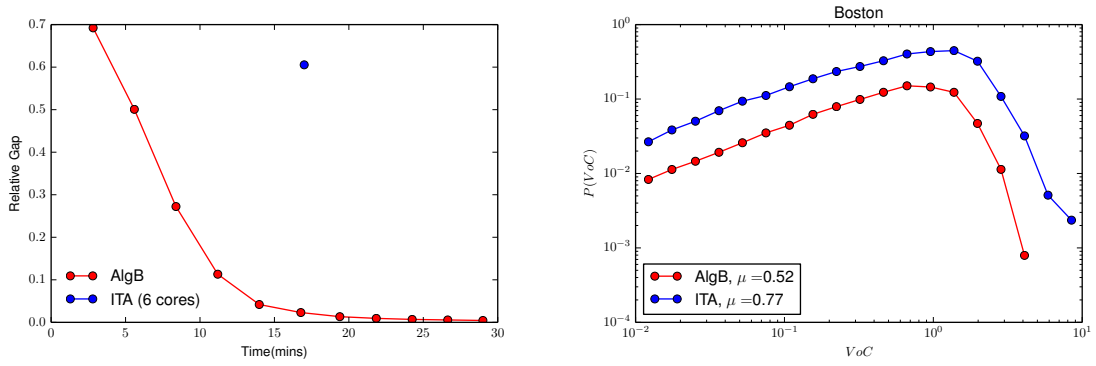


FIG. 3. The performance of algorithm B achieves higeher convergence considerable quickly, as well as returning a considerably better Voc (volume over capacity) distribution for Boston.

Land use classification using call detail records

Kaushalya Madhawa[1], Sriganesh Lokanathan[1], Danaja Maldeniya[1], Rohan Samarajiva[1]

[1]LIRNE*asia*, 12 Balcombe Place, Colombo 8, Sri Lanka

Colombo is the largest city in Sri Lanka by population and economic activity. With the post-war infrastructure boom, land use patterns in Colombo are changing quickly, with residential population continuously decreasing making way for commercial activities. Having an up-to-date overview of land use characteristics is critical in for urban planning, and traditional survey and census-based methods in addition to being expensive cannot give high-frequency insights. Mobile network big data could be utilized as a low-cost and high frequency alternative for understanding changes in land use. Our method potentially goes further than previous efforts, by allowing for a more fine-grained understanding of land use changes.

Whenever a mobile phone is used to make or receive a call or SMS, or to access the Internet, the event is captured via meta-data stored in the operator's logs. This is called a Call Detail Record (CDR). The CDR also contains the location of the Base Transceiver Station (BTS) that services the subscriber, as well as timestamp for the event. Using these records it is possible to get a time-series of activity at each BTS. Building on prior work [1, 2, 3, 4], this usage pattern of BTS activity can be leveraged to understand the characteristics of the land beneath it.

For each BTS, we utilize a month of CDR data to construct a time-series measure of the number of users connected to the BTS at any given time. We then project this data into a 7-day week, excluding data from public holidays since human dynamics change dramatically during holidays. For each BTS we then end up with a time-series measure for each day of the week, where the diurnal activity pattern for that day is an average of all the same weekdays in that month.

Projecting the user data of a month to a week results in a unique 'signature' of mobile usage at each site as seen in Figure 1. A vector of 168 elements represents the time-series signature of a BTS. This signature provides a good measure of the variation in people's behaviors between days and even between the hours of each day.

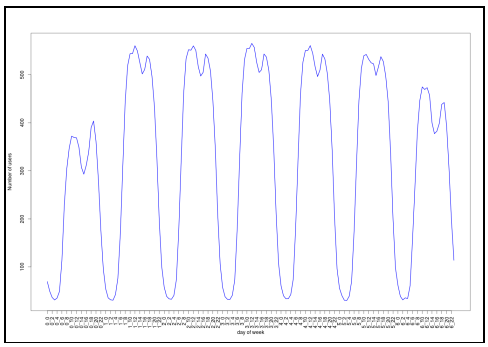


Figure 1: **Hourly user distribution for a week fro a sample BTS**

The actual number of users connected to a BTS at any given time depends on the area it covers and the place it is located at. To remove these biases we normalized the hourly user distribution by taking the z-score. The diurnal signature of each BTS also contains occasional noise that could affect the land use classification. Using methods articulated in [1] to deal with noise in Wi-Fi signals, we remove the random variations in our normalized diurnal data by decomposing the covariance matrix of user time series data into a combination of its principal components.

Each time series vector for a BTS can be represented by a linear combination of its principal eigenvectors as,

(1) $$T_i = C_{i1}V_1 + C_{i2}V_2 + \cdots + C_{in}V_n$$

Here T_i represents the time series of a BTS *i*, and C_{i1} represents the correlation of the first eigenvector V_1 and so on.

The first eigenvector accounts for the general diurnal pattern of the city. The first fifteen eigenvectors of the covariance matrix is capable of representing 95% of the variation in the time series data.

Then the time series distributions are redistributed using only the first fifteen eigenvectors.

Land use classification

The time series of each BTS is then assigned to one of three clusters using a k-means algorithm considering each reconstructed time series as a 168-dimensional vector.

We use the signature of cluster centroids to infer the actual land use category that each cluster belongs to. Whilst prior work [4] had identified more than three clusters, we find that interpretation of the these additional categories is problematic for our data, especially when we validate using known land use characteristics in Colombo.

The centroid pattern of Cluster-3 is consistent with the dynamics of a commercial region, since more users are detected during weekdays than weekends and the daily peak occurs during the daytime. Similarly the centroid pattern of Cluster-2 is consistent with the dynamics of residential regions, with weekday/ weekend variations being minimal and the daily peak occurring during the late evening each day. Cluster-2 is classified as belonging to areas of mixed-use land characteristics, since there are no discernable variations in the time of the daily peaks or between weekdays and weekends.

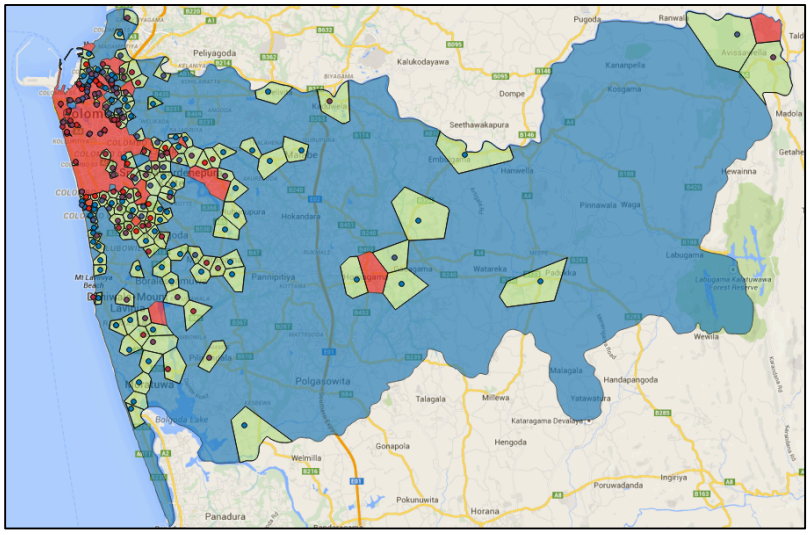


Figure 2: **Distribution of clusters in Colombo district.** Cluster-3 (Commercial) areas are in red, Cluster-2 (residential) areas are in blue, and Cluster-1 (mixed-use) areas are in green. Areas with blue dots signify a BTS signature closer to Cluster-2 than Cluster-3. Areas with red dots signify a BTS signature closer to Cluster-3 than Cluster-2

The spatial distribution of clusters is shown in Figure 2.

Cluster-1 also exhibits the least average silhouette value as compared to the other two clusters. We then calculate the distance from each point in Cluster-1 to the centroids of other two clusters. The ratio of the distance from the centroid of a Cluster-1 area to the centroids Cluster 2, and Cluster 3, is used to understand how the mixed-use regions compare with other two categories.

An analyses of these ratios for each Cluster-1 area (i.e. mixed use), suggests that mix-used regions neighboring a commercial area exhibit a pattern more similar to the commercial pattern than a residential pattern. The inverse seems to hold true for mixed-use areas adjacent to mainly residential areas.

Future Work

We plan to use this technique repetitively over a longer time period to investigate if the classifications of the silhouettes (blue and red dots in Figure 2) show changes. This might reveal a high-frequency measure of the changes in land use than is possible using just the three-cluster classification.

References

[1] F. Calabrese, J. Reades, and C. Ratti, “Eigenplaces : Segmenting Space Through Digital Eigenplaces : Segmenting Space through Digital Signatures,” 2014.

[2] J. Reades, F. Calabrese, and C. Ratti, “Eigenplaces: analysing cities using the space – time structure of the mobile phone network,” *Environment and Planning B: Planning and Design, 36:824–836, 2009*.

[3] G. Andrienko, N. Andrienko, M. Mladenov, M. Mock, and C. Politz, “Discovering bits of place histories from people’s activity traces,” in *2010 IEEE Symposium on Visual Analytics Science and Technology*, 2010, pp. 59–66.

[4] V. Soto and E. Frias-Martinez, “Robust land use characterization of urban landscapes using cell phone data,” *In Proceedings of the 1st Workshop on Pervasive Urban Applications (Purba’11)*, 2011.

Impact of Indoor-Outdoor Context on Crowdsourcing based Mobile Coverage Analysis

Mahesh K. Marina*, Valentin Radu*, Konstantinos Balampekos†
*The University of Edinburgh †Nokia Networks

We consider the crowdsourcing based mobile cellular network measurement paradigm that is becoming increasingly popular. In particular, we aim to study the impact of user indoor/outdoor environment context at time of measurement. Focusing on signal strength as the measurement metric and using a real large crowdsourced measurement dataset for central London area along with estimated environment state (indoor or outdoor), we show that indoor-outdoor context has a significant impact, suggesting that conflating indoor and outdoor measurements can lead to unreliable results. We validate these observations using a set of diverse and controlled measurements with indoor/outdoor ground truth information.

Crowdsourcing [1–10] has recently emerged as a new approach for mobile cellular network measurement and analysis. It exploits smartphones (with built-in cellular network interface and location sensing capabilities) as measurement sensors and the natural mobility of people carrying them for cost-effective, continual and fine-grained spatio-temporal monitoring of mobile networks. The crowdsourcing approach has several advantages over existing approaches like drive testing (e.g., [11]), coverage modelling (e.g., [12]) and network-side passive analysis (e.g., [13]). It captures reality better than the coverage modelling approach; less expensive than the drive testing approach; and unlike the network-based passive monitoring approach, it allows direct measurement at user side including context. Also measurements with the crowdsourcing approach reflect user perceived mobile performance as they are obtained from real end-user devices. The foregoing discussion suggests that the crowdsourcing approach will likely be an integral part of a broader approach to meet the emerging mobile network measurement and monitoring needs.

However the crowdsourcing approach also presents several challenges some of which have been discussed in [14]; one of the challenges mentioned concerns *device and environment context* and is the focus of this paper: ".. it is not always clear where the device is located when the test is made (so it could be indoors or outdoors, in a bag or in the users hand) ..". Gember et al. [5] have in fact partially characterized the impact of some of the relevant contextual factors, especially device position (phone in hand or not), and show that such factors have a significant impact on measurement results. For example, moving the phone from hand to pocket can cause up to 79% difference in measured throughput.

In this paper, we complement the previous work in [5] by showing that whether a mobile user participating in a crowdsourced measurement system is *indoor or outdoor* at the time of measurement matters significantly. This is an environment related aspect of user context that has not received much attention till date. Not surprisingly, existing mobile crowdsourcing systems (e.g., [1, 2]) lack the indoor-outdoor detection capability.

To study the indoor-outdoor impact, we *focus on signal strength (RSSI)* as the measurement metric driven by at least three reasons. Firstly, coverage or signal strength is not only an intuitive metric for users but also the primary metric targeted by operators and regulators; this is clearly stated by the authors of [15] "Without exaggeration, we can say that coverage is the most important and the highest-priority target that has to be achieved by cellular operators." and is also evident from [12, 16]. Secondly, signal strength is recently shown to correlate well with throughput and mobile device battery energy drain [9, 17, 18]. Finally, unlike other performance metrics like throughput that may require active measurement, signal strength can be measured passively with little or no impact of device battery consumption; for this reason, it is also the most widely supported metric across all existing crowdsourced mobile network measurement systems.

In particular, we highlight the *downside of conflating indoor and outdoor measurements*. Such conflation of measurements from different environments and contexts is quite likely to happen in practice because it is typical to aggregate nearby measurements into coarser geographic units such as grid squares or postcodes. It is easy to find examples from real-world practice and research literature where aggregation of measurements over space is done. Ofcom's mobile coverage analysis [12, 16] aggregates coverage predictions into postcodes and 200mx200m grid squares. WiScape [3] partitions the world into zones each around 0.2 sq. km. each. Online coverage checkers of web sites of almost all mobile operators indicate coverage at the postcode level [19].

Our analysis is based on a large crowdsourced measurement dataset, consisting of nearly 8 million measurements spanning over 3 years, from OpenSignal [1] for central London. As the indoor-outdoor context is not embedded in the dataset, we choose to rely on a GPS based method used previously in [20] to infer whether a measurement was collected while indoors or outdoors.

We show that indoor and outdoor measurements have quite different signal strength characteristics (see Figure 1). As existing crowdsourced measurement systems do not differentiate between indoor and outdoor measurements and spatial aggregation of measurements is common, conflation of diverse measurements can thus lead to erroneous conclusions — coverage outdoors is likely to be underestimated, whereas indoor coverage may be overestimated. Imprecision or uncertainty concerning location of a measurement can compound this problem.

Figure 2 shows a sample result that highlights the risk of conflating indoor and outdoor measurements. We have also validated these observations using a diverse set of controlled measurements with *indoor/outdoor ground truth information*.

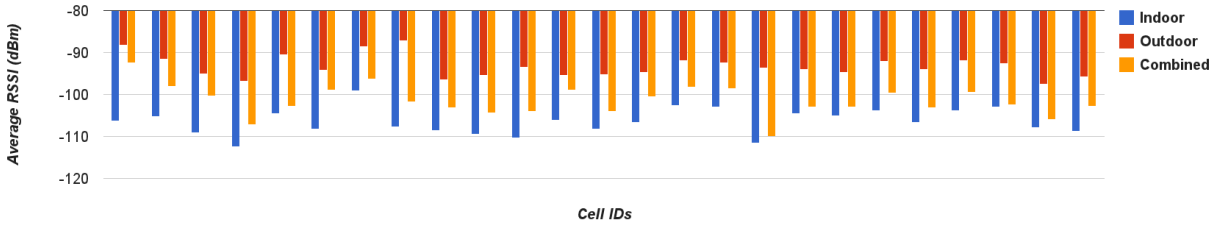


Figure 2: Differences in indoor and outdoor average RSSI values for a selection of 25 cell sectors from the OpenSignal dataset.

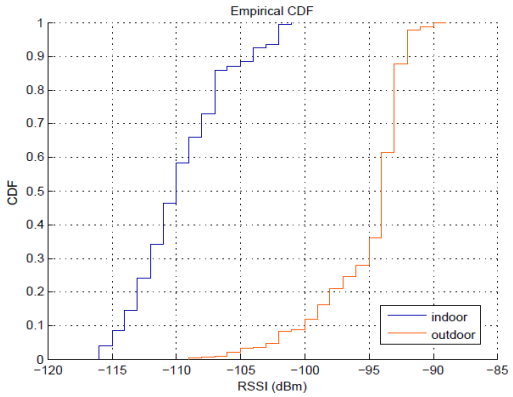


Figure 1: CDF of indoor and outdoor RSSI values for a specific cell sector and operator combination (OpenSignal dataset).

1. REFERENCES

[1] Open Signal Inc. http://opensignal.com/.

[2] MobiPerf. http://www.mobiperf.com/.

[3] S. Sen, J. Yoon, J. Hare, J. Ormont, and S. Banerjee. Can they hear me now?: A Case for a Client-Assisted Approach to Monitoring Wide-Area Wireless Networks. In *Proc. ACM Internet Measurement Conference (IMC)*, 2011.

[4] BBC News. 3G Mobile Data Network Crowd-Sourcing Survey. http://www.bbc.co.uk/news/business-14574816, Aug 2011.

[5] A. Gember, A. Akella, J. Pang, A. Varshavsky, and R. Caceres. Obtaining In-Context Measurements of Cellular Network Performance. In *Proc. ACM Internet Measurement Conference (IMC)*, 2012.

[6] Ookla Speedtest Mobile Apps. http://www.speedtest.net/mobile/.

[7] J. Sommers and P. Barford. Cell vs. WiFi: On the Performance of Metro Area Mobile Connections. In *Proc. ACM Internet Measurement Conference (IMC)*, 2012.

[8] S. Sonntag, J. Manner, and L. Schulte. Netradar - Measuring the Wireless World. In *Proc. 9th International Workshop on Wireless Network Measurements (WiNMee)*, 2013.

[9] A. Nikravesh, D. R. Choffnes, E. Katz-Bassett, Z. M. Mao, and M. Welsh. Mobile Network Performance from User Devices: A Longitudinal, Multidimensional Analysis. In *Proc. Passive and Active Measurement (PAM) Conference*, 2014.

[10] Y. Xu, Z. Wang, W. K. Leong, and B. Leong. An End-to-End Measurement Study of Modern Cellular Data Networks. In *Proc. Passive and Active Measurement (PAM) Conference*, 2014.

[11] C. N. Pitas, A. D. Panagopoulos, and P. Constantinou. Speech and Video Telephony Quality Characterization and Prediction of Live Contemporary Mobile Communication Networks. *Wireless Personal Communications*, 69(1):153174, Mar 2012.

[12] Ofcom. UK Mobile Services Map 2013. http://maps.ofcom.org.uk/mobile-services/.

[13] A. Gerber, J. Pang, O. Spatscheck, and S. Venkataraman. Speed Testing without Speed Tests: Estimating Achievable Download Speed from Passive Measurements. In *Proc. ACM Internet Measurement Conference (IMC)*, 2010.

[14] Ofcom. Measuring Mobile Voice and Data Quality of Experience. http://stakeholders.ofcom.org.uk/consultations/mobile-voice-data-experience/, Jan 2013.

[15] A. Galindo-Serrano, B. Sayrac, S. Ben Jemaa, J. Riihijarvi, and P. Mahonen. Automated Coverage Hole Detection for Cellular Networks Using Radio Environment Maps. In *Proc. 9th International Workshop on Wireless Network Measurements (WiNMee)*, 2013.

[16] Ofcom. Infrastructure Report: 2013 Update. http://stakeholders.ofcom.org.uk/market-data-research/other/telecoms-research/broadband-speeds/infrastructure-report-2013/, Oct 2013.

[17] C. Cheng and P. Hsiu. Extend Your Journey: Considering Signal Strength and Fluctuation in Location-Based Applications. *IEEE/ACM Transactions on Networking*, 22, Feb 2014.

[18] N. Ding, D. Wagner, X. Chen, A. Pathak, Y. Charlie Hu, and A. Rice. Characterizing and Modeling the Impact of Wireless Signal Strength on Smartphone Battery Drain. In *Proc. ACM SIGMETRICS*, 2013.

[19] Mobile Coverage Checkers. http://consumers.ofcom.org.uk/internet/mobile-internet/mobile-coverage-checkers/.

[20] L. Ravindranath, C. Newport, H. Balakrishnan, and S. Madden. Improving Wireless Network Performance Using Sensor Hints. In *Proc. USENIX NSDI*, 2011.

Poster Session 2 :: April 9

Anomaly detection in mobile phone data — Exploratory analysis using Self-Organizing Maps

Veena Mendiratta, Vijay K. Gurbani and Chitra Phadke
Bell Laboratories, Alcatel-Lucent
Email: {veena.mendiratta,vijay.gurbani,chitra.phadke}@alcatel-lucent.com

***Abstract*—Communications traffic on wireless networks generates large amounts of metadata on a continuous basis across the various servers involved in the communication session. The networks are engineered for high reliability and hence, the data from these networks is predominantly normal with a small proportion being anomalous. From an operations perspective however, it is important to detect these anomalies when they occur to correct any vulnerabilities in the network. The objective of our work is anomaly detection in communication networks to improve network performance and reliability. In this paper we explore the use of neural network based Kohonen Self Organizing Maps (SOM) applied to Per Call Measurement Data (PCMD) records from a 4G network for data analysis and anomaly detection.**

I. Introduction

Communications networks have evolved over the years to become large, complex and fault tolerant systems. Various network elements in the system capture and log traffic on a continuous basis. Given the stable nature of the networks, much of this data represents normal operations. However, faults and errors in the network are reflected in data that shows anomalous behavior of the network. The fault tolerant mechanisms in the networks correct most errors and the anomalous data is typically of short duration. However, at times, these errors may be a precursor to a larger failure in the system. It is therefore important to detect such anomalies in a timely fashion to detect vulnerabilities and take corrective measures as necessary.

Anomaly detection algorithms [1], [2] can be employed to detect anomalies in the networks. The need for detecting anomalies in near real time necessitates the application of anomaly detection on streaming data. But anomaly detection for data streams remains a challenging task. Sadik et al enumerate the research issues in anomaly detection for streaming data [3], defining streaming data as an infinite sequence of data points with explicit or implicit timestamps. Since communications network servers are generating data continuously, anomaly detection on data streams would be relevant in this context. Our approach is to first explore anomaly detection algorithms with subsets of our data in batch mode and then extend the work to streaming data. In this paper we focus on techniques for anomaly detection when anomalies are transient and rare. The focus is on unsupervised learning techniques since some small numbers of errors are expected in the system — considered normal — and not every error is deemed an anomaly.

The objective of our work is outlier detection in communication networks to improve network performance and reliability using performance metrics from the network, particularly the Per Call Measurement Data (PCMD) of an 4G LTE system which is a very rich source of data. To this end, in this paper we describe work in progress which includes exploratory data analysis; clustering and outlier detection based on neural network based Kohonen Self Organizing Map (SOM) [4].

The rest of this paper is organized as follows. Section II gives a brief overview of the data that we used in this analysis. Section III presents the analysis on the data using SOM. Related work is presented Section IV and conclusions and future work are in Section V.

II. Dataset and exploratory analysis

PCMD provides call measurement data on a per-procedure basis for the important procedure interactions that a user device (UE) has with the mobile network for voice, data and SMS sessions in a 4G network. Examples of common procedures triggered by the UE are: requests to register with the network; service requests that trigger activation of radio bearers; handovers to handle mobility and requests for release of a session. Each PCMD record captures key data fields such as service type, session length, setup latency, signal quality, data throughput, data related to handovers, the session result code, additional levels of detail on the result code, the sequence of intermediate procedure steps also known as Procedure Markers (PM) required for the session, etc., thereby providing a view into network performance. A single record is typically comprised of a large number (hundreds) of populated fields.

For our exploratory analysis we look at the PCMD records generated at a single server from a network that experienced a failure. The time period of the data collection spans a few minutes and includes: a stable period, the occurrence of the failure event followed by a gradual recovery period, and back to a stable state. The dominant procedures (as shown in Figure 1) were Procedure 6 (release), Procedure 10 (paging) and Procedure 11 (service). Due to lack of space and for illustrative purposes we will limit our discussion in this paper to Procedure 6.

Our dataset consists of the primary and secondary fields of PCMD records, about 250 fields in total. Guided by PCMD domain experts, we chose the following six fields for the analysis and modeling of anomaly detection :Duration of Procedure; Connect Code (indicating final success/failure of the procedure); Qualifier Primary and Qualifier Secondary, which provide further details on the resultant connect codes; and the Sequence Index. The selection was based on the usefulness of the information in the field for determining

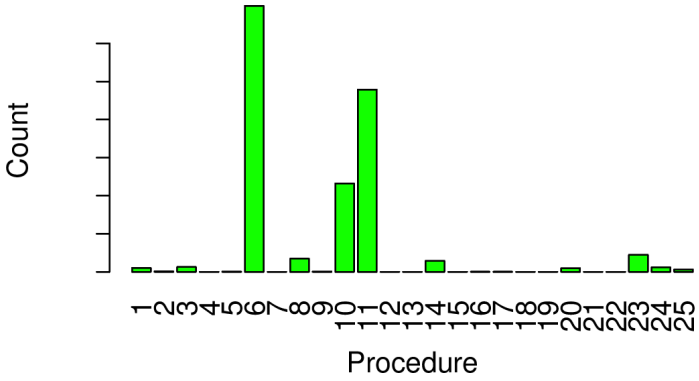


Figure 1: Number of observations by Procedure Type

the cause of the anomalous behavior. In addition, the timing of the record as the minute relative to the start of the data collection record was retained. This allows for differentiation of records between when the network was in a stable state versus when the network was in a failure or recovery state. The field Sequence Index deserves some explanation since by itself it does not occur in a PCMD record, rather it is derived based on artifacts in the record.

Among other fields, each PCMD record contains up to 20 distinct Procedure Markers (PM) which are represented as integers corresponding to the codes of intermediate events that occur during the execution of a procedure. Based on prior field experience with addressing failures, it is known that certain PM sequences are likely indicators of anomalous conditions in the network. Chandola et al. [5] indicates that often anomalies can be found in repeating patterns (or discrete sequences) whereby anomalies may be hidden in a pattern of sequences whose frequency of occurrence is anomalous. Therefore, the PMs in a record were concatenated to produce a sequence. Each such unique PM sequence was indexed and a frequency distribution was produced for each unique PM sequence. Procedure 6 was found to have about 100 unique sequences though only a very small number of sequences make up the majority of records.

III. Analysis of Results

We investigated various clustering algorithms for unsupervised learning. Given the objective of anomaly detection but still being in the exploratory phase, we chose to work with the neural network based Kohonen Self Organizing Map (SOM) [4]. It is particularly suited to discovering input values that are novel and for the visualization of otherwise difficult to interpret data. A key characteristic of the SOM is its topology preserving ability to map a multi-dimensional input into a two dimensional form.

Figure 2 shows the results of applying the SOM algorithm (using the R Kohonen package [6]) to our dataset for Procedure 6 using the 6 features described above. The figure shows the mapping of observations to the 10x15 grid where each node in the grid is created such that the distance to the neighbor node is minimized. The graphs in the figure display the heatmaps for the 6 selected variables. Based on the characteristics of the data, certain nodes may have no observations mapped to them — in our case about 30% of the nodes contain no observations. Given the large number of observations and based on the data distribution, we chose a grid with enough nodes (10x15) such that each node is distinct enough and has an optimal number of observations in it for clustering.

Next, the SOM clustering algorithm was applied to the grid of nodes. Based on the within-cluster sum of squares (WCSS) metric for k-means for different clustering sizes, between 8 to 12 clusters showed good results. For the case of 8 clusters, Figure 3 shows the cluster boundaries on the grid as well as the average value of each variable by node. The results are interesting for anomaly detection — there is one large cluster, as expected, and 7 small clusters of 3 nodes or 1 node each, of which 2 clusters contain no observations. This gives 5 clusters to analyze and to study the characteristics of the observations to determine the causes of the anomalous behavior. The sum of the observations in these 5 clusters represents less than 1% of the total observations and range between 19 and 1691 observations (412, 19, 1175, 98, 1059, 1691, 292) per cluster. However, for real-world large-scale streaming data applications of anomaly detection it is not practical, nor timely, to visually detect outliers. To address this issue, future work will investigate techniques for the automatic extraction of rules from SOMs [7].

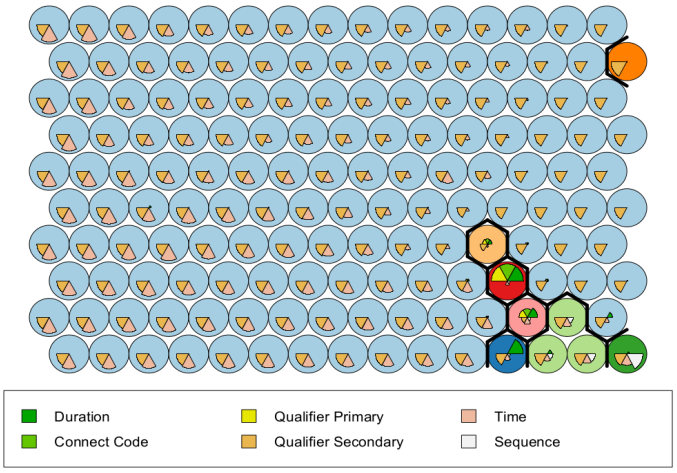


Figure 3: Procedure ID 6 with 8 clusters

IV. Related Work

Anomaly detection is a mature area of research. Chandola et al. [2] provide a comprehensive taxonomy of anomaly detection techniques across varying application domains. PCMD, a rich data source for network state information has been used for different objectives such as estimation of subscriber locations [8], using that information further to improve the efficiency of paging mechanisms [9] and for load balancing [10]. Vaidyanathan [11] used PCMD to analyze the effects of caller traffic movement in response to emergency and non-emergency events (rugby match).

Our work, by contrast, is focused on understanding the availability and reliability of the network by detection of anomalies using unsupervised outlier detection learners [12] and clustering techniques.

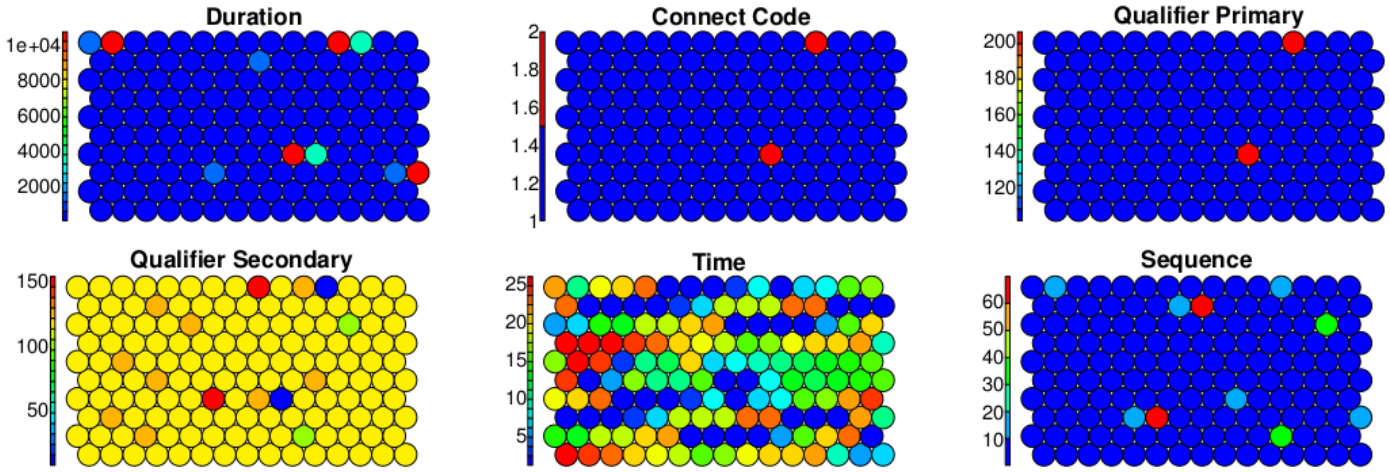


Figure 2: SOM applied to Procedure 6

V. Conclusion and Future Work

In this paper we have applied self-organizing maps for initial exploratory analysis and as a clustering and anomaly detection tool. One of the advantages of SOM is that it enables one to visualize a multi-dimensional data in a two dimensional grid form to understand the data distribution. A clustering on the SOM nodes enables one to quickly analyze those nodes which are minority clusters. Anomalous SOM nodes detected via this technique can then be further investigated to derive rules on the underlying data. The large volume of data generate by communications network precludes the store-and-process paradigm and streaming analytics and anomaly detection methodologies for data streams need to be applied in practice. We plan to also exploit the nature of the sequences of the Procedure Markers in a PCMD record to get more details into the interworkings in the network to perform root cause analysis and flag anomalies. These will be the focus of future work in this area.

References

[1] A. Patcha and J.-M. Park, "An overview of anomaly detection techniques: Existing solutions and latest technological trends," *Computer Networks*, vol. 51, no. 12, pp. 3448—3470, August 2007.

[2] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 15:1–15:58, Jul. 2009. [Online]. Available: http://doi.acm.org/10.1145/1541880.1541882

[3] S. Sadik and L. Gruenwald, "Research issues in outlier detection for data streams," *ACM SIGKDD Exploration Newsletter*, vol. 15, no. 1, pp. 33—40, June 2013.

[4] R. Wehrens and L. M. Buydens, "Self- and super-organizing maps in r: The kohonen package," *Journal of Statistical Software*, vol. 21, no. 5, pp. 1—19, October 2007.

[5] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection for discrete sequences: A survey," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 24, no. 5, pp. 823–839, May 2012.

[6] (2014, December) R Package — Kohonen. [Online]. Available: http://cran.r-project.org/web/packages/kohonen/kohonen.pdf

[7] J. Malone, K. McGarry, S. Mermter, and C. Bowerman, "Data mining using rule extraction from Kohonen self- organizing maps," *Neural Computing and Applications*, vol. 15, no. 1, pp. 9—17, August 2005.

[8] M. J. Flanagan, L. M. Drabeck, L. A. Cohen, A. H. Diaz, and J. Srinivasan, "Wireless Network Analysis using Per-Call Measurement Data," *Bell Labs Technical Journal*, vol. 11, no. 4, pp. 307—313, 2007.

[9] H. Zang and J. C. Boloy, "Mining call and mobility data to improve paging efficiency in cellular networks," in *Proceedings of the 13th Annual ACM International Conference on Mobile Computing and Networking*. ACM, 2007, pp. 123—134.

[10] A. Vaidyanathan, W. Wong, M. Billinghurst, and H. Sirisena, "Understanding directional load balancing using per call measurement data," in *Performance Evaluation of Computer Telecommunication Systems, 2009. SPECTS 2009. International Symposium on*, vol. 41, July 2009, pp. 213–220.

[11] A. Vaidyanathan, "Manikarnika: Proactive crowd-sourcing for location services," Ph.D. dissertation, University of Canterbury, New Zealand, 2010.

[12] A. Zimek, R. J. Campello, and J. Sander, "Ensembles for unsupervised outlier detection: Challenges and research questions," *ACM SIGKDD Exploration Newsletter*, vol. 15, no. 1, pp. 11—22, June 2013.

Mobile Phone Data as a Means of Studying Activity Space Segregation at Scale

Robert Manduca, Bradley Sturt, Marta González

Few topics in sociology are as well studied as segregation. Hundreds of papers over the past several decades have sought to determine both the causes and effects of the pronounced segregation that still exists today across the United States. However, the vast majority of these studies have focused on residential segregation, despite the fact that most of the impacts of segregation are felt through the environments to which people are exposed during the day—their "activity spaces" (Kwan, 2013).

In this paper, we utilize cell phone call detail record (CDR) data from metropolitan Boston to examine segregation in daytime activity spaces. We articulate a definition of activity space that is operable on CDR data, then measure social distances between census tracts based on the amount of overlap between the activity spaces of their residents. We find notable clusters of adjacent, strongly connected tracts, particularly in outlying parts of the metropolitan area.

Residential segregation has attracted enormous amounts of scholarly attention over the past fifty years as both an outcome of interest and a cause of disparities in life outcomes (Bruch & Mare, 2006; Massey & Denton, 1993; Sampson, 2012; Schelling, 1971; Wilson, 2012). In recent years, however, an increasing number of researchers in sociology, geography, and public health have noted that many of the processes thought to be affected by segregation occur primarily during the day, when most people are not at home. These researchers have begun to move beyond simple residence to a more complete conception of activity space in their work on the impact of environmental factors on social, economic, and health outcomes (Crowder & South, 2011; Inagami, Cohen, & Finch, 2007; Kwan, 2012, 2013; Matthews, 2011). Unfortunately, thus far there have been few studies of segregation in activity spaces with a large enough sample to draw conclusions at scales smaller than the overall metropolitan area. Previous studies in activity space segregation have used data from travel surveys (Wang, Li, & Chai, 2012; Wong & Shaw, 2011), cell phone GPS traces (Palmer, 2013), and census data (Ellis, Wright, & Parks, 2004), all of which are limited by either a sample of not more than a few thousand individuals or a limited definition of activity space centered only on home and work.

CDR data offers the possibility of collection rich activity space information for samples of hundreds of thousands of people, and we exploit this potential in the present study. We begin by extracting mobility patterns from CDRs in metropolitan Boston over two months in spring 2010 in the manner described by Alexander et al. (2014). For each individual in the dataset, this method provides an estimated home location, work location, and other 'stay points' from which they made multiple calls during the study period. We define an individual's activity space from their set of stay points, selecting those points that are visited more than a threshold number of times (the threshold is not theoretically determined, and we investigate several values of it, ranging from only including home and work to including the full set of stay points). These activity spaces encompass substantially more than individuals' tracts of residence, as indicated by Figure 1.

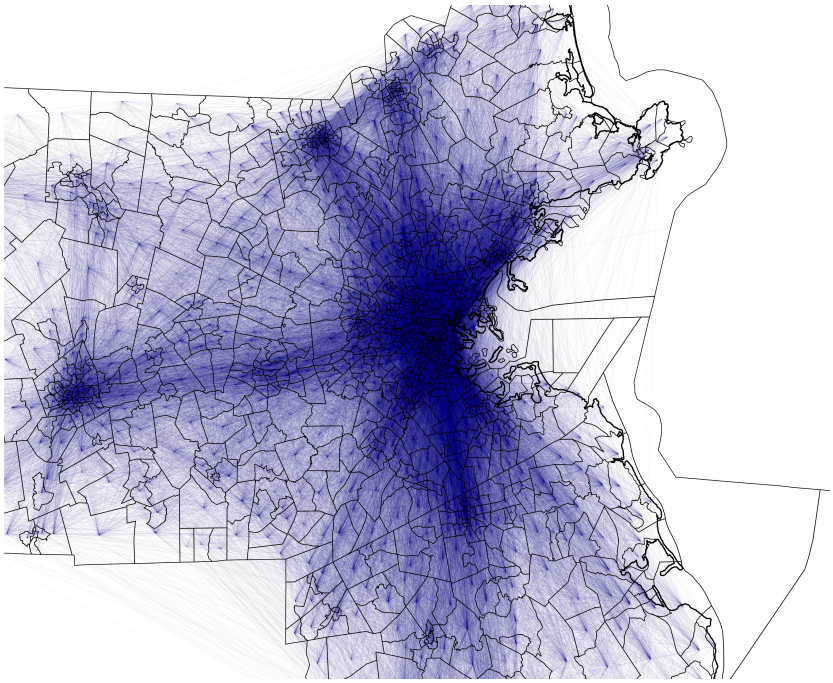


Figure 1: Daily movement patterns in metropolitan Boston

With activity spaces calculated for each individual, we examine the amount of overlap in the activity spaces of census tracts. We calculate a social distance metric for census tracts based on the percentage of their residents' activity spaces that overlap with each other. There is some relation between this social distance and physical distance, but the relationship is far from one-to-one (see Figure 2). The strongest ties are found between tracts in outlying cities such as Worcester, Lowell, and Lawrence. Within central Boston the ties are weaker and physical adjacency seems to be less correlated with activity space overlap. Analysis at this level of geographical detail is made feasible by the large size of the CDR sample and has not be possible in previous work on activity space segregation.

We also examine the social characteristics of connected tracts. Tracts with substantial activity space overlap do appear to have greater similarity in social characteristics, although the relationship is not one to one (see Figure 3). Based on activity spaces we calculate the expected exposure to demographically dissimilar populations for residents of each census tract.

Further work to be presented at the conference will probe more deeply the relationship between physical proximity and activity space overlap, investigate the similarity of overlapping tracts on a number of demographic characteristics, and attempt to locate boundaries in the Boston metropolitan area that define groups of tracts with substantial overlap from one another. Each of these analyses will be novel to the activity space segregation literature because they will delve beyond a simple metropolitan segregation index into the specific geographic patterns that create daytime segregation.

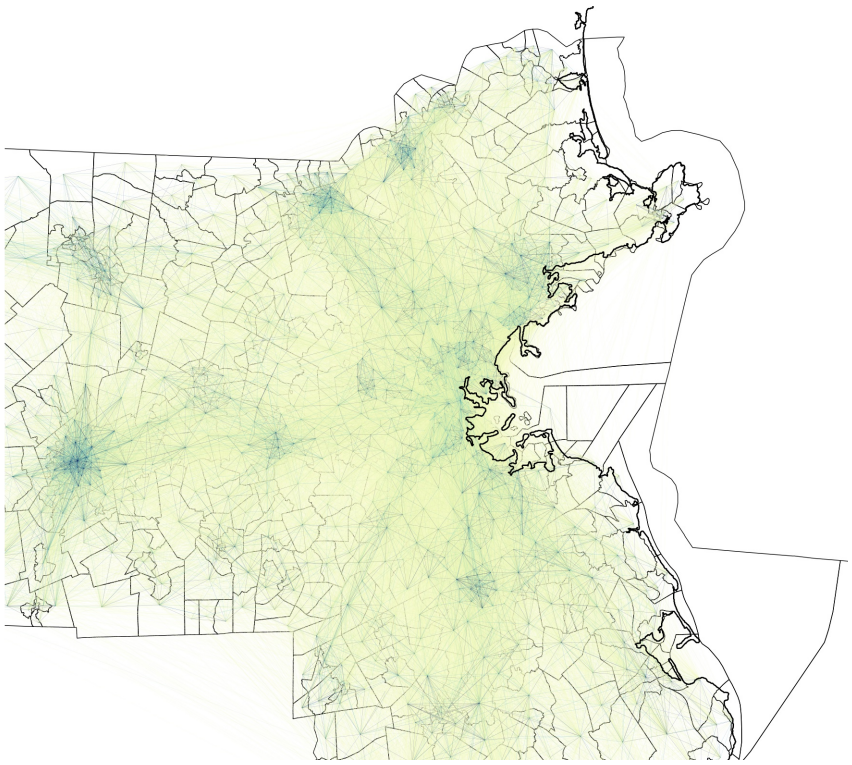


Figure 2: Activity space overlap between census tracts (blue indicates higher levels of overlap)

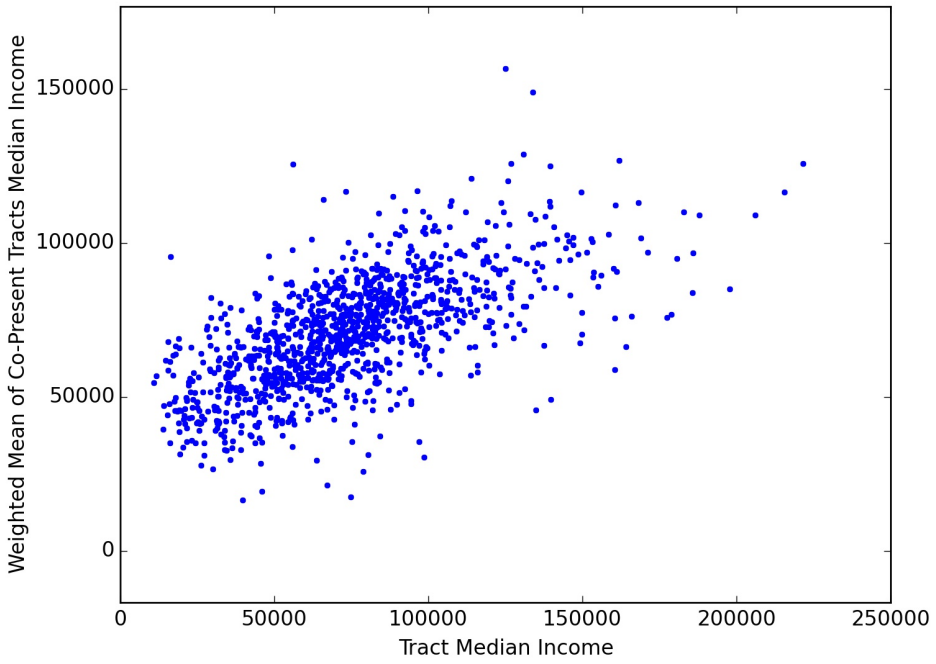


Figure 3: Tract median income compared to median income of tracts with overlapping activity spaces

References

Alexander, L., Jiang, S., Murga, M., & González, M. C. (2014). *Validation of origin-destination trips by purpose and time of day inferred from mobile phone data*. Retrieved from http://humnetlab.mit.edu/wordpress/wp-content/uploads/2010/10/TRC_BigData_612014.pdf

Bruch, E. E., & Mare, R. D. (2006). Neighborhood Choice and Neighborhood Change. *American Journal of Sociology*, *112*(3), 667–709. doi:10.1086/507850

Crowder, K., & South, S. J. (2011). Spatial and temporal dimensions of neighborhood effects on high school graduation. *Social Science Research*, *40*(1), 87–106. doi:10.1016/j.ssresearch.2010.04.013

Ellis, M., Wright, R., & Parks, V. (2004). Work Together, Live Apart? Geographies of Racial and Ethnic Segregation at Home and at Work. *Annals of the Association of American Geographers*, *94*(3), 620–637. doi:10.1111/j.1467-8306.2004.00417.x

Inagami, S., Cohen, D. A., & Finch, B. K. (2007). Non-residential neighborhood exposures suppress neighborhood effects on self-rated health. *Social Science & Medicine*, *65*(8), 1779–1791. doi:10.1016/j.socscimed.2007.05.051

Kwan, M.-P. (2012). The Uncertain Geographic Context Problem. *Annals of the Association of American Geographers*, *102*(5), 958–968. doi:10.1080/00045608.2012.687349

Kwan, M.-P. (2013). Beyond Space (As We Knew It): Toward Temporally Integrated Geographies of Segregation, Health, and Accessibility: Space–Time Integration in Geography and GIScience. *Annals of the Association of American Geographers*, *103*(5), 1078–1086.

Massey, D. S., & Denton, N. A. (1993). *American apartheid: Segregation and the making of the underclass*. Harvard University Press. Retrieved from http://books.google.com.ezp-prod1.hul.harvard.edu/books?hl=en&lr=&id=uGslMsIBNBsC&oi=fnd&pg=PR8&dq=american+apartheid&ots=I55Db-I6X8&sig=GLwpw7AwrfPCyUn7Kv7rH1JvM2k

Matthews, S. A. (2011). Spatial polygamy and the heterogeneity of place: studying people and place via egocentric methods. In *Communities, Neighborhoods, and Health* (pp. 35–55). Springer. Retrieved from http://link.springer.com.ezp-prod1.hul.harvard.edu/chapter/10.1007/978-1-4419-7482-2_3

Palmer, J. R. (2013). *Activity-space segregation: Understanding social divisions in space and time*. Princeton University, Princeton, NJ.

Sampson, R. J. (2012). *Great American city: Chicago and the enduring neighborhood effect*. University of Chicago Press. Retrieved from http://books.google.com.ezp-prod1.hul.harvard.edu/books?hl=en&lr=&id=POs5iroB7PsC&oi=fnd&pg=PR5&dq=great+american+city&ots=GjCMmtFemg&sig=i3660i6vg1HmHxIwnKI6nQZhyzw

Schelling, T. C. (1971). Dynamic models of segregation†. *Journal of Mathematical Sociology*, *1*(2), 143–186.

Wang, D., Li, F., & Chai, Y. (2012). Activity spaces and sociospatial segregation in Beijing. *Urban Geography*, *33*(2), 256–277.

Wilson, W. J. (2012). *The truly disadvantaged: The inner city, the underclass, and public policy*. University of Chicago Press. Retrieved from http://books.google.com.ezp-prod1.hul.harvard.edu/books?hl=en&lr=&id=N_tVVPXLfMIC&oi=fnd&pg=PR5&dq=truly+disadvantaged&ots=gGjfzmR6h8&sig=L8_z6fxGDGB-sfSOWHuXCzjnpnk

Wong, D. W. S., & Shaw, S.-L. (2011). Measuring segregation: an activity space approach. *Journal of Geographical Systems*, *13*(2), 127–145. doi:10.1007/s10109-010-0112-x

Gender-based Characterization of Communication Bel

Riyadh Alnasser*, Faisal Aleissa*, Fahad Alhasoun†, Abdullah Almaatouq†, Anas Alfaris†, Marta González†
* King Abdulaziz City for Science and Technology KACST
{ralnasser, fsaleissa}@kacst.edu.sa
† Massachusetts Institute of Technology MIT
{fha,amaatouq,anas,martag}@mit.edu

The pervasiveness of cell phones opens new horizons for understanding and studying human behavior. Mobile phones are powerful sensors, and they are quickly becoming the main source for social and behavioral data. In this paper we present a study aimed at identifying the gender of cell phone users based on their visited locations then investigate variations of usage behavior given predicted gender information. The analysis was carried out using CDRs (Call Detail Records) on 10,000 cell phone users. Our contributions can be summarized in (1) understanding the communication characteristics pertaining to gender; and (2) Identifying the difference in mobility habits between males and females. The results highlight the behavioral differences between males and females in based on their mobile phone usage.

I. Introduction

The rise of smart phones, apps, and mobile Internet makes the cell phone a key battleground in the fight for customer attention by the telecom companies. Basic demographic attributes, including gender, may be registered by the companies, where they may use this information in user profiling and marketing. However, it is common that the registered information doesn't represent the actual users of the service. Call Detail Records (CDRs) provide a convenient data source that can be used to infer different attributes about the mobile phone users [1], [3]–[5], [7], [8]. Saudi Arabia is highly gender segregated society. Acts of gender segregation manifest in many different forms; for example, separate educational facilities for men and women only, female-only shopping malls or shopping malls with access restricted to families only. Also, many entertainment or leisure facilities are built specifically for either gender. For example, attending football matches at various stadiums in the Kingdom is limited to men only.

II. Data Description

The dataset consists of one full month of records for an entire country, with 3 billion mobile activities to over 10 thousands unique cell towers, provided by a single telecom service provider.

In order to study the influence of gender in communication behavior, we extracted a sample ground truth sample (i.e., labels of female and male), by exploiting the gender segregation in Saudi Arabia. Specifically, since the CDRs used in this study has no gender information for the cell phone users, we have developed a methodology for gender identification. We were able to identify the gender of 10,000 cell phone users (5000 female students and 5000 male) students. The first step in the identification process was to locate Point of Interests (POIs) that restrict access based on gender. After locating these POIs, we extracted the cell towers situated around each POI that provide radio coverage over that area. Then, we distinguished the people in stay versus people in move (i.e., who is passing by and who was actually there). Finally, we identify the gender of the users based on the number of distinct gender specific locations visited.

III. Communication Pattern

In this section we attempt to understand the calling behavior of females and males using the CDRs. We focused on examining the 5,000 male and 5,000 female students activities during different days of the week and the month. We denote the normalized activity levels of females and males as $VFemale$ and $VMale$, respectively.

A. Monthly scale

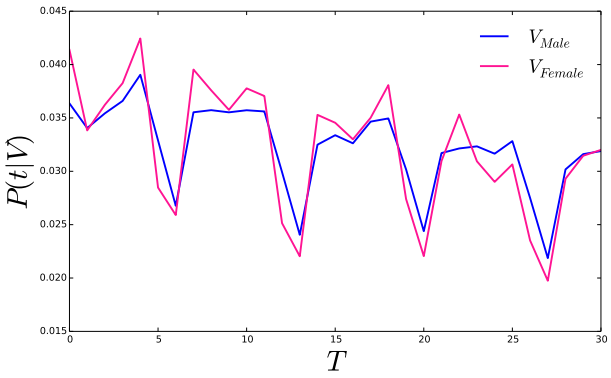


Figure 1: Daily Male and Female Activities across the month

Fig 1 depicts the female and male activities across the entire month, where each point represents the total number of activities in a day. Interestingly, we observe a drop in the activities for both females and males during weekends. Female activities during the weekend shows a larger decrease; this might indicate that women tend to use cell phones for professional activities or work related tasks. On the other hand, males might favor social and personal calls. We also observe a slight increase in the mobile phone activities during the first and the last day of the weekdays, and it appears more regularly for females than males. In addition, as the month progresses we observe a decrease in the cell phone activities. Further investigation shows that the last two weeks coincided with the final exams period.

B. Daily scale

In order to have a better look at the daily activity, we decompose the daily mobile phone activities into 288 five minutes interval. This gives us a better understanding of how the activities change within the day. Fig 2 shows the average calling activities of females and males in 5 minutes interval. On the left figure, we observe a peak in female activities at 6:30 am and 2:30 pm. Our hypothesis is that since women do not drive in Saudi Arabia, these peaks correspond to time at which they call the drivers to come and drive them to the university, or in case of 2:30 pm, drive them back home. To further investigate this, on the right figure we can see females 5-minute interval activities during weekends. As expected, these two peaks disappear on weekends, thus these sharp increases in the activities during a weekday are females calling their drivers to get them around. The influence of the daily prayers can be observed in females and males daily activities. Fig 2 shows significant drops occurring four times across the day. Previous analysis has showed that the drops coincide with daily prayer time [2]. An interesting observation is that women tend to not pray Duhr (12pm) on time during the weekdays; instead, we observe a slight increase in cell phone activities at this time. An interpretation for this observation could be that women

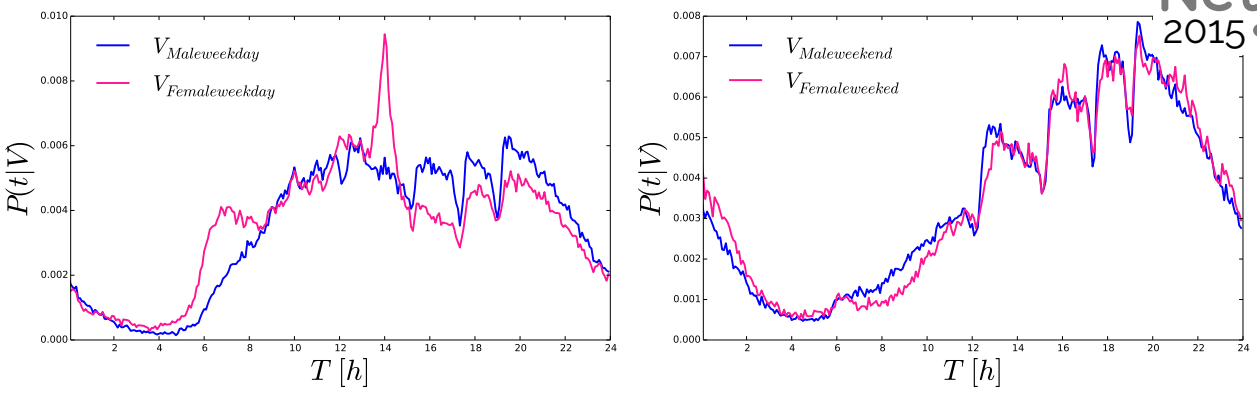


Figure 2: 5 min interval activities of males and females in 24 hours window

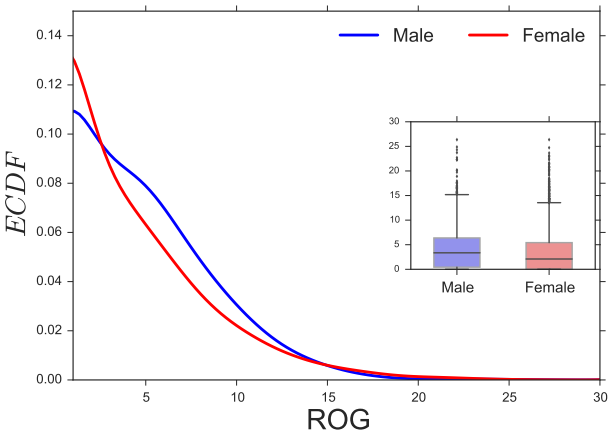


Figure 3: Radius of Gyration (ROG) for Males and Females

postpone this prayer until they get home. However during weekends, women pray Duhr on time. Also, the mobile phone activities for females and males during weekends shows that they have similar calling behavior. During weekends, males tend to be less active in the first half of the day, and generate more mobile phone activities during the second half of the day. The opposite pattern is observed during weekdays. Females show the same pattern in the change of calling behavior in weekdays versus weekends. We have observed that in fact, as it may be the cultural belief, women spend more time on the phone than men. However, females tend to make less number of phone calls in the second half of the day. We found that females tend to make fewer, but longer phone calls. We calculated the Empirical Cumulative Distribution Function (ECDF) of the calls duration made by females and males. We find that nearly 50% of our sample of females had calls duration greater than 40,000 seconds as opposed to their male counterparts where 50% of them had calls duration greater than 20,000 seconds.

IV. Mobility Discrepancy

In this section, we attempt to examine the discrepancies of mobility patterns between the female and male population. Namely, we quantify mobility by measuring the Radius of Gyration *ROG* (i.e., a measure of how far from the center of mass the mass is), as implemented by the CDRs analysis package Bandicoot [3]. The Radius of Gyration has been used previously in the literature to quantify individual trajectory tracked from CDRs, and researchers have shown that it has a strong impact on travel distance distributions over all users [5], [6]. Figure 3, shows the Kernel Density Estimation (KDE) of the *ROG*. Although the difference in the size of the *ROG* between the male and female population is subtle ($\mu(male) = 4.056674$ and $\mu(female) = 3.366997$), using the two-sample Kolmogorov-Smirnov (*KS*) test, we find the difference to be statistically significant ($KS = 0.15, p < 0.001$) .

V. Conclusion

In this work, we show that in some cases it is possible to infer the gender of mobile phone users through the analysis of CDRs coupled with POIs. The main purpose of this work is to investigate the existence of differences in usage patterns in the CDRs given the gender of a user. This study suggests that females and males use their cell phones to communicate in a different manner and shows that there exists a considerable variation in terms of phone usage given the gender of a user. We acknowledge that some of the explanations and conclusions proposed in this work might lack rigorous validations and this is due to the nature of the CDRs where it lacks sufficient granularity in space and time. We also lack the ground truth validation for our proposed method of labeling. However, we believe that our analysis can describe well the trends and discrepancies of gender specific communication and can be leveraged for several applications. Future work will involve the improvement of the statistical analysis and the investigation of additional factors that influences usage of mobile phones such as the social structure of females and males users. In addition, we plan to use such extracted features in training a classifier that can predict gender using CDRs.

References

[1] F. Alhasoun, A. Almaatouq, K. Greco, R. Campari, A. Alfaris, and C. Ratti. The city browser: Utilizing massive call data to infer city mobility dynamics. In *The 3rd International Workshop on Urban Computing (UrbComp 2014)*, 2014.

[2] A. Almaatouq, F. Alhasoun, R. Campari, and A. Alfaris. The influence of social norms on synchronous versus asynchronous communication technologies. In *Proceedings of the 1st ACM International Workshop on Personal Data Meets Distributed Multimedia*, PDM '13, New York, NY, USA, 2013. ACM.

[3] Y.-A. de Montjoye, J. Quoidbach, F. Robic, and A. S. Pentland. Predicting personality using novel mobile phone-based metrics. In *Social Computing, Behavioral-Cultural Modeling and Prediction*. Springer, 2013.

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

[5] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. 2008.

[6] 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. IEEE, 2013.

[7] S. Jiang, J. F. Jr., and M. C. Gonzalez. Discovering urban spatial-temporal structure from human activity patterns. pages 95–102, 2012.

[8] J. R. Kwapisz, G. M. Weiss, and S. A. Moore. 12(2).

DYNAMICS OF SOCIAL AND SPATIAL SEGREGATION USING MOBILE PHONE METADATA

Johannes Bjelland, Bjørn-Atle Reme, Pål Sundsøy

Telenor Group Research, Norway

{ johannes.bjelland,bjorn-atle.reme,pal-roe.sundsoy, }@telenor.com

Abstract

In this paper we study segregation between several ethnic groups in a European city using anonymized mobile phone metadata collected from the major mobile phone operator. In particular we look at social segregation (patterns of communication), spatial segregation (patterns of movement) and how these relate to each other.

Over the last decades there has been an increased focus on the lack of social integration of immigrants in western societies. A policy tool often considered to mitigate this problem is resettlement programs enforcing spatial integration. This policy relies on implicit assumptions regarding the relationship and causality between social and spatial segregation. Even though this paper is not able to answer questions about the causal links between social and spatial segregation, we present evidence on the empirical relationship between these variables for different ethnic groups. We believe that a better understanding of the relationship between these variables, and the mechanisms involved, is crucial to developing more effective integration policies.

Increasing the social integration between ethnic groups is often an important objective for policy makers. From the call data we obtain a proxy for ethnicity. This allows us to shed new light on social integration by providing an in-depth analysis of the communication patterns of different ethnic groups in the city. In particular we derive the extent of across-group and within-group communication for the different ethnic groups in the city. Using Hofstede's cultural dimension theory we derive a measure of cultural distance between the different ethnicities in the sample and find how much of the social integrations and across-group communication can be explained by this measure.

When studying the relationship between spatial and social integration we make use of a "benchmark integration level" inspired by Blumenstock and Fratamico (2013). The benchmark is defined by what the structure of the social network would be assuming random pairing of nodes within a geographical area. We then compare this benchmark with the structure of the actual social networks. Based on this analysis we develop the "spatial-social-integration-matrix" which provides an overall picture of the extent to which the different groups are more/less integrated than what random pairing suggests.

References

Blumenstock, JE and Fratamico L (2013). *Social and Spatial Ethnic Segregation: A Framework for Analyzing Segregation With Large-Scale Spatial Network Data*, The 4th Annual ACM Symposium on Computing for Development (DEV '13).

Behavioral Modeling for Churn Prediction: Early Indicators and Accurate Predictors of Custom Defection and Loyalty

Muhammad Raza Khan[1], Joshua Manoj[1], Anikate Singh[1], Joshua Blumenstock[1]
[1]Information School, University of Washington, Seattle, WA, USA

Churn prediction, or the task of identifying customers who are likely to discontinue use of a service, is an important and lucrative concern of firms in many different industries. As these firms collect an increasing amount of large-scale, heterogeneous data on the characteristics and behaviors of customers, new methods become possible for predicting churn. In this paper, we present a unified analytic framework for detecting the early warning signs of churn, and assigning a "Churn Score" to each customer that indicates the likelihood that the particular individual will churn within a predefined amount of time. This framework employs a brute force approach to feature engineering, then winnows the set of relevant attributes via feature selection, before feeding the final feature-set into a suite of supervised learning algorithms. Using several terabytes of data from a large mobile phone network, our method identifies several intuitive - and a few surprising - early warning signs of churn, and our best model predicts whether a subscriber will churn with 89.4% accuracy.

This paper describes a data driven quantitative and computational framework that can be used to find leading indicators of churn, and identify individuals with a high likelihood of churning. By using simple machine learning algorithms to mine historical transaction records, this method discovers behavioral patterns that are empirically correlated with the propensity to churn. This is in contrast to the more traditional approach taken by many companies, where churn strategies are determined by the intuition of key individuals, or through direct feedback from customers.

The framework has two primary components. The first is designed to identify early warning signs of churn by isolating specific and easily-measured behavioral patterns that are highly correlated with churn. One such pattern we find, which is indeed highly correlated with churn, is the total amount of observed activity of a customer. It should come as no surprise that customers with low levels of activity are more likely to churn than customers with high levels of activity. However, this particular metric is one among thousands of other metrics that are correlated with churn, and the aim of the framework is to zero in on the most predictive metrics. The method we develop is a semi-supervised, brute force approach to feature engineering, in which our algorithm first constructs tens of thousands of features through combinatoric feature generation, then uses established techniques for feature selection to prune the long list down to the most predictive behavioral traits.

We test and calibrate this framework on a large dataset from a mobile phone operator in South Asia. Starting with a raw dataset of several billion transactions, spanning roughly ten million prepaid mobile phone subscribers over a period of multiple years, we extract a calibration dataset consisting of all network-based communication for roughly 100,000 subscribers over 6 months. On this dataset, where the natural churn rate during our evaluation period is roughly 24 percent, our method is able to predict customer churn with just under 90 percent accuracy.

Understanding why customers terminate relationships has been a focus of marketing research for several decades (cf., Jain & Singh, 2002). In recent years, as data on customer activities and characteristics becomes increasingly available to companies, more sophisticated metrics have evolved to describe customer behavior (cf., Gupta & Zeithaml, 2006). Churn prediction has received recent attention from the applied machine learning community. These approaches have tested a battery of models including expert systems (Wei, Chiu, & others, 2002), support vector machines (Archaux, Martin, & Khenchaf, 2004), and bagging and boosting (Lemmens & Croux, 2006), to name just a few. They further vary in terms of the approach to the data, with some focusing on customer profiles and features (Qian, Jiang, & Tsui, 2006), and others concerned primarily with the importance of social ties and social structure (Dasgupta et al., 2008; Bonchi, Castillo, Gionis, & Jaimes, 2011; Karnstedt, Rowe, Chan, Alani, & Hayes, 2011). Zhang, Zhu, Xu, and Wan (2012) provide a recent overview of the different types of subscriber attributes used to model and predict customer churn in prior work, and Verbeke, Dejaeger, Martens, Hur, and Baesens (2012) benchmark several classification techniques for prediction. Neslin, Gupta, Kamakura, Lu, and Mason (2006) discuss the importance of different methods for predicting churn in the context of a public tournament between 33 different competitors. In these and related studies, the behavioral traits are pre-computed most of the times (cf., Neslin et al. 2006). By contrast, we focus on the process of generating these predictor variables from the raw transactional records.

Furthermore, majority of studies of churn in telecommunications focus on the post-paid network while our method deals with the pre-paid users.

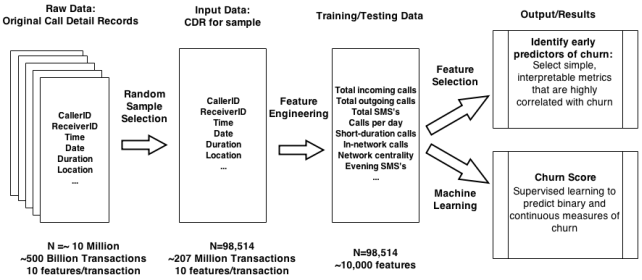


Figure 1 Overview of Approach

Figure 1 shows the overview of our approach. We begin by randomly selecting a subsample of roughly 100,000 subscribers from the full mobile phone subscriber base, and extract all transactions in which they are involved. For this subset of subscribers, we then generate a large number of aggregated metrics that describe a wide range of inferred behavioral characteristics (Feature Engineering). With this matrix, we then separately isolate the handful of metrics that are most predictive of churn (Feature Selection), and develop a Churn Score that indicates the likelihood that a subscriber will churn (Machine Learning)

After generating an extensive list of behavioral features using combinatoric approach we employ standard methods of feature selection that determine which features are most correlated with customer churn. In addition to using statistical significance tests to evaluate the individual predictive performance of each feature we also use a tree-based method for feature selection that allows us to estimate the conditional ability of each additional feature to improve the overall accuracy of a joint classifier (Geurts, Ernst, & Wehenkel, 2006; Hastie et al., 2009).

We quantify churn in two ways: first, as a binary condition that is true if the subscriber is completely inactive in the testing phase. In total, 26 % of our subscribers fit this stringent definition of churn. Second, we define a more flexible version of churn as the percentage of days on which no activity is observed. The list of top predictive features used in our final predictive model is shown in Table 1. Unsurprisingly, we find that "Percent of inactive days" – a feature indicating the fraction of days during the training period when the subscriber had zero transactions – is highly predictive of future churn. The next highest ranked feature is perhaps less obvious: we find that high variance in call activity (measured as the maximum month-to-month change in the ratio of incoming to outgoing calls), is strongly predictive of churn.

While the set of features listed in Table 1 are all unconditionally highly correlated with churn, these features are also correlated with one another. Thus, while the "Percent of inactive days" feature may be the best single linear discriminant between churners and non-churners, it is not necessarily the case that the ensemble of 10 features in Panel A will be the best joint predictor of churn. Thus, in Panel B of Table 1, we provide the 10 features that are, taken together, the best joint predictors of subscriber churn. For comparison with Panel A, we also list the R^2 from the unconditional (univariate) regression for each of the features, though it is important to note that this is the unconditional R^2 and is different from the criteria used to rank-order features in Panel B. Two patterns can be seen in Table 1. First, the unconditionally predictive features (Panel A) generally reflect aggregate metrics of activity. However, the conditionally predictive features (Panel B) tend to be micro-aggregates which may elude even the sharpest marketing director.

Rank	Feature	R^2
	Panel A: Tested individually	
1.	Percent of inactive days (during training period)	0.66
2.	Maximum monthly Δ in incoming calls / Total incoming calls	0.44
3.	Maximum monthly Δ in incoming calls / Total outgoing calls	0.42
4.	Outbound network degree (most recent month)	0.35
5.	Incoming text messages received from competitor's network	0.33
6.	Average number of calls to Information Portal per active day	0.33
7.	Unique weekend contacts per active day	0.33
8.	Average daily text messages received from competitor's network	0.33
9.	Number of Inactive Days in the First Month of the Training Period	0.31
10.	Daytime degree (voice calls)	0.25
	Panel B: Tested jointly	
1.	Outgoing degree (most recent month)	0.35
2.	Outgoing degree (first month) / SMS Degree	0.09
3.	Incoming degree (second month) / Total incoming calls	0.14
4.	SMS to Mobile Money Service / Total days of inactivity	0.01
5.	Short-duration calls (first month) / Total incoming calls	0.24
6.	Incoming calls / Incoming events	0.12
7.	Calls to Mobile Money servce (first month) / Total days of inactivity	0.32
8.	Outgoing events (first month) / Call degree	0.22
9.	Total incoming int'l SMS (weekends) / Incoming degree	0.18
10.	Total outgoing degree (first month) / Call degree	0.22

Table 1 Top Predictors of Churn

For evaluating predictive performance of our machine learning models we use "Percent of inactive days" as a baseline feature, and build a linear discriminant model based on that feature. We find that using the empirically found threshold of 76% for inactive days our prediction is correct in 83.9% percent of cases.

Depending on the algorithm used to predict churn, we achieve accuracy rates of roughly 88.5-89.5 percent. This represents a modest improvement of roughly 6 percent over the single-feature baseline, or approximately 14 percent over the majority-class baseline. A variety of performance characteristics of each model, as well as the linear discriminant baseline, are given in Table 2.

Model	Accuracy	Precision	Recall	F-Score	AUC
SVM	89.4	0.89	0.89	0.89	0.91
Random Forest	88.4	0.88	0.88	0.88	0.92
KNN	88.2	0.88	0.88	0.89	0.89
AdaBoost	89.2	0.89	0.89	0.89	0.94
Logistic Regression	89.3	0.89	0.89	0.89	0.93
Baseline Model	83.9	0.83	0.84	0.83	0.89

Table 2 Churn Prediction Performance

While these initial empirical results are promising, we see the primary contribution of this paper being the description of a systematic framework that can be used to generate interpretable features and predict customer outcomes. Several of the modelling assumptions we have made, such as the axes and dimensions used to generate features, are quite arbitrary and it is likely that more careful design of these behavioral metrics could yield more intuitive predictors and more accurate predictions.

References

Archaux, C., Martin, A., & Khenchaf, A. (2004). An SVM based churn detector in prepaid mobile telephony. In Information and communication technologies: From theory to applications, 2004. Proceedings. 2004 international conference on (pp. 459–460). IEEE. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=13070

Bonchi, F., Castillo, C., Gionis, A., & Jaimes, A. (2011). Social network analysis and mining for business applications. ACM Transactions on Intelligent Systems and Technology (TIST), 2(3), 22. Retrieved from http://dl.acm.org/citation.cfm?id=1961194

Dasgupta, K., Singh, R., Viswanathan, B., Chakraborty, D., Mukherjea, S., Nanavati, A. A., & Joshi, A. (2008). Social ties and their relevance to churn in mobile telecom networks. In Proceedings of the 11th international conference on extending database technology: Advances in database technology (pp.668–677). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=1353424

Dwyer, F. R. (1997, September). Customer lifetime valuation to support marketing decision making. J. Direct Mark., 11(4), 6–13.

Geurts, P., Ernst, D., & Wehenkel, L. (2006, April). Extremely randomized trees. Mach. Learn., 63(1), 3–42. Retrieved from http://dx.doi.org/10.1007/s10994-006-6226-1

Gupta, S., & Zeithaml, V. (2006, November). Customer metrics and their impact on financial performance. Marketing Science, 25(6), 718–739. Retrieved 2014-09-05, from http://pubsonline.informs.org/doi/abs/10.1287/mksc.1060.0221

Hart, C., Heskett, J., & Sasser, W. E., Jr. (1990, July). The profitable art of service recovery. Harvard Business Review. Retrieved from http://hbr.org/1990/07/the-profitable-art-of-service-recovery/ar/1

Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Friedman, J., & Tibshirani, R. (2009). The elements of statistical learning (Vol. 2) (No.1). Springer. Retrieved from http://link.springer.com/content/pdf/10.1007/978-0-387-84858-7.pdf

Jain, D., & Singh, S. S. (2002, March). Customer lifetime value research in marketing: A review and future directions. J. Interactive Mark., 16(2), 34–46. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/dir.10032/abstract

Karnstedt, M., Rowe, M., Chan, J., Alani, H., & Hayes, C. (2011). The effect of user features on churn in social networks. In Proceedings of the 3rd international web science conference (p. 23). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=2527051

Lemmens, A., & Croux, C. (2006, May). Bagging and boosting classification trees to predict churn. Journal of Marketing Research, 43(2), 276–286. Retrieved from http://journals.ama.org/doi/abs/10.1509/jmkr.43.2.276

Neslin, S. A., Gupta, S., Kamakura, W., Lu, J., & Mason, C. H. (2006). Defection detection: Measuring and understanding the predictive accuracy of customer churn models. Journal of marketing research, 43(2), 204–211. Retrieved from http://journals.ama.org/doi/abs/10.1509/jmkr.43.2.204

Qian, Z., Jiang, W., & Tsui, K.-L. (2006). Churn detection via customer profile modelling. International Journal of Production Research, 44(14), 2913–2933. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/00207540600632240

Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. European Journal of Operational Research, 218(1), 211–229. Retrieved from http://www.sciencedirect.com/science/article/pii/S0377221711008599

Wei, C.-P., Chiu, I., & others. (2002). Turning telecommunications call details to churn prediction: a data mining approach. Expert systems with applications, 23(2), 103–112. Retrieved from http://www.sciencedirect.com/science/article/pii/S0957417402000301

Zhang, X., Zhu, J., Xu, S., & Wan, Y. (2012). Predicting customer churn through interpersonal influence. Knowledge-Based Systems, 28, 97–104. Retrieved from http://www.sciencedirect.com/science/article/pii/S0950705111002693

Places and Mobility: the Influence of Attractions on People's Movement

Fahad Alhasoun*, May Alhazzani*, Faisal Aleissa *, Riyadh Alnasser*, Marta González*
* Massachusetts Institute of Technology
{fha, mayh, riyadh, aleissa, martag}@mit.edu

Extensive theoretical work attempts to address the predictability of human mobility in cities by means of understanding the periodicity of people movement as well as incorporating social networks features to improve the predictions. Using mobile phone data coupled with city structure data of the city of Riyadh in Saudi Arabia, we examine the potentials of better predicting the mobility of users using data pertaining to attraction areas within a city. We find that people movement around a city is highly correlated with the distribution of places around the city depending on the time of the day. We found a correlation factor of 0.86 between the number of POIs and the number of phone calls per location in the city. We further explored the correlation of different POIs to different times of the day to better understand the correlations during regular business hours versus night times.

I. Introduction

With the rapid adoption of pervasive technologies, a significant portion of the worlds population utilizes mobile phones, emails and social media (e.g., Twitter, Facebook...etc) forming a platform for people to exchange information, broadcast thoughts and convey feelings. Researchers today are using data generated from such technologies to better understand human behavior at unprecedented scales. Such data driven research unveiled statistical patterns that provide understanding of how people communicate, feel, move and so forth. Extensive research investigated the periodicity of human mobility and showed that humans can be highly predictable [3], [5]. Such understanding of the predictability helped significantly improve predicating social links within the social networks [4], [6]. Additionally, researchers were able to improve the prediction of human mobility through social networks features and investigated whether certain trips are motivated by social influences [1], [2]. In this paper, we present preliminary insights of the relationship of human mobility to the number of places around the city and the possibility of using such data to improve mobility predictions. Exiting research have used POI data as a proxy to the structure of cities in order to better understand how different areas within a city attract people [7].

II. Dataset

The data set used is mobile phone billing information, also known as Call Detail Records (CDRs), the dataset holds calling information including caller and receiver identifiers as well as the locations where the call was made. First, the dataset was filtered for users within the bounds of the city of Riyadh in Saudi Arabia. After the initial filtration process, the data set withholds around 3 million unique users and around 32 million unique social ties. (Peeking into the structure of the social network, figure (a) shows a subset of the social network in Riyadh. In this social network, there are around 23 thousands nodes and 93 thousands edges.) Along with the CDRs, we are using data pertaining to places of interests (POIs) in the city of Riyadh. We have around 12 thousands POIs in the city of Riyadh; each POI entry contains its location and its type (i.e. restaurant, store, bank...etc), where there are 96 types to tag POIs.

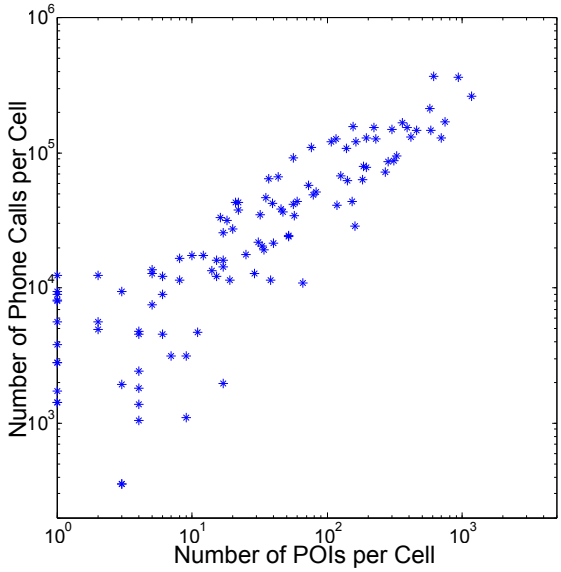


Figure 1: The number of phone records versus the number of POIs

III. Spatial Decomposition of POIs versus Phone Activity

In order to quantify the relationships of the number of POIs and the level of activity across the whole day, we split the city into squared cells of equal areas. Then, we correlate the aggregation of phone calling activity to the aggregation of number of POIs in each cell. Figure 1 shows the plot of the number of corresponding quantities for each cell where we have a correlation factor of 0.86 indicating very high correlation between the places people visit and the number of POIs in a place. Spatially visualizing the heat map of the POIs in the city of Riyadh versus the phone activity, figure 2 shows mobile phone activity on the left versus density of POIs on the right. The high correlation between cellular activity and the POIs of the city of Riyadh are visually evident; we can see that both heat maps overlap indicating similar spatial distribution. This hints out the possibility of using such correlating quantities in predicting human mobility around the city.

A. Time of Day Influence on Correlations

Although the spatial distribution of activity correlates highly with the number of POIs in a given cell, the correlation coefficient shows varying values depending on the time of the day. In order to gain insights, we filter phone calling activity during the day versus the activity during the night. Table I shows correlation coefficients between night and day activity versus the number of POIs on each cell. The table suggests that during the day people tend to be located around areas where the density of POIs is high having a correlation factor is 0.825. The Table also suggests that during the night people tend to be located in places that have POIs as well but with a 0.487 correlation coefficient.

Figure 2: Mobile activity (left) and POI density (right) decomposed spatially in Riyadh where the they have a spatial correlation of 0.86

	POI Count	Day Activity	Night Activity
POI Count	1	0.825	0.487
Day Activity	0.825	1	0.733
Night Activity	0.487	0.733	1

Table I: Correlation matrix of POIs versus mobile activity

B. Most Attractive POI Types

In order to better investigate whether certain locations are more preferred than others, we correlate the night and day activity to the POIs given their types. That way we can examine if certain POIs are more preferred to be visited during the day versus the night. Table II shows the types of POIs that correlates the most with phone activity during night and day. We can see that places with shopping malls correlate the most across both day and night times, but they seem more attractive during the day. We can also see that the most attractive POIs overlap greatly.

Day Activity	Night Activity
shopping mall (0.895)	shopping mall (0.772)
finance (0.89)	bank (0.753)
bank (0.86)	museum (0.749)
establishment (0.857)	finance (0.725)

Table II: POI types correlation rank versus mobile activity

IV. Future directions towards better predicting Mobility

Table II shows that there isn't much of a variation in the preference of certain types of POIs across day and night. Therefore, we aim to investigate the preferences of POIs by spatially correlating the POIs to the home and work locations of the users. In order to quantify the home and work locations, we assume a number of latent states (locations) we are going to predict. We aim to predict two major states: home and work. We assume that the spatial distribution of people around home and work are 2 time-invariant Gaussian distribution [1]. The distributions are centered around the home/work locations respectively. Also, we assume that user's phone activity times follows normal distribution. Figure 3 shows the traces of a user modeled as a two-components Gaussian mixture model, where each represents the underlying hidden state. The user's visits are classified either as home or work based on the time of the user's phone activity according to the temporal normal distribution. Then, based on the predicted state, the geographic

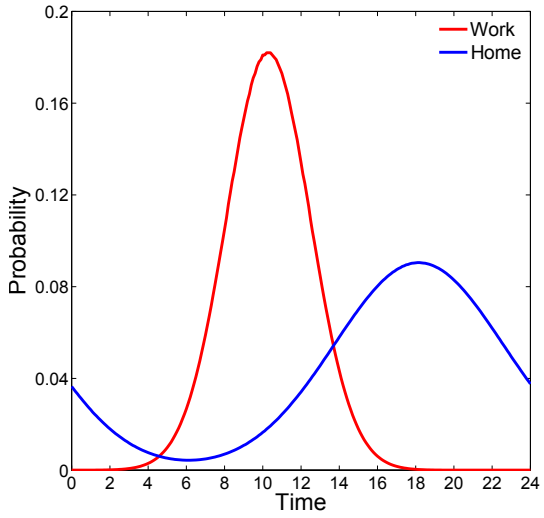


Figure 3: Gaussian mixture model for inferring home/work traces

location can be predicted from the location distribution. In addition to investigating the correlations of home/work locations with POIs, this prediction technique will help us identify visits that don't belong to one's home or work locations. 10% to 30% of such outlier traces can be explained by incorporating the social network information [1]. We intend to explore the possibility of explaining more about the outlier traces through incorporating POIs. Potential research direction includes incorporate an existing Periodicity Mobility Model (PMM) that can capture home and work for each user.

References

[1] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1082–1090. ACM, 2011.

[2] M. De Domenico, A. Lima, and M. Musolesi. Interdependence and predictability of human mobility and social interactions. *Pervasive and Mobile Computing*, 9(6):798–807, 2013.

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

[4] S. Scellato, A. Noulas, and C. Mascolo. Exploiting place features in link prediction on location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1046–1054. ACM, 2011.

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

[6] D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A.-L. Barabasi. Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1100–1108. ACM, 2011.

[7] J. Yuan, Y. Zheng, and X. Xie. Discovering regions of different functions in a city using human mobility and pois. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 186–194. ACM, 2012.

Classification of Human Population in Hospitals Using Mobile Phone Data

Guanning Dong dongguanning@tuoming.com
Xianfeng Song xfsong@ucas.ac.cn
Xinhai Liu xinhai.liu@foxmail.com

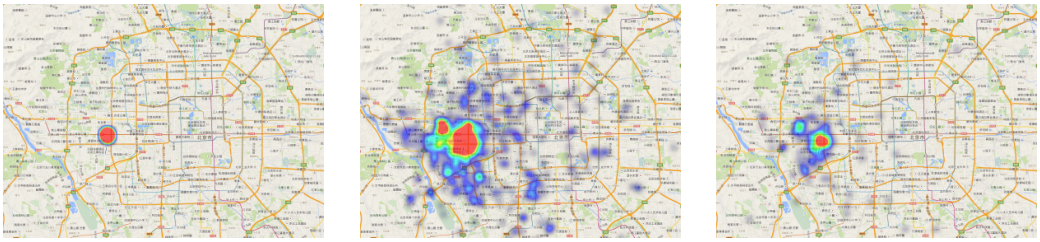
INTRODUCTION

As the most important destinations of cities, hospitals usually have irrational arrangement. For example, the patients may crowd into the top three hospitals, increasing the burden of medical workers, in the meantime, resulting in a waste of medical resources of other hospitals. When anyone is in hospital, communication with family and friends becomes an essential element. The generated mobile phone data offer us a chance to study the behavior pattern of human population in different level hospitals, which will help to improve the existing medical facilities and service and adjust the quantity and capacity of hospitals according the real-time demand of patients.

Some important conclusions have been reached about human pattern based on mobile phone data[1]. Mobile phone trajectories represent a reasonable proxy for individual mobility and can provide useful insights into intra-urban mobility patterns[2]. Experimental studies have explored the relationship between human behavior and mobile phone datasets, and identify the home-work location[3]. In this paper, focusing on further application of the mobile phone data, we propose an analytical process aimed at classifying human population in hospitals.

METHODS

We first hypothesize that human population in hospitals are mainly composed of healthcare workers and patients. Both of them have characteristic behavior model, thus mobile phone traces. Specifically, healthcare workers follow a relatively obvious home-hospital model. The patients, who divide into outpatients and inpatients, have different patterns. Inpatients may also communicate in hospital in working hours and at home in off-working hours, but the situation will not last long. The datasets of outpatients can be generated in hospital at any time of a day. The pattern is shown in Figure 1.



(a) Inpatients (b) Outpatients (c) Healthcare workers

Figure 1: Spatial distribution of human population in hospitals

	Healthcare workers	Inpatients	Outpatients
Working time	In hospital	In hospital	In hospital
Off-working time	At home	In hospital	At home

Table 1: Feature of human population in hospitals

	Healthcare workers	Inpatients	Outpatients
Population	173	2126	1053

Table 2: Classification of human population in hospitals

We then classify the healthcare workers and patients using the Support Vector Machines (SVM). For modeling, three groups of training samples were used in concert with four different kernel functions. Table 2 is the preliminary classification results of The General Hospital of People's Liberation Army (known as 301 hospital) on 15th December, 2014.

Moreover, we obtain the spatial distribution of healthcare workers and patients by cluster the mobile phone datasets, as is presented in figure 1(301 hospital). After that, we analyze the original destination matrix to understand the distance human spend on the way to hospital. Finally, based on aided-data, we research the correlation between the quantity of patients and indicator factors, such as age and gender.

RESULTS

The experimental finding importantly evaluates the potential feasibility of using the large scale of mobile phone data to research on the healthcare allocation. The contribution is that we classified the human in hospital into groups, including the healthcare workers, outpatient and inpatient. This is based on the temporal and spatial feature among them. Furthermore, the systematic study on patients' medical tendencies indicates that the patients of different level hospitals have certain distinct features. Afterwards, further analysis and applications will be explored to provide evidence and advice for healthcare facility and resource reallocation and management, aiming to improve the service level and efficiency of healthcare system in Beijing and other main cities of China.

REFERENCES

1. Hoteit S, Secci S, Sobolevsky S, Ratti C, Pujolle G. Estimating human trajectories and hotspots through mobile phone data. Computer Networks. 2014; 64(0): 296-307.
2. Calabrese F, Diao M, Di Lorenzo G, Ferreira Jr J, Ratti C. Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. Transportation Research Part C: Emerging Technologies. 2013; 26(0): 301-13.
3. Csáji BC, Browet A, Traag VA, Delvenne J-C, Huens E, Van Dooren P, et al. Exploring the mobility of mobile phone users. Physica A: Statistical Mechanics and its Applications. 2013; 392(6): 1459-73.

Campaign Optimization through Mobility Network Analysis

Yaniv Altshuler[1], Erez Shmueli[2], Guy Zyskind[3], Oder Lederman[3], Nuria Oliver[4], Sandy Pentland[3]
[1]Athena Wisdom, Israel
[2]Department of Industrial Engineering, Tel-Aviv University, Israel
[3]MIT Media Lab, USA
[4]Telefónica Research, Spain

In a world of limited resources, behavior change campaigns (*e.g.* marketing, service provision, political or homeland security) can rely on creativity and attractiveness up to a certain point. The success of a campaign can generally be defined as the product of *reach* (portion of the population exposed to the campaign messages) and *value* of a single interaction (the capacity of a message to induce a certain behavior in an exposed audience) [1]. Hence, campaign managers typically distribute their budget between content enhancement (to increase the value a single interaction) and wide reach. Yet, to date it seems that the optim trade-off between these two factors is found as a result of "intuition" rather than based on well established analysis.

In this work, we propose a novel mathematical method that, given the characteristics of the target audience and its ability to be persuaded, generates an optimized campaign strategy in terms of: (a) the quantity of interacting units, also referred to as *insertions* and (b) the monetary allocation to each unit. The model takes into account the population's mobility in an urban environment as it can be inferred from real data received from a large mobile phone carrier. Even though different populations located in different environments would be tailored with different campaign strategies, the optimality of each strategy would be maintained.

A major contribution in our optimization model is the use of network analysis methods to approximate the reach of a campaign. More specifically, given the network of mobility between the different geographic locations, and a subset of locations, we use the *Group Betweenness Centrality* (GBC) [2] – a network measure that calculates the percentage of shortest paths among all pairs of network nodes that pass through a pre-defined sub-set of the network's nodes – to approximate the reach of this subset of locations. We then demonstrate that this function can be approximated using a smooth and easily analyzed *Gompertz* function. This tackles the main limitation of works on campaign optimization hitherto – efficiently estimating the campaign reach as a function of the number of units and their locations.

We have validated our results by using two comprehensive real-world geo-spatial datasets. The first dataset included a large number of mobile phone records, from which we have produced a mobility pattern model which was in turn analyzed in order to derive an optimized campaign for the region in question. The second dataset comprised of a large number of taxi rides in the city of New York. While analyzing these dataset we have first shown a way to analytically calculate the exact optimal cost for units in a certain campaign, of generic nature. We then demonstrated how the optimal number of such units can be produced, that would guarantee a maximal utilization of a campaign's budget.

Finally, we have discussed several campaign scenarios, involving various utilization schemes, demonstrating the usability of the techniques presented in this chapter for real world use.

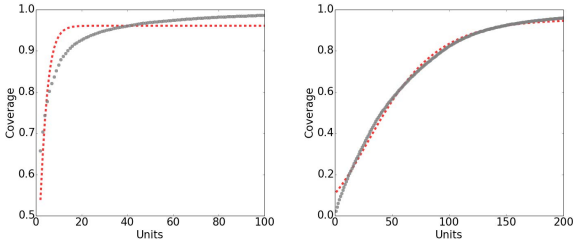


Figure 1: GBC Deployment fort the two mobility datasets: (a) CDR (left) and (b) taxi rides (right). The appropriate Gompertz fit of the curves is also included.

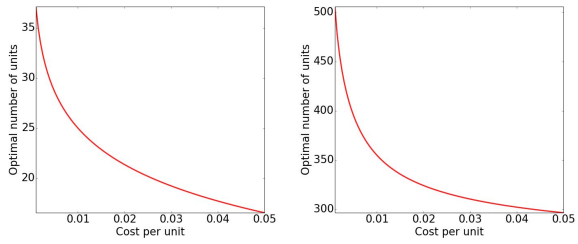


Figure 2: The optimal number of units, as a function of the ratio between the cost of a unit and the cost of the best unit available in the two mobility datasets. For example, in the CDR dataset (left), when a single unit costs 1% of the maximal campaign impact, the optimal number of campaign units would be 25, whereas if cheaper units are used (such as units that cost merely $\frac{1}{2}$% of the maximal campaign impact) the optimal number of units would be 28. Similarly, in the taxi rides dataset (right), when a single unit costs 1% of the maximal campaign impact, the optimal number of campaign units would be 350, whereas if cheaper units are used (such as units that cost merely $\frac{1}{2}$% of the maximal campaign impact) the optimal number of units would be 380.

1. REFERENCES

[1] Peter J Danaher and Roland T Rust. Determining the optimal level of media spending. *Journal of Advertising Research*, 34(1):28–34, 1994.
[2] M. G. Everett and S. P. Borgatti. The centrality of groups and classes. *Mathematical Sociology*, 23(3):181–201, 1999.