Editors:

Francesco Calabrese
Esteban Moro
Vincent Blondel
Alex 'Sandy' Pentland

5 – 7 APRIL 2017
VODAFONE VILLAGE
MILAN

NetMob

BOOK OF ABSTRACTS

ORAL

vodafone
MIT
Universidad Carlos III de Madrid
UCL Université catholique de Louvain

INDEX

ORAL

SESSION 1: SOCIO – ECONOMIC

SESSION 2: MOBILITY

SESSION 3: SOCIAL NETWORK

SESSION 4: SOCIAL GOOD

SESSION 5: HEALTH

SESSION 6: DATA QUALITY AND PRIVACY

SESSION 7: SOCIAL NETWORK

SESSION 8: MOBILITY

SESSION 9: SOCIAL GOOD

SESSION 10: TELCO APPLICATIONS

SESSION 1
SOCIO – ECONOMIC

Higher-order correlations of consumption patterns in social-economic networks

Yannick Leo*, Márton Karsai*, Carlos Sarraute† and Eric Fleury*
*Univ Lyon, ENS de Lyon, Inria, CNRS, UCB Lyon 1, LIP UMR 5668, IXXI, F-69342, Lyon, France
†Grandata Labs, 550 15th Street, San Francisco, CA, USA
yannick.leo@ens-lyon.fr, marton.karsai@inria.fr, charles@grandata.com, eric.fleury@inria.fr

Introduction

The consumption of goods and services is a crucial element of human welfare. The uneven distribution of consumption power among individuals goes hand in hand with the emergence and reservation of socioeconomic inequalities in general. Individual financial capacities restrict personal consumer behaviour, arguably correlate with one's purchasing preferences [1], [5], and play indisputable roles in determining the socioeconomic position of an ego in the larger society [1]–[4]. Consumption behaviour has been addressed from various angles considering e.g. environmental effects, socioeconomic position, or social influence coming from connected peers [1]. However, large data-driven studies combining information about individual purchasing and interaction patterns in a society large population is still rare, although questions addressing higher-order correlation between consumption and social behaviour are at utmost interest.

In this study [9] we address these crucial problems via the analysis of a dataset, which simultaneously records the mobile-phone communication, bank transaction history, and purchase sequences over several months of millions of individuals living in a developing country. This corpus, one among the firsts at this scale and details, allows us to infer the socioeconomic status, consumption habits, and the underlying social structure of millions of connected individuals. Using this information our overall goal is to identify people with certain financial capacities, and to understand *how much money they spend, on what they spend, and whether they spend like their friends?* More precisely, we formulate our study around two main research questions:

- Can one associate typical consumption patterns to people and to their peers belonging to the same or different socioeconomic classes, and if yes how much such patterns vary between individuals or different classes?
- Can one draw relations between commonly purchased goods or services in order to understand better individual consumption behaviour?

Data description

Communication data used in our study records the temporal sequence of $\sim$ 8B call and SMS interactions of $\sim$ 112M anonymized mobile phone users for 21 months. Using call detailed records (CDRs) we constructed a large social network where nodes are users (whether clients or not of the actual provider), while links are drawn between any two users if they interacted (via call or SMS) at least once during the observation period. We also insure that each node has at least one incoming and outgoing call or SMS and remove non-human and commercial actors from the network. At the same time to calculate individual economic estimators we used a dataset which records financial details of $\sim$ 6M people assigned with unique anonymized identifiers over 8 months. The data provides time varying customer variables as the amount of their debit card purchases and demographic attributes such as age and gender. A subset of IDs of the anonymized bank and mobile phone customers were matched. The combined data of the bank and mobile datasets (from now called DS1) contained $\sim$ 1M people connected by $\sim$ 2M links with communication events and detailed bank records available.

To study consumption behaviour we used purchase sequences recording the time, amount, merchant category code of each purchase event of each individual during the observation period of 8 months. Each purchase events are linked to one of the 281 merchant category codes (MCC) indicating the type of the actual purchase (food restaurants, airlines, gas stations). MCCs were grouped into 28 purchase category groups (PCGs) as in [7]. We decided to remove 11 of the 28 PCGs with extremely low activity (see the 17 remaining categories in Fig.1a). At the same time we analyze group k_1 *Service Providers* separately from the remaining $K_{2\text{-}17} = K_{17} \backslash \{k_1\}$ set

of 17 groups as it corresponds to cash retrievals and money transfers (70% of the total purchases) and since we have no further information how the withdrawn cash was spent.

Using only bank transaction traces we build a second data set (DS2), which collects data about the age and gender of ~ 3.6M individuals together with their purchase sequence recording the time, amount, and MCC of each debit card purchase. To estimate the personal economic status we used average monthly purchase (AMP) of individuals. More precisely, in case of an ego u who spent $m_u(t)$ amount in month t we calculated the AMP as $P_u = \sum_{t \in T} m_u(t)/|T|_u$ where $|T|_u$ corresponds to the number of active months of user u (with at least one purchase in each month). After sorting people by their AMP values we computed the $C_P(f)$ normalized cumulative distribution function of P_u as a function of f fraction of people. We used the $C_P(f)$ function to assign egos into nine s_j $(j = 1 \dots 9)$ socioeconomic classes (smaller numbers assigning lower classes) such that the sum of AMP in each class s_j was the same.

SOCIOECONOMIC CORRELATIONS IN PURCHASE PATTERNS

In order to address our first research question, for each socioeconomic class s_j we take every user $u \in s_j$ and calculate the m_u^k total amount of purchases they spent on a purchase category group $k \in K_{17}$ and measure the $r_{\%}(k, s_j)$ fractional distribution of spending for each PCGs. In Fig.1a each line shows the $r_{\%}(k, s_j)$ distributions for a PCG as the function of s_j social classes, and lines are sorted (from top to bottom) by the total amount of money spent on the actual PCG. Interestingly, people from lower socioeconomic classes are spending more on PCGs associated to essential needs (*Retail Stores (St.)*, *Gas Stations*, *Service Providers*) while richer people spend more on extra needs (*High Risk Personal Retail*, *Automobiles*, *Hotels* and *Airlines*).

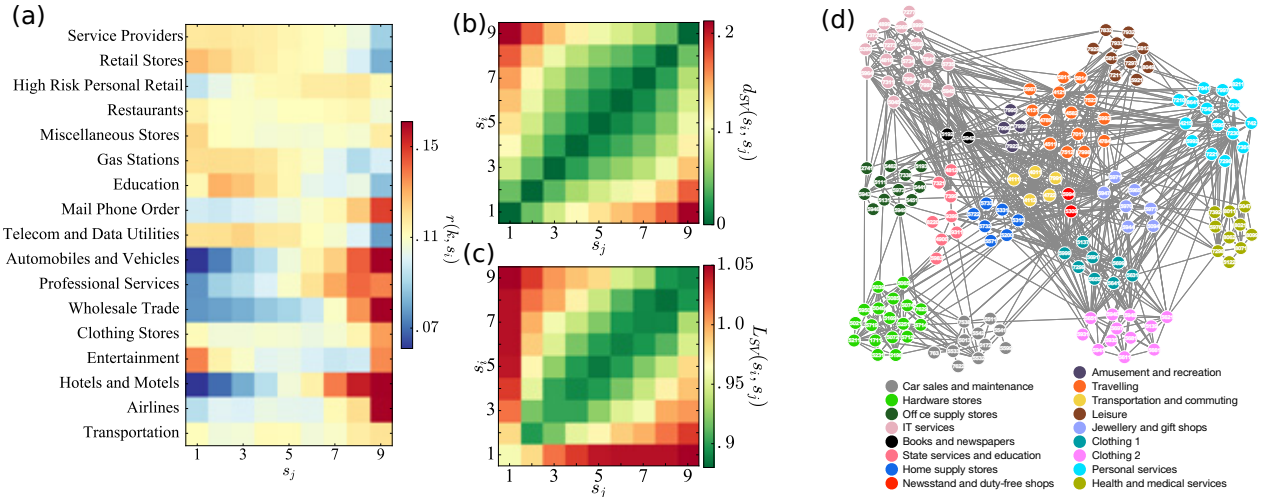


Fig. 1. **Consumption correlations in the socioeconomic network (a)** $r_{\%}(k, s_i)$ distribution of spending in a given purchase category group $k \in K_{17}$ by different classes s_j. Distributions are normalized such that they sum up to 1 for each category. **(b)** Heat-map matrix representation of $d_{\overline{SV}}(s_i, s_j)$ distances between the average spending vectors of pairs of socioeconomic classes considering PCGs in $K_{2\text{-}17}$. **(c)** Heat-map matrix representation of the average $L_{\overline{SV}}(s_i, s_j)$ (resp. $L_{k_1}(s_i, s_j)$) measure between pairs of socioeconomic classes considering PCGs in $K_{2\text{-}17}$. **(d)** Weighted $G_\rho^>$ correlation graph with nodes annotated with MCCs. Colors assign 17 communities of merchant categories with representative names summarized in the figure legend.

To further investigate these dissimilarities between socioeconomic classes, we build a spending vector $\overline{SV}(u)$ for each ego u where each item $SV_k(u)$ assigns the fraction of money what a user u spent on a category $k \in K_{2\text{-}17}$ out of his/her $m_u = \sum_{k \in K} m_u^k$ total amount of purchases. Using these individual spending vectors we calculate the $\overline{SV}(s_j)$ average spending vector of a given socioeconomic class s_j and use it to quantify differences between distinct socioeconomic classes. Results in Fig.1b show the pairwise differences between each class as a heat-map which appears with a strong diagonal green component and with distinct off-diagonal values. It indicates that the average spending behaviour of a given class is the most similar to the class itself and neighboring classes, while dissimilarities increase with the gap between socioeconomic classes. Note that we observed the same behaviour for the single PCG k_1 (results not shown here), and also observed that richer people tends to be more similar in terms of their purchase behaviour while they used to allocate their purchases in more PCGs.

To continue our investigation we characterize the effects of social relationships on the purchase habits of individuals through the analysis of DS1. We address this problem through an overall measure quantifying differences between individual purchase vectors of connected egos positioned in the same or different socioeconomic classes. More precisely, we consider each social tie $(u, v) \in E$ connecting individuals $u \in s_i$ and $v \in s_j$, and for each purchase category k we calculate the average absolute difference of their purchase vectors items as $d^k(s_i, s_j) = \langle |SV_k(u) - SV_k(v)| \rangle_{u \in s_i, v \in s_j}$. In order to quantify the effect of the social network we calculate the average of the same quantity $d^k_{rn}(s_i, s_j)$ measured on corresponding configuration networks. Finally we take simply the $L_{\overline{SV}}(s_i, s_j)$ average of their ratio $L_k(s_i, s_j) = d^k(s_i, s_j)/\langle d^k_{rn} \rangle(s_i, s_j)$, averaged over each category group $k \in K_{2\text{-}17}$ or respectively k_1. This measure shows whether connected people have more similar purchase patterns than one would expect by chance without considering any effect of homophily [6], social influence or structural correlations [4]. Results depicted in Fig.1c for $L_{\overline{SV}}(s_i, s_j)$ indicates that the purchase patterns of individuals connected in the original structure are actually more similar than expected by chance if the link connects egos from similar socioeconomic groups (diagonal component). Note that we found the same correlation trends in cash purchase patterns (not presented). These observations do not clearly assign whether homophily or social influence induce the observed similarities in purchasing habits but undoubtedly clarifies that social ties (i.e. the neighbors of an ego) and socioeconomic status play deterministic roles in the emerging similarities in consumption behaviour.

Finally our aim is to obtain an overall picture of the consumption structure at the level of merchant categories and to understand precisely how personal and socioeconomical features correlate with the spending behaviour of individuals and with the overall consumption structure. Using the spending vectors of individuals we define a $\rho(c_i, c_j)$ overall correlation measure between categories. This symmetric measure quantifies how much people spend on a category c_i if they spend on another c_j category or vice versa. Using the obtained correlation matrix we take into account only positively correlated frequently purchased category pairs and we define a weighted correlation network. This network appears with well observable purchase category communities, what we visualize in Fig.1d by using the Louvain community detection method [8]. Interestingly, each of these communities group a homogeneous set of merchant categories, which could be assigned to similar types of purchasing activities, while inter-community links show how different groups of purchase categories are connected together.

In this study, from individual level, we manage to draw relations between commonly purchased goods or services in order to obtain a global consumption picture with 17 merchant category groups (see Fig. 1d). Thanks to the large scale corpus, offering to infer the socioeconomic status, consumption habits, and the underlying social structure of millions of connected individuals we derive results that give us the intuition of how consumption patterns are organized in a society with heterogeneous distribution of wealth and purchase capacities. Whereas we do not present the results in this abstract, we also studied correlations between average age, SEG and gender of each MCC category. Interestingly we observe the presence of anecdotic purchase patterns characterizing the two gender groups, and that wealthier people tend to be older, males, and used to purchase from more expensive categories in accordance with our intuition.

REFERENCES

[1] A. Deaton, Understanding Consumption *Clarendon Press* (1992).
[2] T. Piketti, Capital in the Twenty-First Century. (*Harvard University Press*, 2014).
[3] P. West, Conspicuous Compassion: Why Sometimes It Really Is Cruel To Be Kind. *Civitas, Institute for the Study of Civil Society* (London) (2004).
[4] Y. Leo, E. Fleury, J. I. Alvarez-Hamelin, C. Sarraute, M Karsai, Socioeconomic correlations and stratification in social-communication networks *J. R. Soc. Interface* **13**:125 20160598 (2016).
[5] J. E. Fisher, Social Class and Consumer Behavior: the Relevance of Class and Status", in Advances in Consumer Research Vol. 14, eds. M. Wallendorf and P. Anderson, Provo, UT : Association for Consumer Research, pp 492–496 (1987) .
[6] M. McPherson, L. Smith-Lovin, J. M. Cook, Birds of a Feather: Homophily in Social Networks. *Ann. Rev. Sociol.* **27** 415–444 (2001).
[7] Merchant Category Codes and Groups Directory. *American Express @ Work Reporting Reference* (http://tinyurl.com/hne9ct5) (2008) (date of access: 2/3/2016).
[8] V. Blondel, J-L Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks *J. Stat.l Mech: theory and experiment* P10008 (2008).
[9] Y. Leo, M. Karsai, C. Sarraute, E. Fleury, Correlations of consumption patterns in social-economic networks *IEEE/ACM ASONAM* p. 500-507, San Francisco CA, August 18-21 (2016).

Mobile phone-based Credit Scoring

Skyler Speakman, Eric Mibuari, Isaac Markus, Felix Kwizera
IBM Research – Africa
skyler@ke.ibm.com

INTRODUCTION

Mobile Money platforms are gaining traction across developing markets as a convenient way of sending and receiving money over mobile phones. These systems operate on low-cost feature phones and do not require users to have a bank account. These low barriers-to-entry make mobile money platforms excellent tools for financial inclusion of the poor. Additional financial services, such as saving and lending products, are now being offered over these mobile money channels. In this work, we demonstrate how boosted decision trees (Adaboost) may be used to create credit scores (probability of repaying a low-value, short-term loan) for under-banked populations, allowing them to access credit that was previously unavailable due to a lack of financial data. The boosted tree model demonstrated significant results over the original model used by the bank. We show a 55% reduction in default rates while simultaneously offering credit opportunities to a million customers that were given a 0 credit limit in the bank's original model.

The mobile-based saving and lending product is a joint collaboration between a mobile network operator (MNO) and a regional bank in East Africa. The MNO provides access to the customer through their mobile money platform as well as data on recent customer activity. This data is shared with the banking partner when a customer signs up for the savings and loan product in order to help determine risk. The features are broken down into three large categories. We provide a non-exhaustive list. Demographic: Age, Gender, Account age, etc. Network: Airtime usage, Top-up amounts, Active days, Airtime borrowing, etc.

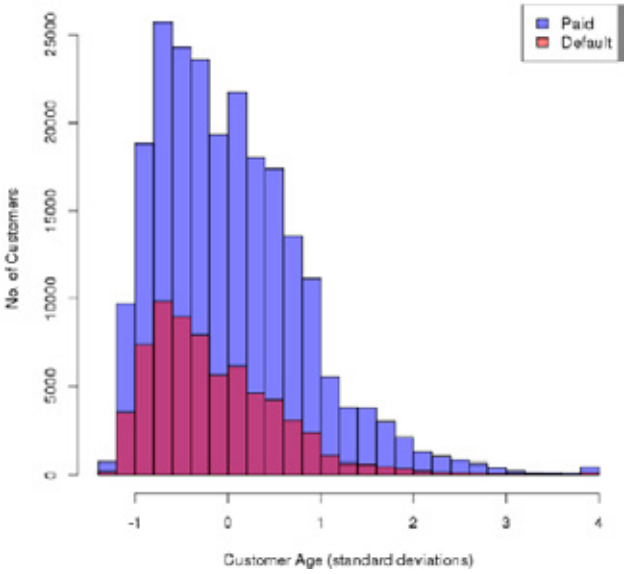


Figure 1: Histogram of customer age and payment class.

Mobile Money: Amount(s) sent, received, deposited, withdrawn, etc.

On top of this rich feature space the bank provided labeled data of customers who have repaid their mobile loans and those who have defaulted. In particular, the data set contained 295,926 labeled examples with over 30 categorical and real-valued features. These records cover customer activity in 2013 and early 2014. There was a slight class imbalance with 23.1% of customers defaulting on a mobile loan.

Figure 1 provides a histogram showing the distributions of "Paid" and "Default" customers across the Customer Age feature for this data set. The x-axis is in standard deviations away from the mean. Therefore, 0 represents the mean age of borrowing customers. Age is positively skewed and capped at 4 standard deviations for display purposes. As expected,

younger borrowers are riskier than older borrowers. However a single feature (histogram) cannot account for the interactions involved in calculating an accurate credit score. We therefore turn to one of the best off-the-shelf classifiers to assist in this textbook supervised learning problem.

BOOSTED DECISION TREES

Decision trees have been a workhorse of data analytics for decades. Their ability to incorporate real and categorical attributes for both classification and regression problems make them applicable to a wide range of applications. Additionally they are able to gracefully handle missing values that are common in real-world data sets. In order to increase the accuracy of the binary classification (default or repaid) we trained multiple, simpler decision trees in sequence. Each consecutive tree attempts to correct the mistakes made by the preceding ones. This process is known as boosting [1,2] and we use the Adaboost implementation in R (gbm package) [3].

Boosting introduces two additional parameters to the training process: the depth (complexity) of each individual tree and the number of trees to create in the sequence. We set the value of these parameters through 10-fold cross validation with the goal of maximizing the area under the receiver operator characteristic curve (also known as AUC). The model is then built using these parameters. Finally, the values provided from the boosting process are turned into class probabilities using a logistic correction [4].

Figure 2 shows the AUC values for various combinations of tree depth and number of trees used in the boosting sequence. 180 trees of depth 2 maximized the AUC around 0.764. Logistic regression on the same data set resulted in an AUC of 0.74. We start to see overfitting with large amounts of depth 3 trees. Other depth values were ommitted for display purposes.

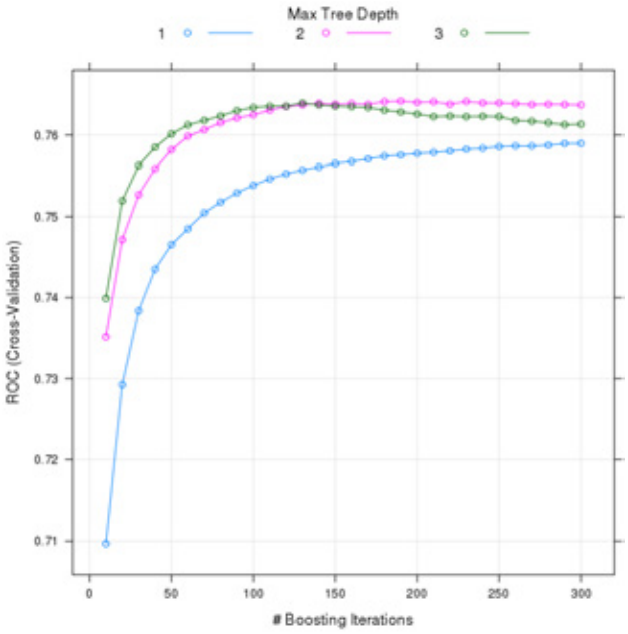


Figure 2: Area-under-curve results for cross validation.

EMPIRICAL RESULTS & DISCUSSION

The method described above takes a customer's features (as provided by the MNO) as input and produces a "probability of repaying" based on this information. We now discuss two straight forward business implications of increasing the classification accuracy over the original method used by the bank.

For discussion purposes we assume that a risk manager at the bank has assigned a threshold of 0.7 for allowing new customers to borrow money on the loan product. Therefore anyone with a probability of repaying higher than 0.7 would be given a credit limit whereas customers below that probability would only be allowed to save money on the product. The vertical black bar on Figures 3 and 4 represents this cutoff value.

With this assumption in place, we demonstrate a 55% reduction in default rate. In other words, the new boosted tree model would have correctly excluded 55% of customers who went on to default on a mobile loan. This 55% reduction in default rate may also be viewed as the sensitivity or recall of the model. The same threshold retains 83% of the paying customers on the product. This 83% retention rate may be viewed as the specificity of the model.

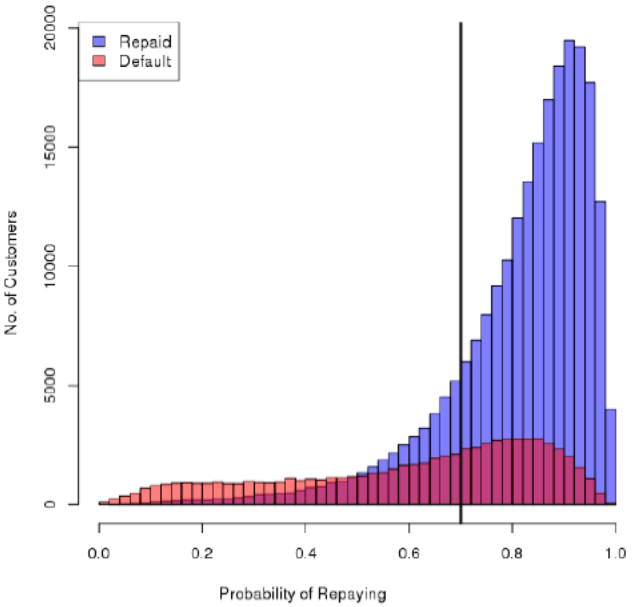


Figure 3: Histogram of probability of repayment and payment class.

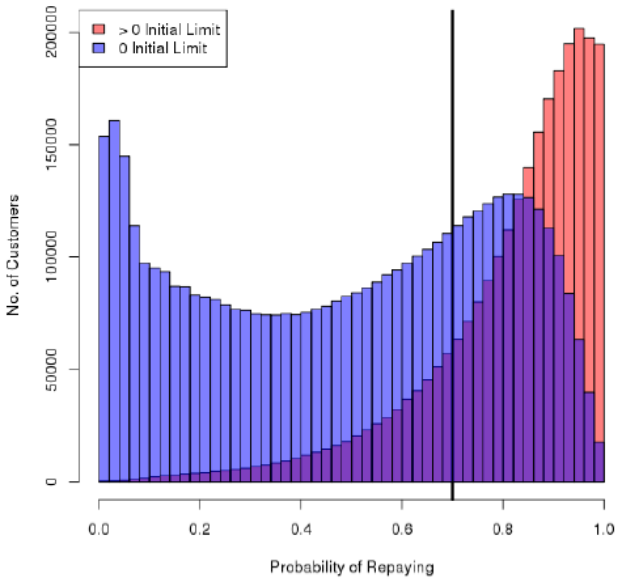


Figure 4: Histogram of probability of repayment and initial inclusion on the loan product.

The precision is 49% which means roughly half of the people excluded from the credit portion of the product were defaulters. Each of these numbers vary with the choice of threshold set by the risk manager.

Figure 3 is a histogram showing the distribution of customers over their probability of repaying for payers (blue) and defaulters (red). The x-axis is the probability of repaying the loan and the y-axis is the number of customers within each bin. 55% of the defaulting customers fall below the threshold value and would have been correctly excluded from borrowing money.

A reduction in default rate is not the only gain of the model. We also consider the larger population of customers on the savings and loan product including those that were initially excluded by the original model used by the bank. The original bank model allowed 35% of their total customers on the credit product initially. The boosted tree model (and assumed 0.7 threshold) allws 49% of the customer on the product; an increase of over 1 million customers.

Figure 4 visually demonstrates this increase in revenue potential for the bank as well as increased financial inclusion for the society. The red bars in the histogram of Figure 4 represent the customers who were initially allowed on the credit portion of the savings and loan product. The boosted tree model suggests the original model was too conservative and that many more customers (blue to the right of the 0.7 threshold) could also have been allowed to initially borrow on the mobile credit product.

The boosted tree model has also produced human-interperable relationships between features of cell phone usage and mobile money volumes that will help create additional mobile and banking products that support sustainable finanical inclusion.

REFERENCES

[1] Y. Freund and R.E. Schapire (1997). "A decision-theoretic generalization of online learning and an application to boosting," Journal of Computer and System Sciences, 55(1):119-139.

[2] T. Hastie; R. Tibshirani; J. Friedman (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer.

[3] G. Ridgeway (2005). "Generalized boosted models: A guide to the gbm package".

[4] A. Niculescu-Mizil and R. Caruana. (2005). "Obtaining calibrated probabilities from boosting." In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence* (UAI'05).

Mobile Wallet Usage Study

Apply CDR models to increase Mobile Wallet Adoption & Activity

IFC, Cignifi & Airtel[1]

This study is the result of a collaborative project between the International Finance Corporation (IFC), part of the World Bank Group; Airtel Uganda, Uganda's second-largest mobile operator; and Cignifi, a big-data analytics company. The developmental objective of the project was to expand financial inclusion in Uganda. Due to multiple factors such as poor infrastructure, lack of collateral, lack of proper financial education over existing alternatives, a majority of people over the age of 15 in Uganda do not have access to formal financial services. Among the population aged 15+ only 17% have a formal savings account and 16% have access to formal borrowing (Findex 2014). Cell phone penetration is high at over 55% (Uganda Communications Commission, 2014), which establishes a market for expanding financial access through digital financial services. To this end we used predictive analytics to help Airtel, a mobile network operator, identify customers most likely to become regular users of Airtel Money—a digital financial service offered by Airtel that allows customers to transfer money, pay bills, and engage in various other financial transactions. When poor people have access to formal financial services, they can earn more, build their assets and cushion themselves against external shocks. Access to financial services can support entrepreneurs and microenterprises, which in turn promote job growth and shared prosperity for others in the community to earn income as well. That is why the UN's Sustainable Development Goals framework explicitly recognizes the importance of promoting financial inclusion in to achieve the Sustainable Development Goals (SDGs) by 2030.

The project was based on an analysis of anonymized customer data provided by Airtel Uganda. This data included customers' detailed calling data (voice calls, text messages, data sessions), customer payment (or "recharge") data for their mobile accounts, and mobile money transactions for all Airtel Money users. The data, approximately one terabyte, covered a six-month period from November 2014 to April 2015. It did not include any

[1] Contact Qiuyan Xu, qxu@cinifi.com or Soren Heitmann, sheitmann@ifc.org for further information.

personally identifying data (such as customer names, addresses, or voice or text content). Cignifi worked with IFC to define predictive models, identify the primary user groups, and set a project road map. Cignifi then proceeded to extract, process, and normalize the mobile data and develop analytics models that segmented Airtel customers into three primary groups:

- *Highly Active Mobile Money Users*
- *Active Mobile Money Users*
- *Non-active Mobile Money Users*

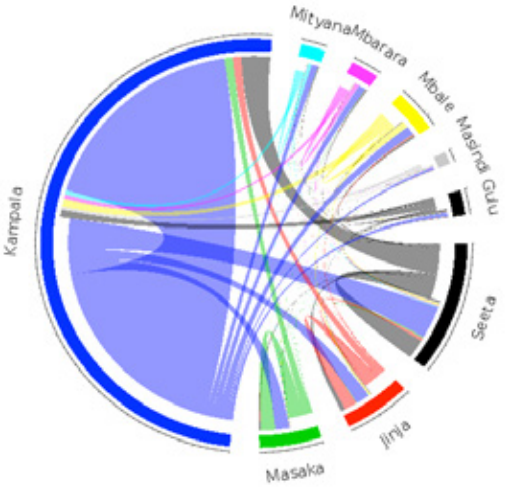
Working with the predictive model to divide the subscriber base into three user profiles—highly active, active and non-active—enables a clearer understanding of the factors influencing mobile money activity to emerge. These factors can help Mobile Money service providers identify subscribers more likely to be active mobile money users in order to increase the efficiency of marketing and customer targeting efforts. Social network analysis and geospatial analysis were applied on call and mobile wallet network to strengthen the findings and to support strategic growth beyond urban areas to advance financial inclusion. The developed model allows to identify and predict probable active Airtel Money subscribers from its GSM base with an accuracy rate of at least 85%.

This study identified three main indicators of a customer's likelihood to be an active mobile wallet user:

- ***GSM Activity**: A user's level of GSM activity strongly predicts their likelihood to be actively using Mobile Money (MM) Voice transactions in particular emerged as the most powerful transaction type for discriminating between profile groups. As technology leaders who make most frequent use of their phones and make both more frequent and larger transactions with Airtel Money, this user group is also more likely to be smartphone owners. This increases the ease of use for MM since the app version can be downloaded instead of the USSD version available for feature phones.*
- ***Revenue Generation**: Those who generate higher levels of revenue are also associated with higher levels of MM activity. Using revenue generation as a proxy for a customer's income, the study suggests that those users who may generate relatively higher incomes have higher participation in MM services. This makes sense because the number of uses for MM and frequency of applicable use increases as users have more money and use it for a variety of transactions. Additionally, similar to GSM activity, the probability that this segment contains a larger percentage of smartphone users is high, as affluence increases so does buying power.*
- ***Connectedness**: How connected a user is to their network, and the size of that network, also appear to be significant factors in determining propensity for MM activity. As users talk to and message more unique counterparties, their propensity increases. The behavior of a user's social network also provides significant discriminating power between the segments. The more active a user's network the more likely the user is to also be active. This also makes sense due to the nature of the MM product. The more connections a user has already using*

the product, the more useful it will be to the user when they adopt and the more likely they are to use it regularly.

Visualization of Mobile Money Flow amongst Clusters



The geospatial and cluster analysis of P2P (person to person money transfer services) flows further allowed to identify established and nascent transaction channels. See graph on the left in which circle segments represent origin, destination as well as the size of P2P transaction flows.

We see a clear concentration of transfer volumes in the capital city of Kampala. They represent almost half of all money sent via interpersonal transfers. In nearly every case, net money flows go from more populated to less populated areas. Access to mobile phones, the proximity of mobile money agents and a more robust economy in more urban areas might all play a role in explaining these results.

Our project's overall finding - that indicates middle upper income segments as the highest propensity users - calls into question the development premise on mobile money and DFS (digital financial services) as a financial inclusion driver amongst customers at the bottom of the pyramid. However, results from the social network analysis and geospatial analysis also highlight how DFS can be a driver of financial inclusion by combining high propensity user models with geospatial analysis that helps to identify where strategic segments are; and where users in strategic segments may already be transacting.

The developmental value of the project resides precisely in these findings: as long as Mobile Network Operators rely on their usual business tools for acquiring clients, DFS will expand but will include non-key segments of their potential client base. By including CDR analysis and have it inform business decisions, strategic areas and population segments can be identified, which can maximize development impact.

Longitudinal Analysis of Mobile Savings Accounts

Isaac M. Markus*, Skyler Speakman, Isaac Waweru, Abdigani Diriye, Jonathan Lenchner
IBM – Africa Research Center
Nairobi, Kenya
ismarkus@ke.ibm.com

The growth of microcredit in developing nations has accelerated over the last decade with the wide adoption of smart and feature phones. In conjunction with these mobile loan products several savings products have been deployed, expanding the availability of formal financial instruments to previously unbanked individuals. In this work we examine the utilization of such mobile savings products in eastern Africa, examining correlations between savings and credit usage. K-means clustering of savings data reveals the degree of stratification between users, with typical users (98%) having on average less than 1000Ksh ($10) in their account, and the remaining 2% holding 30% of the total capital in the mobile savings accounts. Further longitudinal clustering reveals a noticeable balance decrease preceding a loan request, providing an early indicator of a pending loan request that is applicable for a large group of users.

I. Introduction

The penetration of feature phones in Sub-Saharan African, followed by the advent of mobile money has provided critical infrastructure, the absence of which had previously hampered financial inclusion efforts on the continent [1,2]. With the adoption of mobile money, large telcos and banks have circumvented several of the barriers that prevented offering formal financial services to large swaths of the population [3,4]. By leveraging cell phone networks, brick and mortar banking locations are no longer required, which has reduced service cost and access limitations, particularly in rural areas. Furthermore, mobile phone data has facilitated risk assessment of previously unbanked individuals, allowing micro finance insitutions (MFIs) and banks to provide small loans with a stronger foundational understanding of credit worthiness [5].

In eastern Africa the process of applying and qualifying for a mobile based loan has been streamlined, with clear requirements that are anchored to establishing a transaction history with mobile money. As part of the qualification process customers are required to have a positive balance in a mobile savings account [6]. Although a positive balance is required for assessing credit risk, this financial product is for many customers the first opportunity to have an interest yielding savings account. As an innovative product itself, examining the behavior of customers in regards to their savings accounts can provide insights into how traditionally unbanked individuals will utilize formal banking product, such as a deposit account.

This paper explores cluster behavior of such account holders over a period of two years. The sample size includes an overlapping population that also received a mobile based loan. For the individuals having both savings and credit products, further analysis of their mobile money usage and repayment rates was performed. The work seeks to identify activity levels of deposit accounts, including analysis before and after loan disbursement in order to correlate with credit worthiness. Long term goals of this project include providing dynamic credit scoring capabilities that consider an individual's savings history and pattern of deposits and withdrawals leading up to a loan request.

II. Methodology

The dataset analyzed covers the time window from October 2012 to July 2014, including 4.1 million customers from east Africa that had signed up for a mobile money deposit account. Individual customers did not have the same time windows in which their accounts were open, with the main difference being at what time during the analyzed window was the account open. The original dataset was formatted as a set of transactions linked to a member ID, including initial deposit, and subsequent debits, credits, and interest yields; with each transaction accompanied by an action date and a clearance date. From the initial preprocessing step, the data for each member ID was updated on the clearance date, requiring that the balance for days with no transaction be filled using the last known balance. The running balance was calculated as a preprocessing step employing the clearance date using a custom workflow in SPSS modeler. Subsequent data processing and analysis was performed using the KNIME analytics platform.

Initial analysis of the data was performed employing K-means clustering, by using the average savings per account for each month over the sample window. Because individuals had differing time windows in which their accounts were open from October 2012 to July 2014, missing information was assigned a balance of 0. The dataset was cross-referenced with a separate dataset of customers that had also taken out a mobile based loan, which further included 30 features relating to their mobile money and airtime usage, for example the number of mobile money transactions or the average airtime top up amount. The total number of sample having both a loan product and a mobile savings account totaled 866229 customers over the two-year window. For the population that took out loans the K-means clustering for average monthly savings was repeated including 30 of the features from the mobile money usage in order to examine behavior related to default rates. This set of features was selected since previous work by our group had found them to be good descriptors for credit worthiness.

In order to avoid further artificial distortion of results due to varying time windows, the overlapping population sample was reexamined using the loan disbursement date as a reference point, examining 90 days before and after loan disbursement. Clustering was performed initially using the average daily balance. A separate analysis was also performed employing a longitudinal K-means algorithm, where every day in the time series constituted its own feature. Subsequent examinations for the longitudinal data are to be performed using latent class mixed model packages as implemented in R. This latter method allows for the clustering of a population along prechosen patterns, facilitating comparison for users that might have similar behavior but significantly different absolute account values.

III. Results

Figure 1 shows the trajectory for 5 calculated clusters, indicating the different levels of stratification that exist in the data and the number of users belonging to each cluster. The initial K-means clustering over the two-year period reveals that the majority of users are centered around 1000KsH ($10) in their account. The data indicates that the typical user will keep low levels of capital in this type of account, with less than 5% maintaining balances in excess of $100, and the combined savings of this minority representing 30% of the total capital of in the mobile savings accounts. The initial increase in balance observed for the clusters is mainly attributed to the inclusion of new users signing up to the mobile savings accounts.

The initial modeling results clearly indicate that users of this type of accounts have limited resources invested in them. Although this can be due to a preference for keeping cash in hand or in other investments, the fact that 90% of the regional population subsist on less than $10/day suggest that typical users have limited resources to begin with. It is also important to note that this type of account is mandatory for any person that intends to request a mobile based loan; having to maintain a minimum deposit of 1KSh.

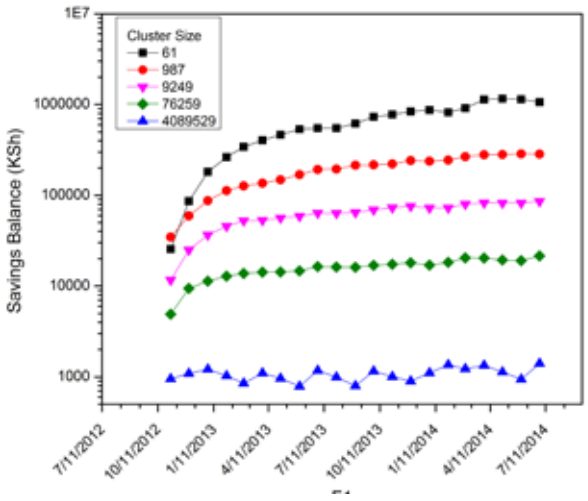


Figure 1 Average monthly balance for clusters calculated using K-means. The number of users per cluster is shown in the legend.

Figure 2 shows the result of calculating K-means clusters for the overlapping sample population that had both savings accounts and had taken out a loan during the same time window. The results are calculated using not just the monthly average saving for each individual, but the full set of 30 features related to airtime and mobile money usage. The inclusion of these features reveals that there is a correlation between the running balance and the likelihood of defaulting on a loan, with lower balance holders being almost twice as likely to default as the highest balance account holders. Plotting the data again reveals that the k-means algorithm simply stratifies the data, with the majority of users belonging to the low balance clusters.

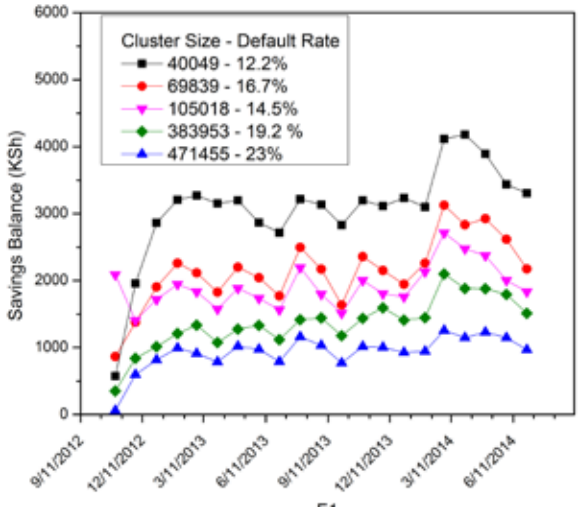


Figure 2 Average monthly balance for clusters calculated using k-means with 30 added features relating to mobile money and airtime usage. The number of users and default rate for each cluster is shown in the legend.

Because the prior modeling relied on average monthly savings with no distinction for the length of time that accounts were open for, the loan disbursement date was used a reference point for further analysis (Figure 3 and 4). Initially k-means clustering was performed employing the daily average savings for the 90 days prior to taking out a loan. Using this reference point reduced the differences in history lengths, but the results still indicate that a significant number of users do not have an established deposit account of the requisite 90 days. This lack of established deposit account is identifiable as the steep difference in the balance 90 days prior to having taken out a loan before a relative plateau is reached. From figure 3, it is important to note that prior to taking out a loan the account balance of all clusters decreases. With the time to onset and sharpness of decline potentially giving banks an early warning that a customer will take out a loan. This clustering was performed without considering the additional features of mobile money usage, but post cluster analysis indicates that the difference in default rates was not as marked as in Figure 2, with the exception that the lower balance clusters still have the highest default rates.

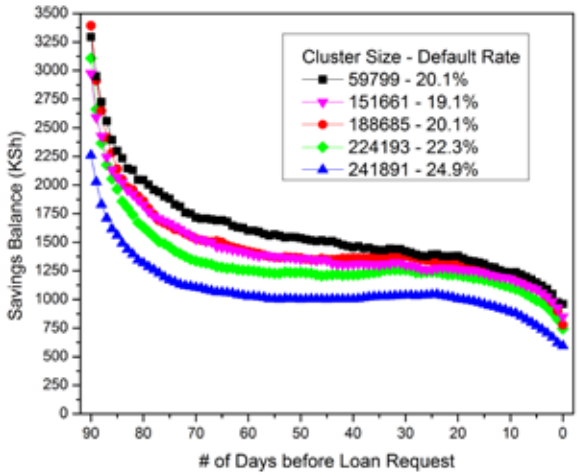


Figure 3 Average daily balance for the 90 days prior to taking out a mobile based loan for clusters calculated using k-means. The legend shows the cluster size and default rate for each cluster.

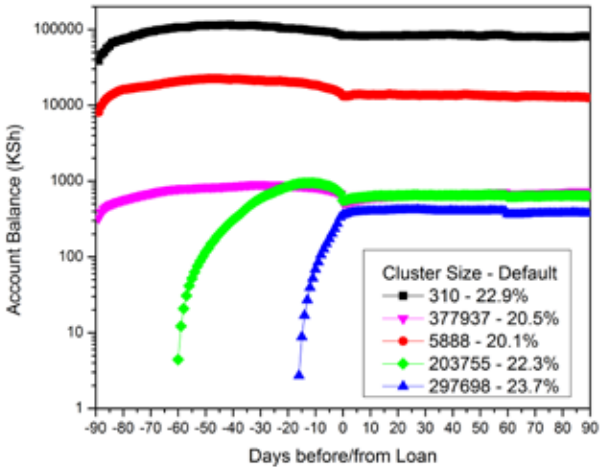


Figure 4 Longitudinal k-means clustering for 90 days before and after a loan disbursement of mobile deposit account holders. Associated cluster size and default rates are displayed in the legend.

K-means longitudinal analysis spanning 90 days before and after a loan disbursement further clarify the savings behavior of individuals that took out a loan. Figure 4 shows that low balance account holders can be separated into three main clusters with varying history lengths. The analysis indicates that individuals with the shortest deposit history prior to taking out a loan tend to maintain the lowest account balances, while also having the highest default rates. Furthermore, the results clarify that the balance dip prior to taking out a loan is linked to users that have over 40 days of account history. Examining the balance after loan disbursement indicates that on average account activity is minimal, with no significant change observed for any of the clusters. Lastly, exclusion of features related to mobile money or airtime usage results in small differences in the default rate between clusters. This suggest that rigorous modeling of credit worthiness still depends on a lengthier list of features about customer behavior, with saving accounts usage providing some weaker insights for default rates.

Ongoing work focuses on understanding how savings data can be leveraged for predicting credit worthiness and associated credit scores. The longitudinal k-means method is to be repeated with additional inclusion of mobile money and airtime usage. Further work will also explore patterns of behavior based on relative account balance changes, instead of the absolute account balances and will employ additional longitudinal methods, specifically latent class mixed models.

REFERENCES

[1] Nick Hughes and Susie Lonie "M-PESA: Mobile Money for the "Unbanked" Turning Cellphonrd inyo 24-hour Tellers in Kenya." Innovations: Technology, Governance, Globalization 2007 2:1-2, 63-81

[2] Etzo, Sebastiana, and Guy Collender. "The mobile phone 'revolution'in africa: Rhetoric or realty?." *African affairs* (2010): adq045.

[3] Porteous, David. "The enabling environment for mobile banking in Africa." (2006).

[4] Asongu, Simplice A. "How has mobile phone penetration stimulated financial development in Africa?." *Journal of African Business* 14.1 (2013): 7-18.

[5] Donovan, Kevin. "Mobile money for financial inclusion." *Information and Communications for Development* 61 (2012): 61-73.

[6] Cook, Tamara, and Claudia McKay. "How M-Shwari works: The story so far." *Consultative Group to Assist the Poor (CGAP) and Financial Sector Deepening (FSD)* (2015).

SESSION 2
MOBILITY

Route flow estimation using trips extracted from cellular network data

Nils Breyer[a], David Gundlegård[a], Clas Rydergren[a]

[a] *Department of Science and Technology (ITN)*
Linköping University
601 74 Norrköping, Sweden

*Corresponding author. Email: nils.breyer@liu.se, Phone +4611-36 32 80

Abstract. The signalling data in cellular networks provide means for analyzing the use of the transportation system. We propose methods that aim to reconstruct the used route through a transportation network from call detail records (CDRs) which are spatially and temporally sparse.

The route estimation methods are compared regarding the characteristics of the routes estimated. We also investigate the effect that the use of different route estimation methods has when used in a complete network assignment for a larger city. Using CDR data for Dakar, Senegal we show that the choice of the route estimation method can have a significant impact on resulting link flows. The data was made available by Orange/Sonatel within the framework of the D4D challenge (de Montjoye et al., 2014).

Keywords: cellular network data; route estimation; network assignment

1 Introduction

Using cellular network signaling data like Call Detail Records (CDRs) is a promising way of getting insights about the usage of the transportation system including all modes of travel. In order to gain insights about travel patterns, the raw data has to be filtered and processed. A big challenge with CDR data is that a record including a timestamp and the basestation that the user was connected to is only saved when the user actively makes a call or writes or receives a text message.

As only the movements of users are of interest, a necessary first step is to extracting trips from the data. The information which cells a user connects to during a trip gives an indication about which mode and which route that has been used. To infer the route for road-bound traffic, Fillekes (2014) discovered problems with using simple map-matching techniques to recover routes from cellular network data. For example just matching to a road because it is close to a base station tower that a user connected to is very unrealistic given the big area that base stations cover.

The evaluation of route inferring algorithms is usually made for individual routes or single OD-pairs only. For practical applications, like estimating infrastructure use from cellular network data, it is important to understand the impact of route estimation algorithms when embedded into a complete network assignment process. We therefore propose a pipeline to estimate flows for the individual links in a transportation network from cellular network data. We use the pipeline to evaluate the effects of different algorithms for estimating the used routes in a transport network from CDRs. We compare the characteristics of the estimated routes and the impact of different route estimation algorithms on traffic flows.

2 Computational pipeline

Several steps are necessary in order to estimate transport infrastructure use from cellular network data. In Gundlegård et al. (2016) a computational pipeline was proposed similar to the one in Figure 1. The first step is to extract trips from the data. Each trip has an origin and destination cell and a *cellpath*. The cellpath $p = (b_1, ..., b_n)$ is the list of base stations that a user connected to during the trip. The trip extraction method used for the experiments in Section 4 is the same as described in Gundlegård et al. (2016). The extracted trips are used in order to calculate a travel demand in each pair of cells, which is stored in an origin-destination matrix (OD-matrix). Currently, we use an unscaled OD-matrix generated by aggre-

gating the detected trips for one day. In order to estimate the actual flows in terms of vehicles or people, scaling would be necessary assuming that the data does not cover the whole population.

The cellpath routing estimates a route through the transportation network for a given cellpath. In the final flow distribution step, the travel demand given in the OD-matrix is distributed among the estimated routes and assigned to the transportation network. The cellpath-routing is implemented for road-bound traffic, though similar algorithms might be applicable to other modes as well. The result of the pipeline can be expressed either in form of route flows or link flows.

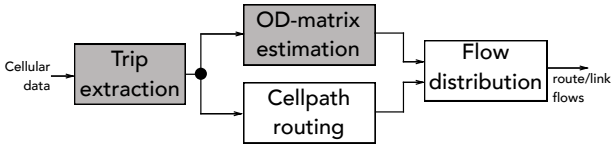


Figure 1: Computational pipeline to estimate traffic flows from cellular network data. The gray parts are not covered in this article.

3 Cellpath routing

We formulate the *cellpath routing problem* as follows: Given a *cellpath* $p = (b_1, ..., b_n)$ consisting of the base stations that a user connected to along a trip, find the route through the road network, that the user most probably took. Furthermore, the base station postions are known. As no other information about the base station configuration is available, a Voronoi diagram is used as an estimation of the coverage area of each base station. The link costs on the road network correspond to the free-flow travel time. The route estimation can be divided into two steps: First, a startnode s end an endnode t for the route inside the Voronoi cells of b_1 respectively b_n is selected. Second, a route through the road network is estimated between s and t. The selection of the start- and endnode for a route is the same for all algorithms presented in this section. We propose the following 4 algorithms that differ in the way the route between s and t is calculated:

- *Direct routing*: The route estimation between startnode s and endnode t is a single shortest path calculation on the given transportation network.
- *Waypoint Routing*: By selecting a waypoint in each Voronoi cell that appears in the cellpath, the route has to visit each cell that appears in the cellpath. The waypoints are selected such that the impact on the cost of the total travel time is minimized.
- *Magnetic Routing*: The link costs are lowered for links close to cells in the cellpath. The route between s and t is a shortest path calculation but using the modified link costs, which makes the route more likely to follow the cellpath. This algorithm is a simplified version of the idea proposed by Leontiadis et al. (2014).
- *Magnetic Waypoint Routing*: A waypoint is placed in a few selected cells of the cellpath to ensure that the route follows the shape of the cellpath approximately. Between waypoints Magnetic Routing is used in order to make the route likely to follow the cellpath between waypoints.

Table 1 shows a summary of the characteristics of the different algorithms. For the algorithms that require parameters, two parameter setups have been tested. Variant 1 heavily reduces the link costs within cells that appear in the cellpath but not around the cells in the cellpath, while variant 2 only cuts the cost by half and lowers the cost around the cells slightly.

4 Experiments & results

To compare the cellpath routing algorithms described in Section 3, we used CDR data for the city of Dakar, Senegal from the D4D challenge (de Montjoye et al., 2014) provided by Orange. Direct Routing generates the cheapest and also shortest routes (see Figure 2). This is because Direct Routing uses the shortest path between start- and endpoint, while all other algorithms potentially prolong the route in order to match the cellpath. Most impact on both route cost and route length had Waypoint Routing. Compared to Direct Routing, the average travel time increased by 26.7 %. Comparing a network assignment using

Table 1: Characteristics of the cellpath routing algorithms. "No. SPs" refers to the number of shortest path calculations. n is the number of cells in the cellpath for which the route is calculated.

	Direct Routing	Waypoint R.	Magnetic Routing	Magn. Wayp. R.
Waypoints	First/last cell	In every cell	First/last cell	"important" cells
Link costs	Original	Original	Lower in visited cells	Lower in visited cells
No. SPs	1	n	1	no. waypoints - 1

Magnetic Routing, variant 1 (see Figure 3) with Direct Routing, the most significant flow difference on a link in the city center (in the south), where a lot of traffic is shifted to another equivalent route in the grid-style road network.

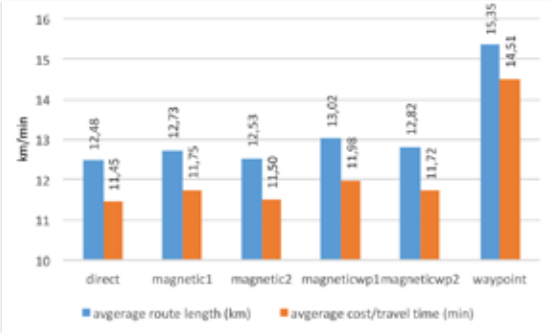


Figure 2: Average length and travel time of all routes calculated.

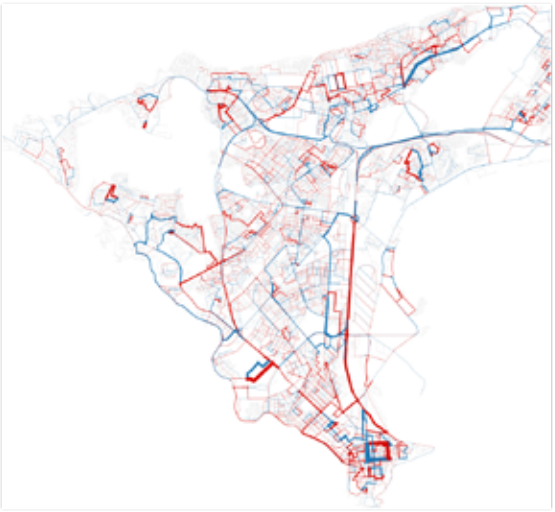


Figure 3: Absolute flow difference between a network assignment using Magnetic Routing, variant 1 and Direct Routing. Red: higher flow than i Direct Routing, blue: lower flow than in Direct Routing. The width indicates the difference.

5 Conclusions

Using the pipeline in Section 2, we can compare algorithms for each of the steps of the pipeline not only individually, but also regarding their impact on link flows in a network assignment for a whole city. A key component of the pipeline is the estimation of used routes in each OD-relation. The results show that fully relying on CDR data for routing (Waypoint Routing) is likely to create unrealistic detours that increase the flow estimations in the network significantly. Magnetic Routing has the potential to be more realistic, even though validation using other data is needed.

The experiments in Section 4 use a home/work based trip extraction algorithm which is described in Gundlegård et al. (2016). Many studies using cellular network data for traffic analysis use some kind of trip extraction procedure as the first step, but seldom a comparison of different methods and analysis of systematic properties of different approaches is made. Our aim is to include preliminary results of a comparison of different trip extraction methods in the conference presentation.

References

Y. de Montjoye, Z. Smoreda, R. Trinquart, C. Ziemlicki, and V. D. Blondel. D4d-senegal: The second mobile phone data for development challenge. *CoRR*, abs/1407.4885, 2014.

M. Fillekes. Reconstructing trajectories from sparse call detail records. Master's thesis, University of Tartu, 2014.

D. Gundlegård, C. Rydergren, N. Breyer, and B. Rajna. Travel demand estimation and network assignment based on cellular network data. *Computer Communications*, 95:29 – 42, 2016.

I. Leontiadis, A. Lima, H. Kwak, R. Stanojevic, D. Wetherall, and K. Papagiannaki. From cells to streets: Estimating mobile paths with cellular-side data. In *Proceedings of the 10th ACM International on Conference on emerging Networking Experiments and Technologies*, pages 121–132. ACM, ACM, 2014.

Improving human mobility prediction with geo-tagged pictures

M. G. Beiró, A. Panisson, M. Tizzoni, C. Cattuto[1]

1. ISI Foundation, Turin, Italy

1. Introduction

Predicting human mobility flows at different spatial scales and aggregation levels has interesting applications in areas such as urban planning, resource allocation and disease spreading control, for example. In this sense, the vast amounts of geolocated data currently generated by mobile devices as cellphones or digital cameras provide a unique opportunity for improving our understanding of human mobility processes, with immediate benefits.

In this work we analyze a set of 18 million timestamped, georeferenced pictures from Flickr, taken by 40,000 users in the U.S, which are part of the Yahoo Flickr Creative Commons 100M public dataset [1]. We show that by assimilating these data into well-established mobility models as the gravity model or the radiation model we can improve their predictions significantly.

2. Methodology and results

We use the Flickr georeferenced pictures to model two different types of human mobility flows: (a) daily commuting at the county level and; (b) flows in the air-transportation network. For each case, we have real flow patterns provided by the U.S. Census Bureau [2] (for the commuting flows), and the U.S. Bureau of Transport Statistics [3] (for the air-travel flows), which we use as ground-truth for our models.

We process the geotagged pictures of each user, ordered by the time in which they were taken. Each of these pictures is assigned to a geographical basin according to its coordinates. In the commuting network the basins are the counties, while in the air-travel network we define the basins from a Voronoi tessellation of the airports network. Then, we consider that a user makes a trip between two basins i and j whenever he takes two consecutive pictures in those basins. After this process, the network of Flickr users' flows involves 350,000 trips between airport basins and 520,000 trips between counties.

Now, we evaluate two well-known models of human mobility: (a) the radiation model [4], which considers human movements as diffusion processes that depend on the population distribution over the space, reproducing Stouffer's theory of intervening opportunities and; (b) the gravity model [5], in which the amount of people travelling between two locations is computed as directly proportional to some power of their populations, and decays as some power of the distance between them. We combine these models and our Flickr users flows into a hybrid model, using a stacked regression procedure expressed as:

$$\mathbf{H}(\alpha, \beta, \gamma, A, B; \mathcal{P}) = A \cdot \mathbf{B}(\alpha, \beta, \gamma; \mathcal{P}) + B \cdot \mathbf{F},$$

where $\mathbf{H} = (h_{ij})$ represents the matrix of flows between basins in our hybrid model, $\mathbf{B}$ represents the flows in the base model (gravity or radiation) and $\mathbf{F}$ represents our Flickr users flows matrix; $\alpha, \beta, \gamma, A, B$ are real-valued fitting parameters (α, β, γ only apply for the gravity model). We will fit the model by minimizing the following loss function L:

$$L = \| \mathbf{Y} - \mathbf{H}(\alpha, \beta, \gamma, A, B; \mathcal{P}) \|,$$

where $\| \cdot \|$ denotes the Frobenius norm.

In order to evaluate the predictions we adopted a machine learning approach. The model was fitted and cross-validated using a 10-fold cross validation procedure. In Table 1 we showed the performance of our

hybrid model compare to that of the base gravity or radiation models, in terms of the Pearson correlation between the real and predicted flows, and in terms of the determination coefficient.

Model	**Commuting**		**Air travel**	
	ρ	r^2	ρ	r^2
Gravity model	0.69	0.41	0.68	0.40
Radiation model (*)	0.78	0.60	0.47	-0.21
Flickr model	0.69	0.47	0.78	0.62
Hybrid model (Gravity+Flickr)	0.79	0.62	0.84	0.72
Hybrid model (Radiation+Flickr)	0.85	0.73	0.80	0.64

Cuadro 1: **Cross-validated model performance (10-fold cross-validation).** The table shows the performance of the hybrid models in terms of the Pearson correlation coefficient and the determination coefficient. We also display the results for the gravity model, the radiation model and the Flickr model alone. All the values were produced under a 10-fold cross-validation scheme, except for the radiation model (as noted by the asterisk), which is parameter free.

We observe that the assimilation of the Flickr traces into a hybrid model produced a substantial increase in the predictive performance both for daily commuting as for air travel in the U.S., as shown in the last two rows of Table 1. In particular, we observed that the incorporation of Flickr traces solved one biases of the gravity model, i.e., the underestimation of large flows across distant cities, as previously observed in [6]. A comparison between the real and estimated flows for the gravity model and the hybrid model assimilating the Flickr traces is shown in Figure 1.

3. Conclusions

In this work we investigated the quantitative effect that the use of geolocalized traces may have in the prediction of human mobility flows at different scales. We showed that the assimilation of geolocalized traces, as those collected from Flickr, can significantly improve the predictions of well-established human mobility models as the gravity model and the radiation model.

Additionally, we evaluated the out-of-sample true predictive power of the models by using a learning approach, and we assessed their applicability in two different spatial scales.

4. References

[1] Thomee B, Shamma DA, Friedland G, Elizalde B, Ni K, Poland D, Borth D, and Li L (2016) Yfcc100m: The new data in multimedia research., Comm.ACM 59(2), 64-73.
[2] Bureau of Transportation Statistics. Commuting (journey to work).
URL http://www.rita.dot.gov/bts/data$_a nd_s tatistics/index.html, Date of access$: 11/04/2016.
[3] US Census Bureau. Airline origin and destination survey.
URL http://www.census.gov/hhes/commuting/, Date of access: 11/04/2016.
[4] Stouffer SA (1940) Intervening opportunities: A theory relating mobility and distance. Am Soc Rev 5(6), 845-867.
[5] Alonso W (1976) A Theory of Movements: I, Introduction. Working paper No. 266. Institute of Urban Regional Development, University of California, Berkeley.
[6] Simini F, González MC, Maritan A, Barabási A-L (2012) A universal model for mobility and migration patterns. Nature 484(7392), 96-100.

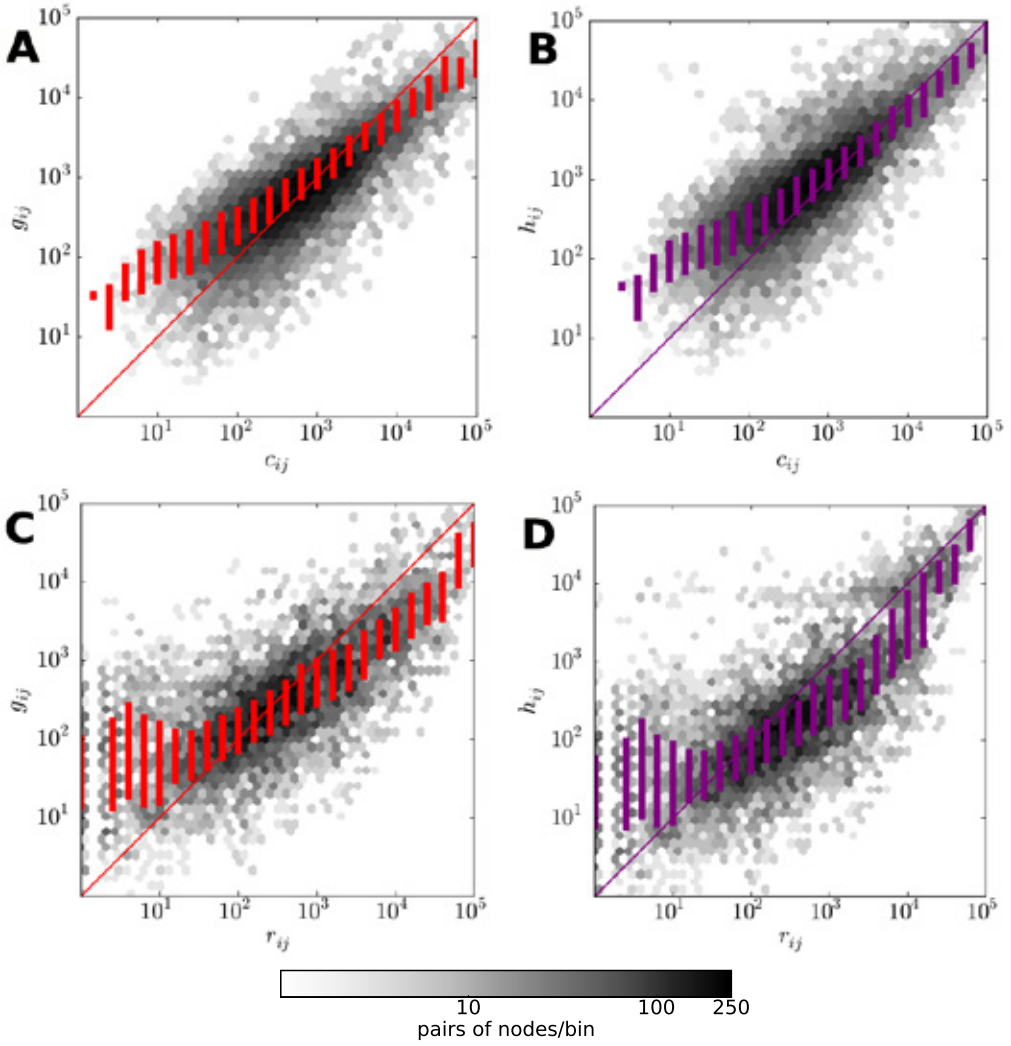


Figura 1: **Predictions for the gravity model and the hybrid (gravity+Flickr) model**. These 2D-histograms compare the predicted values against the real flow values for **(A, C)** the gravity model; **(B, D)** the hybrid model. Each point represents flows with some real/estimated flow value relation. Points color in a gray scale represents the frequency values. The boxplots correspond to the interquartile range within a bin. **(A, B)** U.S. Commuting network. **(C, D)** U.S. air travel network.

Inferring Commuting Statistics in Greater Jakarta from Social Media Locational Information from Mobile Devices

Imaduddin Amin, Ni Luh Putu Satyaning P.P, Yulistina Riyadi, and Jong Gun Lee
Pulse Lab Jakarta - United Nations Global Pulse
Email: {imaduddin.amin, ni.paramita, yulistina.riyadi, jonggun.lee}@un.or.id
Setia Pramana and Robert Kurniawan
Institute of Statistics, Jakarta, Indonesia
Email: {setia.pramana, robert}@stis.ac.id

ABSTRACT

Jakarta is the biggest city in Indonesia with a population of more than 10 million and a megacity with 1.38 million people commuting around Greater Jakarta [1]. In order for the Government of Indonesia to understand the commuting behaviors of citizens and better plan the transportation system in Greater Jakarta, the Indonesian Central Bureau of Statistics first conducted its commuting survey in 2014. In this paper, we produce commuting statistics from locational information on social media and show that the information from mobile phone apps is a promising source of data. It implies that the commuting behaviors from big data can allow the public sector more frequent statistics.

1 INTRODUCTION

It is known that about 1.38 million people commute daily in Greater Jakarta or Jabodetabek, which includes Jakarta, Bogor, Depok, Tangerang and Bekasi. The scale of the commuting population along with the population of Jakarta (about 10 million) has caused Greater Jakarta to face many urban challenges, including transportation. This led the Government of Indonesia to conduct the first commuting survey by the Indonesian Bureau of Statistics (Badan Pusat Statistik or BPS) in 2014.

Among many development issues, transportation is one of the key issues that directly affect citizens' daily lives. In many cases, a survey-based data collection method is widely used but one of its inherent drawbacks is that such an approach takes some time from the design of a survey to the release of its result. For instance, the first commuting survey in Greater Jakarta took longer than one year. Consequently, there are many attempts to use other types of data to produce similar statistics using different kinds of geo-located information such as GPS devices, sensors, social media and mobile phone data, in a shorter cycle [2, 3, 4, 5, 7, 8, 9].

In Indonesia, social media is recognized as a promising data source to understand the behavioral patterns of people, specifically as about 88.1 million people use the Internet and among them about 79 million people use the Internet for accessing social media[1]. Jakarta is often named as the Twitter capital of the world with 10 million tweets every day. As another example, in Greater Jakarta, between January and May 2014 for five months, about 40 million tweets were posted with GPS information.

In this paper, we test how the locational information from social media on mobile devices can reveal commuting patterns in Greater Jakarta. First we produce Origin - Destination statistics given 10 cities in Greater Jakarta from the entire GPS-stamped tweets posted by mobile devices, by identifying a set of people who commute in these areas. Second, we calibrate the initial result based on the population distribution and Twitter user distribution. Finally, we verify the result with the official commuting statistics. Our result confirms that geo-located tweets has the potential to complement official commuting statistics to fill information gaps.

2 DATA SET

2.1 Official Commuting Statistics

The Indonesian Central Bureau Statistics (BPS) conducted the first commuting survey in household level covering 13,120 households from 1,312 census blocks. This survey aims to understand the behavior of commuting users in 13 cities or regencies in Greater Jakarta which includes five Jakarta cities (Central, South, North, East, and West Jakarta), five satellite cities (Bogor, Bekasi, Depok, Tangerang, and South Tangerang), and three regencies (Bogor, Bekasi, and Tangerang). Greater Jakarta is a loosely defined term with few definitions. For instance, one defines it with 13 cities and regencies but another understands Greater Jakarta only with 10 cities. We use the latter, which is widely perceived by citizens, excluding three

[1] "E-Government Indonesia" (Source: Ministry of Communication and ICT) http://www.detiknas.go.id/wp-content/uploads/2016/08/PAPARAN-SNGIDC-2016-Bambang-Dwi-Anggono.pdf

regencies, i.e., Kabupaten (Regency) Bekasi, Kabupaten Tangerang, Kabupaten Bogor, which rather connect other close provinces such as Banten and West Java.

Table 1 shows a part of the commuting survey results particularly showing commuters from the 10 cities in Greater Jakarta to five cities in Jakarta. For instance, among the total number of commuters to the five cities in Jakarta, 2.31% of people commute from South Jakarta to East Jakarta.

ORIGIN	DESTINATION				
	SJ	EJ	CJ	WJ	NJ
South Jakarta (SJ)	-	2.31%	4.69%	1.88%	0.97%
East Jakarta (EJ)	5.33%	-	4.66%	1.62%	3.34%
Central Jakarta (CJ)	1.83%	0.86%	-	1.70%	1.20%
West Jakarta (WJ)	2.65%	0.46%	4.65%	-	4.19%
North Jakarta (NJ)	0.87%	1.14%	3.17%	1.82%	-
Bogor (Bo)	0.34%	0.27%	0.39%	0.25%	0.05%
Bekasi (Be)	3.36%	7.27%	3.48%	1.10%	1.73%
Depok (De)	7.40%	2.35%	2.37%	0.72%	0.54%
Tangerang (Tg)	2.62%	0.17%	1.62%	4.23%	0.45%
South Tangerang (ST)	6.13%	0.33%	1.89%	1.30%	0.27%

Table 1 Commuting Survey 2014 - Statistics (%) from 10 cities in Greater Jakarta to 5 cities in Jakarta

2.2 Twitter Data

We collect all GPS-stamped tweets posted in Greater Jakarta from a data firehose and specifically for this study, we use a set of tweets posted between January 1st 2014 and May 30th 2014, for five months, considering that the official commuting survey was conducted during the first quarter of 2014. Given all tweets in the Greater Jakarta area, we first filter tweets posted using mobile devices only and then map and use the locations of tweets to the administrative resolution. Finally, we use 38,491,430 tweets from 1,456,927 unique users.

3 METHODOLOGY

3.1 Inferring Origin-Destination Locations

First, per user, we infer two locations, Origin (Home) and Destination. It is worth noting that even though we finally produce city-level statistics, we process the locations of one's tweets at sub-district level, when determining the Origin and Destination cities.

- **Home Location** is inferred as the most tweeted sub-district location between 9pm and 7am in order to avoid using the locational information during commutes.
- **Destination Location** is determined as the most tweeted sub-district location during weekdays, excluding Home Location.

Using this approach, among 1,456,927 unique users who posted tweets in Greater Jakarta for five months from January 2014, we find the origin information of 877,054 users, and the origin and destination information of 305,761 users in sub-district level, which is about 2.8% (14%) of the whole population (the commuting population, respectively) in Greater Jakarta.

3.2 Calibrating the Initial Result

Once we map the Origin-Destination information at sub-district level to city level, we do a calibration based on the population data from 10 cities, due to the unequal penetration rates of Twitter in the cities. For instance, people in South Jakarta post tweets much more than people who are based in other cities in Greater Jakarta. After the calibration with Twitter penetration information[2], the cross-correlation score between two forms of statistics, official statistics and the statistics from our approach improved from 0.92 to 0.97.

4 RESULTS

The final result is presented in Table 3(a) and its chord diagram is shown in Figure 1. In Table 3(b) we show the rank difference between Table 2 and Table 3(a). For instance, the value for SJ ⇒ CJ is calculated as '0' because two ranks from two statistics are same as people in South Jakarta most commute to Central Jakarta than other three cities. Table 3(b) shows that our simple approach produces good results without significant differences. Even though three origin cities have three different rank pairs, it is worth mentioning that the values in the official statistics are originally close each other, and our results also present similar values. For instance, the values of De ⇒ EJ and De ⇒ CJ are 2.35 % and 2.37% in the official statistics (2.46% and 2.39%, respectively and in our results).

5 SUMMARY

We have shown that social media is a promising source of data to infer commuting statistics in Greater Jakarta. For future work, we will improve our simple approach with other calibration methods utilizing further information, such as demographic information.

[2] It is worth mentioning that we tried a calibration based on a basic gravity model but it does not produce a significant improvement.

ORIGIN	DESTINATION				
	SJ	EJ	CJ	WJ	NJ
South Jakarta (SJ)	-	2.85%	4.69%	1.42%	0.67%
East Jakarta (EJ)	5.49%	-	4.77%	0.85%	2.18%
Central Jakarta (CJ)	1.89%	0.87%	-	0.93%	0.87%
West Jakarta (WJ)	3.53%	0.66%	5.07%	-	2.49%
North Jakarta (NJ)	1.52%	2.05%	3.88%	2.21%	-
Bogor (Bo)	1.47%	0.74%	2.16%	0.48%	0.39%
Bekasi (Be)	3.64%	6.48%	4.26%	0.96%	1.53%
Depok (De)	6.06%	2.46%	2.39%	0.58%	0.41%
Tangerang (Tg)	2.99%	0.64%	3.01%	3.56%	0.69%
South Tangerang (ST)	3.99%	0.40%	1.64%	0.77%	0.25%

ORIGIN	DESTINATION				
	SJ	EJ	CJ	WJ	NJ
South Jakarta (SJ)	0	0	0	0	0
East Jakarta (EJ)	0	0	0	0	0
Central Jakarta (CJ)	0	0	0	0	0
West Jakarta (WJ)	-1	0	0	0	+1
North Jakarta (NJ)	0	0	0	0	0
Bogor (Bo)	0	0	0	0	0
Bekasi (Be)	0	0	0	0	0
Depok (De)	0	+1	-1	0	0
Tangerang (Tg)	-1	0	+1	0	0
South Tangerang (ST)	0	0	0	0	0

Table 2 (a) Commuting Statistics from Social Media (left) (b) Rank Comparison with Official Statitics

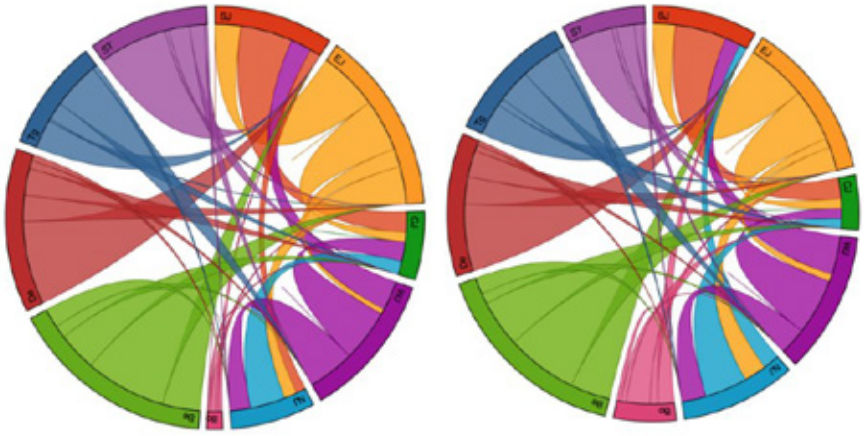


Figure 1 [Best Shown in Color] Official Commuting Flows (Left) and Statistics from Twitter (Right)

REFERENCES

[1] Jakarta Commuting Survey http://www.bps.go.id/index.php/publikasi/4416
[2] Palchykov V, Mitrović M, Jo HH, Saramaki J, Pan RK. Inferring human mobility using communication patterns. Scientific reports. 2014;4. doi: 10.1038/srep06174. pmid:25146347
[3] A Kurkcu, K Ozbay, EF Morgul, "Evaluating the Usability of Geo-Located Twitter as a Tool for Human Activity and Mobility Patterns: A Case Study for New York City"
[4] B Hawelka, I Sitko, E Beinat, S Sobolevsky, P Kazakopoulos, C Ratti, Geo-located Twitter as proxy for global mobility patterns
[5] Liu, W., Zamal, F.A., & Ruths, D., (2012) Using social media to infer gender composition of commuter populations.
[6] Swier, N., Komarniczky, B., & Clapperton, B. (2015) Using geolocated Twitter traces to infer residence and mobility. GSS Methodology Series No 41
[7] Gonzalez MC, Hidalgo CA, Barabási A-L. Understanding individual human mobility patterns. Nature. 2008;453(7196):779–782. doi: 10.1038/nature06958. pmid:18528393
[8] Palchykov V, Mitrović M, Jo HH, Saramaki J, Pan RK. Inferring human mobility using communication patterns. Scientific reports. 2014;4. doi: 10.1038/srep06174. pmid:25146347
[9] Jiang S, Fiore GA, Yang Y, Ferreira F Jr, Frazzoli E, González MC. A review of urban computing for mobile phone traces: current methods, challenges and opportunities. In Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing (p. 2). ACM. (2013, August).

An emergent strategy for characterizing urban hotspot dynamics via GPS data.

Antonio L. Alfeo[1], Mario G.C.A. Cimino[1], Sara Egidi[1], Bruno Lepri[2] and Gigliola Vaglini[1]
[1]*Department of Information Engineering, Università di Pisa, largo Lazzarino 1, Pisa, Italy*
[1]*Bruno Kessler Foundation, via S. Croce, 77, Trento, Italy*
luca.alfeo@ing.unipi.it, mario.cimino@unipi.it, s.egidi1@studenti.unipi.it, lepri@fbk.eu, gigliola.vaglini@unipi.it.

Keywords: Urban mobility, taxi-GPS traces, stigmergy, emergent paradigm, hotspot.

INTRODUCTION AND MOTIVATION

The increasing volume of urban human mobility data arises unprecedented opportunities to monitor and understand city dynamics. Identifying events which do not conform to the expected patterns can enhance the awareness of decision makers for a variety of purposes, such as the management of social events or extreme weather situations [1]. For this purpose GPS-equipped vehicles provide huge amount of reliable data about urban dynamics, exhibiting correlation with human activities, events and city structure [2]. For example, in [3] the impact of a social event is evaluated by analyzing taxi traces data. Here, the authors model typical passenger flow in an area, in order to compute the probability that an event happens. Then, the event impact is measured by analyzing abnormal traffic flows in the area via Discrete Fourier Transform. In [4] GPS trajectories are mapped through an Interactive Voting-based Map Matching Algorithm. This mapping is used for off-line characterization of normal drivers' behavior and real-time anomalies detection. Furthermore, the cause of the anomalies is found exploiting social network data. In [5] the authors employ a Multiscale Principal Component Analysis to analyze Taxi GPS data in order to detect traffic anomalies. The most of the methods in the literature can be grouped into four categories: distance-based, cluster-based, classification-based, and statistics-based [6]. Typically, due to the complexity of this kind of data, the modeling and comparison of their dynamics over time are hard to manage and parametrize [7]. In this paper, we present an innovative technique aimed to handle such complexity, providing a study of urban hotspot dynamics.

APPROACH DESCRIPTION

The developed approach is based on stigmergy, a mechanism belonging to the emergent paradigms. Emergent paradigms allow to avoid the explicit modeling of a system, which works only under the assumption formulated by the designer. Emergent paradigms offer model-free computational approach, characterized by adaptation, autonomy and self-organization of data [8]. In particular, with the emergent mechanism based on computational stigmergy, each sample position is associated to a digital pheromone deposit (mark). Marks are defined in a three-dimensional space, and characterized by *evaporation* over time, i.e. a progressive decay of mark intensity. Marks aggregate according to their spatio-temporal proximity, forming a stigmergic trail. As an effect, the evaporation is counteracted while new marks arrive in a given region. Thus, aggregation and evaporation can produce an emerging mechanism which acts as an agglomerative spatio-temporal clustering with historical memory. Employing the principle of stigmergy, we exploit positioning data to identify high-density areas (*hotspots*) within a city and to analyze their activity (presence of people) over time. Fig. 1 shows the positional stigmergy applied to hotspot identification. First, input data undergo the *smoothing* process, which focuses the analysis, highlighting relevant dynamics while removing insignificant activity levels. At a periodic time instant, a mark is released in a computer-simulated spatial environment (stigmergic space). The mark intensity is proportional to the number of people. Marks aggregation forms a trail, which progressively evaporates. For a given threshold of trail intensity, areas with intensity higher than the threshold correspond to hotspots. As an example, Fig. 2 shows the hotspots identified in Manhattan (New York). Their locations correspond to: East Harlem - Upper East Side (A), Midtown East (B), Broadway (C), East Village -Gramercy - MurrayHill (D), Soho - Tribeca (E), Chelsea (F) and Time Square – Midtown West - Garmet (G).

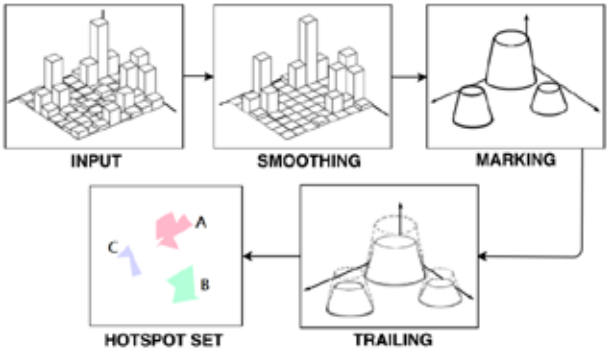


Figure 1. Positional stigmergy for hotspot identification.

Figure 2: Hotspots identified in Manhattan (2015).

In order to characterize the dynamics of the identified hotspots, the activity of each hotspot generates a mono-dimensional time series for every observation day. In a given time series, what is actually interesting is not the continuous variation of the activity over time, but the transition from one type of behavior to another. Each type of hotspot activity behavior should be defined by an expert in the field, in order to be general and reusable for many hotspots and many cities. More formally, each type is called an *archetype*, since it is an ideal time series segment defining a behavioral class of the hotspot activity. An example of archetype is *rising activity*, which means that the hotspot is moving from an ordinary activity level to its highest activity level. Fig. 3a and Fig. 3a' show such archetype and a real example of time series, respectively.

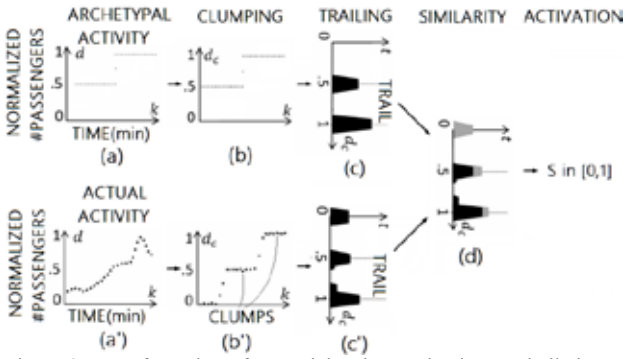


Figure 3. Transformation of an activity time series into a similarity time series with respect to an archetype.

In order to assess the match between an archetypal behavior and the current hotspot behavior, the time series are processed in sequential stages (Fig. 3). The initial samples first undergo the *clumping* process, which is a soft discretization of samples with respect to a set of levels of interest for any archetype (Fig. 3b and Fig. 3b'). Subsequently, in correspondence of each sample value, a trapezoidal *mark* is released (Fig. 3c and Fig. 3c'). Marks aggregate over time in a *trail*. Trail intensity is subject to evaporation. As an effect, a trail captures the spatiotemporal behavior of the time series in the short-term. A degree of similarity between the archetype trail and actual activity trail is then computed (Fig. 3d). Finally, to better sharpen the similarity value against the other archetypes, an activation function is applied to enhance only relevant similarity values, while removing insignificant values.

The overall processing schema is called *stigmergic receptive field* (receptive field for short), because it is receptive to a specific archetype, and it takes inspiration from the neurocomputing domain. A receptive field should be properly parameterized to effectively assess the archetypal behavior. For example, short-life marks evaporate too fast, preventing aggregation and pattern reinforcement, whereas long-life marks cause early activation. For this purpose, the receptive field is equipped with a parametric adaptation mechanism, based on the differential evolution algorithm [9]: parameters are adjusted by minimizing the mean square error over a set of annotated sample signals.

Since any real signal is usually similar to more than one archetype, a collection of receptive fields specialized on different archetypes is arranged to make a *stigmergic perceptron*, i.e., a connectionist topology whose architecture is represented on the left of Fig. 4. More specifically, the archetypes are ordered for increasing activity of the hotspot: a) *asleep*, b) *falling*, c) *awakening*, d) *flow*, e) *chill*, f) *rising*, and g) *rush-hour* hotspot activity. The stigmergic perceptron combines linearly the similarity values provided by all receptive fields, providing a value between zero and *N*-1, where *N* is the number of archetypes, called *activity level.*

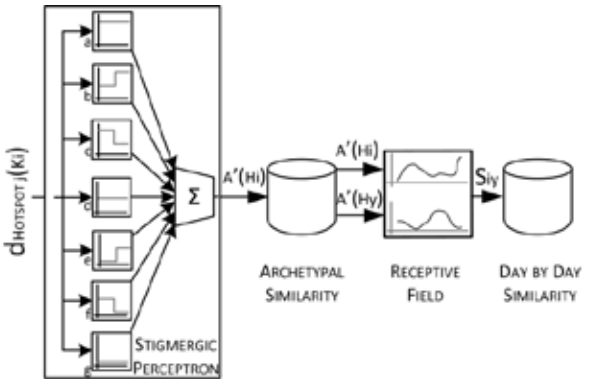


Figure 4: System architectural overview.

As a result, the stigmergic perceptron provides a new time series of activity levels for a given time series of activity samples. In order to compute the overall similarity between two days of hotspot activity, a further receptive field is used. Such receptive field compares the time series of two activity levels corresponding to two days. This measure of similarity is used to identify anomalies in the hotspot activity. For this purpose, we assume that during days characterized by an anomaly, the hotspot activity level is dissimilar with the activity level of equivalent days. Two days are considered *equivalent* if they fall on the same day of the week and in the same month. In order to detect the anomaly, we define the *normality index* of a given day as the average similarity of that day with its equivalent days. The normality index is 1 (or 0) if the day is perfectly similar (dissimilar) to its equivalent days.

The overall architecture allows the incremental detection of perturbations on the city routine that can involve the hotspots for a variety of reasons. As an example, the next section shows some experimental results on anomalies caused by adverse weather conditions.

EXPERIMENTAL RESULTS AND DISCUSSION

The analysis is based on taxis OD (Origin-Destination) traces provided by Taxi and Limousine Commission of New York City [10], which contains information about all medallion taxi trips from 2009 to 2016. Data are spatio-temporally discretized in bins characterized by 10 ft. wide and 5 minutes duration. In order to focus on urban hotspot dynamics the number of passengers in each bin, in terms of both pick-up and drop-off is first considered. Subsequently, the min-max normalization is applied. As a case study we focus on anomalies caused by an adverse weather condition. Specifically, we analyze Manhattan in January 2015. In January 26-27, 2015 many mobility issues, caused by a blizzard, were reported [11]. We refer this knowledge as the ground truth. In Fig. 5 the activity level and the corresponding stigmergic trail have been generated analyzing the activity in hotspot G in every Tuesday of January 2015. Here, it is apparent that the January 27 is characterized by an anomaly, for the notable differences with the other days both in terms of activity level (left side) and stigmergic trail (right side). A measure of these differences is based on the similarity values supplied by the receptive field, used to calculate the normality index of each hotspot. In Fig. 6 and Fig.7, a radar chart shows the normality indexes for all hotspots in each Tuesday and Monday of January, respectively.

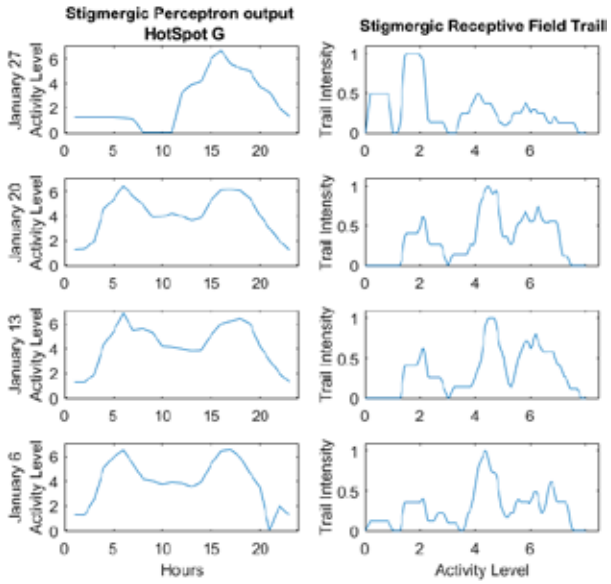


Figure 5. The activity levels on Hotspot G (on the left) during each Tuesday of January 2015 and the corresponding trails (on the right) which feed by the second level receptive field.

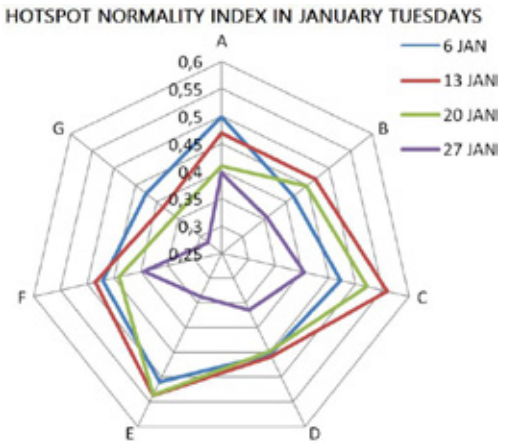


Figure 6. Normality index of January Tuesday for each Hotspot.

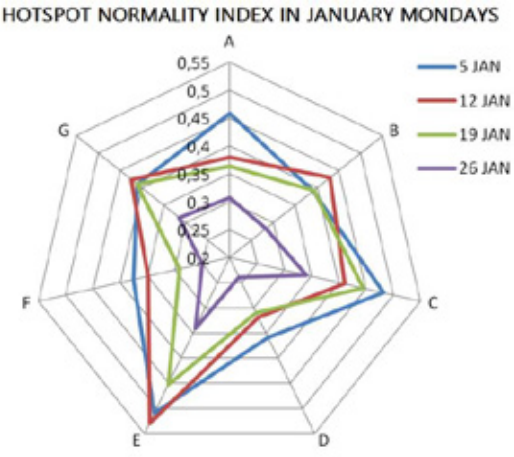


Figure 7. Normality index of January Mondays for each Hotspot.

Since a blizzard is a city-wide event which affects urban activity in the whole city, any hotspot during 26 and 27 January (purple line) results in the lowest normality index values. Table 1 shows, for equivalent days, the average normality index calculated between all hotspots. Again, it is apparent that the lowest values are related to the blizzard.

Table 1. The average of the normality index, over all city hotspots.

Mondays	05 Jan	12 Jan	19 Jan	26 Jan
Average Normality	0,45	0,42	0,38	**0,28**

Tuesdays	06 Jan	13 Jan	20 Jan	27 Jan
Average Normality	0,46	0,48	0,45	**0,35**

Although the study focuses on a blizzard, the analysis can be carried out on other events. Moreover, our approach provides a similarity measure that may be used with clustering techniques. Clustering can discover a structure in data and may generate data-driven prototypes of normal days, which are more effective than calendar-driven days. For this reason, to adopt a clustering technique is considered a key investigation task for future work.

REFERENCES

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

[2] Veloso, M., Phithakkitnukoon, S., & Bento, C. (2011, September). Urban mobility study using taxi traces. In Proceedings of the 2011 international workshop on Trajectory data mining and analysis (pp. 23-30). ACM.

[3] Zhang, W., Qi, G., Pan, G., Lu, H., Li, S., & Wu, Z. (2015). City-scale social event detection and evaluation with taxi traces. ACM Transactions on Intelligent Systems and Technology (TIST), 6(3), 40.

[4] Pan, B., Zheng, Y., Wilkie, D., & Shahabi, C. (2013, November). Crowd sensing of traffic anomalies based on human mobility and social media. In Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (pp. 344-353). ACM.

[5] Kuang, W., An, S., & Jiang, H. (2015). Detecting traffic anomalies in urban areas using taxi GPS data. Mathematical Problems in Engineering, 2015.

[6] Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM computing surveys (CSUR), 41(3), 15.

[7] Castro, P. S., Zhang, D., Chen, C., Li, S., & Pan, G. (2013). From taxi GPS traces to social and community dynamics: A survey. ACM Computing Surveys (CSUR), 46(2), 17.

[8] Vernon, D., Metta, G., & Sandini, G. (2007), A Survey of Artificial Cognitive Systems: Implications for the Autonomous Development of Mental Capabilities in Computational Agents. IEEE Transactions on Evolutionary Computation, 11(2), 151-180.

[9] Cimino, M. G., Lazzeri, A., & Vaglini, G. (2015, June). Improving the analysis of context-aware information via marker-based stigmergy and differential evolution. In International Conference on Artificial Intelligence and Soft Computing (pp. 341-352). Springer International Publishing.

[10] *NYC.gov*, the official website of the City of New York: Taxi and Limousine Commission (TLC) Trip Record Data, http://www.nyc.gov/html/ tlc/html/about/trip_record_data.shtml

[11] Weather NYC: Thousands of transatlantic travellers face serious disruption caused by New York winter storm 'Juno'. The Independent. January 26, 2015.

Inferring customer visitors by means of WiFi crumbles

Fabio Pinelli, Carlo Ceriotti and Emilia Traficante Cloud4wi.com Inc., 363 Clementina Street
San Francisco, CA 94103
Email: {fpinelli,cceriotti,etraficante} @cloud4wi.com

In the online world, it is easy for businesses to gather valuable information about their customers, such as: who they are, what theyre searching for, and how much time they spend browsing. Brick-and-mortar locations have traditionally been at a disadvantage when it comes to knowing customers, however, because its nearly impossible to know each person that walks into a venue. But as the demand for guest Wi-Fi grows, businesses are realizing they can achieve insights comparable to what theyd get online through analytics collected from Wi-Fi networks something that was never before possible. Businesses can make data-driven operational and strategic decisions and cultivate customer relationships by analyzing foot traffic, evaluating customer behavior, and understanding where visitors spend most of their time and how they move within the physical space of a venue. By having this information, businesses can:

- Monitor marketing campaign performance
- Position displays and products where there is significant foot traffic
- Optimize business hours
- Allocate support and security services more effectively

For these reasons, it is becoming very crucial to enhance the analysis of the data gathered from a guest Wi-Fi network infrastructure. Once this has be done, then retailers and store managers can deliver new accurate services to the right customers at right time. In particular, among all the detected devices is necessary to understand which ones belong to a real customer and which ones belong to a passerby. In this context, we propose a new machine Learning framework, Smart Presence ANalyicts (SPAN) that leverages the Wi-Fi data logs in order to distinguish *visitors* from *passersby*, so that new services can be designed and adopted using the insights extracted by means of SPAN.

Several technologies are already available to count and track people in store, such as Monocular (Single Lens) and 3D Stereo Video, Thermal Imaging, Infrared Beams, and BLE beacons. However, the Wi-Fi opens an unique opportunity for retailers to establish a direct communication channel with their customers.

Each technology has its own set of challenges and benefits. The difference between the solutions also includes the relationship between the capture technology and data integration requirements. Moreover, each solution has its own unique concepts of accuracy and data quality.

Monocular devices capture images through a single lens camera. The image can be processed within the sensor itself, sent to an in-store server for analysis, or processed at a central server. Due to costs and bandwidth requirements, most counters are At-the-Edge (computer and camera) devices and send only metric data to the central server. Some methods adopting this methodology are discussed in the following papers: [6], [7], [9], [2].

The principle behind 3D Stereo Vision is the assumption that accuracy starts with capturing the most high-resolution camera and creating a three-dimensional view of the tracked object. The key challenge for stereo vision is cost, since the counting solution requires special devices and cannot use standard cameras. Some works investigating such technology are presented in [8], [10], [1].

Thermal imaging works by detecting emissions from moving targets, by locking on to the targets, and then tracking them within the sensors field of view. Thermal imaging ignores the background features and focuses only on the moving object. As a result, thermal technology is not sensitive to light and allows the sensors to function well in challenging conditions such as the fluctuations from darkness to bright light. Examples of methods are described in: [4], [3], [5].

Infrared beams are mounted either sideways on the gate door or top-down from the ceiling. These low-cost devices are simple to install and setup. The device sends a direct infrared beam, and counts the person moving across the doorway once the beam is broken. Unfortunately, beam counters suffer from accuracy challenges [11].

Differently from these methods, the insights can be extracted from the Wi-Fi data gathered costless by the network infrastructure. This opens an unique opportunity for retailers, and store managers to enable a direct communication channel with their customers.

In our case study, data are collected by several access points (APs) installed inside the venue. Each request probe, received by an AP, is logged into the system with this information: MAC address of the sending device, the signal strength, the x,y estimated location, the status of the device: associated or not. The association is the process that enables the client the actual access to the WLAN network. It is the same like plugging the cable into the wired network. When a device is associated to one AP, it does not indicate necessarily that the user is logged into the Guest Wi-Fi network. Notice that the system receives as raw data log separately from each AP, therefore, an aggregation step is needed to recover information generated at device level.

The framework work-flow is reported in Figure 1. SPAN

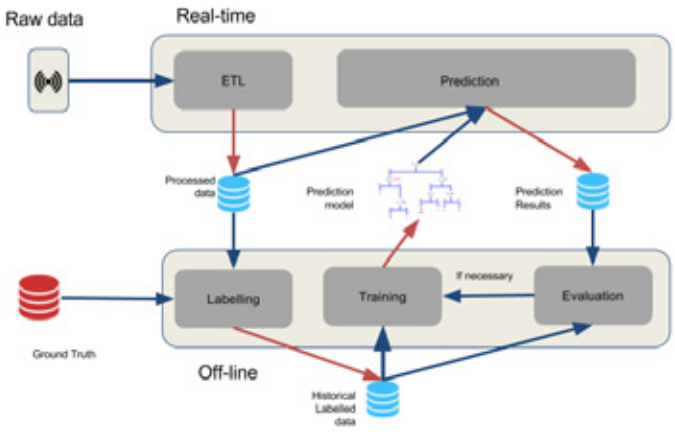


Fig. 1. The SPAN data work-flow, from the raw data to the predictions based on models extracted from historical data.

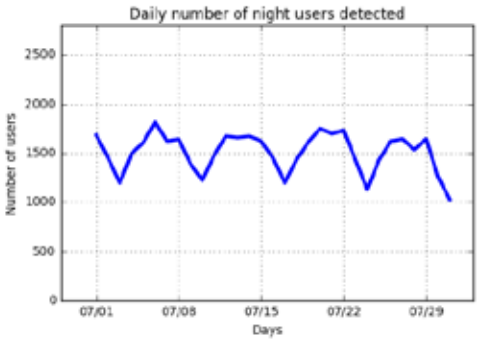
is composed of two main components, one working on (near)real-time and the other working in a batch mode. The former receives logs from the APs and it performs an ETL step. It aggregates probes sent by a certain MAC received contemporary by more than one AP and it extracts the features used by the machine learning module. For each aggregated probe, we extract cumulative and instantaneous features. The cumulative intend to describe the behavior of a specific MAC until that instant. For example, the number of probes sent by that MAC, the number of probes with associated status, etc. In this way, we want to model the historical behavior of the users in a sort of visit sessions. Instead, the set of instantaneous features gives a snapshot of the current status of the device, the APs which have received the probe, the RSSI level, etc. The aggregated data with relative features are stored in the Processed Data database. In the real-time component, the processed data are given in input to the prediction step. For each new aggregated probe, the classification model is applied to classify the probe as *visitor* or *passerby*.

The SPAN batch component includes 3 steps: labeling, learning, and evaluating. The labeling takes in input the processed data (aggregated and with features) and it uses the ground truth to label – if possible – each probe as passerby or visitor. In the ground truth we know which devices are passersby, visitors or stationaries. The ground truth is not give, but it is based on some assumptions. The stationary devices are the ones with probes for 10 hours in at least half of the days in the historical dataset. We consider the best archetype of passersby the users with probes sent only during the closing time of the store, i.e. from 0 to 7 and from 21 to 24. All the probes of these devices are labeled as passerby. Instead, as visitor prototype we use the behavior of users connected to the guest Wi-Fi. We labeled as visitor the device probe requests sent during a Guest Wi-Fi session. The definition of visitor is quite restrictive, and it covers few users but it represents a starting point to evaluate the effectiveness of the methodology. The labeling defined as such generates an unbalance datasets. Moreover, not all the received probes get a label, only the ones that belong to night, and guest Wi-Fi devices. The labeled data are given as input to a learning loop that includes the training and the evaluation steps. The training phase creates prediction models. In this paper we used decision trees. The evaluation step computes some quality measures of the predictions, for instance: $precision$, $recall$, $f1\_score$, among the others. When the performance of the model is too poor then another training session is executed.

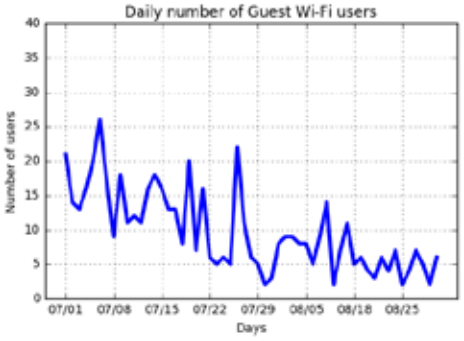
For our tests, we used a dataset of a retail store with 7 distinct Access Points. The store is located in the city center. The dataset covers a period of two months: July and August 2016. During this periods, we counted 5966731 aggregated probes sent by 750176 devices. In the experiments we used July as training set and August as test set.

We detect 15 stationary devices, such as devices observed in at least 50% of the days for at least 10 hours each day. We extracted the users observed only during the night with a daily basis. The results of this process are reported in Figure 2a where we can observe a sort of weekly pattern with evident drops on Sunday nights. All the request probes of night users are labeled as passerby.

The guest Wi-Fi logged-in users are reported in Figure 2b. It is easy to see the different order of magnitude of the two sets. We recall that the probes of a user falling into a guest Wi-Fi session are labeled as visitor, passerby otherwise. This difference is reflected into an unbalanced training set. In fact, we have 10522 probes labeled as visitor, 153339 as passerby, and 2881709 probes without any label. To deal with the unbalance nature of the dataset, we tested different down-samples of the passersby probe set proportionally to the number of visitor probes. The input parameters of the training phase are: the proportion between the number of visitors and passerby; the minimum depth of the decision tree; and the number of features that can be used contemporary to split the tree. Preliminary results are shown in Figures 3a where we reported the $precision$, $recall$ and $f1\_score$ values for the

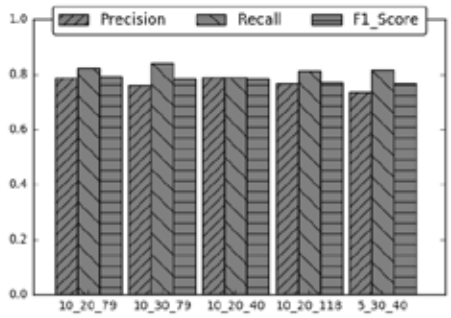


(a) Night users

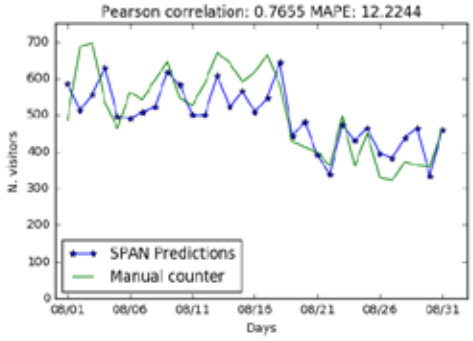


(b) Guest Wi-Fi users

Fig. 2. The figure a) reports the daily number of night users detected to compute the ground truth. The second shows the number of users using the guest Wi-Fi in each day.



(a)



(b)

Fig. 3. The figure a) reports the performance of different models in terms of $precision$, $recall$, and $f1\_score$. Figure b) shows the daily number of users detected with our method and a manual counter.

top 5 models in terms of $f1\_score$ testing the models against labeled data (i.e. only the part of the data getting a label). The input parameters of the model are reported as labels on the graph. Moreover, we evaluate the performance of best model on the totality of the data. We compute the Pearson correlation and Median Average Percentage Error (MAPE) of the daily number of users with at least one probe predicted as visitor, against a manual counter. The results are reported in Figure 3b. The MAPE has been computed as follow: i) we computed an average ratio between the predicted values and the real ones; i) we scale the predicted values accordingly to the ratio; iii) we computed the MAPE. We obtained a Pearson Correlation equal to 0.76 and a MAPE equal to 12.12%.

As general trend, figures show very promising results considering the simple adopted model and the loose definition of visitors.

The presented methodology leverages Wi-Fi probe request data to distinguish passersby customers from real visitor. The results confirm that goodness of the approach. Nonetheless, several directions are still open to be investigated: i) adopt a more complex machine learning model; ii) adopt a more accurate definition of visitors, iii) test the methodology on multiple locations.

References

[1] S. Bahadori, L. Iocchi, G. R. Leone, D. Nardi, and L. Scozzafava. Real-time people localization and tracking through fixed stereo vision. *Applied Intelligence*, 26:83–97, 2007.

[2] Homa Foroughi, Nilanjan Ray, and Hong Zhang. Robust people counting using sparse representation and random projection. *Pattern Recogn.*, 48(10):3038–3052, October 2015.

[3] Weiming Hu, Tieniu Tan, Liang Wang, and S. Maybank. A survey on visual surveillance of object motion and behaviors. *Trans. Sys. Man Cyber Part C*, 34(3):334–352, August 2004.

[4] Miklas S. Kristoffersen, Jacob V. Dueholm, Rikke Gade, and Thomas B. Moeslund. Pedestrian counting with occlusion handling using stereo thermal cameras. *Sensors*, 16(1):62, 2016.

[5] Suren Kumar, Tim K. Marks, and Michael Jones. Improving person tracking using an inexpensive thermal infrared sensor. *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 00(undefined):217–224, 2014.

[6] Damien Lefloch, Faouzi A. Cheikh, Jon Y. Hardeberg, Pierre Gouton, and Romain Picot-Clemente. Real-time people counting system using a single video camera, 2008.

[7] Bin Li, Jian Zhang, Zheng Zhang, and Yong Xu. A people counting method based on head detection and tracking. *2014 International Conference on Smart Computing (SMARTCOMP)*, 00(undefined):136–141, 2014.

[8] Xiaobai Liu. Multi-view 3d human tracking in crowded scenes, 2016.

[9] Venkatesh Bala Subburaman, Adrien Descamps, and Cyril Carincotte. Counting people in the crowd using a generic head detector. In *Proceedings of the 2012 IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance*, AVSS '12, pages 470–475, Washington, DC, USA, 2012. IEEE Computer Society.

[10] Xiaogang Wang. Intelligent multi-camera video surveillance: A review. *Pattern Recogn. Lett.*, 34(1):3–19, January 2013.

[11] Hong Yang, Kaan Ozbay, and Bekir Bartin. Investigating the performance of automatic counting sensors for pedestrian traffic data collection. In *Proceedings of the 12th World Conference on Transport Research, Lisbon, Portugal*, volume 1115, 2010.

SESSION 3
SOCIAL NETWORK

Age disparities in ethnic segregation: a study of activity spaces using the Call Detail Record dataset

Siiri Silm, Veronika Mooses, Rein Ahas

Department of Geography, University of Tartu, Estonia

E-mail of corresponding author: Siiri.Silm@ut.ee

Ethnic segregation or, in other words, the spatial separation of ethnic groups is a widespread phenomenon. The emergence and persistence of segregation is influenced by various geographical, cultural, political, and economic factors (Musterd 2005). In theory, the differences between a minority group and the majority population should decrease with time and the minority group should assimilate into the surrounding society. The excessive attention paid to segregation studies has concentrated on places of residence and workplaces, while in recent years the interest has been in expanding into people's entire activity space (Farber et al, 2012; Järv et al, 2014; Silm and Ahas, 2014; Wang and Liu, 2016; Tammaru and van Ham, 2016).

The aim of this study is to determine the differences in ethnic segregation between generations; the differences between the levels of segregation for different generations should express the progress of assimilation over time. We use census data and passive mobile positioning data (CDR) from Estonia. Two dimensions are used for the assessment of segregation level in different age groups. Firstly, we use the place-based approach using segregation indexes and density patterns for different ethnic groups (Massey and Denton, 1988; Johnston et al, 2004). We differentiate between places of residence, workplaces, and out-of-home non-employment activity places. This approach offers the opportunity of being able to discover which place offers the highest integration potential. Nonetheless, the place-space approach ignores the personal experience, geographical context, and spatial mobility (Kwan 2012). In order to counter this, we also use the activity space approach (Farber et al, 2012; Järv et al, 2014; Silm and Ahas, 2014; Wang and Liu, 2016), in which ethnic differences in space-time behaviour have been assessed with individual activity space measures. We measure the number, randomness, and ethnicity of visited districts, and the extent of the activity space.

The study area is Estonia, which is an interesting country when it comes to studying the temporal dimension of ethnic segregation. The migrant population here is formed by the Russian-speaking minority, whose massive immigration ended in 1991 when Estonia was liberated from Soviet occupation. Since the number of new immigrants from Russian-speaking countries has been very low since 1991, we have been able to observe the results of the assimilation process for immigrants of the Soviet era across different generations. This also provides a unique experimental picture of the potential assimilation process for immigrants and also a picture of the changes in segregation following the end of immigration.

The place-based approach shows that segregation in places of residence and out-of-home non-employment activity districts is higher in younger age groups than it is in older age groups. The geographical distribution of places of residence and out-of-home non-employment activity districts across different age groups is quite similar both in Estonians and Russian-speakers. Greater differences can be seen in the distribution of workplaces in the 30-39 age group. The activity space approach shows that the difference between the activity space measures for Estonians and Russian-speakers generally decreases with age, which refers to the higher ethnic segregation in younger age groups. Younger people are more connected with districts which are more heavily populated by their own ethnic group. Thanks to these findings, the result of the study may show that the spatial behaviour of Russian-speakers does not became similar over time to that of Estonians. Our results for ethnic differences in activity spaces by age groups does not follow the pattern described in the spatial assimilation theory, which dictates that minority group disparities decrease with time.

The rippling effect of social influence on phone communication network

Yan Leng,[1] Xiaowen Dong,[1] Esteban Moro,[1,2] and Alex "Sandy" Pentland[1]

1. MIT Media Lab, Cambridge, MA, USA
2. Departamento de Matematicas & GISC, Universidad Carlos III de Madrid, 28911 Leganes, Spain

***Abstract*—What is the reach of the social influence of our actions? Although many works have studied how influence diffuses within direct neighbors at local level and mostly on online social networks, little is known of the fingerprints that our actions leave in the network as a whole. Leveraging the data set of mobile phone records, we use state-of-the-art methods to detect social influence and distinguish it from homophily. We find that although most influence disappears after one hop of distance in the social network, it persists up to six degrees of separation. Our results suggest that people's actions yield to a network-wide ripple effect which could have important implications in understanding phenomena such as mobilization, marketing campaigns, or opinion formation.**

We live in a connected world and are increasingly closer to each other thanks to the emerging information technologies [1]. From "six degrees of separation" discovered by Milgram in 1967 [2], it has been recently found that the average degree of separation between two people via Facebook is reduced to 4.74 [3]. Furthermore, individuals are not merely connected; as a series of experiments in domains such as obesity and happiness has demonstrated, connectivity also indicates behavioral similarities of up to three degrees of separation [4].

The recent availability of large-scale communication and networked data, such as emails, phone records and social media, enables, at an unprecedented level of detail, resolution, and scale, studies of not only information diffusion and correlations of adoption behaviors, but also social contagion processes [5], [6], [7]. In particular, the understanding of the phenomenon of and the mechanism that drives the social contagion process helps to promote behavioral change for commerce, public health, politics and social mobilization at both local and global scales [8], [9], [10].

However, most of the previous works focus on online social networks, and measure influence between direct contacts involving either long-term habits or low-cost decision-making in virtual space. In this study, we are interested in how social influence propagates over a nation-wide offline communication network and how it manifests in short-time decision-making and social mobilization that are more costly than merely information diffusion or online production adoption. More importantly, while most literature pays attention to social influence between immediate neighbors, we are particularly interested in the subtle and invisible influence that extends beyond direct contact.

Leveraging a data set of mobile phone records with high resolution in the European country of Andorra, we construct a nation-wide communication network and mirror the contagion process of social influence, which is measured by the likelihood of attending a large-scale international cultural event in the capital city. We found that the degree of separation in phone communication from attendees of the event is a strong indicator of the likelihood of attendance. In order to control for the bias caused by homophily and quantify the impact of social influence on decision-making, we utilize a matching method to mimic random assignment of treatments [8], [11]. Our results show that influence decays across social distances from initial attendees, but persists, surprisingly, up to six degrees of separation.

Problem formulation

Mobile phone logs provide a proxy for human mobility and social interactions at a societal scale [5], [12]. In this study, we use the countrywide mobile phone logs in Andorra to study how the likelihood of an individual attending a local Cirque Du Soleil performance, which was held repetitively on July, 2016, is affected if someone in his social circle receives phone calls from past attendees of the event.

We assume that people who connected to a nearby cell tower of the performance venue, as shown in the left panel of Figure 1, during the performance period ($\pm$ 30 min as buffer time) attended the events and are named as **attendees**. We construct **influence cascades**, as shown in the right panel of Figure 1, by adding links between the caller and receiver if: 1) at least one of them are linked directly or indirectly with the attendees. 2) and the calls happened within 24 hours after the performance started. We use **hop** to capture the nearest social distance to the attendee on the influence cascade. We observed 16,043 attendees across the observational periods. 161,857 individuals are connected directly or indirectly to the attendees via the communication network. Meanwhile, 71,337 users are not connected to the attendees via the influence cascades.

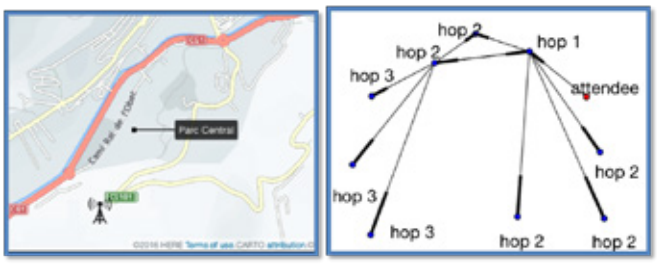


Fig. 1. Illustration of attendee, influence cascade and hop

METHOD

In order to quantify social influence, the key challenge is to distinguish social influence and homophily to avoid overestimation. We use matched sample estimation to mimic the assignment of treatment as in a randomized experiment. Conditioning matches on a vector of historical mobility patterns yields influence estimates. In particular, the matching results establish an upper bound to which extent social influence, rather than homophily, explains the attendance [8].

In order to purge homophily, which creates upward selection bias, and quantify social influence through phone calls, we quantify the similarities among individuals in the influence cascades based on their behavioral characteristics, specifically the historical mobility patterns [13]. Then, we segment individuals into two groups: the treatment group and the control group. One individual in the treatment group is paired with another individual in the control group who are most similar in terms of interests. This constructs the case where the differences between paired individuals are due to the treatment of social influence [14]. Pairs in the control and treatment groups are matched based on nearest Mahalanobis distance, calculated as:

$$md(X_j, X_k) = [(X_j - X_k)^T S^{-1} (X_j - X_k)]^{1/2}, \quad (1)$$

where X_i and X_j are the covariates for individual j and individual k and S is the sample covariance matrix for X. In our study, we use the first fifteen principle components of the mobility frequency matrix, capturing the historical mobility patterns during the weekends for the past six months, as the observed covariates.

The difference in the attendance rates of the two groups is the average treatment effect of social influence in our setting:

$$ATE_g = E(Y_{ig} - Y_{ic}), \quad (2)$$

where ATE_g is the average treatment effect of the treatment group on hop g. Y_{ig} and Y_{ic} are the outcome for matched pair i in the treatment group g and in the control group respectively.

RESULTS AND DISCUSSIONS

In this study, we quantify how our decision-making are influenced by, and how the social network propagates our influence to, people that are several degrees away from us in the communication network by analyzing the pattern of attendance of a large-scale cultural event in Andorra using nation-wide mobile phone data. The blue and red dash lines in Figure 2 represent the average treatment effect of social influence of Mahalanobis Distance Matching and random matching across different hops. The general decay pattern of social influence originated from the attendees reveals the existence of a "ripple effect" across information cascades in the communication network. Specifically, the average treatment effect of social influence is 11% on the first hop, drops dramatically to a half at the second hop and decays slowly afterwards. Yet, surprisingly, our results show that social influence persists up to six degrees of separation. This result highlights the hidden relationship and influences between individuals that are not necessarily acquaintances to each other.

The ripple effect demonstrated through our study has applications that are important for viral marketing, public health and social mobilization via offline communication. Recent works have demonstrated the success of social mobilization via Internet-based services [15], but also shown that such mechanisms are not without limitations [16]. Our findings suggest that an alternative would be to exploit the hidden and often overlooked influence between people that are caused by chains of offline communication, thus leading to more effective strategies in marketing or political campaigns.

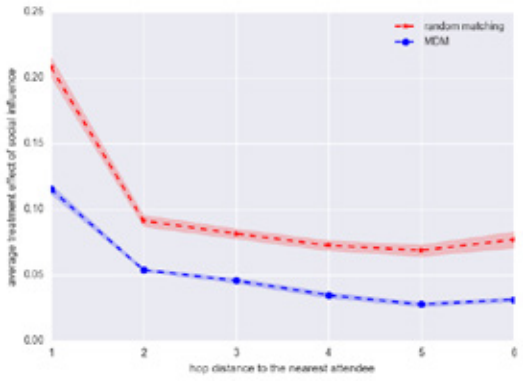


Fig. 2. Average treatment effects of social influence. X-axis is hop distance to the nearest attendee and Y-axis is the average treatment effect of social influence. The blue and red dash lines represent Mahalanobis Distance Matching and random matching. The shaded regions represent the 95% confidence intervals. The decay pattern of social influence exhibits "ripple effect".

REFERENCES

[1] D. J. Watts and S. H. Strogatz, "Collective dynamics of small-worldnetworks," *nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[2] S. Milgram, "The small world problem," *Psychology Today*, 1967.

[3] L. Backstrom, P. Boldi, M. Rosa, J. Ugander, and S. Vigna, "Four degrees of separation," in *Proceedings of the 4th Annual ACM Web Science Conference*, pp. 33–42, ACM, 2012.

[4] N. Christakis and J. Fowler, "Connected: The surprising power of our social networks and how they shape our lives," 2009.

[5] P. Deville, C. Song, N. Eagle, V. D. Blondel, A.-L. Barabási, and D. Wang, "Scaling identity connects human mobility and social interactions," *Proceedings of the National Academy of Sciences*, p. 201525443, 2016.

[6] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A.-L. Barabási, "Structure and tie strengths in mobile communication networks," *Proceedings of the National Academy of Sciences*, vol. 104, no. 18, pp. 7332–7336, 2007.

[7] G. Miritello, E. Moro, and R. Lara, "Dynamical strength of social ties in information spreading," *Physical Review E*, vol. 83, no. 4, p. 045102, 2011.

[8] S. Aral, "Social science: Poked to vote," *Nature*, vol. 489, no. 7415, pp. 212–214, 2012.

[9] J. Ugander, L. Backstrom, C. Marlow, and J. Kleinberg, "Structural diversity in social contagion," *Proceedings of the National Academy of Sciences*, vol. 109, no. 16, pp. 5962–5966, 2012.

[10] R. M. Bond, C. J. Fariss, J. J. Jones, A. D. Kramer, C. Marlow, J. E. Settle, and J. H. Fowler, "A 61-million-person experiment in social influence and political mobilization," *Nature*, vol. 489, no. 7415, pp. 295–298, 2012.

[11] G. King and R. Nielsen, "Why propensity scores should not be used for matching,"

[12] J. L. Toole, C. Herrera-Yaqüe, C. M. Schneider, and M. C. González, "Coupling human mobility and social ties," *Journal of The Royal Society Interface*, vol. 12, no. 105, p. 20141128, 2015.

[13] Y. Leng, L. Rudolph, A. Pentland, J. Zhao, and H. N. Koutsopolous, "Managing travel demand: Location recommendation for system efficiency based on mobile phone data," *CoRR*, vol. abs/1610.06825, 2016.

[14] E. A. Stuart, "Matching methods for causal inference: A review and a look forward," *Statistical science: a review journal of the Institute of Mathematical Statistics*, vol. 25, no. 1, p. 1, 2010.

[15] G. Pickard, W. Pan, I. Rahwan, M. Cebrian, R. Crane, A. Madan, and A. Pentland, "Time-critical social mobilization," *Science*, vol. 334, pp. 509–512, 2011.

[16] A. Rutherford, M. Cebrian, S. Dsouzaa, E. Moro, A. Pentland, and I. Rahwan, "Limits of social mobilization," *Proceedings of the National Academy of Sciences*, vol. 110, no. 16, pp. 6281–6286, 2013.

Cohesive groups in urban area: characterization of p-cliques in mobile phone graph

Matteo Zignani, Christian Quadri, Sabrina Gaito, and Gian Paolo Rossi

Department of Computer Science, University of Milan, Milan, Italy
firstname.lastname@unimi.it

Social networks built on top of mobile phone data are one of the main sources facilitating the understanding of the interplay between people's sociality and proximity. At the dyad level, it is well established that people connected by a social tie are more likely to share city's locations than unrelated ones [1]. On the contrary, more complex social structures such as communities remain quite loose organizations that only sporadically gather in a place to carry common activities out [2].

In this work we focus on cohesive groups expressed through calls and text-messages and on their spatial characterization. To this aim we exploit the localization provided by the phone data to reconstruct the proximity patterns of the group members. In particular, we proceed by *i*) identifying and characterizing cohesive groups through the extraction of maximal pseudo-cliques (p-cliques); *ii*) studying the different member roles (leaders or followers); and finally *iii*) mapping a p-clique to the places it visits.

Dataset. Our mobile phone dataset [3] is based on Call Detail Records (CDRs) registering voice-call, text and data activities of 1 million mobile subscribers in the Milan metropolitan area over 67 days in the spring of 2012, before the era of OTT services, when text messages were significantly used by mobile users. For each issued activity, a CDR entry is created as a 6-ple $t_{CDR} = \langle s, r, t_{start}, d, cell, area \rangle$, where s and r represent the sender and the receiver of the call/SMS, t_{start} is the start time of the activity, d is its duration and $cell$ is the serving cell the user s is attached to. The field $area$ indicates the city's zone representing a coarse grain division of the city region. Unlike previous studies where cell tower may cover a zone as wide as a few kilometers [4], the dataset reports data about cell towers inside a city space where small coverage radius is adopted. This characteristic, combined with the knowledge of p-cliques, is a powerful enabler to study the off-line social life of tight-knit groups.

Methods. Starting from the mobile phone data, we construct an undirected graph $G(N,E)$ where N is the set of the mobile operator subscribers and E is the set of ties between them. In order to discard occasional interactions, we required that users had at least 3 interactions (calls or texts) or a total calls duration above 1 minute. In this work, the notion of cohesive subgroup corresponds to the definition of p-clique presented in [5]. Using the concept of p-clique instead of clique allows us to better capture the structure of close groups by relaxing the constrain of the existence of all communication edges. To this aim, we apply the Pseudo Clique Enumeration (PCE) algorithm [5] to find all the maximal p-cliques in an undirected graph. We use 0.8 as value of edge density threshold θ and we impose 5 as minimum p-clique size. Moreover, we exploit the location information to evaluate the co-location of the p-clique. We adopt the following methodology to detect clique co-location. For each clique we reconstruct the mobility trace of each user starting from the CDRs, and we transform it as follows: we convert

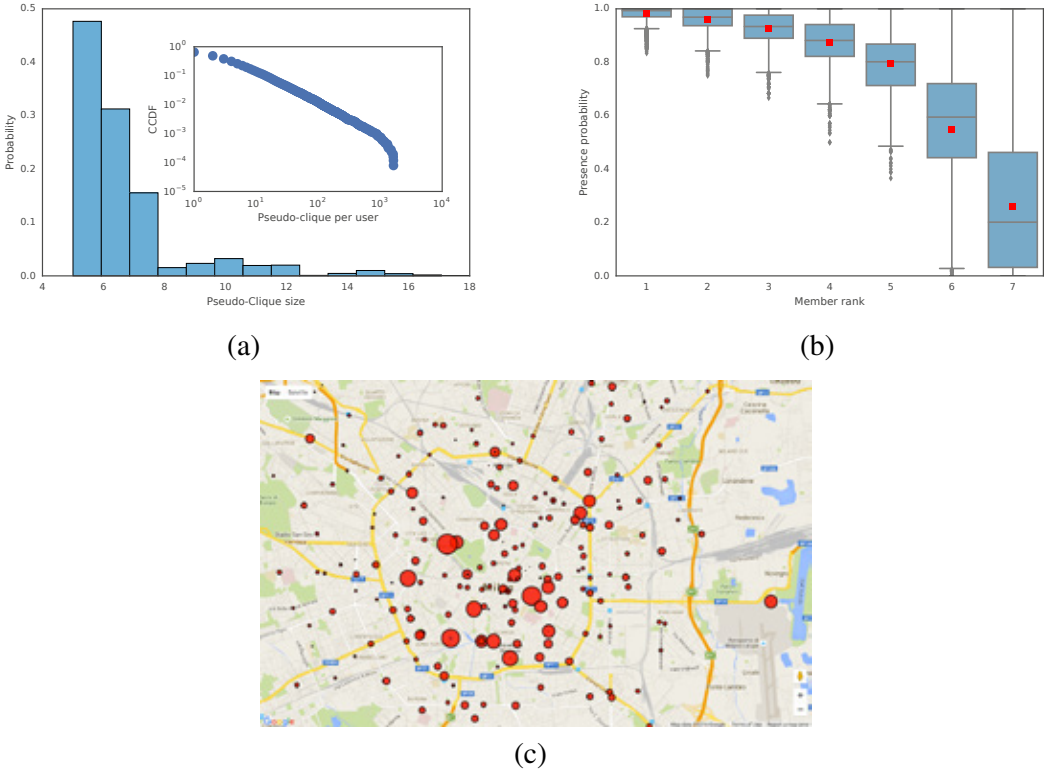


Fig. 1. a) The histogram of the p-clique size (k). The inset figure shows the distribution (CCDF) of the number of p-clique per user. b) The boxplot of the presence probability of each member for p-cliques of size 7. c) The city map with the places where p-cliques meet.

each point, identified by triplet $\langle t, cell, area \rangle$, in a time interval assuming that if the user was in a cell at time t she/he was in that cell from $t - \Delta$ until $t + \Delta$. In this paper we use Δ = 30 minutes in line with [6]. Each time interval maintains the same location attributes (cell and areas) of the trace point from which is generated. Then we merge all the traces of the members of the clique, by retaining only the time overlapping intervals that share the same location attributes. In the following we only consider clique co-location in area. We consider a p-clique to be co-located if at least the 60% of its members are co-located.

Results. We observed 43,646 maximal p-cliques whose size (k) is in the range 5 and 18. 25,422 people, 8.8% of all users, are involved in at least one clique, a very high percentage if we mind the limits of mobile phone data. The same holds for links since the 15% of network ties are intra-p-cliques. This result confirms the presence of strongly cohesive groups in mobile phone graphs. Small p-cliques ($k = 5, 6, 7$) are predominant, Fig.1(a), and we can also observe that most people belong to few p-cliques (mean and median values are 10.7 and 2.0 respectively), as shown in the inset of Fig.1(a). As for the localization of the p-cliques, our findings show that 83.4% of p-cliques meet at least once in the considered time frame. Also for cohesive groups, the result confirms that there is a strong correlation between on-phone interactions and off-line encounters. We also observed that the probability that a member is present at a meeting is not equally distributed across all members; in Fig.1(b) we report the box plot of the presence probability aggregated on all p-cliques of size 7 (similar results can be observed for all the

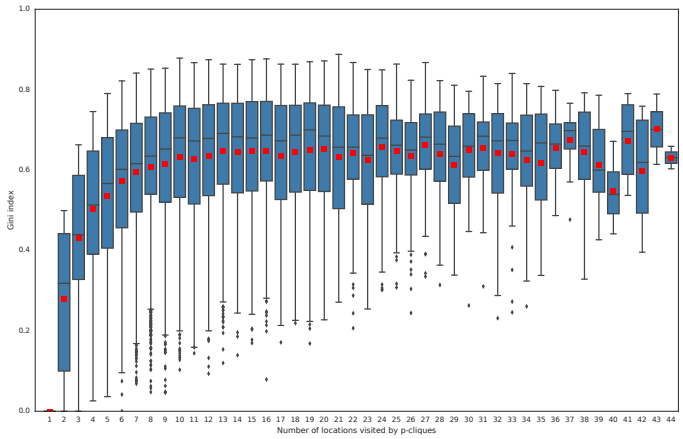


Fig. 2. The boxplot of Gini-index of visited areas by *p*-cliques.

other *p*-clique size). This result suggests the existence of two kind of members, the *leaders* who are almost always present when the *p*-clique meets and represents the core of the *p*-clique and the *followers* who participate less frequently to the off-line social aggregation of the *p*-clique. The map in Fig.1(c) highlights the fact that some city locations are preferred than others for *p*-clique encounters and therefore, indirectly, that cities have places more social than others. Finally, we observe that *p*-cliques meet in few favorite locations. Fig. 2 reports the Gini index distribution grouped by the number of locations visited by *p*-cliques. As we can observe, the values of Gini index are far from 0 (equality condition), meaning that almost all *p*-cliques prefer to meet in few favorite locations.

References

1. Calabrese, F., Smoreda, Z., Blondel, V.D., Ratti, C.: Interplay between telecommunications and face-to-face interactions: A study using mobile phone data. PloS one **6**(7) (2011) e20814
2. Zignani, M., Quadri, C., Gaito, S., Rossi, G.P.: Calling, texting, and moving: multidimensional interactions of mobile phone users. Computational Social Networks **2**(1) (2015) 1
3. Quadri, C., Zignani, M., Capra, L., Gaito, S., Rossi, G.P.: Multidimensional human dynamics in mobile phone communications. PLoS ONE **9**(7) (07 2014) 1–12
4. Nanavati, A.A., Singh, R., Chakraborty, D., Dasgupta, K., Mukherjea, S., Das, G., Gurumurthy, S., Joshi, A.: Analyzing the structure and evolution of massive telecom graphs. Knowledge and Data Engineering, IEEE Transactions on **20**(5) (2008) 703–718
5. Uno, T.: An efficient algorithm for solving pseudo clique enumeration problem. Algorithmica **56**(1) (2010) 3–16
6. Wang, D., Pedreschi, D., Song, C., Giannotti, F., Barabasi, A.L.: Human mobility, social ties, and link prediction. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '11, New York, NY, USA, ACM (2011) 1100–1108

Personality Traits and Ego-Network Dynamics

Simone Centellegher, Eduardo López, Jari Saramäki, Bruno Lepri

09 January 2017

1 Abstract

We interact with a wide network of people on a daily basis, and these social relationships play an important functional role in our lives. A large number of studies has shown that having strong and supportive relationships is essential for health and subjective well-being [1]. However, there are costs to their maintenance [2, 3], resulting in a tradeoff between quality and quantity, a typical strategy being to put a lot of effort on a few high-intensity relationships while maintaining larger numbers of less close relationships. It has also been shown that there are persistent individual differences in this pattern; some individuals allocate their efforts more uniformly across their networks, while others strongly focus on their closest relationships [4, 5, 6]. Furthermore, some individuals maintain more stable networks than others. Here, we focus on how personality traits of individuals affect this picture, using data from the Mobile Territorial Lab study [7]. In particular, we look at the relationship between personality traits, alter rank dynamics, and *social signatures* [8] that measure how communication efforts are allocated to personal network members.

We observe that some traits have effects on the stability and homogeneity of social signatures as well as network turnover and ego-networks rank dynamics. As an example, individuals who score highly on the Openness to Experience trait tend to have more diverse social signature shapes and higher levels of network turnover, with open to experience individuals showing higher rank variations with respect individuals more closed to experience (Figure 1). On broader terms, our study provides one possible explanation for the uniqueness and stability of the individuals' social signatures. As pointed out by Saramäki *et al.*, [8] social signatures' characteristics reflect the fact that ego networks are typically layered into a series of hierarchically inclusive subsets of relationships of different quality. One of the constraints shaping the social signatures seems to be the one arising from differences in personality traits, with some individuals preferring to have a few, intense, and stable relationships and others preferring more diverse, but less intense ones [9].

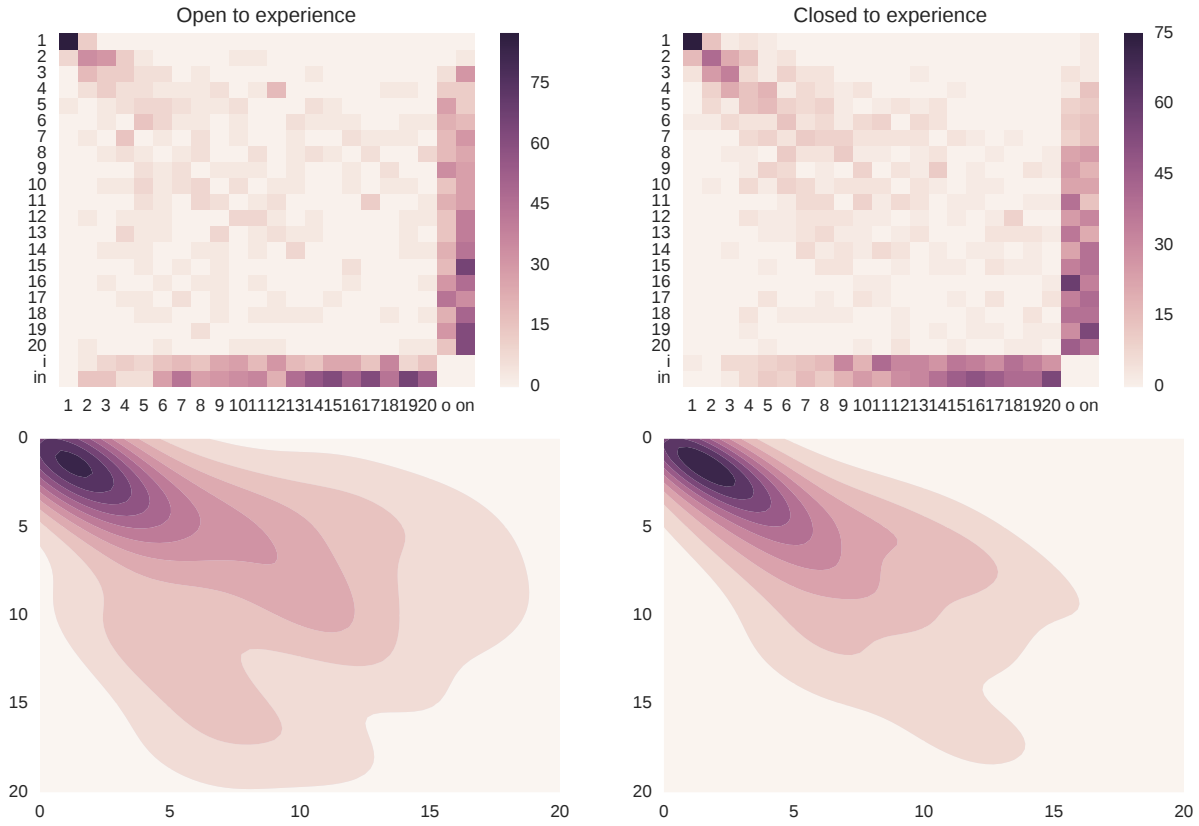


Figure 1: **Rank dynamics for the Open to Experience trait.** Transition matrices for the subgroups of individuals open and closed to experience. It is possible to observe that the open-to-experience subgroup shows a higher spread (large rank variation) with respect to the closed-to-experience subgroup where the "heat" is more concentrated around the diagonal (lower rank variation).

References

[1] Lyubomirsky S, King L, Diener E. The benefits of frequent positive affect: Does happiness lead to success? Psychological Bulletin. 2005;131(6):803–855.

[2] Winship C. The allocation of time among individuals. In: Schuessler K, editor. Sociological Methodology. vol. 9. Wiley; 1978. p. 75–100.

[3] Dunbar RIM. Neocortex size as a constraint on group size in primates. Journal of Human Evolution. 1992;22(6):469–493.

[4] Gonçalves B, Perra N, Vespignani A. Modeling users' activity on twitter networks: Validation of Dunbar's number. PloS one. 2011;6(8):e22656.

[5] Miritello G, Lara R, Cebrian M, Moro E. Limited communication capacity unveils strategies for human interaction. Scientific reports. 2013;3.

[6] Miritello G, Moro E, Lara R, Martínez-López R, Belchamber J, Roberts SG, et al. Time as a limited resource: Communication strategy in mobile phone networks. Social Networks. 2013;35(1):89–95.

[7] Centellegher S, De Nadai M, Caraviello M, Leonardi C, Vescovi M, Ramadian Y, et al. The Mobile Territorial Lab: a multilayered and dynamic view on parents' daily lives. EPJ Data Science. 2016;5(1):1.

[8] Saramäki J, Leicht EA, López E, Roberts SG, Reed-Tsochas F, Dunbar RI. Persistence of social signatures in human communication. Proceedings of the National Academy of Sciences. 2014;111(3):942–947.

[9] Swickert RJ, Rosentreter CJ, Hittner JB, Mushrush JE. Extraversion, social support processes, and stress. Personality and Individual Differences. 2002;32(5):877–891.

Why people stop calling? The temporal weakness of decaying ties

Henry Navarro,[1] Giovanna Miritello,[2] and Esteban Moro[1,*]

[1] *Departamento de Matemáticas & GISC, Universidad Carlos III de Madrid, 28911 Leganés, Spain*
[2] *Telefónica Research, 28050 Madrid, Spain*

Social networks are made out of strong and weak links having very different structural and dynamical properties. Social scientists have long recognized the value of weak links in information discovery but also their relative structural weakness that makes them more likely to decay in the future. What what features of human interaction build a strong tie? Here we approach this question from an practical way by finding what are the properties of social interactions that make ties more persistent and thus stronger to maintain social interactions in the future. Using a large longitudinal mobile phone database we build a predictive model of tie persistence based on intensity, structural and temporal patterns of social interaction. While our results confirm that structural (embeddedness) and intensity (number of calls) are correlated with tie persistence, we find that daily temporal features of communication events in a tie are better and more efficient predictors for tie persistence. Specifically, although communication between links is always bursty we find that links that are more bursty than the average are more likely to decay, signaling that the strength of the tie is not only reflected in the intensity or topology of the network, but also on how we distribute in time our interactions with our relationships. Our results not only are important to understand the strength of social relationships but also to unveil the entanglement between the different temporal scales in networks, from microscopic tie burstiness to network evolution.

Social networks are dynamic objects, they grow and change over time through the addition of new ties or the removal of old ones, leading to an ongoing appearance and disappearance of interactions in the underlying social structure [1, 2]. Identifying the different mechanisms by which a link form or decay is a fundamental and challenging question of individual human behavior, but also it can unravel the processes behind group, community and network dynamics that shape our social fabric and, in turn, how that network evolution impact important processes in our society like economy, cooperation or information diffusion. On the other hand, understanding under what condition a tie is more or less likely to decay may shed light on the circumstances under which an observed tie can be actually considered a genuine social relationship [3, 4].

Link evolution is usually studied in the context of the much more general problem of *link prediction* in data science [5]. In general, link prediction in social networks seeks to determine what are the endogenous and exogenous factors to the network that predict the appearance and persistence of a link, or even to infer the existence of a link when missing or anomalous data about the network is given [6]. In fact, endogenous factors, i.e. those properties that can be extrapolated from the network itself, are very good predictors for tie creation, persistence or decay. They reflect the inherent mechanisms that lead people to tie together and/or maintain a social relationships or to destroy it, such as triadic closure, homophily or geographical proximity [7]. For this reason, the majority of approaches to this issue focus to identify tie properties such as common friends or communities properties similarity among individuals or frequency of interaction, which have been observed to capture the emotional or trust intensity of a tie [8]. Another conclusion that emerges from previous studies is that ties without this structural similarity and/or intensity are more likely to decay [9, 10]. In the context of Granovetter's theory of *strength of weak ties*, strong ties are those which are more likely to persist, since they are structurally embedded (common friends) are are more intense (number of interactions), while bridges between communities are weak and, as Burt found [9], they are more likely to decay in the future.

Despite this understanding of some processes behind social interactions has helped to build very good models for link formation, the problem of link persistence has not received the same attention, largely due to the lack of quality data. Although some online social networks have explicit mechanisms to "*unfollow*" (Twitter) [11] or '*unfriending*" (Facebook) [12] other users, access to structural or intensity data is limited. On the other hand, most studies infer link decay from absence of tie activity in large databases [3, 10]. This is a potential problem, since the burstiness of human interaction could render most of the detected decay events are simply large inactivity periods of the interaction. Thus, although these studies of link decay agree on the general importance of structural embeddedness, intensity or reciprocity of a tie to predict its future persistence, they still provide an incomplete picture of what are the main tie properties that make them strong (persistent) and if, as was done in the problem of link prediction, we can build an efficient model based on endogenous properties of links to predict if a social relationship is bound to decay.

* Corresponding author emoro@math.uc3m.es

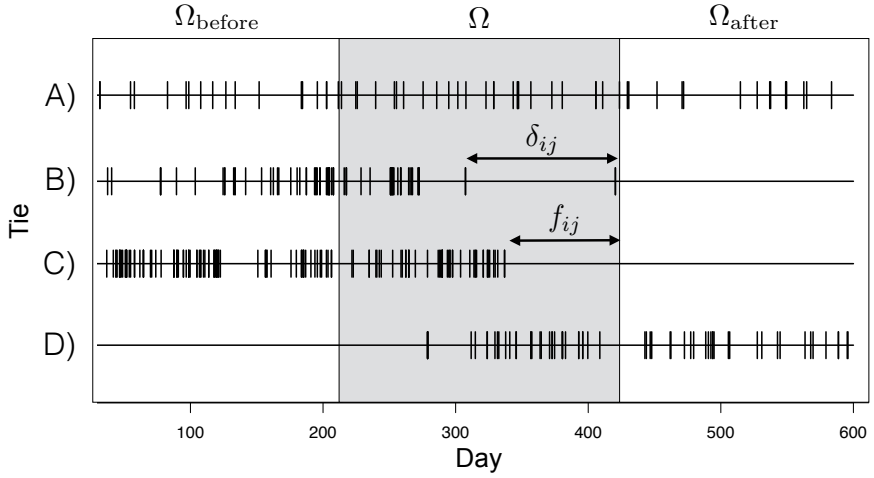


FIG. 1. Definition of observation periods and examples of call activity for 4 given ties. Any vertical segment is a call between the users in a particular tie. Our 19 months database is divided in three periods, where the 7 months in the middle Ω is our observation period where all the tie features will be measured. The period Ω_{after} is used to asses if ties are persistent, i.e. if there is activity in the tie. For example, ties A) and D) are persistent, while ties B) and C) are said to have decayed in Ω_{after}. All links have similar values of number of calls in the observation period with $w_{ij} \in [30, 40]$. We also show specific examples of one inter-event time δ_{ij} (tie B) and freshness f_{ij} (tie C).

In this communication we overcome some of these difficulties by studying tie persistence in human communication using a large longitudinal database of 19 month of mobile phone calls. The large duration of the database allow us to accurate assess the presence of a tie by using the method introduce by Miritello et al. [13] which splits the observation period in different time windows and use each of them to characterize and assess the presence of the tie (see figure 1). But more importantly, having a detailed and large longitudinal database for human communication allow us to characterize better the patterns of communication within a tie and see if temporal properties of human interaction are predictors of tie persistence in the future. Although simple temporal properties have been considered before in the problem of tie prediction [14] and strength estimation [10, 15], here we show that the persistence of a tie is also encoded in the bursty patterns of communication between people. Furthermore, by building a high accurate predictive model based on different tie features (structural, intensity, intimacy and temporal) we are able to show that temporal properties are indeed as important as intensity and much more than structural properties in predicting tie persistence (see figure 2). Our results show that it is possible to build simple predictive models of network evolution based on temporal and intensity properties of the human interaction.

[1] J. Saramaki and E. Moro, The European Physical Journal B **88** (2015).
[2] P. Holme and J. Saramaki, Physics reports **519**, 97 (2012).
[3] C. A. Hidalgo and C. Rodriguez-Sickert, Physica A: Statistical Mechanics and its ... **387**, 3017 (2008).
[4] G. Kossinets and D. J. Watts, Science **311**, 88 (2006).
[5] P. Wang, B. Xu, Y. Wu, and X. Zhou, CoRR **abs/1411.5118** (2014).
[6] D. Liben Nowell and J. Kleinberg, Journal of the American Society for Information Science and Technology **58**, 1019 (2007).
[7] M. T. Rivera, S. B. Soderstrom, and B. Uzzi, Annual Review of Sociology **36**, 91 (2010).
[8] J. Saramaki, E. A. Leicht, E. López, S. G. B. Roberts, F. Reed-Tsochas, and R. I. M. Dunbar, pnas.org (2012).
[9] R. S. Burt, Social Networks **22**, 1 (2000).
[10] T. Raeder, O. Lizardo, D. Hachen, and N. V. Chawla, Social Networks **33**, 245 (2011).
[11] H. Kwak, S. Moon, and W. Lee, "More of a receiver than a giver: Why do people unfollow in twitter?" (2012).
[12] D. Quercia, M. Bodaghi, and J. Crowcroft, in *Proceedings of the 4th Annual ACM Web Science Conference*, WebSci '12 (ACM, New York, NY, USA, 2012) pp. 251–254.
[13] G. Miritello, *Temporal Patterns of Communication in Social Networks*, Springer Theses (Springer, 2013).
[14] L. Tabourier, A.-S. Libert, and R. Lambiotte, EPJ Data Science **5**, 1 (2016).
[15] E. Gilbert and K. Karahalios, in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09 (ACM, New York, NY, USA, 2009) pp. 211–220.

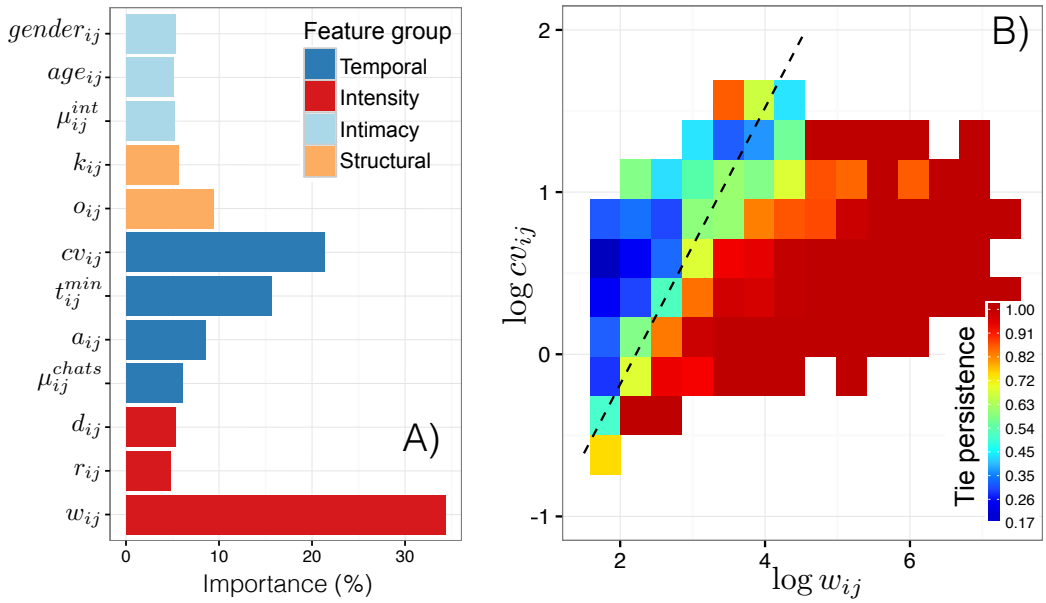


FIG. 2. A) Importance of the variables in the model to predict tie persistence. Importance is measured as the normalized % of the t-statistics for each model parameter. B) Density plot of the average persistence of ties as a function of the two most important variables in A), namely, the coefficient of variation (that measures the burstiness of ties) and total number of calls. The dashed line shows the 1/2 probability for the simplified models in table.

SESSION 4
SOCIAL GOOD

A Tool for Estimating and Visualizing Poverty Maps

Myeong Lee[1], Rachael Dottle[2], Carlos Espino[2], Imam Subkhan[3], Ariel Rokem[3], Afra Mashhadi[3]
[1]iSchool, University of Maryland, College Park, USA
[2]Columbia University, New York, USA
[3]University of Washington, Seattle, USA
myeong@umd.edu, rcd2127@barnard.edu, carlos.espino@columbia.edu, imams@uw.edu, arokem@gmail.com, afra.mashhadi@gmail.com

ABSTRACT
"Poverty maps" are designed to simultaneously display the spatial distribution of welfare and different dimensions of poverty determinants. The plotting of such information on maps heavily relies on data that is collected through infrequent national household surveys and censuses. However, due to the high cost associated with this type of data collection process, poverty maps are often inaccurate in capturing the current deprivation status. In this paper, we address this challenge by means of a methodology that relies on alternative data sources from which to derive up-to-date poverty indicators, at a very fine level of spatial granularity. We validate our methodology for the city of Milano and demonstrate how it could be used to implement a poverty mapping tool for policy makers.

INTRODUCTION
Small area estimation *poverty maps* are a recent innovation that provide detailed estimates of poverty levels in highly disaggregated geographical units [2]. The visualization of deprivation information in this form has been shown to be extremely effective in empowering policy makers and local municipalities to identify those areas in most need of interventions and revitalization programs. Poverty maps not only improve readability compared to traditional tabular data format by simultaneously preserving the spatial distribution of welfare, but also are powerful tools for capturing relations between deprivation and geographical factors such as city infrastructure and offerings.

In order for poverty maps to be impactful in determining and designing interventions, the poverty data ought to be up-to-date and presented in disaggregate level of granularity. However, due to the high costs associated with the household survey and censuses, National Statistical Institutes and the like often possess social and economic well being information that is out-of-date (i.e., collected infrequently) and only inclusive of a rather small sample of the population.

Within the remit of 'Data for Development' there have been a number of promising recent works, whereby researchers have relied on alternative sources of data to estimate deprivation. For example, models exploiting Call Detail Records (CDRs) from mobile phones have shown to be good indicators of the spatial distribution of socio-economic status in developing countries [7, 4]. Other sources of data including readily available open datasets such as those of Volunteer Geographical Information (VGI) have been shown to successfully predict the poverty level based on the offering advantages of the cities [8]. While both these alternative sources have been shown to successfully estimate deprivation level, their results have only been presented in isolation rather than in comparison. Furthermore, each has various shortcomings. The CDR are often hard to obtain due to their commercial and privacy sensitive nature. The Open Data sources on the other hand are readily available but could suffer from biases in coverage [5].

In this paper, we propose a methodology that leverages both open source and proprietorial datasets to compute and offer a more complete spatial distribution of deprivation at the city scale. In so doing, we extract features corresponding to the functional offerings and connectedness of the urban areas from OpenStreetMap (OSM), and network and activity related features from the CDR for the city of Milano. We quantitatively draw a comparison between the poverty estimation offered by each source with respect to poverty indicators derived from costly census data (ground truth). Our results, based on a Random Forest Classifier, indicate that the combinations of features that were extracted from CDR and OpenStreetMap data predicted poverty level better than the baseline models and complement each other. Based on the proposed methodology we built a visualization tool that displays the estimated poverty level for each area of the city as well as the determinants that contribute to the predicted poverty level.

FEATURE EXTRACTION AND BASELINES
For the purpose of this study we used CDR dataset made available by Telecom Italia [1] and freely available OSM data. We extracted features from these datasets following the measures that are often considered as proxy of the poverty level. Insights for the dynamics of a city from CDR data are usually based on individual-level mobility, people's communication network, and the amount of activity. Due to the anonymization of the CDR data, however, it is not possible to extract mobility information. We thus used network advantages and activity signatures as major predictors from the dataset. Network advantages include (1) call volume (the extend to which people are active in each region); (2) introversion (the amount of calls made inside of a region divide by the outgoing call, which is related to access to resources and social capital outside of a neighborhood); (3) PageRank and eigenvector centrality (relative popularity based on communication network structure); and (4) entropy (the diversity of a region that may imply the potential capability to reach diverse resources). Furthermore, we generated different CDR activity signatures by separating weekdays and weekends due to different human behavioral patterns. As a result, a vector with 288 elements (144 for weekdays and 144 for weekends in 10-minute interval) was

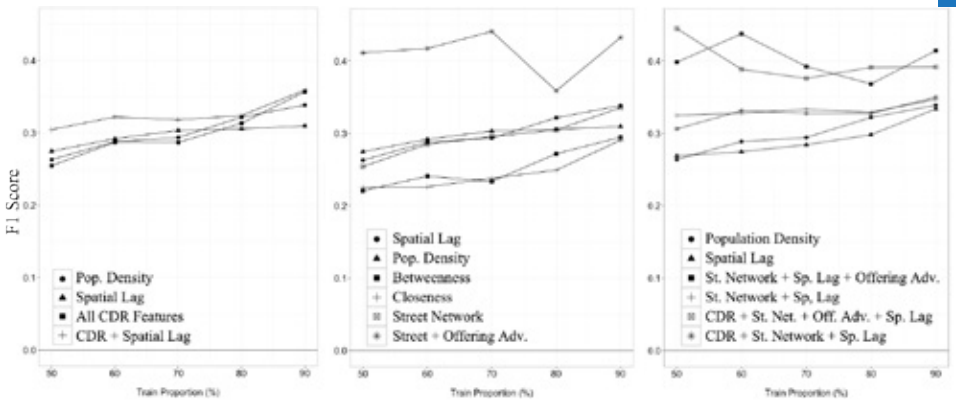


Figure 1. Average F1-score for models based on CDR features (left), OSM features (middle) and combined features (right) for varying train/test set.

constructed for each region. By conducting K-means clustering for these vectors, we classified every census tract into three categories, i.e., business, residential, and mixed regions. The number of clusters, K, was determined by using Davies-Bouldin cluster index [3] ($K = 3$).

While CDR provides information about hidden communication structure and dynamic human activities, OSM allows to extract features about physical resources (i.e., offering advantage) and access structure (i.e., street network) in a city. For offering advantage, we extracted amenity nodes for the city of Milano and grouped them by their categories. We then calculated the *offering advantage* metric for each sub-category tag based on the distribution of the city. The offering advantage metric was utilized to understand what PoIs are present in a given neighborhood, distinct from other neighborhoods [8]. This metric weighs each category by its presence, so that categories that are not very popular are more significant in the analysis of a neighborhood compared to a category that occurs frequently. These features enable a greater understanding of the distribution of resources across the city, and the extent to which areas of the city have less access to such amenities as hospital, police station, and bicycle parking, as opposed to other PoIs, like bars and fast food restaurants. Additionally, street layers were extracted for the city of Milano alongside OSM amenity data. These street layers were converted into graph networks, representing the intersections as nodes and street segments as edges. Two centrality measures, closeness and betweenness, were calculated to understand each urban area's global and local centrality, which is an indicator of that space's social, economic and spatial prosperity and accessibility. These features further contribute to our model as indicative of the spatial distribution of resources and access in the city.

In order to compare the performance of our models against benchmarks, we borrow from the methodology proposed by [6] and implement the two baselines: popularity density and spatial-lag based on past poverty level. Population density is a well-known indicator of poverty as it has been shown to negatively correlate with socio-economic level, especially in developing countries. That is, the populated areas in a city are more likely to exhibit a lower welfare. This measure is calculated from the population of a SEZ region, a smallest census block in Milano, divided by its area. For the second baseline and as an alternative to the above, we target scenarios where no current census information exists and we are limited to the older information (e.g., previous census). For this baseline we exploit the poverty information provided by ISTAT corresponding to 2001 Italian census. While not suitable for prediction due to the changes of census block boundaries as well as urban landscapes, the past poverty data could help us understand the spatial autocorrelation in a given city as poverty often contains a strong degree of spatial autocorrelation. We thus create a second baseline model based on the spatial-lag of this independent variable.

PREDICTION MODEL

In order to test the predictability, we first examined the correlation between poverty level of Milano and each extracted feature by using linear regression. The Spearman's Rho correlation value and Mean Absolute Error (MAE) suggest that each feature by itself is very weakly correlated with the poverty level. Furthermore the baseline models also suffer from low accuracy ($rho = 0.35$ for population density, $rho = 0.15$ for spatial lag regarding 2001 poverty level). This observation indicates that prediction of poverty in such a fine grain level of granularity in a developed city is fundamentally a difficult problem even when we possess knowledge about how the poverty is spatially distributed (spatial-lag baseline) and strong determinant of the poverty (population density). We thus treat this problem as a classification problem where rather than predicting exact numerical poverty level, we aim to accurately classify which areas fall within different poverty levels. For this purpose we categorise the poverty distribution into 7 bins based on standard deviation, σ. We then use Random Forest classification with cross-fold validation. To ensure the robustness of our models, we run each model 20 iterations of random train/test splits with varying training proportion.

We used $F1$ score to measure the performance of the models. By combining the concepts of precision and recall, it provides an intuitive measure for prediction power. Figure 1 reports the $F1$ scores for models based on different features against varying proportions of train sets. As it can be observed, the performance of models are consistently better than the baselines but with varied differences depending on features and combinations. Based on these figures we can observe that

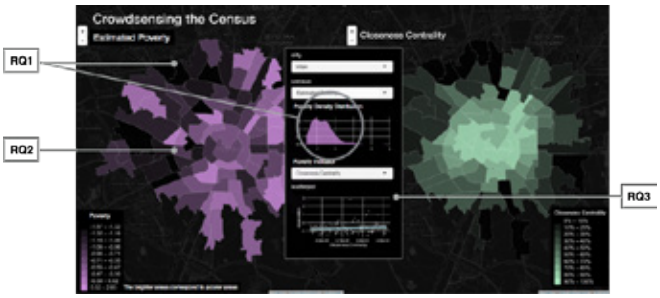


Figure 2. A screenshot of the Poverty Mapping Tool illustrating disaggregated distribution of poverty estimated by our model (left) and the determinants contributing to the estimated poverty (right).

while CDR and OSM features individually provide an improvement over the baseline prediction, they best perform when combined together and boosted with the knowledge of poverty autocorrelation (i.e., spatial lag).

POVERTY MAPPING VISUALIZATION TOOL

In order to put our methodology into actual use, we collected a set of requirements through discussions with the Global Pulse, UN research lab. These requirements led us to the design and development of a poverty mapping tool that leverages our prediction model. Figure 2 (RQ1) presents a screenshot of this tool which enables the user to select a city and view the estimated poverty level presented by a simple chloropleth color ramp. Furthermore, it visualizes the density distribution of poverty across the city as a whole by a bar chart. This representation captures the requirement of simultaneously preserving the spatial distribution of welfare, and provide this information in a highly disaggregated geographical unit. Therefore it enables easy interpretation and comparison across different areas and in relation with the geographical distribution of the areas (e.g., suburban areas vs central). The tool also allows the user to view the poverty level at various spatial granularity catering for diverse information needs of different users. For example, a district commissioner would be interested in sub-district poverty and require a more abstract view of the poverty map. This is illustrated in Figure 2 (RQ2) where poverty is displayed in the sub-municipal area level, in contrast to the finest resolution, SEZ level.

Furthermore the tool also caters for the need for transparency. Indeed many policy makers in the past have viewed the poverty estimation methodology as a black box [2], and thus have often lacked trust in the poverty mapping algorithms behind the scene. Based on this observation we have created a transparent design where the determinants and indicators that contribute to the estimated poverty values are clearly communicated to the user. As illustrated in Figure 2 our tool allows the user to select a poverty determinant from the drop down menu (e.g., community services) and explore the spatial distribution of the selected determinant across the city (right map).

Finally, the last design requirement that our tool caters for is the presentation of the urban infrastructure and elements, allowing the stakeholders to interpret the potential impact of their policies in relation to the existing elements. This information could help with the placement of amenities that

Figure 3. A screenshot of our poverty mapping tool visualizing the urban elements and infrastructure of a given area.

are vital but missing in a neighborhood. For example, as part of gentrification interventions one could decide where to create third places that would encourage a sense of community in isolated neighborhoods. Our tool allows the user to select an area and display the spatial distribution of existing PoIs and street connectivity for that selection. Figure 3 illustrates this feature for one of the poorer areas of Milano, where the map indicates the sheer presence of bars and fast food places, resembling the previous findings reported in [8].

CONCLUSION

We proposed a methodology for estimating poverty levels that relies on alternative data sources than census and thus is relatively low in cost. We have demonstrated how these poverty estimates could be presented as poverty maps and offer spatially fine grained and up-to-date information that is easy to interpret. Our results indicate that our model is able to predict poverty more accurately than the known baselines.

REFERENCES

1. Gianni Barlacchi, Marco De Nadai, Roberto Larcher, Antonio Casella, Cristiana Chitic, Giovanni Torrisi, Fabrizio Antonelli, Alessandro Vespignani, Alex Pentland, and Bruno Lepri. 2015. A multi-source dataset of urban life in the city of Milan and the Province of Trentino. *Sci Data* 2 (27 Oct. 2015), 150055.
2. Tara Bedi, Aline Coudouel, and Kenneth Simler. 2007. *More than a pretty picture: using poverty maps to design better policies and interventions*. World Bank Publications.
3. David L Davies and Donald W Bouldin. 1979. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence* 2 (1979), 224–227.
4. Vanessa Frias-Martinez and Jesus Virseda. 2012. On the relationship between socio-economic factors and cell phone usage. In *Proceedings of the fifth international conference on information and communication technologies and development*. ACM, 76–84.
5. Giovanni Quattrone, Afra Mashhadi, and Licia Capra. 2014. Mind the map: the impact of culture and economic affluence on crowd-mapping behaviours. In *Proceedings of the 17th CSCW*. ACM, 934–944.
6. Chris Smith-Clarke and Licia Capra. 2016. Beyond the Baseline: Establishing the Value in Mobile Phone Based Poverty Estimates. In *Proceedings of the 25th International Conference on World Wide Web*. 425–434.
7. Christopher Smith-Clarke, Afra Mashhadi, and Licia Capra. 2014. Poverty on the cheap: Estimating poverty maps using aggregated mobile communication networks. In *Proceedings of CHI*. ACM, 511–520.
8. Alessandro Venerandi, Giovanni Quattrone, Licia Capra, Daniele Quercia, and Diego Saez-Trumper. 2015. Measuring Urban Deprivation from User Generated Content. In *Proceedings of the 18th ACM CSCW*. ACM, 254–264.

Inside Out: to understand crime mechanisms look at urban fabric first

Marco De Nadai*[1,2], Emmanuel Letouzé†[4], Marta C. González‡[3] and Bruno Lepri§[2]

[1]University of Trento. [2]FBK. [3]MIT. [4]Data–Pop Alliance

1 Introduction

The rapid growth of cities and the increase of population mobility have challenged our ability to understand how to understand crime. The primary focus of criminology research has been on people that commit crimes, and the reason they are involved. For crime to happen three conditions are supposed to be present and interact: the presence of a motivated offender who is willing to commit a crime, a suitable target, and the absence of guardians that would otherwise prevent the crime from taking place [2]. In this equation crime offenders, victims and guardians are all affected by socio-economic conditions, the social disorganization (e.g. unemployment) of community [4] and the place where they intersect. Thus, place matters.

Environmental criminology suggests that place not only is logically required, but also influences the likelihood of becoming a crime hotspot through its physical characteristics. Accordingly, place is one of the five necessary and sufficient components that constitute a criminal incident, namely place, time, law, offender and victim [1]. Thus, environmental criminologists are interested in land use, street design, traffic patterns and daily activities of people. However, scholars have virtually ignored other theories (e.g. social disorganization), and bounded their discussion to macro-areas of few cities.

Urban planners and sociologists argue that cities are not a mere artificial construction that group people. A city is a vital process of the people who compose it; and its neighborhoods are the elementary form of cohesion in urban life. One of the seminal books in city planning is Jane Jacobs' The Death and Life of Great American Cities [6]. In this book she introduced the concept of eyes-on-the-street, which suggests that safety can be maintained by citizens through urban surveillance. For this to work, some physical qualities need to be present in the neighborhoods (i.e. a mix of residential, commercial and recreational land uses) to guarantee the diversity and continuous presence of people throughout the day. It is thus clear the tight coupling of environmental criminology and urban planning theories.

Traditional approaches on describing crime have failed to provide a clear and broad description of the desirable characteristics the different parts of the city should possess to keep crime events low.

In the present study we seek to fill this gap, formalizing the hypothesis that physical characteristics of the city not only are related to better life conditions and vitality [3], but also greatly influence crime. We leverage open data, web data and call detail records (CDRs), which are used to extract mobility networks. Then, we create two types of models. One is focused on *describing* how physical characteristics influence crime in each part of the city. The other *predicts* crime events in a city from the structural features, and answers to the question "can we predict crime from the physical characteristics of the city?". Thanks to several new sources of data and a Negative Binomial model, we model crime through physical characteristics and socio-economic, spatial and routine-activity information.

2 Methods and results

Crime is considered a rare occurring phenomenon: a small proportion of people are victimized, with also few unreported and undiscovered crimes. Crime offenses are not distributed evenly in space; they tend to cluster in parts of the city that exhibits similar characteristics, favorable to the spread of crime. Rare events, expressed through count variables, have been widely modeled

*denadai@fbk.eu
†eletouze@datapopalliance.org
‡martag@mit.edu
§lepri@fbk.eu

on Ordinary Least Squares (OLS) through a logarithmic and square root transformation of the response variable. However, this introduces extra estimation bias, homoschedability assumptions and impossible predictions. Moreover, spatial auto-correlation is rarely accounted for. For this reason, we use a Negative Binomial regression (NB) with an eigenvector spatial filter (ESF) [5] that introduces a set of independent variables that account for the spatial relationship of the variables.

Our hypothesis is that Jane Jacobs' diversity conditions (e.g. small blocks, concentration of diversified enterprises, non-standardized buildings, land use mix) have an impact on crime measures. Thus, we consider the metrics we created and validated [3] to operationalize the Jacobs' theory. Thus, for each neighborhood we measure the entropy of land uses, the average block size, the population density, the walk-ability and the distance to vacuums (e.g. big parks).

Mobility network. As aforementioned, daily routines are supposed to influence the presence of offenders, victims and guardians in a place. Moreover, frequent trips between two places are supposed to influence the criminality of these places. Wang *et al.* [8] used the taxi flow network with a spatial lag regression. We instead use the same ESF method used to take into account the spatial-autocorrelation. Firstly we compute the total number of trips made in a typical day between spatial units. Thus, we utilize anonymized CDRs with an observation period of six months (2013-2014) to estimate origin-destination (OD) matrices for a typical day. For each hour and spatial unit we compute the flows to all other spatial units of the city, using the method described by Toole *et al.* [7]. The result is a $N \times N \times 24$ weighted mobility network denoting the number of trips made by people during a typical day, where N is the number of spatial units in the city. We define the symmetric weight matrix W_t from this mobility network, by applying this transformation to the original mobility network W_m as $W_t = W_m^T W_m$. The eigenvectors inserted in the NB model are supposed to be a proxy for the mobility network dependencies between spatial units.

Covariates. The number of committed crimes is mainly influenced by the number of residents, their social disorganization and routine activity. Social disorganization is the inability of the neighborhood to maintain effective social control. Social disorganization is higher in deprived areas, social heterogeneous units and in places with high unemployment rate, which is also a proxy for motivated offenders. Social heterogeneity is accounted by the entropy of the buildings social class.

We first employ a descriptive model for Bogota where it is possible to understand the interactions of each component of the model to describe crime events. Then, we create a predictive model validated with a 5-fold Cross-validation with 1000 repetitions (to avoid overfit). This allows to answer to the question "can I predict crime events from the characteristics of the city?".

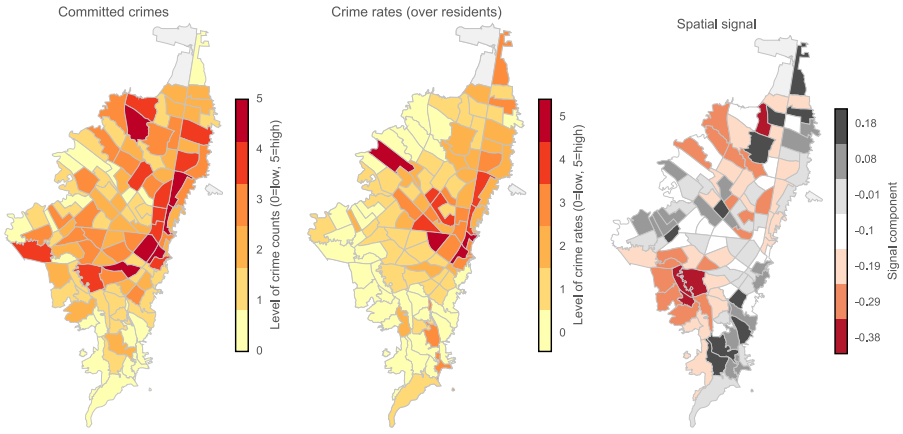
2.1 Descriptive model

From the NB regression fit we observe the β coefficients of the features to understand the importance of each variable, holding the others as constant. The most important variables to describe crime are building density, population density and closeness to daily-use buildings. Particularly, we observe that the higher the building density is, the less crime events happen. On the contrary, population density has a negative correlation with crime events. We also see that, as Jane Jacobs' argued, the distance of buildings from the nearest street has a positive correlation with crime events. This means that her *eyes on the street* has a positive role on the decrease of crime in a neighborhood.

We also find that the number of inhabitants' routine movements between the neighborhoods is closely related to crime, with highly-connected points of the city experiencing a higher number of crimes, as suggested by the routine-activity theory [2].

2.2 Predictive model

Our preliminary findings (see Table 1) indicate that structural characteristics of the city, namely Jacobs' diversity conditions, are a better predictor of the target variables (the number of homicides and robberies) than socio-economic conditions such as unemployment and deprivation. Mobility networks, and thus the routine activity theory, improve the prediction of the model by 15%. The combination of structural and socio-economic information provides better predictions than each on its own. Nevertheless, to avoid the descriptive bias of describing just one city, we plan to apply this method to Bogota, Los Angeles, Boston and Providence, where we have all the data we need. Moreover, we plan to aggregate the smallest spatial-unit in multiple ways using the K-medoid clustering algorithm on the centroids of the units. The evaluation of the *descriptive* and *predictive*



(a) (b) (c)

Figure 1: (a) Crime counts in Bogota for each neighborhood; (b) Crime rate ($|crimes|/(\beta_p|population|)$) in Bogota for each neighborhood; (c) Stochastic signal component B_eE representing the spatial auto-correlation for each neighborhood in Bogota..

model on these multiple and new aggregated spatial units will reduce the MAUP problems previous approaches did not consider.

Together, these observational results suggest that the city structure has a strong connection with crime, and that improving its qualities can discourage criminality.

	S	**C**	**D**	**S+D**	**C+D**	**S+C**	**Full**
McFadderPseudo-R^2 †	0.077	0.113	0.085	0.106	0.120	0.141	**0.143**
RMSE	231.93	145.04	312.70	181.76	133.36	143.35	**127.76**

† This is not a true measure of fit, and not bounded to 1. It indicates the degree to which the model parameters improve upon the prediction of the null model.

Table 1: Negative Binomial regression models that *predict* the number of crime in each spatial unit. The results are average across 1000 iterations of a 5-fold Cross-validation. S: demographic and social disorganization variables only; C: Jane Jacobs' diversity variables only; D: daily routine variables only; Full: model with all the variables.

References

[1] P. J. Brantingham. *Environmental Criminology*. Waveland Press, 1991.

[2] L. E. Cohen and M. Felson. Social change and crime rate trends: A routine activity approach. *American sociological review*, pages 588–608, 1979.

[3] M. De Nadai, J. Staiano, R. Larcher, N. Sebe, D. Quercia, and B. Lepri. The death and life of great italian cities: A mobile phone data perspective. In *WWW*, 2016.

[4] C. Graif, A. S. Gladfelter, and S. A. Matthews. Urban poverty and neighborhood effects on crime: Incorporating spatial and network perspectives. *Sociology Compass*, 8(9):1140–1155, 2014.

[5] D. A. Griffith. *Spatial autocorrelation and spatial filtering: gaining understanding through theory and scientific visualization*. Springer Science & Business Media, 2013.

[6] J. Jacobs. *The death and life of great American cities*. Vintage, 1961.

[7] J. L. Toole, S. Colak, B. Sturt, L. P. Alexander, A. Evsukoff, and M. C. González. The path most traveled: Travel demand estimation using big data resources. *Transportation Research Part C: Emerging Technologies*, 58:162–177, 2015.

[8] H. Wang, Z. Li, D. Kifer, and C. Graif. Crime rate inference with big data. In *KDD*, 2016.

LDA Mapping of Regional Socioeconomic Status

Lingzi Hong
College of Information Studies
University of Maryland
lzhong@umd.edu

Enrique Frias-Martinez
Telefonica Research
Madrid, Spain
enrique.friasmartinez@telefonica.com

Vanessa Frias-Martinez
College of Information Studies
University of Maryland
vfrias@umd.edu

Abstract

The socioeconomic status of a region provides an understanding of the access of its citizens to basic services. While socieconomic maps are key for policy makers, its compilation requires extensive resources and becomes highly expensive. As a result, traditional methods are now using pervasive datasets (such as cell phone traces for example) to infer regional socio-economic characteristics in a cost efficient manner. In this paper we use mobility information derived from cell phone records to identify socioeconomic levels by using a Latent Dirichlet Allocation that extracts recurring patterns of co-occurring behaviors across regions.

Introduction

Socio-economic maps contain information that characterizes various social and economic aspects like the educational level of the citizens or the access to electricity. The accuracy of these maps is critical given that many policy decisions made by governments and international organizations are based upon such information. National Statistical Institutes (NSIs) compute these maps every five to ten years, and typically require a large number of enumerators that carry out interviews gathering information pertaining the main socio-economic characteristics of each household. All these prerequisites make the computation highly expensive, especially for budget-constraint emerging economies. The ubiquitous presence of cell phones worldwide is generating datasets of spatio-temporal data across large groups of individuals. As previous research has shown, cell phone data can offer a detailed picture of how humans move and interact with each other (Becker et al. 2013). Recent results found that cell phone-based behavioral patterns might be correlated to specific socio-economic characteristics (Eagle, Macy, and Claxton 2010; Soto et al. 2011; Frias-Martinez et al. 2013; 2012). For example, higher socio-economic levels have been associated to stronger social networks or longer distances traveled (Blumenstock and Eagle 2010). Framing the problem as a supervised learning setting, these approaches use the spatio-temporal data to compute a set of pre-determined behavioral features per region and attempt to predict the regional socio-economic levels manually collected by the NSIs. Rather than pre-determined features, regions might be better characterized by probabilistic models of latent behaviors not obvious through observation. Such latent recurring patterns might best reflect the complex nature of human behaviors and the impact that geography and time might have on that behavior. We propose a novel approach to the problem of inferring regional socio-economic levels from spatio-temporal data by using a topic modeling framework based on Latent Dirichlet Allocation (LDA).

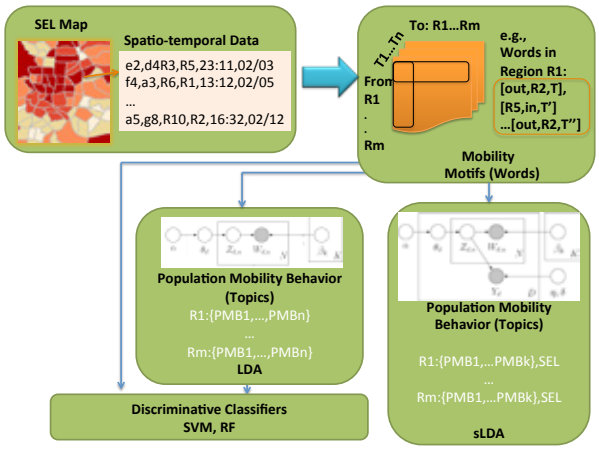


Figure 1: General Approach. sLDA and LDA plate notation from (Mcauliffe and Blei 2008).

Learning Architecture

Figure 1 shows the general approach proposed in this paper. We start with a socio-economic map where each region is represented by a pair $(SEL, Spatio-temporal;;data)$. Although the SEL is a continuous variable, it is often times expressed as a discrete value through a letter (A, B, C, etc.). On the other hand, the spatio-temporal data of a given region contains all calls in that region for a given period of time. Since this information is collected by telecommunication companies for billing purposes, it contains behavioral information about millions of users. As a result, and as shown in previous research (Vieira et al. 2010), the mobility patterns extracted from such data can be representative of the regional population at large, and thus of the underlying socio-economic level.

In this paper, we use the spatio-temporal data as a proxy of the mobility across regions, which will in turn be used as a predictor of regional socio-economic levels. With this approach in mind, we organize the spatio-temporal data into geographic data structures amenable to represent mobility as words. We define the *mobility motifs* as individual transitions containing origin and destination regions together with the time range at which that event happens. As a result, each region will have a set of such motifs every time a transition is observed for the time period under study.

With these regions and their motifs, we propose three approaches: (1) the use of supervised Latent Dirichlet Allocation to do both latent population mobility behaviors (PMB) extraction and SEL prediction over the mobility motifs, ($PMBSEL - sLDA$) (Mcauliffe and Blei 2008); (2) use unsupervised LDA to reveal the population mobility behavior (topic) proportions across regions which are in turn used as input to discriminative regression and classification algorithms to predict the SELs, *PMB-LDA*; and (3) Predetermined Features (PF), which represents each region as a vector of pre-determined mobility motifs where each component shows the number of times a given motif happens. By comparing the accuracy of the three approaches, PMBSEL-sLDA, PMB-LDA and PF, we expect to quantify the impact of using latent topics to predict socio-economic levels from spatio-temporal data.

Mobility Motifs and LDA approaches

The proposed approach uses large-scale spatio-temporal data collected from cell phones to model mobility. Each record collected is of the type $(i, j, R_i, R_j, T, D$ where i, j are encrypted phone numbers, R_z the regions where the individuals were when the phone call was made and T, D are time and date of the call. We use such records to compute the *mobility motifs* of a region R_i as the set of individual continuous transitions that depart or reach that region for a given period of time.

Given two call records from the same individual i, we can build a transition as follows. If i was in region R_i and called individual j in region R_j at time T and date D *i.e.,* (i, j, R_i, R_j, T, D) and next the individual i moved to region R'_i and called to k in region R'_k at time T' and date $D' \geq D$ *i.e.,* $(i, k, R'_i, R'_k, T', D')$ we can extract a mobility motif for region R_i as the tuple $mm = (out, R'_i, T)$ meaning that we observe an individual outgoing transition from region R_i to region R'_i at time T; and a mobility motif for region R'_i as the tuple $mm = (in, R_i, T')$ meaning that we observe an individual incoming transition from region R_i to region R'_i at time T'. The average time between visited regions is $3.2h$, thus, we discretize the time into six four-hour ranges *i.e.,* $T \in \{[0-4), [4-8), ...[20, 24)\}$. Repeating the process for all transitions observed, we can build a collection of regions (documents) each containing a specific set of mobility motifs modeled from the observed calling data as $R_i = \bigcup_{j=1 \in 1...6*R^2} (out, R_j, T) \vee (R_j, in, T)$ where $6 * R^2$ is the size of the vocabulary accounting for all possible bidirectional transitions between any two given regions in the area under study at any four-hour time range.

PMBSEL-sLDA

In this approach, we assume that the mobility motifs in each region arise from a set of latent topics or population mobility behaviors (PMB) at large scale *i.e.,* a set of unknown distributions over the mobility motifs. The set of PMBs is common to all regions, but each region will have a different combination of them.

We propose to use a supervised latent Dirichlet allocation (sLDA) in such a way that the generative process also includes the socio-economic label for each region as part of the model (Mcauliffe and Blei 2008). As a result, the inference is based on model estimates that take into account the socio-economic labels i.e., the empirical PMB frequencies put together and non-exchangeably both mobility motifs and SELs.

PMB-LDA

This second approach focuses on using topic modeling to extract the population mobility behaviors (PMB) in an unsupervised manner *i.e.,* topics are identified with the regions treated as unlabelled. Next, we use the PMBs as features to predict SELs either as continuous values or as classes. In this scenario, the LDA is used for dimensionality reduction.

Pre-determined Features

In this approach, each region R_i is represented by a vector containing all possible mobility motifs as features. We refer to the features as pre-determined because they are defined from behavioral hypothesis about human behavior and socio-economic levels rather than from latent topics directly extracted from the features which add the possibility of finding more complex behaviors. Instead of using population mobility behaviors extracted with LDA, here the regions have hard-coded all possible mobility motifs.

Results

To evaluate the accuracy of the approaches proposed, we use two datasets: a large-scale spatio-temporal dataset containing one month of calling activity for three cities from the same country, and the socio-economic map for those three cities containing regional SEL information. The spatio-temporal dataset contains a total of $134M$ calls and $1.8M$ individuals; while the SEL map contains a total of 186 regions distributed across the three cities.

SEL Inference

We first compute the mobility motifs from the spatio-temporal dataset. We obtain a total of $\approx 4.4M$ mobility motifs across all 186 regions, with an average of $\approx 24K$ motifs per region ($\sigma = \approx 32K$).

Table 1 shows the results for the four approaches using regression to infer SELs as continuous values. The results reported are for 20 topics for PMB-LDA (RF and SVR) and 25 for PMBSEL-sLDA, which turned out to be the number of topics that had the best results in terms of accuracy (R^2) as shown in Figure 2. For Support Vector Regression, we used a Gaussian RBF kernel and the parameters (C, γ, ϵ) were selected using 5-fold cross validation to minimize the

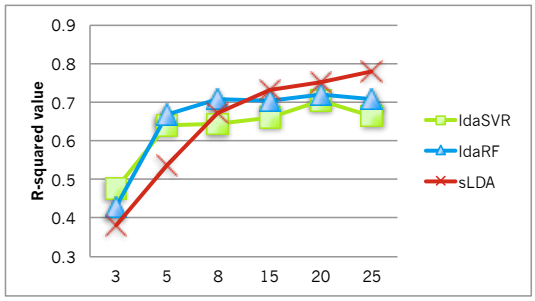


Figure 2: R^2 per number of topics for PMB-LDA (SVR and RF) and PMBSEL-sLDA approaches.

REGRESSION		R^2	**RMSE**
PMBSEL-sLDA		**0.7802**	0.0902
PMB-LDA	SVR	0.7050	0.1088
	RF	**0.7188**	0.1058
PF	SVR	0.2573	0.1731
	RF	**0.6927**	0.1156
PF2	SVR	0.5721	0.1290
	RF	**0.6288**	0.1195

Table 1: Accuracies for Regression with topic models and pre-determined features.

mean squared error. For Random Forest, the results are reported for 8 random trees in PMB-SEL; 146 trees in PF and 14 trees in $PF2$.

We can observe that both topic model approaches, PMBSEL-sLDA and PMB-LDA, have the best R^2 values together with the lowest error. In the case of PMB-LDA the best results are obtained using RF although SVR gave results with R^2 only $\approx$ 1% worse. As hypothesized, the topic model approaches outperform the pre-determined feature approaches by $\approx$ 9% in the best case. In fact, $R^2 = 0.7802$ for PMBSEL-sLDA while $R^2 = 0.6927$ for PF. These results show that the topic models reveal latent population mobility behaviors (PMB) that appear to characterize SELs (and the complex behaviors associated to them) better than the mobility motifs in which the PMBs are based on. Comparing both topic model approaches, the supervised approach gave $\approx 6\%$ better R^2 than the unsupervised approach combined with RF. A similar result was also reported by (Mcauliffe and Blei 2008) in an experiment inferring movie ratings with sLDA.

Table 2 shows the accuracies and F1 scores for all four approaches when SELs are defined as three discrete classes: A, B and C (from high to low socio-economic level). The results are reported for 25 topics for PMBSEL-sLDA and 15 topics for PMB-LDA, which are the topics that gave the highest F1 scores. For SVM, we used an RBF Gaussian Kernel and for RF the number of trees were 8, 146 and 2 for PMB-LDA, PF and PF2, respectively.

In general, the findings and trends are similar to the ones already discussed for the regression results. Here again, we observe that both topic models appear to improve the average F1 score obtained with the pre-determined features approach by $\approx 4\%$ in the best case scenario (PMBSEL-sLDA

CLASSIFICATION		**ACC**	**AVG.F1**	F1		
				A	B	C
PMBSEL-sLDA		0.7565	**0.7526**	0.7273	0.7283	0.8023
PMB-LDA	SVM	0.6237	0.6302	0.6609	0.5519	0.6777
	RF	0.7130	**0.7212**	0.7786	0.6572	0.7276
PF	SVM	0.4522	0.4510	0.7856	0.6283	0.7160
	RF	0.7004	**0.7100**	0.7856	0.6283	0.7160
PF2	SVM	0.6200	0.6374	0.7409	0.5586	0.6128
	RF	0.6440	**0.6567**	0.7468	0.5847	0.6387

Table 2: Accuracy (ACC), average F1 and per-class F1 score with topic models and pre-determined features.

vs. PF). It seems that LDA-based approaches might be doing a better job at extracting more complex population mobility behaviors than just the mobility motifs. Similarly, the pre-determined mobility motifs approach is slightly better than the simpler features of PF2 when RF are used. Moving on to the per-class F1 scores, we observe that the results across classes are quite balanced, specially when topic models are used. Interestingly, this fact reveals that regions are not simply being classified as the *most frequent class* which would be B in this case.

References

Becker, R.; Caceres, R.; Hanson, K.; Isaacman, S.; Loh, J.; Martonosi, M.; Rowland, J.; Urbanek, S.; Varshavsky, A.; and Volinsky, C. 2013. Human Mobility Characterization from Cellular Network Data. In *CACM*.

Blumenstock, J., and Eagle, N. 2010. Mobile divides: Gender, socioeconomic status, and mobile phone use in rwanda.

Eagle, N.; Macy, M.; and Claxton, R. 2010. Network diversity and econmic development. *Science* 328.

Frias-Martinez, V.; Soto, V.; Virseda, J.; and Frias-Martinez, E. 2012. Computing cost-effective census maps from cell phone traces. *Workshop on Pervasive Urban Applications, PURBA*.

Frias-Martinez, V.; Soguero-Ruiz, C.; Frias-Martinez, E.; and Josephidou, M. 2013. Forecasting socioeconomic trends with cell phone records. *ACM Symposium on Computing for Development*.

Mcauliffe, J. D., and Blei, D. M. 2008. Supervised topic models. In *Advances in neural information processing systems*, 121–128.

Soto, V.; Frias-Martinez, V.; Virseda, J.; and Frias-Martinez, E. 2011. Prediction of socioeconomic levels using cell phone records. In *Proceedings of the Int. Conference on User Modeling, Adaption, and Personalization*.

Vieira, M.; Frias-Martinez, E.; Bakalov, P.; Frias-Martinez, V.; and Tsotras, V. 2010. Querying spatio-temporal patterns in mobile phone-call datasets. *International Conference of Mobile Data Management*.

Rapid Assessments of Population Displacement in the 2015 Nepal Earthquake

Robin Wilson[1,2], **Elisabeth zu Erbach-Schoenberg**[1,2,*], Maximilian Albert[1], Daniel Power[1], Simon Tudge[1], Miguel Gonzalez[1], Sam Guthrie[1], Heather Chamberlain[1,2], Christopher Brooks[1], Christopher Hughes[1], Lenka Pitonakova[1], Caroline Buckee[1,3,4], Xin Lu[1,5,6], Erik Wetter[1,7], Andrew Tatem[1,2], and Linus Bengtsson[1,8]

[1] Flowminder Foundation, Stockholm, Sweden
[2] Geography & Environment, University of Southampton, Southampton, UK
[3] Center for Communicable Disease Dynamics, Harvard T.H. Chan School of Public Health, Boston, USA
[4] Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, USA
[5] Department of Public Health Sciences, Karolinska Institute, Stockholm, Sweden
[6] College of Information System and Management, National University of Defense Technology, Changsha, China
[7] Stockholm School of Economics, Stockholm, Sweden
[8] Dept. of Public Health Sciences, Karolinska Institute, Sweden.
*Corresponding author for this abstract, elisabeth.zu.erbach@flowminder.org

Introduction

Disasters, such as hurricanes, floods, landslides or earthquakes lead to displacement of populations from the areas affected. These populations are vulnerable, as they frequently do not have access to shelter, food, clean water and other necessities. Relief efforts aiming to provide these necessities rely on up to date information on locations and sizes of displaced populations for both assessment of resource needs as well as prioritisation of resource allocation. In the later stages of relief efforts, assessing which areas are recovering well and which are lagging behind in terms of recovery compared to other areas, becomes important.
Information on population movements following disasters is often not available or has poor spatial coverage, due to lack of suitable data sources and interruption of existing data collection systems.

Mobile phone call detail records (CDRs) have been used to study population flows in a variety of research settings [1-6]. They have been shown to allow accurate estimation of displaced population flows in the context of the 2010 Haiti earthquake in a retrospective analysis [6]. Here, we describe how CDRs were used to assess displacement resulting from the 2015 Nepal earthquake, which happened on 25 April 2015. We describe the methods used to estimate magnitude and destinations of population flows related to displacement as well as measure return rates. The methods described were used to provide these estimates in near-real time, which meant key findings could be included in reports by the UN resident Coordinator, adding to the data evidence base available for relief agencies.

Data and Methods

Mobile phone data for this project were provided by Ncell, the leading mobile phone services provider in Nepal with 12.9 Million subscribers. The Flowminder Foundation had signed an agreement with Ncell before the earthquake and was in the process of setting up the technical infrastructure to process CDRs according to the GSMA privacy guidelines at the premises of the operator. This setup had not been completed by the time the earthquake hit, however, Ncell were able to provide access to deidentified records within 6 days of the earthquake.

Preprocessing

Preprocessing included removing unused information from the data, data compression as well as calculating a 'daily location' for each user. We defined the daily location as the location of the last call made by each user. This definition aims at capturing the nighttime location of individuals, which is generally the place of residence or shelter and this method was chosen following preliminary analyses.

The daily locations on tower level, were mapped to admin level locations (admin 3 (district) for the displacement flows and admin 4 level (Village Development Committee, VDC) for the return rates). Admin units were used instead of tower level

estimates to match other information used by aid agencies and to further increase anonymity.

Estimation of population flows

Population flows can be estimated by calculating a user's most frequent location for two non-overlapping, subsequent periods and comparing these two locations. This can be formalised as a matrix of location pairs and associated flows of users from an origin to a destination location, called an origin-destination matrix (OD matrix). This method has been used in a number of settings to measure population flows [7-10]. The aim for this application was more specific: to measure mobility resulting from the disaster, and with mobility under normal conditions being high in Nepal, we have to account for this high baseline mobility to measure movement in excess of this baseline.

To account for baseline mobility, we normalised the post-earthquake flows using pre-earthquake mobility estimates. Pre-earthquake estimates were calculated as an OD matrix using two periods before the earthquake. More specifically we compared a benchmark period from the beginning of the year (1st January until 7th April 2015) to a comparison period immediately before the earthquake (20th-24th April, note that this period is shorter, chosen to exclude population flows resulting from travel around the Nepali New Year festival) to obtain the pre-earthquake mobility estimate.
Post-earthquake flows were calculated as changes in location between the benchmark period (see above) and a focal period, which consisted of the most recent week of data at any point in time during the response period.
This normalisation provides us with a measure of how much traffic volumes following the earthquake deviate from the volumes that are a result of normal mobility.

To provide and allocate aid appropriately, humanitarian agencies require information on population flows. The flows extracted from the CDRs are numbers of users travelling in excess of the flows observed under normal conditions. To estimate population flows from user flows, we calculated scaling factors to be used to transform the flows from user estimates to approximate population estimates. The scaling factors were calculated by relating user and population numbers (using population counts provided by the WorldPop project) (http://www.worldpop.org.uk/) for each district.

Quantifying return rates

At later stages of relief efforts, an important measure is the rate of return to certain areas. Low return rates might indicate that recovery in an area is not sufficient yet to allow return of residents. To estimate return rates, we extract user's usual residence as the most frequent daily location over the benchmark period. A user is considered displaced if they spent a period of at least 7 days away from their usual residence in the two weeks following the earthquake. We can then calculate how many of the users considered displaced return to their place of usual residence in the weeks following.

Results

In the first weeks following the earthquake, large above normal flows were observed going out of the Kathmandu area into the surrounding areas. Figure 1 shows the districts that received above normal flows from Kathmandu for the period of 10-14 May 2015.
Figure 2 shows the lower level admin 4 districts (VDCs) coloured by the percentage of usual residents who remain after being displaced following the earthquake. This highlights that there is variation in recovery of areas.
The full paper [11] including additional results can be found at:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4779046/

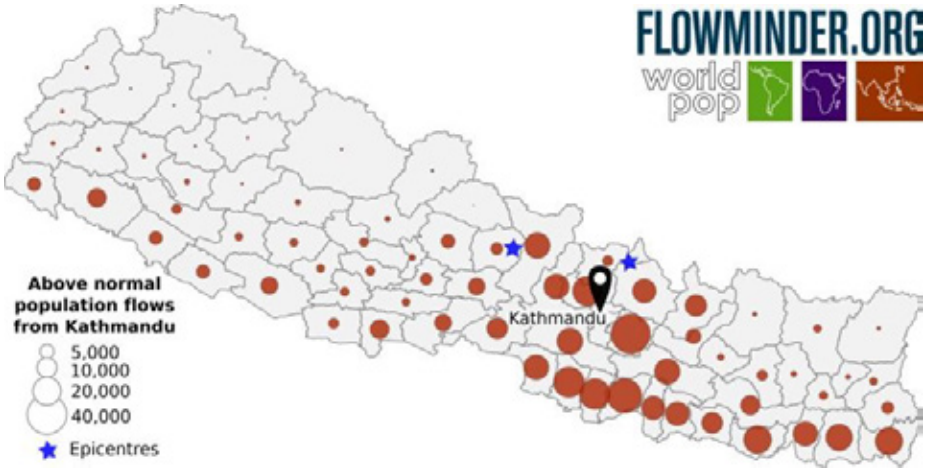


Figure 1: Outflows out of Kathmandu with above normal rates compared to baseline mobility. Boundaries shown are admin 3 (districts)

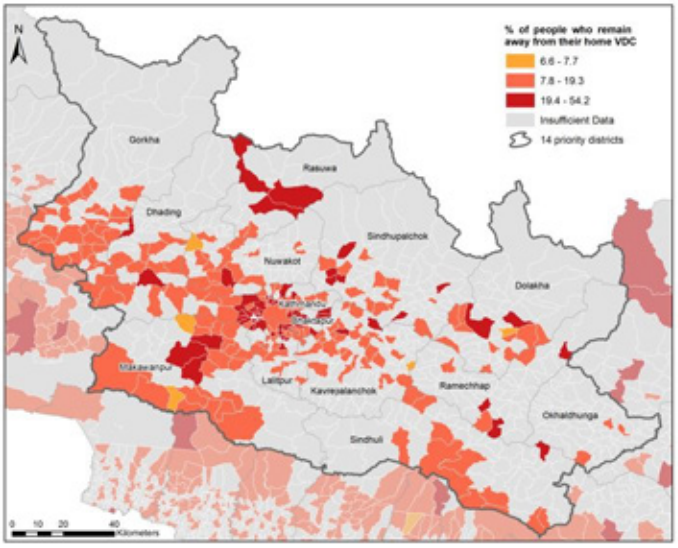


Figure 2: Percentage of people displaced by the earthquake who remain away as of 19 August 2015. District boundaries shown in grey, VDC (admin 4) boundaries shown in white.

References

[1] Le Menach, A., Tatem, A. J., Cohen, J. M., Hay, S. I., Randell, H., Patil, A. P., & Smith, D. L. (2011). Travel risk, malaria importation and malaria transmission in Zanzibar. *Scientific reports*, *1*.

[2] Calabrese, F., Colonna, M., Lovisolo, P., Parata, D., & Ratti, C. (2011). Real-time urban monitoring using cell phones: A case study in Rome. *IEEE Transactions on Intelligent Transportation Systems*, *12*(1), 141-151.

[3] Wesolowski, A., Metcalf, C. J. E., Eagle, N., Kombich, J., Grenfell, B. T., Bjørnstad, O. N., ... & Buckee, C. O. (2015). Quantifying seasonal population fluxes driving rubella transmission dynamics using mobile phone data. *Proceedings of the National Academy of Sciences*, *112*(35), 11114-11119.

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

[5] Sevtsuk, A., & Ratti, C. (2010). Does urban mobility have a daily routine? Learning from the aggregate data of mobile networks. *Journal of Urban Technology*, *17*(1), 41-60.

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

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

[8] Wesolowski, A., Buckee, C. O., Bengtsson, L., Wetter, E., Lu, X., & Tatem, A. J. (2014). Commentary: Containing the Ebola outbreak–the potential and challenge of mobile network data. *PLOS currents outbreaks*.

[9] Calabrese, F., Di Lorenzo, G., Liu, L., & Ratti, C. (2011). Estimating origin-destination flows using mobile phone location data. *IEEE Pervasive Computing*, *10*(4), 0036-44.

[10] Bengtsson, L., Gaudart, J., Lu, X., Moore, S., Wetter, E., Sallah, K., ... & Piarroux, R. (2015). Using mobile phone data to predict the spatial spread of cholera. *Scientific reports*, *5*.

[11] Wilson, R., zu Erbach-Schoenberg, E., Albert, M., Power, D., Tudge, S., Gonzalez, M., ... & Pitonakova, L. (2016). Rapid and near real-time assessments of population displacement using mobile phone data following disasters: the 2015 Nepal earthquake. *PLoS currents*, *8*.

Uncovering the Spread of Chagas Disease in Argentina and Mexico

Juan de Monasterio*, Alejo Salles*†, Carolina Lang*, Diego Weinberg‡, Martin Minnoni§, Matias Travizano§, Carlos Sarraute§
*FCEyN, Universidad de Buenos Aires, Argentina. Email: {*laterio, carolinalang93*}@*gmail.com*
†Instituto de Cálculo and CONICET, Argentina. Email: *alejo@df.uba.ar*
‡Fundación Mundo Sano, Buenos Aires, Argentina. Email: *dweinberg@mundosano.org*
§Grandata Labs, 550 15th Street, San Francisco, CA, USA.
Email: {*martin, mat, charles*}@*grandata.com*

***Abstract*—We use mobile phone records for the analysis of mobility patterns and the detection of possible risk zones of Chagas disease in two Latin American countries. We show that geolocalized call records are rich in social and individual information, which can be used to infer whether an individual has lived in an endemic area. We present two case studies, in Argentina and in Mexico, using data provided by mobile phone companies from each country. The risk maps that we generate can be used by public health campaign managers to target specific areas and allocate resources more effectively. Finally, we show the value of mobile phone records to predict long-term migrations, which play a crucial role in the spread of Chagas disease.**

I. Introduction

Chagas disease is a parasitic disease of global reach, spread mostly across 21 Latin American countries. Caused by the *Trypanosoma cruzi* parasite, its transmission occurs mostly in the American endemic regions via the *Triatoma infestans* insect family. In recent years and due to globalization and migrations, the disease has become an issue in other continents [1], particularly in countries that receive Latin American immigrants such as Spain and the United States.

A crucial characteristic of the infection is that it may last 10 to 30 years in an individual without presenting symptoms [2], which greatly complicates effective detection and treatment. Long-term human mobility (particularly seasonal and permanent rural-urban migration) thus plays a key role in the spread of the epidemic [3].

In this work, we discuss the use of Call Detail Records (CDRs) for the analysis of mobility patterns and the detection of possible risk zones of Chagas disease in two Latin American countries. We generate predictions of population movements between different regions, providing a proxy for the epidemic spread. We present two case studies, in Argentina and in Mexico, using data provided by mobile phone companies from each country.

II. Chagas Disease in Argentina and Mexico

A. Endemic Zone in Argentina

The *Gran Chaco*, situated in the northern part of the country, is hyperendemic for the disease [4]. The ecoregion's low socio-demographic conditions further support the parasite's lifecycle, where domestic interactions between humans, triatomines and animals foster the appearance of new infection cases, particularly among rural and poor areas. This region is considered as the endemic zone E_Z in the analysis described in Section IV.

Recent national estimates indicate that there exist between 1.5 and 2 million individuals carrying the parasite, with more than seven million exposed. National health systems face many difficulties to effectively treat the disease. In Argentina, less than 1% of infected people are treated (the same statistic holds at the world level). Even though governmental programs have been ongoing for years now, data on the issue is scarce or hardly accessible.

B. Endemic Zone in Mexico

The Mexican epidemic area, which includes the states having the top 25% prevalence rates nationwide [5], covers most of the South region of the country and includes the states of Jalisco, Oaxaca, Veracruz, Guerrero, Morelos, Puebla, Hidalgo and Tabasco. This region is considered as the endemic zone E_Z for the Mexican case.

Despite the lack of official reports, an estimate of the number of *Trypanosoma cruzi* infections by state in the country indicates that the number of potentially affected people in Mexico is about 5.5 million [6]. In recent years there has been a focus on treating the disease with two available medications, benznidazole or nifurtimox, with less than 0.5% of infected individuals receiving treatment in Mexico [7].

People from endemic areas of Chagas disease tend to migrate to industrialized cities of the country, mainly Mexico City, in search of jobs [8]. Therefore, the study of long-term mobility is crucial to understand the spread of the Chagas disease in Mexico.

III. Mobile Phone Data Sources

Our data source is anonymized traffic information from two mobile operators. The Argentinian dataset contains CDRs collected over a period of 5 months, from November 2011 to March 2012. The Mexican dataset contains CDRs for a period of 24 months, from January 2014 to December 2015.

For our purposes, each record is represented as a tuple $\langle i, j, t, d, l \rangle$, where user i is the caller, user j is the callee, t is the date and time of the call, d is the direction of the

call (incoming or outgoing), and l is the location of the tower that routed the communication. The dataset does not include personal information from the users.

We aggregate the call records for a five month period into an edge list $(n_i, n_j, w_{i,j})$ where nodes n_i and n_j represent users i and j respectively and $w_{i,j}$ is a boolean value indicating whether these two users have communicated at least once within the five month period. This edge list represents our communication 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. We note that only a subset $\mathcal{N}_C$ of nodes in $\mathcal{N}$ are clients of the mobile operator. Since geolocation information is available only for users in $\mathcal{N}_C$, in the analysis we considered the graph $\mathcal{G}_C = \langle \mathcal{N}_C, \mathcal{E}_C \rangle$ of communications between clients of the operator.

IV. Risk Maps for Chagas Disease

A. Methodology for Risk Map Generation

The first step it do determine the area in which each user lives. For each user $u \in \mathcal{N}_C$, we compute its *home antenna* H_u as the antenna in which user u spends most of the time during weekday nights [9]. The users such that H_u is located in the endemic zone E_Z are considered the *residents of* E_Z.

The second step it to find users highly connected with the residents of the endemic zone E_Z. To do this, we get the list of calls for each user and then determine the set of neighbors in the social graph $\mathcal{G}_C$. For each resident of the endemic zone, we tag all his neighbors as *vulnerable*.

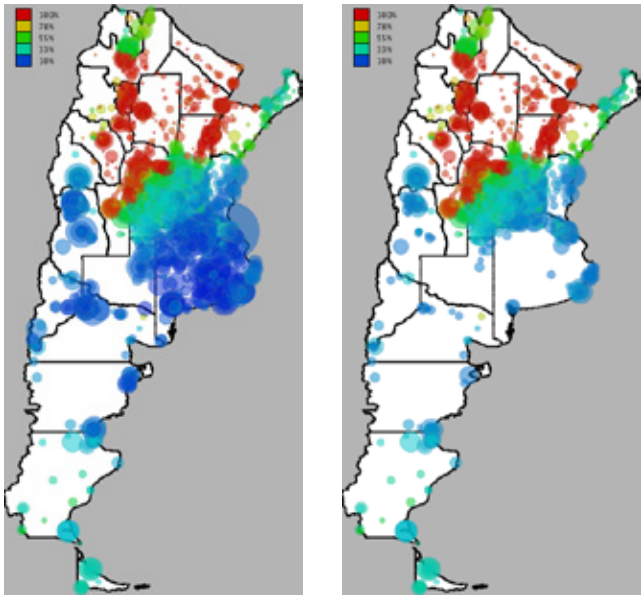
The third step is to aggregate this data for every antenna. For every antenna a, we compute: the total number of residents N_a, the total number of residents which are vulnerable V_a, the total volume of outgoing calls C_a, and the number of outgoing calls whose receiver lives in the endemic area VC_a (*vulnerable calls*).

We generated heatmaps to visualize these antenna indicators, overlapping these heatmaps with political maps of the region taken for study. Each antenna is represented by a circle whose *area* depends on the population living in the antenna N_a and whose *color* depends of the fraction V_a/N_a of vulnerable users living there. We used two filtering parameters: each antenna is plotted if its fraction of vulnerable users is higher than β, and if its population is bigger than m_v.

B. Results and Observations

Fig. 1 shows the risk maps for Argentina, generated with two values for the β parameter and fixing $m_v = 50$ inhabitants per antenna. After filtering with $\beta = 0.15$, we see that large portions of the country harbor potentially vulnerable individuals. Namely, Fig. 1(b) shows antennas where more that 15% of the population has social ties with the endemic region E_Z.

Advised by Mundo Sano Foundation's experts, we then focused on areas whose results were unexpected to the epidemiological experts. Focused areas included the provinces of Tierra del Fuego, Chubut, Santa Cruz and Buenos Aires, with special focus on the metropolitan area of Greater Buenos Aires whose heatmap is shown in Fig. 2.



(a) $\beta = 0.01$ (b) $\beta = 0.15$

Figure 1. Risk map for Argentina, filtered according to β.

High risk antennas were separately listed and manually located in political maps. This information was made available to the Mundo Sano Foundation collaborators who used it as an aid for their campaign planning and for the education of community health workers.

V. Prediction of Long-term Migrations

In this section, we describe our work on the prediction of long-term mobility. The CDR logs available in the Mexican dataset span 24 months, from January 2014 to December 2015, making them suitable for this study.

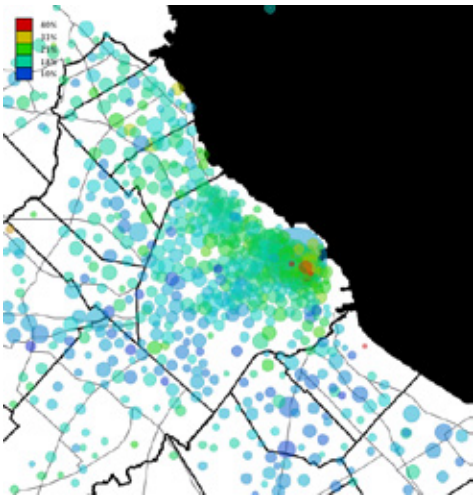


Figure 2. Risk map for the metropolitan area of Buenos Aires, filtered with $\beta = 0.02$.

We divide the available data into two distinct periods: T_0, from January 2014 to July 2015, considered as the "past" in our experiment; and T_1, from August 2015 to December 2015, considered as the "present". Knowing which users live in the endemic region E_Z and how they communicate during period T_1, we want to infer whether they lived in E_Z in the past (period T_0).

A. Model Features

The features constructed reflect calling and mobility patterns. Each week is divided into 3 time periods: (i) *weekday* is from Monday to Friday, on working hours (8hs to 20hs); (ii) *weeknight* is from Monday to Friday, between 20hs and 8hs of the following day; and (iii) *weekend* is Saturday and Sunday. The model consists of the following features, which can be classified in 4 categories:

1) Used and home antennas: For each user $u \in \mathcal{N}_C$, we register the top ten most used antennas, considering all calls or only calls made during the *weeknight* period. Users were tagged as 'endemic' if their home antenna is in the endemic zone E_Z and 'exposed' if any of the top ten antennas is in the risk area.

2) Mobility diameter: The user's logged antennas define a convex hull in space and the radius of the hull is taken to be as the mobility diameter. We generate two values, considering (i) all antennas and (ii) only the antennas used during the *weeknight*.

3) Communications graph: We enrich the social graph $\mathcal{G}_C$ built from the CDRs. For each edge $\langle n_i, n_j \rangle \in \mathcal{E}_C$, we gather the number of calls exchanged, the sum of call durations (in seconds), the direction (incoming or outgoing), segmented according to the periods *weekday*, *weeknight*, and *weekend*.

Since the samples in our dataset are users, we aggregate these variables by grouping interactions at the user level. The combination of different variables amounts to a total of 130 features per user.

We also compute the user's degree and the total count of endemic neighbors, labeling each user i as *vulnerable* whenever he has an edge with another user j who lives in the endemic region E_Z.

4) Validation data: We perform an analysis similar to the home antenna detection previously described, but considering the time period T_0 (from January 2014 to July 2015), in order to determine the home antenna of users during T_0.

B. Supervised Classification

We used most common techniques for this task: Support Vector Machines, Random Forest, Logistic Regression, and Multinomial Naive Bayes. The data was split into 70% for training and 30% for testing.

The Multinomial Bayes classifier has a linear time complexity, and thus serves as a fast benchmark. Support Vector Machines (SVM) and Logistic Regression performed better than Multinomial Bayes. We tuned the standard hyperparameters: $L2$-penalty regularization for Logistic Regression and kernel bandwidth for the Gaussian Kernel SVM. Both learning routines were executed in parallel and in each iteration 5% of the training set was sampled for cross validation. The best model was a Logistic Regression Classifier with an $L2$-penalty value of 0.01. The scores obtained by the selected model on the out-of-sample set are F1-score: 0.964537; accuracy: 0.980670; AUC: 0.991593; precision: 0.970838; recall: 0.958316.

High values across all scoring measures are achieved. These results can be explained by the fact that communication and mobility patterns are in essence highly correlated across time periods. In this case, a user being endemic in T_1 is correlated to being endemic in T_0, and the same holds with a user's interaction with vulnerable neighbors during T_1.

VI. Conclusion

The heatmaps shown in Section IV expose an expected "temperature" descent from the endemic regions outwards. We also found out communities atypical compared to their neighboring region, which stand out for their strong communication ties with the endemic region E_Z. The detection of these communities is of great value to health campaign managers, providing them tools to target specific areas and prioritize resources and calls to action more effectively.

In Section V, we tackled the problem of predicting long-term migrations. In particular, we showed that it is possible to use the mobile phone records of users during a bounded period in order to predict whether they have lived in the endemic zone E_Z in a previous time frame.

To conclude, we showed here the value of generating risk maps in order to prioritize effectively detection and treatment campaigns for the Chagas disease. The results stand as a proof of concept which can be extended to other countries with similar characteristics.

References

[1] Gabriel A Schmunis and Zaida E Yadon. Chagas disease: a Latin American health problem becoming a world health problem. *Acta tropica*, 115(1):14–21, 2010.

[2] Anis Rassi and Joffre Marcondes de Rezende. American trypanosomiasis (Chagas disease). *Infectious disease clinics of North America*, 26(2):275–291, 2012.

[3] Roberto Briceño-León. Chagas disease in the Americas: an ecohealth perspective. *Cadernos de Saúde Pública*, 25:S71–S82, 2009.

[4] OPS. Mapa de Transmisión vectorial del Mal de Chagas. *Organizacion Panamericana de la Salud*, 2014.

[5] A Cruz-Reyes and José M Pickering-López. Chagas disease in Mexico: an analysis of geographical distribution during the past 76 years-A review. *Memorias do Instituto Oswaldo Cruz*, 101(4):345–354, 2006.

[6] Alejandro Carabarin-Lima, María Cristina González-Vázquez, Olivia Rodríguez-Morales, Lidia Baylón-Pacheco, José Luis Rosales-Encina, Pedro Antonio Reyes-López, and Minerva Arce-Fonseca. Chagas disease (american trypanosomiasis) in Mexico: an update. *Acta tropica*, 127(2):126–135, 2013.

[7] Jennifer M Manne, Callae S Snively, Janine M Ramsey, Marco Ocampo Salgado, Till Bärnighausen, and Michael R Reich. Barriers to treatment access for Chagas disease in Mexico. *PLoS Negl Trop Dis*, 7(10):e2488, 2013.

[8] Carmen Guzmán-Bracho. Epidemiology of Chagas disease in Mexico: an update. *TRENDS in Parasitology*, 17(8):372–376, 2001.

[9] Balázs Cs Csáji, Arnaud Browet, VA Traag, Jean-Charles Delvenne, Etienne Huens, Paul Van Dooren, Zbigniew Smoreda, and Vincent D Blondel. Exploring the mobility of mobile phone users. *Physica A: Statistical Mechanics and its Applications*, 2012.

SESSION 5
HEALTH

Spatially explicit modeling of potential Ebola spread in Senegal

Lorenzo Mari*, Francesca Dagostin, Letizia Raffa,
Manuela Ciddio, Lorenzo Righetto, Marino Gatto, Renato Casagrandi*

Politecnico di Milano
Piazza Leonardo da Vinci, 32 – 20133 Milano, Italy

* corresponding authors: lorenzo.mari@polimi.it, renato.casagrandi@polimi.it

Abstract

Ebola virus disease (EVD) is a viral hemorrhagic fever with severe, and often fatal, consequences for the infected human hosts. In 2014, several West African countries were hit by the largest EVD epidemic ever recorded, which exerted a toll of thousands of victims in Liberia, Sierra Leone and Guinea. Only one case occurred in Senegal (a person who traveled by road to Dakar from Guinea), but a prompt sanitary response successfully contained the outbreak and prevented onward transmission. A thorough understanding of the main potential propagation pathways of EVD in each country of the area remains a crucial goal of sanitary policies, inasmuch as it can guide the optimization of the surveillance infrastructure, as well as emergency management should an EVD outbreak occur in the future.

We propose a spatially-explicit model for EVD transmission in Senegal accounting for the dynamics of the human population suitably divided into the compartments of susceptible, exposed, infectious and deceased people. The model is developed at the spatial scale of health districts. It is parameterized with literature data and georeferenced indicators available from demographic surveys describing the spatial heterogeneity of socioeconomic conditions throughout the country. Also, a large dataset of anonymous mobile phone traces obtained through the D4D-Senegal Challenge organized by Orange and Sonatel in 2014 is used to describe human mobility and social links (through aggregated source-destination matrices based on the mobile phone users' communication patterns), which are the key drivers for the propagation of EVD at various spatial scales. Conditions for epidemic outbreaks are determined via the linear stability analysis of the model, while numerical simulations are carried out to outline the ensuing spatiotemporal patterns of disease spread. The analysis of the model is performed under several scenarios describing different outbreak intensities and initial epidemic sources. Also, different strategies for epidemic control (involving e.g. early quarantine and treatment measures) are assessed, and their outcomes are quantitatively evaluated.

We show that the spatial heterogeneity induced by population distribution, human mobility and socioeconomic conditions play nontrivial, important, and possibly counterintuitive roles in determining the likelihood of EVD epidemics in Senegal. The densely populated region of Dakar, the main hub of human mobility within the country, is key in the early phases of an epidemic. Model simulations show in fact that outbreaks originating close to Dakar would be likely to produce a high number of cases, and that prompt and timely localized interventions might reduce the odds of large-scale epidemic spread. We thus conclude that spatially-explicit modeling, usefully assisted by the analysis of mobile phone traces, can elucidate the main propagation pathways of EVD in Senegal and anticipate the potential patterns of spread in the unfortunate case of an epidemic outbreak. Results from our analysis can be used to inform sanitary policies, and to guide the optimization of surveillance infrastructure, resource allocation and emergency management.

Keywords: call detail record analysis, spatially explicit modeling, epidemiology

Anticipatory Monitoring of Depressive States through the Analysis of Multimodal Phone Data

Abhinav Mehrotra
University College London
United Kingdom
a.mehrotra@ucl.ac.uk

Mirco Musolesi
University College London
United Kingdom
m.musolesi@ucl.ac.uk

ABSTRACT
Remarkable advances in smartphone technology, especially in terms of passive sensing, have enabled researchers to passively monitor user behavior in real-time and at a granularity that was not possible just a few years ago. Recently, different approaches have been proposed to investigate the use of different sensing and phone interaction features, including location, call, SMS and overall application usage logs, to infer the depressive state of users. In this paper, we propose an approach for monitoring of depressive states through the analysis of multimodal phone data. More specifically, we present the initial results of an ongoing study to demonstrate the association of depressive states with the indicators extracted from mobile phone traces.

Author Keywords
Mobile Sensing; Depression; Anticipatory Computing.

INTRODUCTION
Today's smartphone comes with an array of sensors and high-performance computing power. They are also carried by their owners all the time. These characteristics have not only enabled researchers to build very effective systems for passively monitoring numerous physical-context modalities such as users' location [5], physical activity [4] and mobile phone interaction [10], but also cognitive context [12], such as mood and well-being states.

However, cognitive context is inferred mostly by employing ESM techniques, according to which users are prompted with a series of questions that are required to be responded repeatedly. Past studies have shown the potential of exploiting mobile sensing data to learn and, potentially, predict the user's cognitive context [1, 2, 3]. For example, Canzian et al. have used mobility data to monitor depressive states [3] and Alvarez-Lozano at al. have exploited application usage logs to monitor patients affected by bipolar disorder [1].

The key limitation of these approaches is the fact that they rely on a single data source for monitoring depressive states. Instead, in this paper, we argue that depressive states should be monitored via the analysis of multimodal phone data. This might also contribute to the improvement of the performance of the machine learning algorithms for predicting depressive states.

Our focus instead is on monitoring *depressive states* that are quantified by means of a PHQ-8 questionnaire [7, 8] asked over a period of 14 days. Moreover, the authors of [9] do not consider the role of *micro-interactions* such as clicks, scrolls and many others, and reactions to notifications. Preliminary results of this study are presented in [11].

MEASURING MOBILE PHONE USAGE AND DEPRESSION SCORES
In order to study the association of users' depressive states on their mobile interaction behavior, we conducted a longitudinal field study. More specifically, we developed an Android app and collected smartphone interaction data from 25 participants for a time period of 30 days. The collected data include logs for notification handling and phone usage.

We use several smartphone interaction metrics to analyze the collected data, capturing various dimensions, from application usage to the number of clicks on the screen. The metrics are summarized in Table 1.

We also collect the responses to the PHQ-8 questionnaire from the users via an ESM approach. This data is then used to compute the depressive scores for each users. Finally, we compute the Kendall's Rank correlation coefficients to analyze the relationship between the severity of the depressive state and phone interaction metrics (based on notification, application and phone usage).

PRELIMINARY RESULTS
Past studies have used mobility, activity, application usage and communication data for inferring depressive state of users [6, 1, 3]. We hypothesize that there are additional features (which can be captured via mobile phones) that are associated with the changes in the user's depressive state. Therefore, in this paper we present the initial findings of our ongoing study to investigate the impact of depressive state on the micro-interaction data including notification and phone interaction data.

Group	Metric	Description
Notifications	Count	Total number of notifications clicked.
	Acceptance %	Percentage of notifications clicked out of total arrived.
	% Handled (Other Device)	Percentage of notifications that are not handled on phone out of total notifications arrived.
	Average Seen Time (ST)	Average of the seen time of all notifications. Here, seen time is the time from the notification arrival until the time the notification was seen by the user.
	Average Decision Time (DT)	Average of the decision time of all notifications. Here, decision time is the time from the moment a user saw a notification until the time they acted upon it (by clicking, launching its corresponding app or swiping to dismiss).
	Average Response Time (RT)	Average of the response time of all notifications. Here, response time is the sum of seen and decision times.
Phone Usage	Launch Count	Number of times applications are launched.
	App Count	Number of applications launched.
	App Usage Time	Time duration for which applications were used.
	Sig Launch Count	Number of times significant applications are launched.
	Sig App Unique Count	Number of significant applications launched.
	Sig App Usage Time	Time duration for which applications were used.
	Non-Sig Launch Count	Number of times non-significant applications are launched.
	Non-Sig App Count	Number of non-significant applications launched.
	Non-Sig App Usage Time	Time duration for which applications were used.
	Phone Usage Time	Time duration for which phone was used.
	Click Count	Number of clicks on the phone screen.
	Long Click Count	Number of long clicks on the phone screen.
	Unlock Count	Number of times the phone was unlocked.

Table 1. Description of phone interaction metrics.

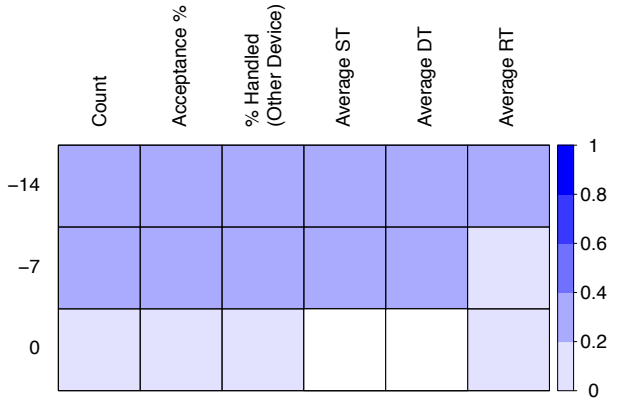


Figure 1. Results for correlation between depressive state and notification metrics.

Interpretation of Correlation Plots

The correlation results are presented as a plot of the correlation matrix. In this matrix the x-axis indicates the notification and phone interaction metrics and y-axis indicates the days for which the metrics are computed. For instance, in Figure 1 the box in the first column (*Count*) and first row (*-14*) presents the absolute value of the coefficient for the correlation of the depression score with the acceptance percentage of notifications, computed by using data from the current day to 14 days before the current day. Here, the current day refers to the day in which users reported their depressive state via PHQ-8 questionnaire. Moreover, we set the significance level α for the correlation results to 0.001 and non-significant correlation coefficients are indicated by the white boxes in the correlation plots.

Depressive State and Notifications

In Figure 1 we show the correlation coefficients that are computed to assess the relationship between depression score and notification metrics. The results show that users' depressive state moderately correlates with all metrics that are computed by using the past 14 days of data. The correlation results are the same for the metrics that are computed with past 7 days of data, except that the average DT has a weak correlation. On the other hand, users' depressive state does not correlate with the average ST and DT of notifications arriving on the current day when the user responded to the PHQ-8 questionnaire. Moreover, other metrics computed with the current day's data have a weak correlation with the depression score.

Depressive State and Phone Usage

In order to quantify the association between users' depressive state and their phone usage pattern we compute the correlation coefficients and present the results in Figure 2. The results show that users' depression score moderately correlates with all the metrics that are computed by using the past 14 days of data. On the other hand, users' depressive state weakly correlates with most of the metrics computed with the data of past 7 days and there is non-significant correlation with the metrics computed with the data of the current day.

SUMMARY

In this paper, we have presented our initial analysis of mobile phone usage traces for monitoring depressive states. The results are based on the data collected from 25 participants for

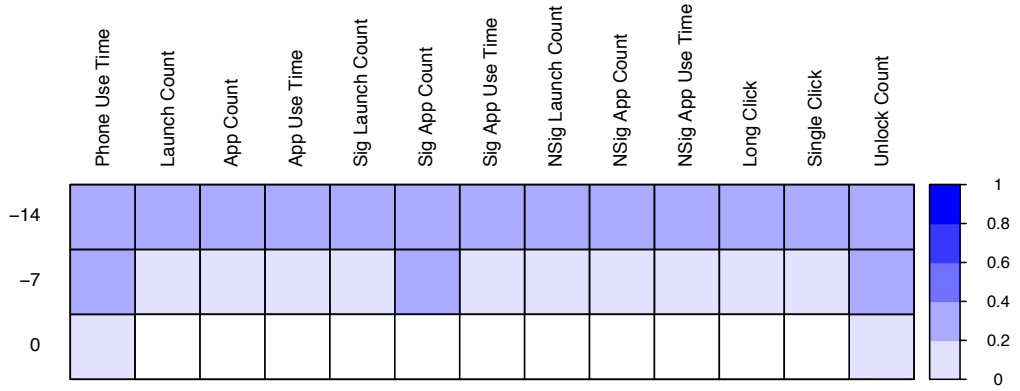


Figure 2. Results for correlation between depressive state and phone usage metrics.

a period of 30 days. Our results suggest that by using sensed data (including both notification and phone interaction data) of the past 14 days, it is possible to improve the accuracy of the prediction of the current day's depression score of a user. We envision this work to contribute in the direction of building a technique to learn and predict users' depressive state through the logs of their interaction with smartphones.

REFERENCES

1. Jorge Alvarez-Lozano, Venet Osmani, Oscar Mayora, Mads Frost, Jakob Bardram, Maria Faurholt-Jepsen, and Lars Vedel Kessing. 2014. Tell me your apps and I will tell you your mood: correlation of apps usage with bipolar disorder state. In *PETRA'14*.
2. Jakob E Bardram, Mads Frost, Károly Szántó, Maria Faurholt-Jepsen, Maj Vinberg, and Lars Vedel Kessing. 2013. Designing Mobile Health Technology for Bipolar Disorder: a Field Trial of the Monarca System. In *CHI'13*.
3. Luca Canzian and Mirco Musolesi. 2015. Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. In *UbiComp'15*.
4. Sunny Consolvo, David W McDonald, Tammy Toscos, Mike Y Chen, Jon Froehlich, Beverly Harrison, Predrag Klasnja, Anthony LaMarca, Louis LeGrand, Ryan Libby, and others. 2008. Activity Sensing in the Wild: a Field Trial of UbiFit Garden. In *CHI'08*.
5. N. Eagle, A. Clauset, and J. Quinn. 2009. Location Segmentation, Inference and Prediction for Anticipatory Computing. In *AAAI'09*.
6. Agnes Grünerbl, Patricia Oleksy, Gernot Bahle, Christian Haring, Jens Weppner, and Paul Lukowicz. 2012. Towards smart phone based monitoring of bipolar disorder. In *mHealthSys'12*.
7. Kurt Kroenke, Robert L Spitzer, Janet B. W. Williams, and Bernd Löwe. 2010. The patient health questionnaire somatic, anxiety, and depressive symptom scales: a systematic review. *General hospital psychiatry* 32, 4 (2010), 345–359.
8. Kurt Kroenke, Tara W. Strine, Robert L Spitzer, Janet B. W. Williams, Joyce T. Berry, and Ali H. Mokdad. 2009. The PHQ-8 as a measure of current depression in the general population. *Journal of Affective Disorders* 114, 1 (2009), 163–173.
9. Robert LiKamWa, Yunxin Liu, Nicholas D Lane, and Lin Zhong. 2013. Moodscope: Building a mood sensor from smartphone usage patterns. In *MobiSys'13*.
10. Abhinav Mehrotra, Robert Hendley, and Mirco Musolesi. 2016a. PrefMiner: Mining User's Preferences for Intelligent Mobile Notification Management. In *UbiComp'16*.
11. Abhinav Mehrotra, Robert Hendley, and Mirco Musolesi. 2016b. Towards Multi-modal Anticipatory Monitoring of Depressive States through the Analysis of Human-Smartphone. In *Adjunct UbiComp'16*. Heidelberg, Germany.
12. Kiran K. Rachuri, Mirco Musolesi, Cecilia Mascolo, Jason Rentfrow, Chris Longworth, and Andrius Aucinas. 2010. EmotionSense: A Mobile Phones based Adaptive Platform for Experimental Social Psychology Research. In *UbiComp'10*.

Assessing the use of mobile phone data to describe recurrent mobility patterns in spatial epidemic models

Cecilia Panigutti[1,2], Michele Tizzoni[2], Paolo Bajardi[3], Zbigniew Smoreda[4] Vittoria Colizza[5,2]
[1] Dipartimento di Fisica, Università degli Studi di Torino, via Giuria 1, Torino 10125, Italy
[2] ISI Foundation, via Alassio 11/C, Torino 10126, Italy
[3] Aizoon Technology Consulting, Str. del Lionetto 6, Torino, Italy
[4] Sociology and Economics of Networks and Services Department, Orange Labs, Issy-les-Moulineaux, France
[5] Sorbonne Universités, UPMC Univ Paris 06, INSERM, Institut Pierre Louis d'Epidémiologie et de Santé Publique (IPLESP, UMR−S 1136), Paris, France
Email: CP: cecilia.panigutti@gmail.com MT: michele.tizzoni@isi.it, PB: paolo.bajardi@aizoongroup.com, ZB: zbigniew.smoreda@orange.com, VC: vittoria.colizza@inserm.fr

Abstract

The recent availability of large-scale call detail record data has substantially improved our ability of quantifying human travel patterns with broad applications in epidemiology [1–6]. Notwithstanding a number of successful case studies, previous works have shown that using different mobility data sources, such as mobile phone data or census surveys, to parameterize infectious disease models can generate divergent outcomes [7–9]. Thus, it remains unclear to what extent epidemic modelling results may vary when using different proxies for human movements.
In this study, we present an extensive side-by-side comparison of simulated epidemics in France based on two commuting networks: one extracted from an official census survey and one from a large-scale mobile phone dataset. We have previously examined the two networks in terms of their statistical features, comparing their topology and distributions of travel flows, and found a good statistical agreement between the two [8]. In contrast, previous results based on simulated epidemics on the two networks have shown that simulation outcomes may vary substantially when using one dataset or the other, depending on the specific outbreak location and disease parameters.
Here, we thoroughly assess the adequacy of the mobile phone network to match epidemic patterns that have been generated by simulations using the census data. Our goal is to test the goodness of the mobile phone mobility network to replace the census survey mobility network, which is explicitly assumed to be the best representation of commuting patterns in France. To this aim, we compare the spatio-temporal properties of 658,000 simulated outbreaks originating from every possible seed of the mobility networks and quantify their similarity in terms of the epidemic invasion tree and arrival time of first infection.
We identify the features characterizing the outbreak seed nodes that best correlate the similarity between epidemic patterns and discuss how these results can help to assess the adequacy of mobile phone data to describe recurrent mobility patterns in spatial epidemic models. We find that similarity of simulated epidemics is significantly correlated to connectivity, traffic and population size of the seeding nodes, suggesting that the adequacy of mobile phone data for infectious disease models becomes higher when epidemics spread between highly connected and

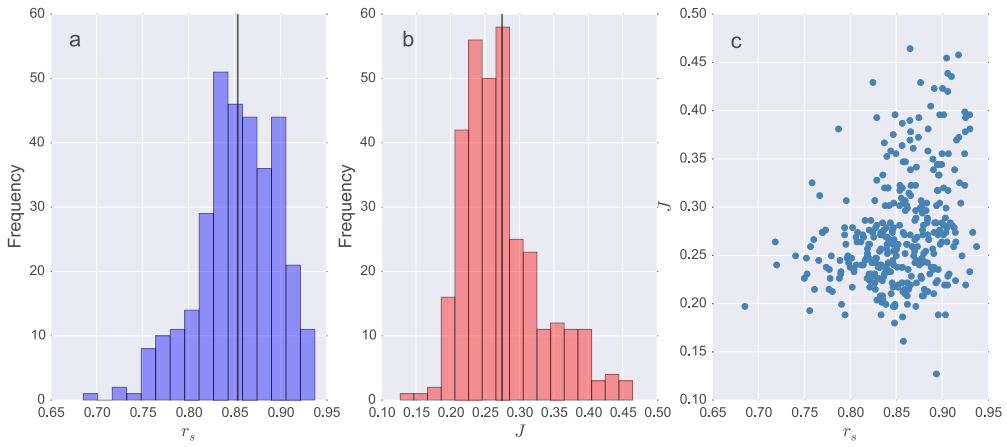


Figure 1: Distributions of similarity measures of epidemic simulations and their correlation. Frequency distributions of the Spearman's rank correlation coefficient measured between arrival times on the two networks (a) and the Jaccard similarity index between the infection trees on the two networks (b). Each value of r_s and $J(s)$ is computed over a statistical ensamble of 1,000 simulations for a given outbreak seed s. Both histograms correspond to 329 binned values, one for each node of the commuting networks, and solid lines indicate the average of the distributions. Panel c shows the relation between $J(s)$ and r_s for each node of the networks.

heavily populated locations, such as large urban areas.

References

[1] A. Wesolowski *et al.*, Proceedings of the National Academy of Sciences **112**, 11114 (2015).
[2] A. Wesolowski *et al.*, Proceedings of the National Academy of Sciences **112**, 11887 (2015).
[3] A. Wesolowski *et al.*, Science **338**, 267 (2012).
[4] L. Bengtsson *et al.*, Scientific Reports **5**, (2015).
[5] F. Finger *et al.*, Proceedings of the National Academy of Sciences 201522305 (2016).
[6] A. Lima, M. De Domenico, V. Pejovic, and M. Musolesi, Scientific reports **5**, 10650 (2015).
[7] A. Wesolowski *et al.*, PloS one **8**, e52971 (2013).
[8] M. Tizzoni *et al.*, PLoS Comp. Biol. **10**, e1003716 (2014).
[9] A. Wesolowski *et al.*, Scientific reports **4**, (2014).

Are you getting sick? Predicting Flu-Like Symptoms Using Human Mobility Behaviours

Gianni Barlacchi
University of Trento,
SKIL, Telecom Italia
Trento, Italy
barlacchi@fbk.eu

Christos Perentis
Fondazione Bruno Kessler
Trento, Italy
perentis@fbk.eu

Abhinav Mehrotra
University College London, UK
London, UK
a.mehrotra@ucl.ac.uk

Mirco Musolesi
University College London, UK
London, UK
m.musolesi@ucl.ac.uk

Bruno Lepri
Fondazione Bruno Kessler
Trento, Italy
lepri@fbk.eu

ABSTRACT

The mobility of individuals is of paramount importance for public health, particularly mobility characterization is key to predict the spatial and temporal diffusion of human-transmitted infections. However, the mobility behavior of a person can also reveal relevant information about her/his health conditions. In this paper, we study the impact of people mobility behaviours for predicting the future presence of flu-like symptoms (i.e. fever, cough and cold). To this end, we collect the mobility traces from mobile phones and the daily self-reported flu-like symptoms. First of all, we demonstrate that daily flu-like symptoms of a user can be predicted by using his/her mobility trace characteristics (e.g. total displacement, radius of gyration, number of unique visited places, etc.). Then, we present and validate models that are able to successfully predict the future presence of flu-like symptoms by analyzing the mobility patterns of our individuals.

Keywords

Computational Health, Human Mobility, Predictive Models

1. INTRODUCTION

The link between the health of a person and its behaviours is a well established connection and previous works have shown the possibility to predict health and well-being conditions using different sources of information from social media and mobile phones. In particular, recent studies have proved that human mobility represents a good proxy for understanding people's mental health conditions like depressive states [1]. These results basically suggests that, similarly to the depression case, human mobility can impact also on the physical health conditions of a person. In our paper, we focus on investigating the effectiveness of individual mobility behaviors for predicting the health conditions of a person. Specifically, we address the challenging problem of predicting future presence of influenza-like symptoms (e.g. fever and cough) by exploring the past mobility activities of a user, thus trying to answer the following question: *can mobility behaviours reveal something about the future health conditions of a person?*.

2. DATA

In this work we use a subset of the data collected during the Mobile Territorial Lab (MTL) study (for a more detailed description of the study, see [2]).

The study was conducted on 29 individuals. They have been selected among the participants in the MTL project (12 males and 17 females), with an age ranged from 28 to 50 years old. We use a combination two different type of location data and survey data with daily information about the health of the user. We focus on data from February 20, 2013 and March 20, 2013 since in this period we have an high presence of influenza-like symptoms.

Location Data. The software installed in the smartphone continuously keeps track of: (i) the communications events (e.g. call and SMS), and (ii) the user's location by means of the Global Positioning System (GPS), which recorded 82% of positions with an accuracy within 20 meters. In addition, to increase the number of location points we also used the cell towers to which the phone is connected. For each location point we kept track of the user ID, the latitude, the longitude, the accuracy and the timestamp.

Daily Symptoms Symptoms data were collected using a daily self-report survey instrument, designed by an experienced epidemiologist. The survey instrument consisted of six questions with yes/no responses for each of the following symptoms: fever, shivers, sore throat, cough, gasping breath, headache, muscular pain, malaise, cold.

3. METHODOLOGY

Our main goal is to study the relationship between mobility behavior and the self-reported flu symptoms. To do so, we need a set of characteristics that systematically describe human mobility behaviour. Canzian et al. [1] have recently introduced mobility metrics able to capture both presence and absence of human mobility. Such features appear to be promising in identifying health conditions, since many of them are related with the nature of the movement. For instance, in [1] they focus on depressive symptoms which could go along with decreased movement patterns and increased spending time at home for a long-term period. In our case, we expect to identify similar signals, but in a short-term context. We computed all the mobility features proposed by Canzian et al. [1]: total distance traveled, standard deviation of the total distance traveled, total displacements, standard deviation of the displacements, maximum displacement between two visited places, radius of gyration of the visited places, maximum displacement from home, number of different places visited, number of different significant places visited, number of moving geo-location points, routine index, unique number of visited places, diversity of the visited places, aggregated mobility features (rolling statistics computed for each of the aforementioned mobility features). Unfortunately, due to space limitations we remind the reader to the original paper [1] for more details.

3.1 Classification Model

We model our problem as a binary classification task, where the target variable is called *Symptoms Presence* and the possible values of the label is {*Yes/No*}, that is if a user has or not the symptoms. Given a target date, our ultimate goal is to understand if a user will present or not flu-like symptoms in the forthcoming days by looking into its very recent mobility behavior. We expect to capture even slight changes in the mobility behavior (e.g. changes in covered distance covered) that can testify an upcoming flu. Formally, given a date t we define:

- t_{hist} as the number of days we go back in user history data from the date t;
- historical time interval as the time interval $[t-t_{hist}, t]$;
- t_{hor} as the number of days ahead we answer our *Symptoms Presence: Yes/Not* question.

To sum up, we classify with the label *Yes* a user who presents flu-like symptoms at time t_{hor}, by using historical data in the interval $[t - t_{hist}, t]$.

Due to the limited size of the dataset, we decide not to build a specific model for each user. Instead, we design a relatively general machine learning framework that can work for every user u. A sample for the model is built using data when more than three consecutive days are available. Thus, given a date t, we consider valid a time window of five days if the following conditions are satisfied: (i) mobility data for $t_{hist} \in [0, 2]$; and (ii) symptoms data for the time t_{hor}. To give an example, with t = *Wednesday* we want to know if a user u will present influenza-like symptoms at *t+2 = Friday* considering her/his previous mobility behaviours from the time interval *t-2 = Monday*, t = *Wednesday*.

In order to carry out our experiments, we split the dataset in two parts: train and test. In the first step, we extract the features described in Section **??**. Due to the high number of features and the limited number of samples, we perform a feature selection step. First, we fit one of the classifier models. Then, we rank the features by the weight assigned in the model. For the classification task, we test four different machine learning models: Logistic Regression (LR), Conditional Random Forest (CRF), Gradient Boosted Trees (GBT) and Support Vector Classification (SVC) with a Gaussian radial basis function kernel. To select the best set of features, we experiment with different feature selection approaches varying the size of the feature set. We evaluate the quality of the feature selection through 10-fold-cross-validation, training the models with the reduced set of features on the training set. At this point, we can proceed with the parameters' optimization for each model by using the selected set of features. In both, feature selection and parameters selection, we choose an optimal set in order to maximize the precision of the algorithm. The last step regards the selection of the best model. Again, through cross-validation, we train each model with its best set of features and the optimal parameters selecting the one that shows the highest precision.

4. RESULTS

In our experiments we compare four different models (LR, CRF, GBT and SVC with a Gaussian radial basis function kernel) to classify if a user will present flu-like symptoms or not (i.e. fever, cough and cold) at a time t_{hor}. To train our models, we used the machine learning library scikit-learn [3]. Due to the unbalanced nature of our dataset, we used well-known metrics for assessing the accuracy of classification systems: (i) Precision and (ii) AUCROC.

4.1 Symptoms classification

In Table 4 we present the classification results in terms of precision and AUCROC. We report the different performances for $t_{hist} \in [-2, 0]$ and $t_{hor} \in [0, 2]$. The results are obtained with 10-fold-cross-validation and using the best setup for each different model.

Firstly, we observe that with $t_{hor} = 0$ (i.e, we consider a longer mobility history) we obtain better classification performance. This is a consequence to the fact that people change their mobility habits in the days before the registration of some influenza-like symptoms, that is, they change the mobility once they start to feel bad. For instance, if a person is getting sick, s/he would likely go home after work instead of doing other activities. Secondly, we can observe that as more days ahead we consider, less is the importance of the history and more difficult it becomes to classify correctly the presence of symptoms by only looking at the mobility behaviours. This reveal an interesting aspect related to the fact that there is a short period (e.g. few days) between feeling bad and reporting the symptoms. In sum, obtained results suggest that mobility behavior can be used for our purpose, but only looking at a short period in the future (e.g., $t_{hor} = 2$) and considering a limited historical period. A long history of mobility data is not relevant.

In overall for all the built models, the following selected features emerged as most important in predicting correctly the

		$t_{hist}=0$		$t_{hist}=1$		$t_{hist}=2$	
		Precision	AUCROC	Precision	AUCROC	Precision	AUCROC
$t_{hor}=0$	LR	65.2	52.2	59.0	57.9	54.1	55.6
	CRF	45.8	52.6	33.3	48.1	56.0	56.5
	SVC	58.8	60.3	45.4	53.4	**81.2**	60.2
	GBT	54.5	60.5	58.6	58.7	68.2	68.3
$t_{hor}=1$	LR	54.1	54.6	61.5	57.9	55.1	56.0
	CRF	54.0	57.0	52.9	55.9	50.0	53.6
	SVC	67.8	59.5	50.0	58.4	62.5	55.1
	GBT	56.5	60.5	**70.4**	69.4	63.4	63.7
$t_{hor}=2$	LR	61.1	59.9	61.5	61.0	61.5	61.0
	CRF	63.6	60.4	41.4	49.2	48.9	54.6
	SVC	65.6	59.9	55.5	53.4	62.5	61.9
	GBT	62.7	65.4	**65.9**	66.4	64.7	66.9

Table 1: Precision and AUCROC of the classifiers with 10-fold-cross-validation and variations of t_{hor} and t_{hist}.

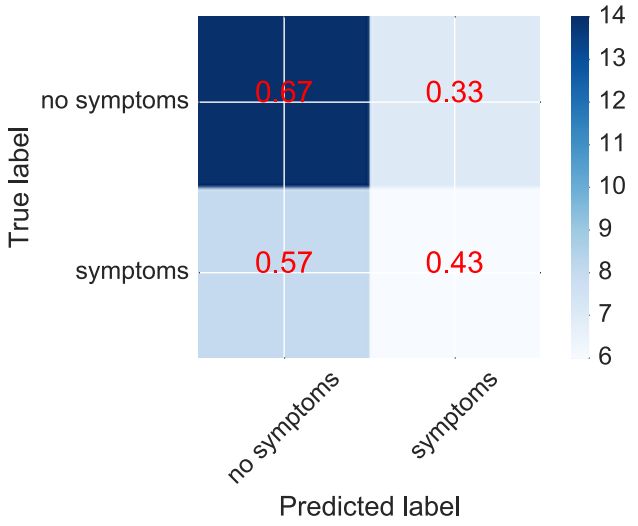


Figure 1: Confusion matrix fot $t_{hor}=1$.

presence of symptoms, namely: (i) the diversity of the visits (ii) the unique number of visited places, (iii) the number of different significant visited places, (iv) the number of moving geo-location points and (v) the aggregated mobility features. To complete our analysis, we also report in Fig. 1 the confusion matrix for the case $t_{hor}=1$. Our model presents good accuracy in classifying no symptoms while, mainly due to the difficult nature of the problem, there is still margin of improvements in the classification of symptoms.

5. CONCLUSIONS

In this work we investigated how to use individuals' mobility behaviour for a novel and challenging and task: predicting the future presence of flu-like symptoms such as fever, cough and cold. To this end, we used the mobility information collected by mobile phones and the daily self-reported flu-like symptoms of 29 individuals in the time interval from February 20 to March 21 of 2013. Previous work has exploited the use of mobility features to predict mental health and well-being dimensions such as positive and negative emotions, stress level, and depression symptoms. To the best of our knowledge, this work represents the first study that utilizes inference algorithms to predict the presence of influenza-like symptoms by only looking at the mobility behaviours of the user. Our results provide evidence supporting our approach and represent a promising starting point for dealing with influenza-like public health issues. The evolution of our proposed methodology could have a societal impact opening the way to customized mobile phone applications, which may detect suggest to the user specific actions in order to prevent disease spreading and minimize the risk of contagion.

6. REFERENCES

[1] L. Canzian and M. Musolesi. Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 1293–1304. ACM, 2015.

[2] S. Centellegher, M. De Nadai, M. Caraviello, C. Leonardi, M. Vescovi, Y. Ramadian, N. Oliver, F. Pianesi, A. Pentland, F. Antonelli, et al. The mobile territorial lab: a multilayered and dynamic view on parentsâĂŹ daily lives. *EPJ Data Science*, 5(1):1, 2016.

[3] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Impact of Human Mobility on Spread of Dengue in Sri Lanka

Lasantha Fernando
LIRNE*asia*, Colombo
Sri Lanka
Email: lasantha@lirneasia.net

Amal Shehan Perera
Dept. of CSE, University of Moratuwa
Sri Lanka
Email: shehan@cse.mrt.ac.lk

Sriganesh Lokanathan
LIRNE*asia*, Colombo
Sri Lanka
Email: sriganesh@lirneasia.net

Azhar Ghouse
Epidemiology Unit, Ministry of Health
Sri Lanka
Email: docazharghouse@gmail.com

Hasitha Tissera
Epidemiology Unit, Ministry of Health
Sri Lanka
Email: dr_korelege@yahoo.co.uk

Rohan Samarajiva
LIRNE*asia*, Colombo
Sri Lanka
Email: rohan@lirneasia.net

***Abstract*—Human mobility plays a significant role in spatio-temporal propagation of infectious diseases. But how much of an impact does human mobility have on propagating a dengue outbreak in a dengue endemic country such as Sri Lanka? We show that a proxy value for human mobility, derived from mobile network big data, has a significant correlation with dengue incidence using past case data. Furthermore, we discuss the applicability of this proxy measure of mobility to increase accuracy of several prediction models based on machine learning techniques. These improved models can be used to do spatio-temporal forecasting of dengue outbreaks which can help medical officers and related governmental or non-governmental agencies execute preemptive measures before the outbreak occurs in actuality.**

I. Introduction

The significance of human mobility in propagation of infectious diseases has been established in multiple works done previously [1, 2]. Until recently, the datasets and methodologies available for estimating human mobility have been limited. However, there has been a significant increase in research done to estimate human mobility using call detail records (CDR) with applications in multiple, wide ranging domains [3–5].

The main focus of our work is in measuring the impact of human mobility in propagating a vector borne tropical disease such as dengue in a dengue endemic country. A recent study from Pakistan [6] developed a model that validates the impact of human mobility in introducing dengue to naive regions as well as regions where dengue incidence is low. However in regions where dengue is already prevalent throughout the year, human mobility would not effect disease propagation in the same manner. The forecasting models developed for regions where the human population is immunologically naive to dengue would not be applicable to a country like Sri Lanka, where dengue is endemic to most regions of the country.

In the next section, we describe the proxy measure for human mobility developed by us, which we use to determine the correlation between that value and dengue incidence. We compare the correlation of this value with other values such as temperature and rainfall (factors previously identified in the literature as contributing factors to disease propagation [7]) to show that there is a significant correlation between human mobility and dengue incidence. In section III, we compare the accuracy of the proxy measure by incorporating the mobility value to multiple prediction models to determine whether mobility plays a significant role in spreading dengue in a dengue endemic country such as Sri Lanka.

II. Mobility Model

Most of the mobility models developed in related work have been based on the number of trips between different regions[6, 8]. Additionally, considering only overnight stays of subscribers in a region other than his or her home, an approach taken for a study on malaria in Namibia [9], is not suitable for modeling dengue propagation. The reason is that the primary vector of dengue, aedes aegypti, is known to be most active during daytime [10]. Our focus was on developing a mobility measure that can be applied to a geographic region so that it can be incorporated directly into machine learning models. In the context of disease forecasting, machine learning methods can be described as working broadly on the principle of identifying a set of attributes that can describe dengue incidence characteristics for a specific spatial region, and training a model to learn the weightage of those attributes that contributes toward disease incidence. For such a forecasting model, we would need a mobility measure that corresponds to each spatial and temporal unit of consideration.

In our project, the smallest spatial unit for which data is available is a Medical Officer of Health (MOH) division, which is an administrative area defined by Sri Lanka's Ministry of Health. Number of dengue cases reported from each MOH division in a given week for the year of 2013 was used to represent dengue incidence. Pseudonymized call detail records from multiple mobile operators in Sri Lanka for the year of 2013 was used to derive the mobility measure. The mobility model derived using the available mobile phone data is described below.

A. Probabilistic Mobility Model

The probabilistic model was developed by building upon available literature on the subject where it is assumed that the number of calls taken or received by a subscriber in a particular region is proportional to the amount of time spent in that region [11]. Our model builds on the same assumption and obtains a normalized mobility value for each MOH division. The mobility value for a subscriber is calculated by the number of calls made by a subscriber who is not a resident of that particular MOH in a given week. This value is normalized by dividing from the total number of calls taken by that subscriber in that week. This derived value per subscriber is aggregated to get a mobility value for an MOH division.

If we consider M as a set of all MOH divisions, and S as a set of all subscribers, our model can be defined as follows:

$CDR(m_i, s_j, w_k)$ = No. of CDR in MOH division m_i, for subscriber s_j during week w_k where $\forall m_i \in M, \forall s_j \in S$

Mobility of subscriber $\mathbf{s_j}$ at MOH $\mathbf{m_i}$ can be defined as

$$mob(m_i, s_j) = \frac{CDR(m_i, s_j, w_k)}{\sum_i^M CDR(m_i, s_j, w_k)} \quad (1)$$

where $\forall m_i \in \{M - Home(s_j)\}, \forall s_j \in S$

Mobility for MOH $\mathbf{m_i}$ can be defined as

$$mob(m_i) = \frac{\sum_j^N mob(s_j)}{N} \quad (2)$$

where N - No. of subscribers travelled to $\mathbf{m_i}$ for that week

III. Results & Prediction Models

The distance correlation was used as a measure to determine the dependence between dengue incidence and mobility. The correlation graph for mobility and other input features is depicted in Fig. 1. This graph shows that mobility has a significant correlation when compared to the other input parameters considered in our predictive models such as temperature, rainfall and the no. of dengue cases in previous weeks.

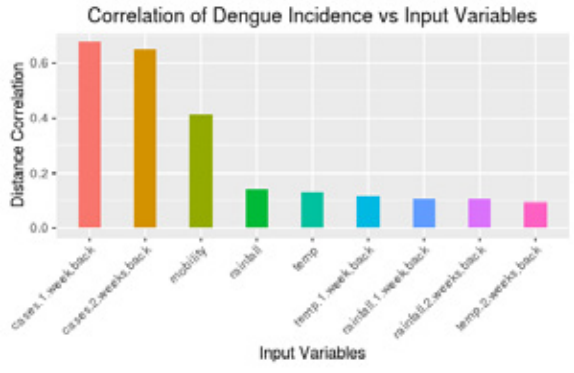


Fig. 1. Correlation of input parameters with dengue incidence

TABLE I
RMSE and r^2 values of different machine learning methods

Model	Mobility Used	RMSE	r^2
Neural Networks	No	10.923	0.194
Neural Networks	Yes	9.867	0.342
XGBoost [12]	No	10.023	0.321
XGBoost	Yes	9.548	0.384
SVR	No	7.445	0.139
SVR	Yes	7.297	0.173

Several different machine learning methods were tried out to come up with a dengue outbreak forecasting model with and without the derived mobility measure. The results from our own preliminary analysis is also given in Table I. The RMSE and the r^2 values for each method improved with the introduction of the mobility measure. Amongst the methods that were tried, Support Vector Regression (SVR) [13] provided the greatest accuracy, reflecting recent literature [14] where the performance of SVR has been shown to be better than Neural Networks. Therefore, SVR was selected as the method to carry out further work on our predictive model. The model was trained for 5 MOH divisions that were identified beforehand and the resultant model was used to predict for the sixth MOH division. After tuning the SVR model, we were able to obtain an RMSE value of 6.463 without mobility and 6.236 with the introduction of mobility. Similarly, the r^2 values were 0.351 and 0.396 for SVR models without and with mobility respectively. The related prediction graphs for Moratuwa MOH division are given in Fig.2 & 3.

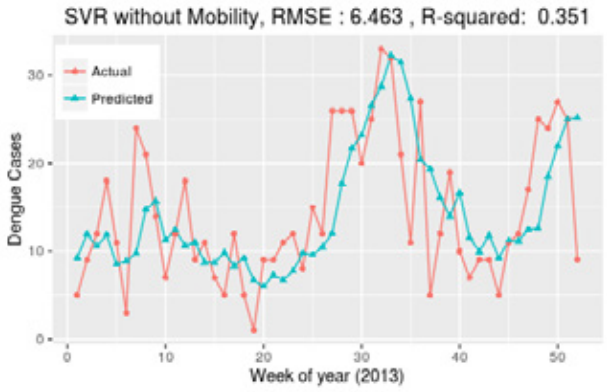


Fig. 2. Predicted vs actual dengue incidence for Moratuwa MOH (without mobility)

IV. Ongoing and Future Work

We are currently working on developing a trip based mobility model where the derived value can be aggregated to an MOH division. This model can be compared with the mobility model described above to determine which approach yields the greatest accuracy and fit. It can also be used as another validation of the observed improvement in predictive accuracy due to the introduction of mobility. Since the mobility measure

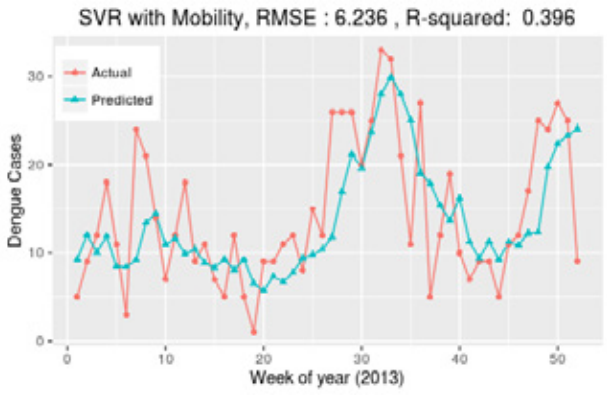


Fig. 3. Predicted vs actual dengue incidence for Moratuwa MOH (with mobility)

is derived independently of the modeling of the disease itself, this measure can be used in forecasting other infectious disease outbreaks as well. Validating the feasibility of using this measure in forecasting other infectious diseases can be done as a part of future work.

V. Conclusion

We have proposed a new methodology to calculate human mobility that provides a mobility measure for each administrative region. The advantage of this measure is it can be directly applied to any machine learning model without having to calculate a separate measure for different machine learning techniques. The results show that this measure is significantly correlated with dengue incidence. Additionally, we were able to obtain an improvement in prediction accuracy of all the considered models by introducing the mobility measure. Based on these results, we can conclude that mobility does have a significant impact in dengue propagation even in a dengue endemic country such as Sri Lanka.

Acknowledgment

The authors would like to thank the Epidemiology Unit, Ministry of Health, Sri Lanka for providing dengue incidence data as well as expert knowledge on the disease dynamics and the epidemiology of dengue. This research was funded through a grant from the International Development Research Centre (IDRC) of Canada and a grant from the Senate Research Committee (SRC) of University of Moratuwa.

References

[1] Michele Tizzoni et al. “On the Use of Human Mobility Proxies for Modeling Epidemics”. In: *PLOS Computational Biology* 10.7 (July 2014), pp. 1–15.

[2] Dirk Brockmann. “Human Mobility and Spatial Disease Dynamics”. In: *Reviews of Nonlinear Dynamics and Complexity* 2 (2010), pp. 1–24.

[3] Enrique Frias-Martinez, Graham Williamson, and Vanessa Frias-Martinez. “An Agent-Based Model of Epidemic Spread using Human Mobility and Social Network Information”. In: *3rd International Conference on Social Computing (SocialCom’11)* (2011), pp. 49–56.

[4] Sibren Isaacman et al. “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)* 6696 LNCS.June (2011), pp. 133–151. ISSN: 03029743.

[5] R Samarajiva et al. “Big data to improve urban planning”. In: *Economic and Political Weekly* 50.22 (2015), pp. 42–48. ISSN: 00129976 (ISSN).

[6] Amy Wesolowski et al. “Impact of human mobility on the emergence of dengue epidemics in Pakistan”. In: *Proceedings of the National Academy of Sciences* 112.38 (2015), pp. 11887–11892.

[7] Suchithra Naish et al. “Climate change and dengue: a critical and systematic review of quantitative modelling approaches.” In: *BMC infectious diseases* 14.1 (2014), p. 167. ISSN: 1471-2334.

[8] Amy Wesolowski et al. “Quantifying the impact of human mobility on malaria.” In: *Science (New York, N.Y.)* 338.6104 (Oct. 2012), pp. 267–70. ISSN: 1095-9203.

[9] Nick W Ruktanonchai et al. “Identifying Malaria Transmission Foci for Elimination Using Human Mobility Data.” In: *PLoS computational biology* 12.4 (2016), e1004846. ISSN: 1553-7358.

[10] Duane J Gubler and Gary G Clark. “Dengue/dengue hemorrhagic fever: the emergence of a global health problem.” In: *Emerging infectious diseases* 1.2 (1995), p. 55.

[11] Flavio Finger et al. “Mobile phone data highlights the role of mass gatherings in the spreading of cholera outbreaks”. In: *Proceedings of the National Academy of Sciences* 113.23 (2016), p. 201522305. ISSN: 0027-8424.

[12] Tianqi Chen and Carlos Guestrin. “XGBoost: A Scalable Tree Boosting System”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’16. San Francisco, California, USA: ACM, 2016, pp. 785–794. ISBN: 978-1-4503-4232-2.

[13] Debasish Basak, Srimanta Pal, and Dipak Chandra Patranabis. “Support vector regression”. In: *Neural Information Processing-Letters and Reviews* 11.10 (2007), pp. 203–224.

[14] Yuhanis Yusof and Zuriani Mustaffa. “Dengue Outbreak Prediction : A Least Squares Support Vector Machines Approach”. In: *International Journal of Computer Theory and Engineering* 3.4 (2011), pp. 489–493. ISSN: 17938201.

SESSION 6
DATA QUALITY AND PRIVACY

Geographical veracity of indicators derived from mobile phone data.

Maarten Vanhoof
Open Lab, Newcastle University, UK
Orange Labs, Paris, FR
M.vanhoof1@newcastle.ac.uk

Thomas Plötz
Open Lab, Newcastle University, UK
School of interactive computing,
Georgia Institute of Technology, USA

Zbigniew Smoreda
Orange Labs, Paris, FR

ABSTRACT

In this contribution we summarize insights on the geographical veracity of using mobile phone data to create (statistical) indicators. We focus on problems that persist with spatial allocation, spatial delineation and spatial aggregation of information obtained from mobile phone data. For each of the cases, we offer insights from our works on a French CDR dataset and propose both short and long term solutions. As such, we aim at offering a list of challenges, and a roadmap for future work on the topic.

1. Introduction

Ever since the first analyses on mobile phone data, individual indicators have been constructed to describe mobility patterns [1], presence patterns [2], aspects of social networks [3] or even personal characteristics [4]. Although originally used to explore the characteristics of large-scale datasets, several applications of mobile phone indicators have been developed over time. And this to a degree that real-time monitoring of socio-economic landscapes based on mobile phone indicators seems now plausible [5], [6].

One logical application of mobile phone indicators is to complement national statistics. As they capture behavior for large populations, the combination with census data (which typically offer contextual information) opens up exciting research opportunities. [6], for example, investigate the relation between deprivation of the neighborhood and the diversity of individual mobility by pairing national statistics with mobile phone data of 20 million users in France. Apart from the obvious confrontation with official statistics, other applications integrate mobile phone indicators for a multitude of purposes. Mobility indicators, for instance, are used to segment users in epidemiological model [7] or to express their 'vulnerability' to wrongful home detection [8]. Indicators on calling behavior can be used to determine social influencers, to segment customers for marketing purposes [9], to annotate social networks [10], etc..

Here, our argument is that, despite many applications, important questions remain on how to evaluate the veracity of indicators produced from mobile phone data. Specifically, we focus on three different spatial aspects that influence uncertainty and error when creating or using mobile phone based indicators: spatial allocation, spatial delineation and special aggregation.

2. Spatial allocation

Often, individual mobile phone users have to be allocated in space in order to enable further analysis of indicators derived from their mobile phone use. Typically, such allocation happens by means of home detection algorithms that incur the most plausible cell-tower to cover a user's home location based on rather simple heuristics (e.g. maximum activities during nighttime). Until now, little work has been done to assess and compare the performance of different heuristics as good validation datasets are hard to find (a.o. because of large population size, differences in cover grids with census, etc.). As a consequence, it is difficult to evaluate the uncertainty for allocating users, and thus indicators, to geographical areas [8].

To address the spatial allocation problem, in [8] we investigated the performance of five simple heuristics for home detection on a French CDR dataset. We applied each heuristic to almost 120 million mobility traces derived from one month of mobile phone data and compared results between them. Additionally, we collaborated with the French National Statistics Office (INSEE) to create a validation dataset describing the French population at the spatial resolution on the cell-tower grid provided by the operator. This allowed us to compare full population numbers with estimated population based on mobile phone data and thus assess performance of the different algorithms both in time as in space (figure 1).

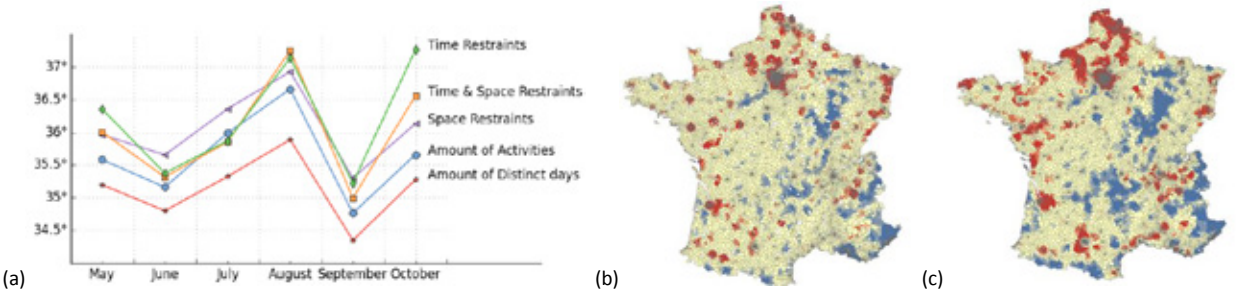


Figure 1: (a) Cosinus Similarity Values (in °) for the comparison between the vectors of (i) the number of detected user homes by the given home detection algorithm and (ii) the total population in the validation dataset based on census data. Vectors are made up of the totals of (i) and (ii) for each cell-tower. A value of 0° denotes the highest possible similarity between both vectors, 90° indicates the lowest similarity in orientation whereas 180° degrees refers to an opposite orientation. (b) and (c) Hotspots (red) and coldspots (blue) defined by the 90+% interval of the Getis-Ord Gi* statistic for (a) the number of detected homes by the amount of activities algorithm in June and (c) the population numbers of the validation dataset based on census data. The map is a made up by the Voronoi tesselation based on cell-tower locations.

3. Spatial delineation

Uneven delineations of space are a second problem when pursuing analysis with mobile phone indicators. Cell-tower cover areas, for example, typically have different boundaries compared to administrative regions, resulting in translation and comparisons problems. Uneven spatial delineations also exist between cell-tower cover areas. Cell-tower density in high population density areas, for instance, is typically higher, resulting in smaller cover areas compared to lower population density areas. Logically, this is directly relevant for population density estimations [11], the creation of mobility indicators [12], or parameter estimation in statistical analysis (e.g. urban scaling laws [13]).

In [12] we address the spatial delineation problem by assessing its effect on the creation of one mobility indicator from mobile phone data. As illustrated in figure 3 (a), we unveiled the calculation of mobility entropy (as proposed in [1], [6]) to be dependent on the density of cell-towers and thus the spatial delineation of cell-tower cover areas. To counter this, we propose a correction of the mobility entropy indicator. Application to a French CDR dataset learns that our proposed Corrected Mobility Entropy is less dependent on cell-tower densities and exhibits a different spatial pattern. We find suburban areas of large cities to depict the highest diversity of movement, compared to city centers for the standard mobility entropy (figure 3 (b) and (c)), as well as a clear decrease in mobility entropy with city size.

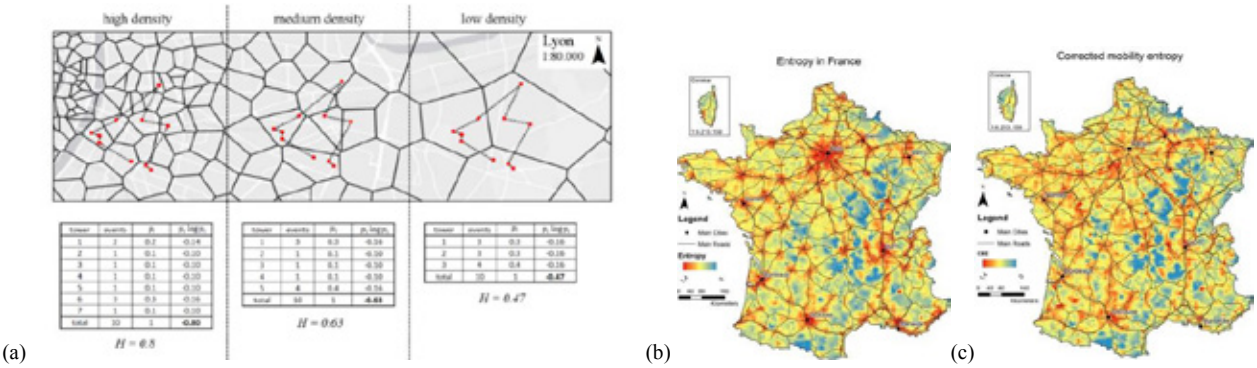


Figure 3: (a) Illustration of the effect of spatial delineation on the calculation of the temporally-uncorrelated entropy (H). The mobility entropy for the same path (dotted line) of a user in three different density settings is calculated and shown to be different. Spatial patterns of mobility entropy values (a) and corrected mobility entropy values (b). Values are calculated for all cell-towers in France as the average value of all users that have a 'home' at these cell-towers.

4. Spatial aggregation

A third problem consists in the spatial aggregation of mobile phone indicators. Although related to the spatial delineation problem, with spatial aggregation we do not mean by which spatial delineation one aggregates. Rather we wonder how the properties of indicators change when aggregating to another spatial level. This is especially relevant for mobile phone based indicators as the choice of scale will imply

unintended selective filtering (e.g. highly active persons, communities), a change in representativeness of population (e.g. unequal market shares for single operators between city and rural areas), a change in human behavior (e.g. short-distance vs. long-distance trips) and infrastructural context (e.g. transport infrastructure); all of which are typically unaddressed in studies using mobile phone data.

Currently, there is little evidence that research is taking into account any (plausible) effects of spatial aggregation. Even though, in our work, we find clear indications that different scales imply different findings like, for example, in the relations between mobile phone indicators, home detection algorithms and census information (table 1). It remains an open question as to what mechanisms are responsible.

Table 1: Correlation coefficients between different indicators at different spatial levels in France. Administrative level is the official IRIS level and municipality level is the commune in French official statistics. Cell-tower level corresponds to estimated cover areas of Orange cell-towers in France. EDI is the European Deprivation Index, Mobility Entropy is as calculated in [12]. Distinct days home detection is also used in figure 1 and described in [8].

Correlation coefficient for	Cell-tower Level	Administrative Level	Municipality Level	Scale difference
Mobility Entropy vs. EDI		-0.03	-0.43	-0.40
Distinct Days Home Detection vs. Population in Census data	0.62	0.92		+0.30

5. Roadmap

Ultimately, we recognize that procedures to create mobile phone based indicators are being institutionalized in, at least, two ways. On the one hand, there is published research and the development of open source software packages like, for example, the python toolbox bandicoot [14]. Such packages allow researchers to easily calculate indicators but their recurrent use might eventually lead to ill-considered applications being legitimized by the toolbox instead of the specific nature of the research case. In this perspective, it's a bit uncanny that current open-source packages offer little consideration to the assessment of uncertainty or error and, as a consequence, to potentially wrongful deployment or interpretation.

On the other hand, national statistic offices are putting considerable efforts to investigate the usability of mobile phone indicators within national statistics. Here, extensive validation of mobile phone indicators is the norm, given that, when ensured by the office, they will directly inform policy decisions. The problem, however, is that the requirements, procedures and measures used to validate and publish official statistics are not (yet) adapted to the nature of current big data sets, including mobile phone data. The assessment of veracity of mobile phone data hence becomes a matter of re-invention of the national statistics services, which likely will take several years before being fully operational.

Within this context, we believe it is necessary for future work to investigate, or at least openly communicate on, the different forms of uncertainties and errors that occur because of spatial allocation, spatial delineation, and/or spatial aggregation. Table 2 offers a more specific roadmap for actions that can be taken in the short and long term to prevent naïve application of mobile phone based indicators.

Table 2: Proposed short term actions and long term solutions to the spatial allocation, delineation and aggregation problem.

Problem	Short term actions	Long term solutions
Spatial allocation	- Investigate errors that come with spatial allocation - Test heuristics for home detection on different databases - Design surveys to gather ground truth at individual level	- Understand how characteristics of current mobile phone use (frequency, social context, etc.) influence spatial allocation - Standardize error assessment for spatial allocation
Spatial delineation	- Assess the influence of spatial delineation on indicators - Develop techniques that allow translation between different spatial delineations	- Develop standard practices that incorporate the effect of spatial delineation on indicators - Reflect on possibilities to standardize spatial delineations like cell-tower areas, administrative boundaries, etc.)
Spatial aggregation	- Develop techniques that define optimal spatial scale for studying a specific process (be it theoretical or empirical) - Develop techniques that express sensitivity of data (interpretation) to spatial aggregation	- Develop techniques that are capable of dealing with changing nature of observations when spatially aggregating - Understand how the problem of spatial aggregation changes in time due to changing mobile phone use and human behavior

References

[1] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of Predictability in Human Mobility," *Science (80-.).*, vol. 327, no. 5968, pp. 1018–1021, Feb. 2010.

[2] M. Janzen, M. Vanhoof, and K. W. Axhausen, "Estimating Long-Distance Travel Demand with Mobile Phone Billing Data," *16th Swiss Transp.*, 2016.

[3] 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*, 2011, pp. 1100–1108.

[4] Y.-A. de Montjoye, J. Quoidbach, F. Robic, and A. S. Pentland, "Predicting personality using novel mobile phone-based metrics," in *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*, 2013, pp. 48–55.

[5] F. Giannotti, D. Pedreschi, A. Pentland, P. Lukowicz, D. Kossmann, J. Crowley, and D. Helbing, "A planetary nervous system for social mining and collective awareness," *Eur. Phys. J. Spec. Top.*, vol. 214, no. 1, pp. 49–75, 2012.

[6] L. Pappalardo, M. Vanhoof, L. Gabrielli, Z. Smoreda, D. Pedreschi, and F. Giannotti, "An analytical framework to nowcast well-being using mobile phone data," *Int. J. Data Sci. Anal.*, Jun. 2016.

[7] A. Lima, V. Pejovic, L. Rossi, M. Musolesi, and M. Gonzalez, "Progmosis: Evaluating Risky Individual Behavior During Epidemics Using Mobile Network Data," *arXiv Prepr. arXiv*, vol. abs/1504.0, 2015.

[8] M. Vanhoof, F. Reis, Z. Smoreda, and T. Plötz, "Investigating Performance and Spatial Uncertainty of Home Detection Criteria for CDR data."

[9] P. Sundsøy, J. Bjelland, A. M. Iqbal, Y.-A. de Montjoye, and others, "Big Data-Driven Marketing: How machine learning outperforms marketers' gut-feeling," in *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*, 2014, pp. 367–374.

[10] J. L. Toole, C. Herrera-Yaqüe, C. M. Schneider, and M. C. González, "Coupling human mobility and social ties," *J. R. Soc. Interface*, vol. 12, no. 105, 2015.

[11] P. Deville, C. Linard, S. Martin, M. Gilbert, F. R. Stevens, A. E. Gaughan, V. D. Blondel, and A. J. Tatem, "Dynamic population mapping using mobile phone data," *Proc. Natl. Acad. Sci.*, vol. 111, no. 45, pp. 15888–15893, Nov. 2014.

[12] M. Vanhoof, W. Schoors, A. Van Rompaey, T. Plötz, and Z. Smoreda, "Correcting Mobility Entropy for Large-Scale Comparison of Individual Movement Patterns."

[13] C. Cottineau, E. Hatna, E. Arcaute, and M. Batty, "Paradoxical interpretations of urban scaling laws," *arXiv Prepr. arXiv1507.07878*, 2015.

[14] Y.-A. de Montjoye, L. Rocher, and A. S. Pentland, "bandicoot: a Python Toolbox for Mobile Phone Metadata," *J. Mach. Learn. Res.*, vol. 17, no. 175, pp. 1–5, 2016.

Preserving Mobile Subscriber Privacy in Open Datasets of Spatiotemporal Trajectories

Marco Gramaglia*, Marco Fiore†, Alberto Tarable†, Albert Banchs*

* IMDEA Networks Institute & Universidad Carlos III de Madrid
Avda. del Mar Mediterraneo, 22
28918 Leganes (Madrid), Spain
Email: name.surname@imdea.org

† CNR-IEIIT
Corso Duca degli Abruzzi, 24
10129 Torino, Italy
Email: name.surname@ieiit.cnr.it

***Abstract*—Disclosure of large-scale datasets of individual trajectories collected by mobile network operations raises significant privacy concerns, yet it offers an unprecedented outlook on the fine-grained activities of whole populations. Technical solutions to the problem of privacy-preserving publishing of spatiotemporal trajectories of mobile subscribers are key to the open circulation of such rich datasets. We introduce an original privacy criterion, named $k^{\tau,\epsilon}$-anonymity, which tackles probabilistic attacks against trajectory data. We also propose a proof-of-concept algorithm that is a first step towards practical $k^{\tau,\epsilon}$-anonymity.**

I. Context

Subscriber trajectory datasets collected by network operators are logs of timestamped, georeferenced events associated to the communication activities of individuals. The analysis of these datasets allows inferring *fine-grained* information about the movements, habits and undertakings of vast user populations. This has many different applications, encompassing both business and research. For instance, they can be monetized via added-value services such as transport analytics [1] or location-based marketing [2]. In addition, large-scale movement data has proven critical to studies in physics, sociology or epidemiology [3].

All these use cases stem from the disclosure of trajectory datasets to third parties. However, the open release of such data is still largely withhold, which hinders potential usages and applications. A major barrier in this sense are privacy concerns: data circulation exposes it to re-identification attacks, and cognition of the movement patterns of de-anonymized individuals may reveal sensitive information about them.

This calls for anonymization techniques. The common practice operators adhere to is replacing personal identifiers (e.g., name, phone number, IMSI) with pseudo-identifiers (i.e., random or non-reversible hash values). Whether this is a sufficient measure is often called into question, especially in relation to the possibility of tracking user movements. What is sure is that pseudo-identifiers have been proven not to protect against trajectory uniqueness, i.e., the fact that subscribers have distinctive travel patterns that make them uniquely recognizable in very large populations [4]–[6].

II. Privacy in spatiotemporal trajectories

We argue that a major threat to spatiotemporal trajectory data is represented by *probabilistic attacks*. The aim of probabilistic attacks is letting an adversary with partial information about an individual enlarge his knowledge on that individual by accessing the target database. The uniqueness of mobile subscriber trajectories is a significant weakness that probabilistic attacks can exploit. Let us imagine a scenario where an adversary knows a small set of spatiotemporal points in the trajectory of a subscriber (because, e.g., he met the victim there), and has access to a dataset of spatiotemporal trajectories of (millions of) inhabitants of the city where the victim lives such as, i.e., a call detail record dataset. The high uniqueness of spatiotemporal trajectory data implies a very high risk that the attacker can use his knowledge to easily pinpoint his victim in the dataset, and retrieve his/her complete mobility. Clearly, revealing the complete movements enables inference of sensitive information about the victim, such as home/work locations, daily routines, visits to healthcare structures or nightclubs.

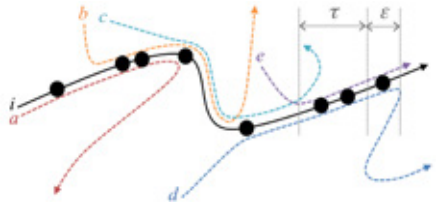


Fig. 1. Illustrative example of $k^{\tau,\epsilon}$-anonymity of user i, with k=2.

The pertinent principle to counter probabilistic attacks on spatiotemporal trajectory data is the so-called *uninformative principle*. It ensures that the difference between the knowledge of the adversary before and after accessing a database is small [7]. In our context, this principle warrants that an attacker who knows some subset of a subscriber's movements cannot extract from the dataset a substantially longer portion of that user's trajectory.

To attain the uninformative principle, we introduce the $k^{\tau,\epsilon}$-*anonymity* privacy criterion. $k^{\tau,\epsilon}$-anonymity can be seen as a variation of k^m-*anonymity*, which establishes that each individual in a dataset must be indistinguishable from at least $k-1$ other users in the same dataset, when limiting the attacker knowledge to any set of m attributes [8]. $k^{\tau,\epsilon}$-anonymity tailors k^m-anonymity to our scenario by defining:

(i) The attacker knowledge τ. It can be any continued sequence of spatiotemporal samples covering a time interval of length at most τ: thus, the m parameter of k^m-anonymity maps to

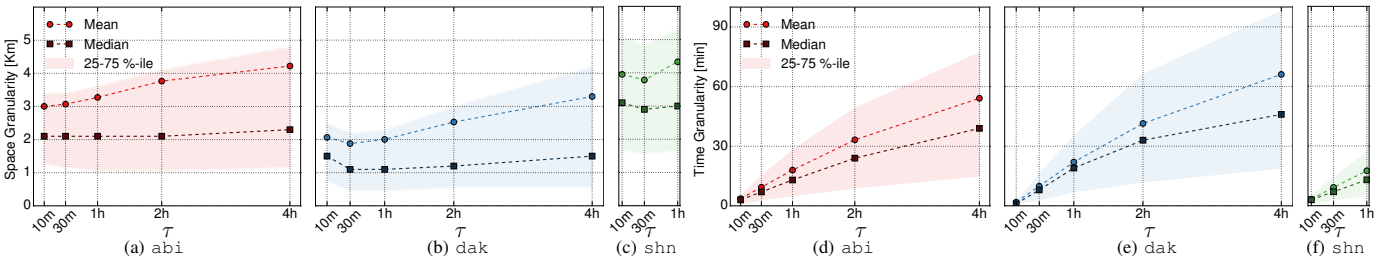


Fig. 2. Spatial (a-c) and temporal (d-f) granularity versus the adversary knowledge τ in the reference datasets. The spatial granularity is expressed as the sum of spans along the Cartesian axes. For instance, 1 km maps to, e.g., a square of side 500 m.

the (variable) set of samples contained in any time period τ. During any such time period, every trajectory in the dataset must be indistinguishable from at least other $k-1$ trajectories. *(ii)* The *leakage* ϵ. It is the maximum additional knowledge that the attacker is allowed to learn. It consists of the spatiotemporal samples of the target user's trajectory contained in a time interval of duration at most ϵ, disjoint from the original τ. In order to fulfill the uninformative principle, the leakage ϵ must be small.

The two requirements above imply alternating in time the $k-1$ trajectories that provide anonymization. An intuitive example of the rationale behind $k^{\tau,\epsilon}$-anonymity is provided in Fig. 1. There, the trajectory of a target user i is $2^{\tau,\epsilon}$-anonymized using those of five other subscribers. The overlapping between the trajectories of a, b, c, d, e and that of i is partial and varied. An adversary knowing a sub-trajectory of i during any time interval of duration τ always finds at least one other user with a movement pattern that is identical to that of i during that interval, but different elsewhere. With this knowledge, the adversary cannot tell apart i from the other subscriber, and thus cannot attribute full trajectories to one user or the other. As this holds no matter where the knowledge interval is shifted to, the attacker can never retrieve the complete movement patterns of i: this achieves the uninformative principle. Still, the adversary can increase its knowledge in some cases. Let us consider the interval τ indicated in the figure: the trajectories of i, d and e are identical for some time after τ, which allows associating to i the movements during ϵ: the opponent learns one additional spatiotemporal sample of i.

III. Achieving $k^{\tau,\epsilon}$-anonymity

We propose algorithms that are a first step towards a complete solution to transform a dataset of spatiotemporal trajectories so that it fulfils the uninformative principle. We refer the reader to the extended version of this abstract [9] for further details. Our algorithms implement, through generalization and suppression, the $k^{\tau,\epsilon}$-anonymity criterion for all trajectories in the dataset, under the assumptions that the attacker can track his victim continuously during any amount of time τ, and learn all spatiotemporal samples in the victim's trajectory over that timespan.

We evaluate our anonymization algorithms with three real-world datasets of mobile subscriber trajectories extracted from call detail records (CDR), and released by Orange within their D4D Challenges, and by the University of Minnesota. They are denoted as `abi`, `dak` and `shn`, and describe the spatiotemporal trajectories of tens of thousands mobile subscribers in the urban regions of Abidjan, Dakar and Shenzhen.

While our algorithms guarantee that $k^{\tau,\epsilon}$-anonymity is achieved by construction, it is interesting to assess the level of accuracy retained in the anonymized data. Fig. 2 portrays the mean, median and first/third quartiles of the sample granularity in the $k^{\tau,\epsilon}$-anonymized citywide datasets `abi`, `dak` and `shn`. The plots show how results vary when the adversary knowledge τ ranges from 10 minutes to 4 hours (except for `shn`, whose limited temporal span prevents us from testing attacks with knowledge τ higher than one hour), and $\epsilon = \tau$.

We remark how the $k^{\tau,\epsilon}$-anonymized datasets retain significant levels of accuracy, with a median granularity in the order of 1-3 km in space and below 45 minutes in time. These levels of precision are largely sufficient for most analyses on mobile subscriber activities, as discussed in, e.g., [10]. The temporal granularity is negatively affected by an increasing adversary knowledge τ, which is expected. Interestingly, however, the spatial granularity is only marginally impacted by τ: protecting the data from a more knowledgeable attacker does not have a significant cost in terms of spatial accuracy.

Overall, these results show that our algorithms attains $k^{\tau,\epsilon}$-anonymity of real-world datasets of mobile traffic, while maintaining a remarkable level of accuracy in the data. More results proved that an equivalent performance is achieved in nationwide datasets, and that the data accuracy is better when most needed, i.e., at daytime, when the majority of relevant human activities take place.

References

[1] Telefonica Smart Steps, http://dynamicinsights.telefonica.com/smart-steps/
[2] Orange Flux Vision, http://www.orange-business.com/fr/produits/flux-vision
[3] D. Naboulsi, M. Fiore, R. Stanica, S. Ribot, *"Large-scale Mobile Traffic Analysis: a Survey,"* IEEE Communications Surveys and Tutorials, 18(1), 2016.
[4] H. Zang, J. Bolot, *"Anonymization of location data does not work: A large-scale measurement study,"* ACM MobiCom, 2011.
[5] Y. de Montjoye, C.A. Hidalgo, M. Verleysen, V. Blondel, *"Unique in the Crowd: The privacy bounds of human mobility,"* Nature Scientific Reports, 3(1376), 2013.
[6] M. Gramaglia, M. Fiore, *"Hiding Mobile Traffic Fingerprints with GLOVE,"* ACM CoNEXT, 2015.
[7] A. Machanavajjhala, D. Kifer, J. Gehrke, M. Venkitasubramaniam, *"l-diversity: Privacy beyond k-anonymity,"* ACM Transactions on Knowledge Discovery from Data, 2007.
[8] M. Terrovitis, N. Mamoulis, P. Kalnis, *"Privacy-preserving Anonymization of Set-valued Data,"* VLDB, 2008.
[9] M. Gramaglia, M. Fiore, A. Tarable, A. Banchs, *"Preserving Mobile Subscriber Privacy in Open Datasets of Spatiotemporal Trajectories,"* IEEE INFOCOM, 2017.
[10] M. Coscia, S. Rinzivillo, F. Giannotti, D. Pedreschi, *"Optimal Spatial Resolution for the Analysis of Human Mobility,"* IEEE/ACM ASONAM, 2012.

Detecting the leakage of personal confidential information from mobile data

Wang Ti[1] Ding Wei[1] Liu Xinhai[2]
[1] China Unicom Network Technology Research Institute, {wt, dingwei}@dimpt.com
[2] the People's Bank of China, xinhai.liu@foxmail.com

Abstract:

So far, voice phishing is quite popular in China, leading to big loss for the public, for example, in 2015, the damage caused by vishing in China was estimated to be about USD 3 billion. Due to its complicate mechanism, fighting vishing attack is hard challenge. Therefore, with mobile data analytics, we propose an approach to detect the pattern of personal information leakage in telecom fraud. The vishing cases in the 40 cities of China are analyzed, which demonstrates that our approach can help detect how privacy is leaked.

Key words: Mobile Data; Voice Phishing; Binary Classification; Information Entropy

1 Introduction

Vishing (Voice phishing) is a variant of phishing. The scammers use phone calls to trick people into divulging financial data or transferring money by masquerading as trusted authorities (for instance, government agency, bank and so on). Moreover the vishing criminals is becoming more and more complicated. Once the scammers have gotten the victims' personal confidential information, especially the certain population groups such as the elderly are more vulnerable to vishing. These factors have led to an increase in vishing attacks. In 2015, the damage due to vishing in china was estimated to be about USD 3 billion. Even though there is an underground market of collecting and reselling the personal confidential information, these information breach still can promote the planning of criminal acts.

Motivated by the challenge of vishing attack, we propose an approach to detect the pattern of personal information leakage in telecom fraud from mobile data. To evaluate the performance of our approach, we analyze the vishing cases in the 40 cities of China. The results demonstrate that our approach can help detect how privacy is leaked.

2 Methodology

It was reported that[1], no matter whether there were specific targets, at the general level the scammer's calling with a phone number would be made for 189 times daily, and one success of telephone fraud costs 357 times calling at average. Meanwhile, based on former research, the successful rate of vishing is about 0.28% [1]. Based on our statistical analysis as shown in Figure 1, the vishing call over mobile

phone has certain evident patterns, such as, high calling frequency(it often occurs during the work day and work time), short talking duration (most less than 2 minutes) , and the called parties is likely hang up the call.

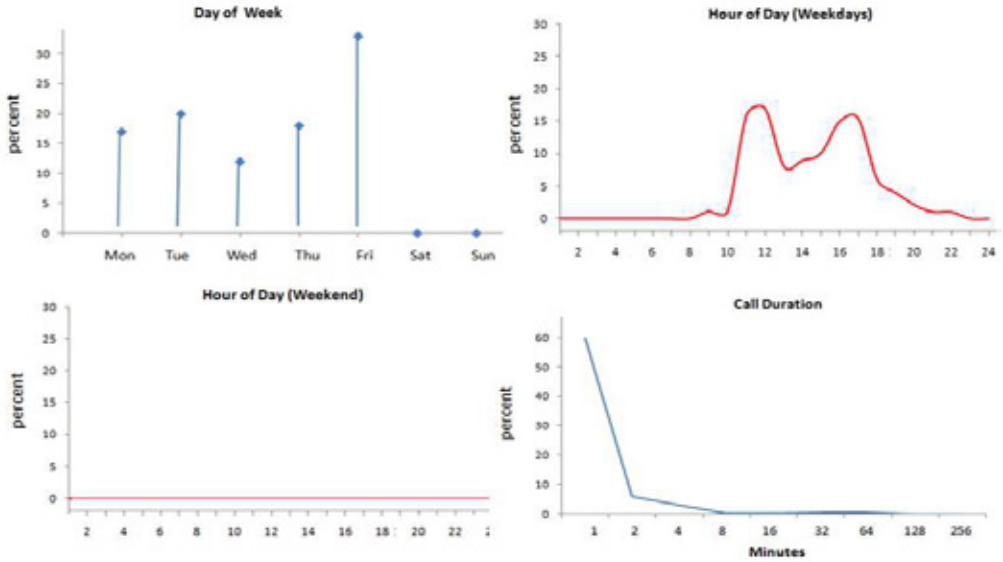


Figure1: The statistical profiles of typical vishing calls

Usually there are not specific targets in some fraud cases, but the situation has changed as both Information Technology and the Internet scenario have made it easy to collect personal confidential information through breaches. In vishing cases, if the phone numbers called in a divided segment are not in digit sequence, except the auspicious phone numbers, the users' personal confidential information probably would be leaked to a large degree. Hence we introduce the discrete rate of called phone number as the index which indicates the disorder degree. We use the information entropy algorithm to evaluate it. Due to the low efficiency of the traditional method called following window, we adopt an elaborate approach called sliding window (see Fig.2), which calculate the index with different window size as well as various calculation manner.

Figure 2: Example of calculating the disorder degree of a phone number by various methods

Since 2015, China Unicom, one of the telecom giants in China, has been carrying out the project to leverage the location technology of cell-sites and map the areas into geo-grids in a proper scale with the functional characteristics. Up to now, the micro grid division of functional areas in 40 major cities has been completed. Then on the basis of consumer's behavior analysis, combined with the real name registration rule of mobile phone user, we can tell the approximate identity of a phone user as shown in Figure 3. The dots indicate the times and locations recorded by the close antennas. In Subfigure A, Trace of an anonymous mobile phone user during a day is recorded. Subfigure B focuses on the two main periods when the user's working and sleeping. While in subfigure C, more precise location information of the working area is available.

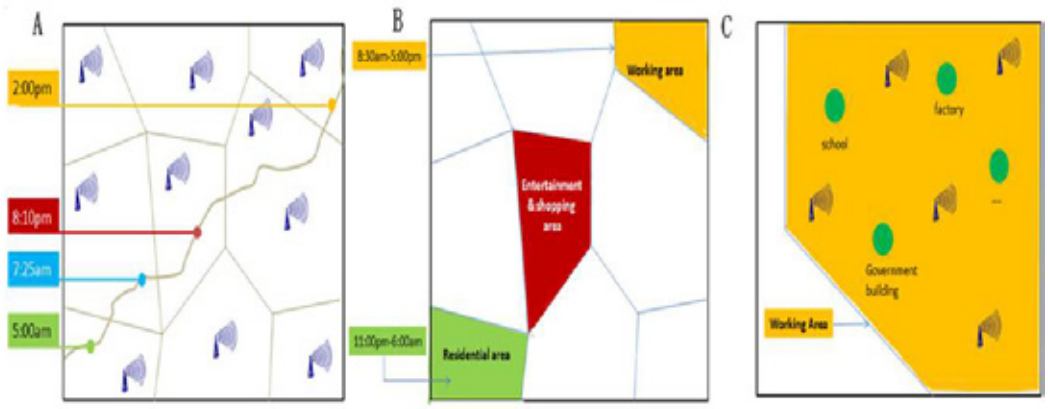


Figure 3: The demonstration of location of a phone user

Herein, we test our approach on a dataset spanning over a period of 6 months and 40 cities in China, where the fraud callers and the called victims are in the same provinces. We use the information entropy algorithm to distinguish the clusters of called number and extract the cluster characteristics (see fig4).

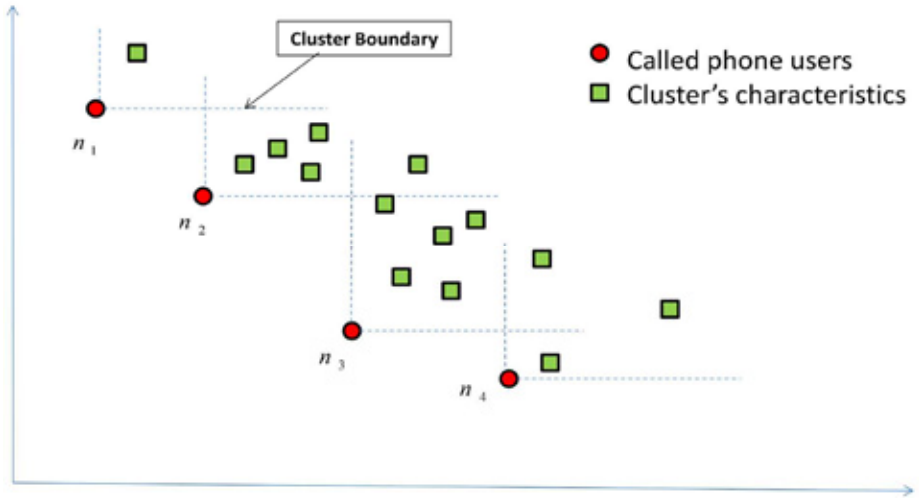


Figure 4: Clusters of called phone users divided by entropy and the characteristics

Typically, we use the available data to analyze the following 4 different scenarios:

Scenario I: The phone users in the special cluster of called parties are usually staying in the proximity area, such as the same school, the same government building, or the same company. So it is possible to divulge the details from the current organization.

Scenario II : If the Scenario I does not exist, one of the user's family members has taken part in some certain activities, for example, the children attended the same extracurricular class, who were often in the same area at some certain time. Then the personal confidential information of the parent was leaked indirectly. (The social relationship of the phone user can be mined through mobile data as well[2]).

Scenario III: if the fraudsters know the phone user's bankcard number and IDs in the vishing cases, and it was reported by the victims to police or the complaint department of service operators, we can query whether they all have taken financial account in the same bank from the credit center of the central bank, which will prove that their private information were leaked from the bank or some e-commerce websites with payment function.

Scenario IV: Beside all above, if the significant correlation among the users of mobile phone cannot be found through big data in the form of the first three scenarios, perhaps it was due to the individual factors of the user's own incaution.

According to the dichotomy principle, we propose three analysis models for the first three scenarios. It means we pose the analysis in the first three scenarios as the binary classification task in the corresponding models. Meanwhile, we take the real facts reported by victims in the vishing cases as the checking dataset to train the models, and demonstrate the feasibility of the method by evaluating its performance through ROC curve. The result shows the areas of the three models are all better than the benchmark (see fig.5).

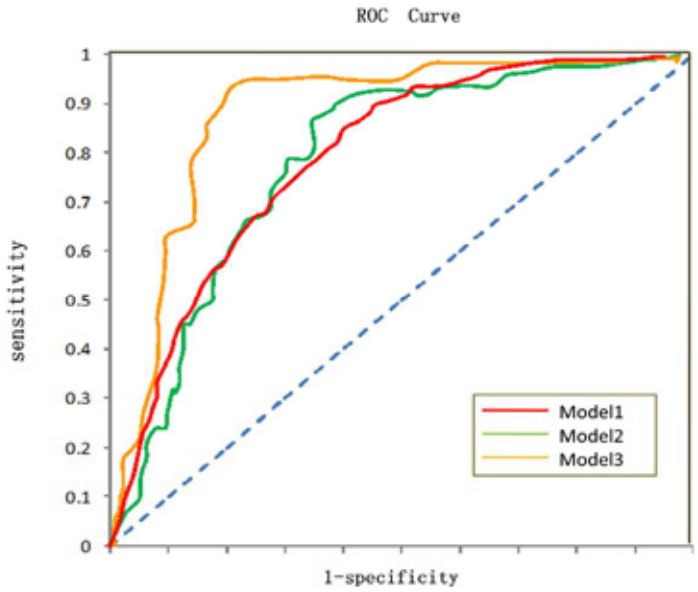


Figure 5: Areas of the classification models under the ROC curve (>0.5)

3 Conclusion

This research explores the framework to detect the pattern of personal confidential information leakage. The approach we present in this paper has several advantages: (i) it has good scalability and so it is suitable for detecting the breach of the confidential information leakage in mobile phone fraud.（ii）it is helpful to safeguard the personal confidential information because it can help the related parties for protection of personal privacy in the criminal acts, no matter malicious intention or inadvertent errors. In some sense,

References

[1] 360 company, Mobile guards Security Report in 2016：the Fraud with phone situation and Characteristics Analysis. http://www.199it.com/archives/545187.html
[2] H Zhang， R Dantu ,Predicting social ties in mobile phone networks. Intelligence and Security Informatics (ISI), 2010 IEEE International Conference

Biases and errors in the temporal sampling of random movements

Riccardo Gallotti
Instituto de Física Interdisciplinar y Sistemas Complejos (IFISC), CSIC-UIB, Campus UIB, ES-07122 Palma de Mallorca, Spain

Rémi Louf
Centre for Advanced Spatial Analysis (CASA), University College London, W1T 4TJ London, United Kingdom

Jean-Marc Luck
Institut de Physique Théorique, CEA, CNRS-URA 2306, F-91191, Gif-sur-Yvette, France.

Marc Barthelemy
Institut de Physique Théorique, CEA, CNRS-URA 2306, F-91191, Gif-sur-Yvette, France. and CAMS (CNRS/EHESS) 190-198, avenue de France, 75244 Paris Cedex 13, France

The new sources of data, available thanks to Information and Communication Technologies, allow to track individual trajectories at an unprecedented scale [1, 2]. In empirical studies, trajectories individuals [3, 4] are sampled in space and time. However, as it is the case for any dataset, these new sources of information have limits and biases [5, 6] that need to be assessed.

Here, we study trajectories alternating rests and moves of random durations [7–9]. Isolate and identifying these intertwined static and dynamic behaviours is an important statistical challenge and a growing array of *segmentation* methods based on spatio-temporal characteristics of the trajectories have been tailored for the specific dataset in question [2, 6]. These procedures are however limited from technological constraints that impose a temporal sampling of the trajectory, as one needs a time Δ between sampled points significantly smaller than the characteristic duration of rests and moves in analysis to reconstruct the trajectory.

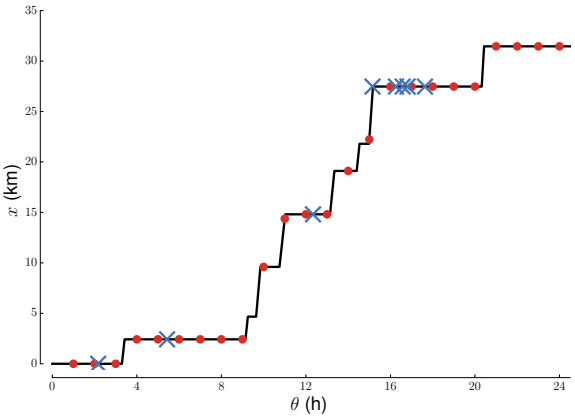


FIG. 1: **Examples of trajectory sampling.** On a trajectory with exponentially distributed rest and move durations, we show the case of constant sampling interval (red circles) and the case of random sampling interval (blue crosses) with $P(\Delta) \propto \Delta^{-1}$ ($\Delta_{\min} = 5$ min, $\Delta_{\max} = 12$ h).

This issue is particularly evident in human mobility data. Currently, the most common sources used are Call Detail Records (CDR) of mobile phone data [11] and geo-located social media accesses [12], where the flaws described above are amplified by random and bursty nature of human communications [3, 13].

In this paper, we discuss the effect of periodical and bursty sampling on the measured properties of random movements. We consider trajectories as an alternating renewal process [14], a generalisation of Poisson processes to arbitrary holding times and to two alternating kinds of events, moves and rests, whose durations t and τ are regarded as independent random variables. The sampling time interval Δ depends on the particular experiment and can be either constant or randomly distributed.

We analytically solve the ideal case of constant sampling and short-tailed distributions of rest and move durations with the naive assumption that every observed displacement is to be associated to a movement. We obtain explicitly the distribution $P(\ell^*)$ of sampled displacements and its first two moments, that also allow us to quantify difference between the real $\ell = vt$ and sampled ℓ^* displacement lengths. Moreover, we are able to provide an optimal sampling time $\hat{\Delta} = 1.96\sqrt{\bar{t}\bar{\tau}}$ maximizing the fraction of correctly sampled movements. We then extend these results numerically, and show that sampling human trajectories in more realistic settings is necessarily worse. Finally, we use high-resolution (spatially and temporally) GPS trajectories [15] to verify our predictions on real data. We find that for real cases, characterized by long-tailed rest durations [4, 10], the fraction of correctly sampled movements is dramatically reduced. We test our results with high-resolution GPS trajectories of human, where constant sampling allows to recover at best 18% of movements, while even idealized methods cannot recover more than 16% of moves from sampling intervals extracted from real communication data [16] .

These figures suggest that in the sampling of of human trajectories alternating rests and movements it is not possible to successfully reconstruct the real moves from the empirical sequence of displacement observed only through the lens of mobile phone communications.

Further studies, taking advantage of the new analytical tools we provide here to evaluate the quality of a sampled individual trajectory, are certainly necessary to assess the bias induced by sampling on statistics aggregated at individual or collective level.

[1] Vespignani, A. Modelling dynamical processes in complex socio-technical systems. *Nat Phys* **8**, 32–39 (2012).
[2] Zheng, Y. Trajectory data mining: An overview. *ACM Transactions on Intelligent Systems and Technology* **6**, 29–41 (2015).
[3] González, M.C., Hidalgo, C.A. & Barabási, A.-L. Understanding individual human mobility patterns. *Nature* **453**, 779–782 (2008).
[4] Song, C., Koren, T., Wang, P. & Barabási, A.-L. Modelling the scaling properties of human mobility. *Nature Phys* **6**, 818–823 (2010).
[5] Williams, N.E, Thomas, T.A., Dunbar, M., Eagle, N., Dobra, A. Measures of human mobility using mobile phone records enhanced with GIS data. *PLoS ONE* **10**(7), e0133630 (2015).
[6] Çolak, S., Alexander, L.P., Alvim, B.G., Mehndiratta, S.R., González, M.C. Analyzing cell phone location data for urban travel. *Transp Res Record* **2526**, 126–135 (2016).
[7] Turchin, P. *Quantitative analysis of movement: measuring and modeling population redistribution in animals and plants.* (Sinauer Associates, Sunderland 1998).
[8] Brockmann, D., Hufnagel, L. & Geisel, T. The scaling laws of human travel. *Nature* **439**, 462–465 (2006).
[9] Gallotti, R., Bazzani, A., Rambaldi, S., Barthelemy, M. A stochastic model of randomly accelerated walkers for human mobility. *Nat Commun* **7**, 12600 (2016).
[10] Gallotti, R., Bazzani, A. & Rambaldi, S. Toward a statistical physics of human mobility. *Int J Mod Phys C* **23**, 1250061 (2012).
[11] Blondel, V.D., Decuyper, A., Krings, G. A survey of results on mobile phone datasets analysis. *EPJ Data Science* **4**, 10 (2015).
[12] Hawelka, B. et al. Geo-located Twitter as proxy for global mobility patterns. *Cartogr Geogr Inform* **41**(3), 260–271 (2014).
[13] Bild, D.R. et al. Aggregate characterization of user behaviour in Twitter and analysis of the retweet graph. *ACM Transactions on Internet Technology* **15**(1), 4 (2015).
[14] Godrèche, C., Luck, J.M. Statistics of the occupation time of renewal processes. *J Stat Phys* **104**, 489–524 (2001).
[15] Zheng, Y., Xie, X., Ma, W.-Y. GeoLife: A Collaborative Social Networking Service among User, location and trajectory. *IEEE Data Engineering Bulletin* **33**(2), 32–40 (2010).
[16] de Montjoye, Y.A., Smoreda, Z., Trinquart, R., Ziemlicki, C., Blondel, V.D. D4D-Senegal: the second mobile phone data for development challenge. arXiv:1407.4885 (2014).

Application of trajectory based models for continuous behavioural user authentication through anomaly detection

Piotr Kałużny, Piotr Jankowiak, Agata Filipowska, Witold Abramowicz[1]

[1]Department of Information Systems, Poznan University of Economics and Business, Poland

***Abstract*—This paper describes how mobility patterns understood as trajectory based models can be applied for an anomaly detection and authentication using telecommunication data. The trajectory based mobility model utilizing stay-point extraction suited for the sparse CDR data is used to describe mobility patterns of a user. In this model, the observed activities are assigned with anomaly scores in three distinctive areas including: geographical, sequential and temporal dimensions. Activities with threat values exceeding the user confidence threshold are identified as anomalies. The model is tested on the sample of Poznan inhabitants. Evaluation of the model performance is based on the similarity classes and the results are presented within the paper.**

I. Introduction

In recent years, due to the ubiquity of cell phones and raise in their penetration rates, mobile phones became not only a basic tool for everyday use, but also a valuable source of information about their users. The value of data generated by those devices has risen significantly over the years [6].

Regardless of value of this data as perceived by a user, mobile devices suffer from a lack of proper protection against unwanted access to this data in case of a theft (of both the device or user identity). Proper authentication techniques and automatic systems are needed to ensure those devices are secure from theft and an unintended use. The traditional methods have their drawbacks caused mostly by users' negligence resulting in 40% of the phones not secured by any means [7]. Due to that fact and drawbacks of point-of-entry traditional approaches, currently existing methods are not enough [1]: sometimes lacking the required security, availability or usability. As a possible answer to this problem a new family of methods is introduced, named behavioural biometry. These methods address unique, non transferable, difficult to reproduce and hard to forget or loose characteristics derived from user behaviour rather than physical traits. The methods can rely on various factors and focus on different aspects of behaviour to find unique patterns of e.g. gait, signature or keystroke dynamics. Those patterns, created while the device is running, can be used to secure it. This allows the authentication process to work continuously and transparently without the user interaction needed. The use of those methods can provide an additional layer of security on top of existing methods without diminishing the usability, e.g. PIN or password would be used only when the behaviour analysis system is not sure about the user's identity. Within those methods there exists a subgroup, referred to as behavioural profiling, which: *identifies people based upon the way in which they interact with services of their mobile device* [13]. The user's identity is determined based upon the comparison of a sample of activities with his profile. If the sample matches the profile, the user will be granted with an access, otherwise he will be refused [17] or an additional proof of identity will have to be brought (e.g. PIN).

Studying behavioural patterns and especially the user mobility had proven to be quite successful in differentiating between users and identifying deviations from a user profile. This is possible due to the fact that cell phone traces closely resemble user trajectories, regardless of the mobility data source being CDR or phone-collected data. Also observations of the human movement conducted by the researchers confirmed the stability of those patterns [9]. The predictability of mobility patterns was proven to be high and stable given historical behaviour of a user [15], [18], [19] regardless of the daily distance travelled [3]. The proxy of BTS (Base Transceiver Station) labelled geographical information, derived mostly from CDR, was proven to be precise enough to study human mobility. Its applications included identifying patterns on a large scale confirming correlation with e.g. population density [2] or transport networks. Those traces also remained precise enough to capture mobility patterns to allow for an individual user analysis based on visited locations and travel models [4], [5], [14], also introducing methods for a better cell dwell time prediction [16], [21]. The users' profiles mostly utilized the semi-structured patterns that can be observed when analysing mobility in a weekly manner in hourly bins [8] even in more frequently generated phone data [12]. Such models can be a source of features for behavioural profiling approach for the anomaly detection model that utilizes multiple methods based on: probability of visiting a location at a given time [22] or sequential characteristics [20] of movement. Those methods gave highly satisfying results on frequent, phone generated data [7].

II. Dataset and approach

Our dataset is the database of Orange covering about 4 million of anonymized users over six months (from February to July) in the 2013. For each record we are given the following information: an anonymized user identification number, a

BTS id which are grouped to obtain locations covering distinct geographical areas, and a timestamp at the initial moment of the phone activity.

The mobility model that we propose utilizes the user focused stay point extraction model mentioned above, with a few assumptions:

- Activities in the same location separated by less than an hour are considered to contain enough information to assume that a user has been in the location during the period.
- Activities which have a *stay time*[1] larger than 30 minutes or are separated by more than 4 hours of inactivity[2] are labeled as stays. Stay locations are the places where users engage in some activity (contrary to the "pass-by" locations).
- All this information is kept in a weekly calendar of a user - a structure divided into hourly bins (timeframes).

The model outcome - profile is treated as a pattern, which is a base for comparison of the activities loaded to detect potential anomalies and frauds. As a result of that comparison, threat scores (which are measures between 0 and 1) are assigned to each activity in three dimensions. They define how much each tested activity varies from a user pattern in the following areas: geography - each activity is tested against the user geographical profile where the geography of a location is compared with a distance to the closest location from a set of important locations[3] compared to the user daily travel range[4]. Time - each activity is assigned a threat metric based on the distance in (hourly) timeframes, when a user is present at a given location and the usual time he is at the location divided by 24 hours[5]. Sequence - by using the trajectories built upon the extracted stays, all of the passed-by BTS are used to construct a mobility trie (TrieRoute) that contains information about frequencies of stations visited by a user when travelling between point A and B on the learning data along with the order. Each activity is assigned a sequence threat depending on the probability of a given point appearing in a sequence.

After defining those measures we conducted an experiment choosing the inhabitants of a Poznan area which is shown in the Figure 1. Home locations were extracted from our model as the longest stay between 7 p.m. and 7 a.m.[6].

The following approach was applied on the data: firstly, the mobility model was built on 24 days of data from March 2013, then 7 day verification period was used to generate a typical threat level for a user. This was used to test how consistent the users are with their patterns, generating threats in three above mentioned perspectives. 90th percentile of those threat values was used to create user confidence interval for each of the target threats. Each tested activity having its threat level above

[1]The difference between a departure and an arrival time from a location based on actions.

[2]To avoid very long movement sequences spanning over multiple days due to the sparse activity data and the uncertainty period of this sparsity [16].

[3]Places that a user visits that comprise at least 5% of the model extracted stay time in any hourly bin during the period of comparison.

[4]The approach is based on [10].

[5]Which is a max. time distance at which threat equals one.

[6]Described in [11].

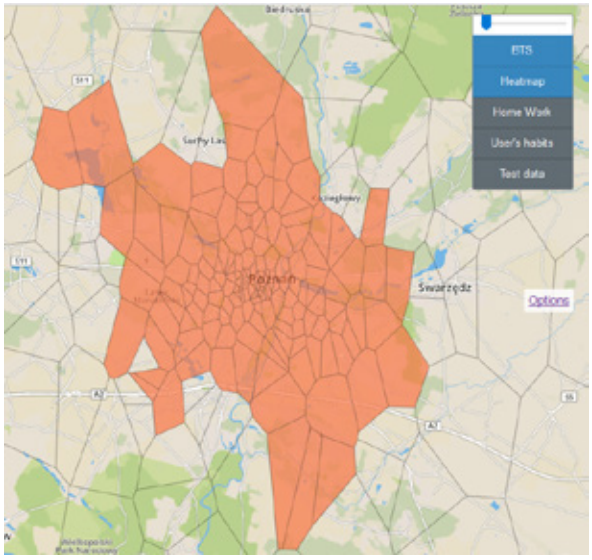


Figure 1. Poznan area - visualization made in Javascript (Mapbox, D3.js, jQuery) to showcase Voronoi cells of BTSs from the chosen area.

the threshold value of the confidence interval is classified as an anomaly in this dimension. Next, based on those upper thresholds for every user, test data from 14 days of April was loaded to test how the model performed in the authentication scenario, evaluating whether the threats generated by other users' data can be used to differentiate between the profiled user and an impostor.

Due to the fact that for the evaluation of mobility authentication models, a random user case is not a valid use case scenario[7], we introduce a class based verification model that utilizes similarity classes. This idea can be compared to simulation of an uninformed and informed attacker case from the security domain (seen also in behavioural biometry cases e.g. [12]). Therefore, the test data is divided into five classes, listed considering the expected raising similarity to the base user and difficulty of the model adjustment: random user, user living in the same town, user having home in the same BTS, user having same home and work locations in respective BTS and finally the same user - with the data loaded from the testing period. Based on this comparison a model is run on the user base, without any prior filtering besides two aspects: all users from respective classes need to be found for the base user and each of the test class users needs to have at least 5 activities in the testing period.

III. Results

The experiment was run on a sample of 1000 users, for which we were able to find corresponding users in all of test classes. We excluded users with unstable mobility patterns or sparsity of the data within CDRs. This fraction accounted for about 0.8% of the sample. The average level of threat generated was highly dependent on the test class as it is shown

[7]Model detecting anomalies for a random user appearing in a different part of country may achieve high accuracy but be practically unusable.

in the Figure 2. This proved that such division is useful in evaluating method accuracy.

Next, an anomaly/impostor scenario was prepared. Each loaded set of consecutive activities (an activity window of 3 activities was used[8]) was classified as an anomaly or normal user behavior. The anomaly was detected when the activities threat values exceeded the user confidence threshold in at least two dimensions (as it was tested to have the best anomaly detection accuracy). Additionally, an algorithm iteratively decreasing the threat percentile to achieve best results was used to improve the accuracy of the method. The average fractions of properly detected anomalies over the sample were as follows:

- 98,97% for a random test user class,
- 91,1% for a test user living in the same town as a base user class,
- 53,32% for a test user having the same home location as a base user class,
- 31,84% for a test user having the same home and work location as a base user class.

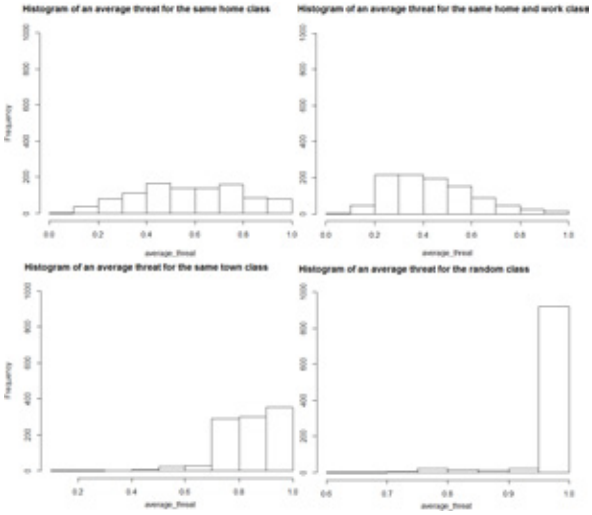


Figure 2. The distribution of average threat levels (considering three threats mentioned in the model) for all of the users.

The false rejection rate, interpreted as a portion of situations where valid user data from the test period was classified as an anomaly (compared to a profile), averaged to 13,71% over our sample. Presented results show that our model has a very high accuracy when applying a testing methodology used in the literature (random user) - about 99%. This showcases the usability of the model performance even on a sparse dataset like the event based CDR logs. However, we found that when comparing a user with another user having a very similar profile (living or working in similar areas), the accuracy is much lower. This indicates that a potential theft of a mobile phone by a thief, who has a similar mobility behaviour profile as the victim may be significantly harder to detect. The division in testing classes also introduces a new methodology for evaluation of methods' performance in both the authentication scenario (e.g. phone theft) and user pattern differentiation (distinguishing between patterns of similar users like e.g. family members).

[8]Meaning an average threat value for three activities was used for classifying whether the data belongs to a user or an impostor.

References

[1] Beyond the password: The future of account security. https://www.telesign.com/wp-content/uploads/2016/06/Telesign-Report-Beyond-the-Password-June-2016-1.pdf. Accessed: 2016-09-10.

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

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

[4] F. Calabrese, G. Di Lorenzo, L. Liu, and C. Ratti. Estimating origin-destination flows using mobile phone location data. *IEEE Pervasive Computing*, 10(4):0036–44, 2011.

[5] S. Çolak, L. P. Alexander, B. G. Alvim, S. R. Mehndiratta, and M. C. González. Analyzing cell phone location data for urban travel: current methods, limitations, and opportunities. *Transportation Research Record: Journal of the Transportation Research Board*, (2526):126–135, 2015.

[6] B. Fox, R. van den Dam, and R. Shockley. Analytics: Real-world use of big data in telecommunications. *IBM Institute for Business Value*, 2013.

[7] L. Fridman, S. Weber, R. Greenstadt, and M. Kam. Active authentication on mobile devices via stylometry, application usage, web browsing, and gps location. 2015.

[8] B. Furletti, L. Gabrielli, C. Renso, and S. Rinzivillo. Analysis of GSM calls data for understanding user mobility behavior. *Proceedings - 2013 IEEE International Conference on Big Data, Big Data 2013*, pages 550–555, 2013.

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

[10] S. Isaacman, R. Becker, R. Caceres, S. Kobourov, M. Martonosi, J. Rowland, and A. Varshavsky. Ranges of human mobility in Los Angeles and New York. *2011 IEEE International Conference on Pervasive Computing and Communications Workshops, PERCOM Workshops 2011*, pages 88–93, 2011.

[11] P. Jankowiak and P. Kaluzny. Human mobility profiling based on Call Detail Records analysis. Bachelor thesis, Poznan University of Economics and Business, Poznan, 2015.

[12] H. G. Kayacik, M. Just, L. Baillie, D. Aspinall, and N. Micallef. Data driven authentication: On the effectiveness of user behaviour modelling with mobile device sensors. *arXiv preprint arXiv:1410.7743*, 2014.

[13] F. Li, N. Clarke, M. Papadaki, and P. Dowland. Active authentication for mobile devices utilising behaviour profiling. *International journal of information security*, 13(3):229–244, 2014.

[14] F. Liu, D. Janssens, J. Cui, Y. Wang, G. Wets, and M. Cools. Building a validation measure for activity-based transportation models based on mobile phone data. *Expert Systems with Applications*, 41(14):6174–6189, 2014.

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

[16] M. Picornell, T. Ruiz, M. Lenormand, J. J. Ramasco, T. Dubernet, and E. Frías-Martínez. Exploring the potential of phone call data to characterize the relationship between social network and travel behavior. *Transportation*, 42(4):647–668, 2015.

[17] H. Saevanee, N. Clarke, S. Furnell, and V. Biscione. Continuous user authentication using multi-modal biometrics. *Computers & Security*, 53:234–246, 2015.

[18] C. M. Schneider, V. Belik, T. Couronne, Z. Smoreda, and M. C. Gonzalez. Unravelling Daily Human Mobility Motifs. *Journal of The Royal Society Interface*, 10(84):20130246(1–8), 2013.

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

[20] G. Tandon and P. K. Chan. Tracking user mobility to detect suspicious behavior. In *SDM*, pages 871–882. SIAM, 2009.

[21] P. Widhalm, Y. Yang, M. Ulm, S. Athavale, and M. C. González. Discovering urban activity patterns in cell phone data. *Transportation*, 42(4):597–623, 2015.

[22] S. Yazji, P. Scheuermann, R. P. Dick, G. Trajcevski, and R. Jin. Efficient location aware intrusion detection to protect mobile devices. *Personal and Ubiquitous Computing*, 18(1):143–162, 2014.

Time Accuracy Analysis of Post-Mediation Packet-Switched Charging Data Records for Urban Mobility Applications

Oscar F. Peredo
Telefónica I+D, Chile

Romain Deschamps
Telefónica I+D, Chile

***Abstract*—Telecommunication data is being used increasingly in urban mobility applications around the world. Despite its ubiquity and usefulness, technical difficulties arise when using Packet-Switched Charging Data Records (CDR), since its main purpose was not intended for this kind of applications. Due to its particular nature, a trade-off must be considered between accessibility and time accuracy when using this data. On the one hand, to obtain highly accurate timestamps, huge amounts of network-level CDR must be extracted and stored. This task is very difficult and expensive since highly critical network node applications can be compromised in the data extraction and storage. On the other hand, post-mediation CDR can be easily accessed since no network node application is involved in its analysis. The pay-off is in the lower accurate timestamps recorded, since several aggregations and filtering is performed in previous steps of the charging pipelines. In this work, a detailed description of the timestamp error problem using post-mediation CDR is presented, together with a methodology to analyze error time series collected in each network cell.**

I. Introduction

In the context of mobile telecommunications, *Charging Data Records* (CDR) are one of the most essential datasets generated and processed by a service provider. Circuit and packet switched events (Voice, SMS, IP, VoIP and similar) are registered by many components of the core network, generating a wide range of CDR types for different purposes. In terms of volume, the amount of CDR generated by packet switched events (related with Internet traffic and services) is several orders of magnitude larger than the circuit switched counterparts (Voice, SMS). According to [1], a simplified diagram of the core network components and connections involved in the packet switched traffic is depicted in Fig. 1. Internal gateway nodes (*SGSN*, *GGSN*, *S-GW* and *P-GW*) handle internet traffic between each subscriber device and applications server, and also generate charging data handled by special network functions (*Charging Gateway Function*). As the generated events can be geo-located using the latitude/longitude of the corresponding network cell (*BTS*, *NodeB* and *eNodeB*), packet switched events can be potentially used in the context of urban mobility applications.

Two main difficulties arise in order to use these events in urban mobility project. The first one is related with the technical complexity involved in the data generation, which requires a considerable amount of *know-how* of the daily operation, business processes and network behaviour. Unlike

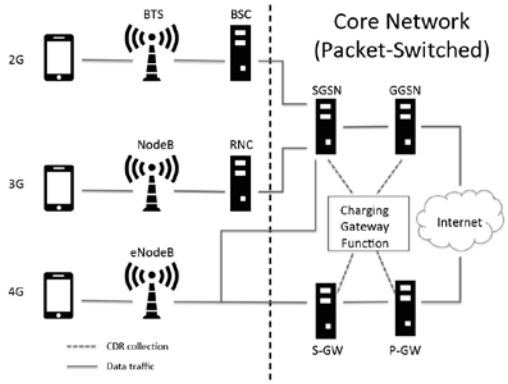


Fig. 1. Core network simplified diagram [1].

circuit switched events, the internet traffic charging processes apply filters and aggregations in several CDR at different nodes of the network ([2] Sect. 10.2 and [3] Sect 13.5). The main problem with this behaviour is the addition of potential inaccuracies in the event timestamps, which is a critical feature in urban mobility. The second one is related with the technical complexity involved in the data storage and access, since the amount of daily data is too large. Big Data persistent storages must be used to keep all CDR for posterior analysis in this case. Additionally, it could happen that the network/operations departments of the mobile operator can refuse to deploy automatic Extraction-Transformation-Load (ETL) processes in the core network nodes. The reason of this rejection is the possibility that the quality of the service to be deteriorated as consequence of the ETL processes resource consumption.

To avoid part of the previous difficulties, post-mediation internet traffic CDR are a cost-effective alternative to use in urban mobility applications (see [4] and [5]). These kinds of registers are generated by the *Billing System*, as depicted in Fig. 2. An internal core network functionality, called *Charging Gateway function*, transfers charging information from the core network gateway nodes to the *Billing System*, which is an external core network component. This system is in charge

of calculating the cost of the services used by each subscriber, based on current tariffs. It is provided by special vendors, often different from well known telecommunication-infrastructure vendors. Since this kind of CDR is usually stored in a mid-term persistent storage far from the network and operation nodes (typically a data warehouse used by business processes), ETLs can be applied to the dataset without further critical monitoring. The remaining problem that must be solved is the time accuracy of the events.

A methodology to obtain insights on the distribution of the time accuracy of post-mediation CDR is proposed. We use network traffic events as ground-truth in order to obtain timely error measurements for the CDR. Using these error measurements, time-series analysis techniques are applied to infer similarities in the error behaviour for each *BTS*, *Node* and *eNode* in the service provider's network.

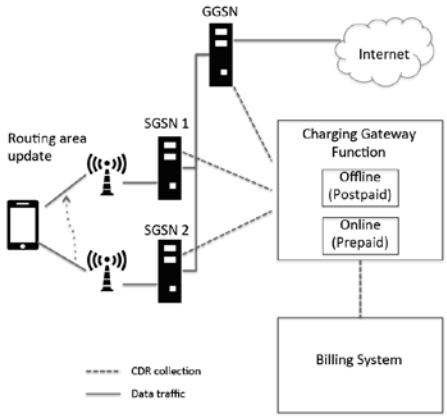


Fig. 2. Charging-related processes [2]. A single subscriber can trigger the generation of CDR in different SGSN nodes. In order to synchronize the session and register traffic usage, the GGSN node also generates CDR. The *Billing System* receives all CDR and calculates the costs of the services used by each subscriber.

II. Case Study

Our case study is based on anonymized data collected from prepaid and postpaid post-mediation CDR. The data is provided by Movistar - Telefónica Chile, with approximately 33% of the market share [6], with more than 7 million subscribers and 40,000 network cells [7].

A. Error measurement

As ground-truth we use data provided by Huawei Smartcare SEQ Analyst [8], a network event measurement tool which collects various signaling/protocol data from the core network. Particularly, each internet traffic session is measured with its corresponding initial/final timestamps, byte consumption and cell ID of origin, among other information. Originally intended to help in quality of service measurement and operations efficiency, in our case we compare this data against post-mediation CDR. We identify the last network event registered in a particular network cell which is closest in time to each CDR event. The comparison is performed backwards in time, since for each CDR event there is at least one network event which happened in the near past in the same cell. In Table I a sample event trace for an anonymous subscriber is depicted. For each CDR event, a backward search is performed until a network event is registered in the same cell. For instance, at time 72198 (20:03:18) the CDR event has an associated network event at time 69406 (19:16:46 at cell A2). We measure the error for all anonymous subscribers during 5 working days of May 2016, aggregating the errors by cell, technology (2G, 3G and 4G) and charging type, using anonymized Cell IDs.

TABLE I
Error measurement using CDR and network events. The closest past network event in the same cell is associated to each CDR.

Time (s)	CellID/Tech	Type	Charging	Error (s)
69405	A1/2G	Network	-	-
69406	A2/4G	Network	-	-
69499	A1/2G	CDR	Prepaid	94
72198	A2/4G	CDR	Postpaid	2792

B. Error analysis

Various analysis can be performed to quantify and understand the behavior of the CDR-Network time errors. In this section, we only focus on the daily variability and the cell to cell similarities. This allow us to investigate both the temporal and spatial behaviors.

1) Daily variability: To assess the variability of the CDR-Network time errors along the day, we split the daily error distributions into time-bins. We then compute the mean value of time errors for each bin. These two steps are repeated for the set of antenna cells in the La Serena-Coquimbo region (500,000 inhabitants - data from May 10^{th}; see Fig. 3). Both 3G and 4G technologies distributions vary along the day with error larger at night (bin [0h - 7h[). The error then progressively diminishes along the day. These distributions are best represented by an exponentially modified normal law. We note that the 3G technology induces smaller errors that the 4G one (see table II). However, the scarce 2G data does not allow us to draw conclusions at this moment. The same behavior are found for both prepaid (not shown) and postpaid.

2) Cell to cell similarities: We used the same time-of-the-day bin splitting to compare the cells. For each of these bins, we set up the histograms of error. We then compare the time-bin histograms between cells thanks to a Kolmogorov-Smirnov test and combine the p-values of the tests using Fisher's method. This results in a correlation index that quantifies the behavior similarities between two cells. We end up creating the full matrix of correlation indices as depicted in Fig. 4. For clarity, in this figure, we only present a random selection of 100 antenna cells. The matrix is ordered according to the sum of all of the cell-to-cell correlation indices. Cells having a close behavior are segregated in the top left of the matrix. This matrix can help to distinguish generic cell behaviors. For

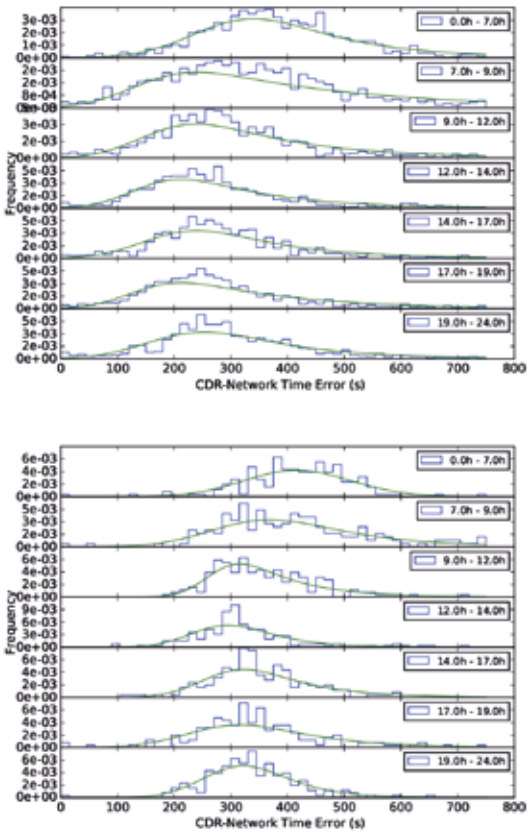


Fig. 3. Normed distribution of the CDR-Network time errors for different time-bins. Only cells from La Serena-Coquimbo region were selected, and the error were separated with respect to technologies (2G, 3G - top Panel, 4G - bottom panel) and types of contract (prepaid not shown, postpaid). The green line is the best fit (exponentially modified normal law).

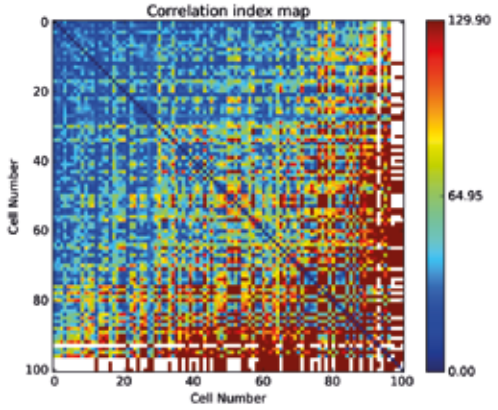


Fig. 4. Cell-cell correlation index (as explained in Sect. II-B2) matrix of CDR time errors for a random sample of 100 antenna cells. The lower the index (blue color), the better the correlation.

TABLE II
MEAN AND STD. DEV. FOR EACH BIN AND EACH CELL TECHNOLOGY.

		2G		3G		4G	
	time bins	mean (s)	std (s)	mean (s)	std (s)	mean (s)	std (s)
Prepaid	[0h - 7h[	479.17	273.26	399.81	330.84	419.83	95.46
	[7h - 9h[	706.36	1377.55	412.66	560.52	420.76	154.17
	[9h - 12h[	523.53	791.01	330.09	247.35	364.24	108.46
	[12h - 14h[	501.32	655.23	291.51	215.58	321.89	94.69
	[14h - 17h[	382.50	307.71	309.88	199.30	366.76	146.57
	[17h - 19h[	446.99	447.46	320.68	288.98	362.92	154.60
	[19h - 24h[	551.35	764.46	320.39	196.56	341.96	93.65
Postpaid	[0h - 7h[	477.34	652.55	343.86	866.33	389.69	205.59
	[7h - 9h[	638.48	2161.95	474.68	751.63	475.75	480.66
	[9h - 12h[	477.67	649.28	377.38	345.09	450.79	709.37
	[12h - 14h[	423.49	571.16	364.22	339.66	397.20	380.28
	[14h - 17h[	691.26	3523.42	394.98	386.74	382.21	237.71
	[17h - 19h[	726.62	1962.85	412.13	540.58	419.53	387.26
	[19h - 24h[	510.77	615.65	376.42	417.74	543.97	1229.22

instance, when ordered, we clearly observe a blue region (same behavior between cells) and a red region (different behavior between cells). A first generic group combines cell from the blue region. Moreover, the presence of blue points in the red region indicates that these exotic cells may form a second group. We finally note the presence of white points that refer to cells without a large enough number of events.

III. CONCLUSION

A methodology to analyze time accuracy in post-mediation CDR is presented. Exploratory analysis shows that different behaviours arise when studying disaggregated data. Particularly between postpaid 3G and postpaid 4G, by analyzing a sample data from a mid-size city. Further analysis is needed, but the overall objective of the methodology can lead to accurate error models for this kind of datasets.

ACKNOWLEDGMENT

The authors would like to thank Movistar - Telefónica Chile and Chilean government initiative CORFO 13CEE2-21592 (2013-21592-1-INNOVA_PRODUCCION2013-21592-1).

REFERENCES

[1] U. Boquist, "The Ericsson SGSN-MME: Over a decade of Erlang success," http://www.erlang-factory.com/upload/presentations/597/sgsn.pdf (last visited 2017-01-15), 2012, Erlang User Conference 2012.
[2] J. Korhonen, *Introduction to 3G Mobile Communications*, ser. Artech House mobile communications series. Artech House, 2003.
[3] C. Cox, *An Introduction to LTE: LTE, LTE-Advanced, SAE, VoLTE and 4G Mobile Communications*. Wiley, 2014. [Online]. Available: https://books.google.cl/books?id=NnGgAwAAQBAJ
[4] F. Calabrese, L. Ferrari, and V. D. Blondel, "Urban sensing using mobile phone network data: A survey of research," *ACM Comput. Surv.*, vol. 47, no. 2, pp. 25:1–25:20, 2014.
[5] V. D. Blondel, A. Decuyper, and G. Krings, "A survey of results on mobile phone datasets analysis," *EPJ Data Science*, vol. 4, no. 1, p. 10, 2015.
[6] SUBTEL, "Mobile Subscribers Market Share in Chile," Technical Report, 2017, Chilean Telecommunications Regulator.
[7] ——, "Active Antennas by Mobile Operator in Chile," Technical Report, 2017, Chilean Telecommunications Regulator.
[8] Huawei, "Huawei Smartcare SEQ Analyst," http://www.huawei.com/us/products/core-network/smartcare/seq-analyst/ (last visited 2017-01-15), 2017.

SESSION 7
SOCIAL NETWORK

What comes first? Social strength or common friends?

Giovanna Miritello,[1] Manuel Cebrián,[2] and Esteban Moro[3,*]

[1] *Telefónica Research, 28050 Madrid, Spain*
[2] *Data61 Unit, CSIRO, Melbourne, Victoria, Australia*
[3] *Departamento de Matemáticas & GISC, Universidad Carlos III de Madrid, 28911 Leganés, Spain*

A famous result in social network theory is the *weak ties hypothesis* by Mark Granovetter by which the strength of a social ties correlates with the social embeddedness of the tie. Strong ties are surrounding by common friends, while weak ties happen between communities. Although this hypothesis was confirmed in a number of online and offline networks, little is know about how this correlation is built and/or destroyed in time, that is, how common friends and tie strength evolve together when a tie is created or destroyed. By analyzing the mobile phone communication network of about 20 million people over a long period of time of 19 months, we found that once that a tie is created is reaches almost instantaneously its strength while its embeddedness slowly growths even months after tie formation. The opposite is find when a tie is destroyed: tie strength last until the very last minute although common friends started to disappear months before. Our results highlight that the Granovetter *weak ties hypothesis* is a dynamical process that happen at a very slow time scale, showing the intricate evolutionary dynamics of network interactions and structure. .

Understanding the very dynamics which regulates the process of why an individual decides to add or remove a social tie is a very complex process which has been of interest of many studies [1, 4–6]. However, since tie creation/removal processes alter the structure of the network in which the individual participates, we may wonder whether changes in the topology of the local network surrounding a tie can tell us something about the fact that a new connection is going to appear or has just been removed. In particular, what we want to answer here is: to what extent individuals choice to create a new tie is related to the their social network before the new connection has been created? Does the topology of their network change after the establishment of a new tie? The same questions can be applied when a connection is removed, instead than created.

To address this we first introduce a new methodology to investigate temporal networks [7], an issue which is convoluted with the way social networks are observed and modeled, and which has been recently been pointed out as a problem in the field of social networks [8]. By means of this methodology we are able to disentangle tie activation/deactivation from their bursty activity. This method allows us to determine, for each tie in our dataset, the set of their common neighbors that are active at any time instant. Moreover, for any given tie, we can identify with high precision the time at which it has been created or destroyed.

As mentioned before, we are interested in understanding whether changes in the local topology are related to the creation/removal of a new/old social connection. For this reason, for each of these two categories of ties, we study the evolution of the topological overlap in a time window which spans a period going from before to after the tie has been created or removed. Accordingly to the weak tie hypothesis, ties between individuals who have many common friends (large overlap), are stronger than the ones between people which few common friends (small overlap), who instead act as bridges between different tight groups Since our analysis allows us to assess when a tie has been created or removed and, at the same time, to analyze the instantaneous contact network, we take advantage of this and try to investigate the *dynamical Granovetter effect*. Specifically, we also separate the ties in different groups according to their strength (total number of calls during the whole 19 months period) and we analyze the temporal evolution of their neighboorhood overlap for each of these groups. For comparison, we also show the average overlap between pairs of nodes randomly chosen from the whole population, and the average overlap of ties that do not form or decay within the observation window.

Our contribution shows a number of important results. According to the weak ties hypothesis (see figure 1), we observe that topological overlap is strictly related to the intensity or weight of a social relation, meaning that the stronger is the relationship between two persons, the more friends they have in common, in line with previous results [1, 2]. Moreover, the overlap between two individuals who form (remove) a connection at some time during their lifetime, is significantly higher than the one observed between any random pair of individuals in the whole population months before (after) the link has been established (removed). This constitutes a more clear evidence of why the topological overlap is usually a very good feature in the prediction of tie creation [3, 9, 10].

More interestingly, we find that the process that drives two individuals to link together is highly dynamical and that, locally, it entails the change of the underlying topology of the network. We observe, in fact, a large overlap many days before the connection has been established,

* Corresponding author emoro@math.uc3m.es

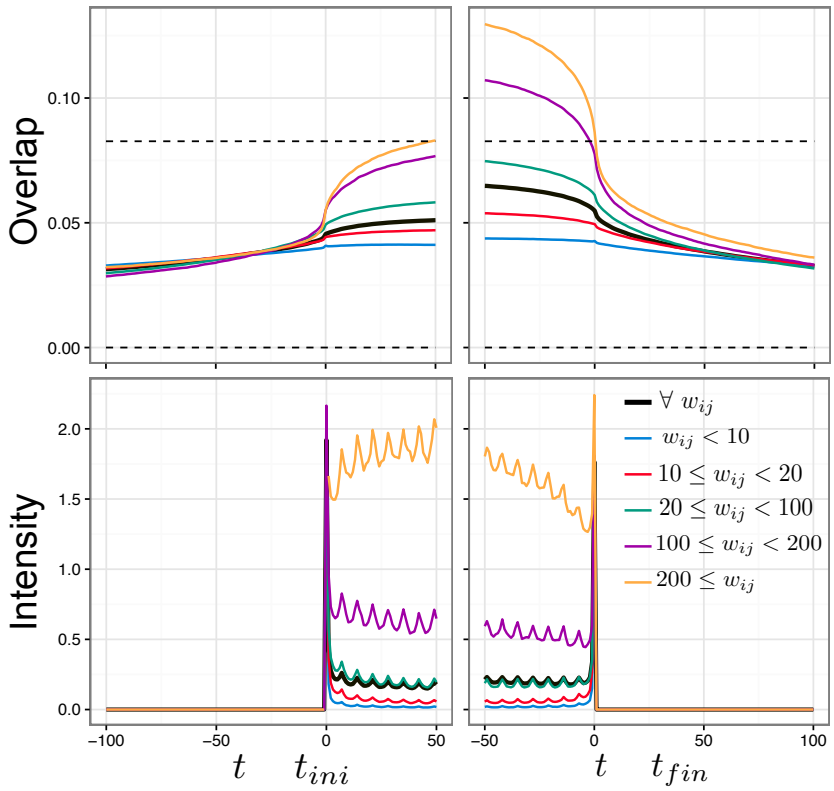


FIG. 1. Co-evolution of the topological overlap (Jaccard coefficient, top) and intensity of communications (number of calls per day, bottom) before and after the creation (left) and destruction of a link (right) for different intensity of the links. Dashed lines in the top panel correspond to the average overlap of two randomly selected nodes in the newtork (bottom) and the average overlap for all the links in the network.

and it continues to increase over time after the tie forms. A similar result is found when a tie is removed: the topological overlap between two individuals starts to decrease much before the connection is removed and it keeps decreasing in time after the breakage, although very slowly. This allows us to reply to the following question: what comes first, tie strength or common neighbors? Our results indicate that the connection comes first, and only after the correlation between tie strength and topological overlap starts to form, suggesting indeed a sort of dynamical Granovetter effect that, to the best of our knowledge, has not been investigated before.

These results indicate that the properties of the local network between two individuals contain important information about their relationship even if no interaction between them is observed. They also have outstanding interest from a sociological and anthropological point of view since they shed more light on the way in which humans establish and remove social connections. We have seen, in fact, that the Granovetter effect is not just a correlation observed in the aggregated contact network, but a dynamical process that happens at a very slow time scale. The correlation between the number of common friends between two individuals and their strength is in fact observable within a time period significantly longer than the lifetime of the social relationship and acts as a "fingerprint" of the social relation itself.

[1] Granovetter M (1973) The strength of weak ties. Am J Sociol 78:13601380.
[2] Onnela J-P, Saramki J, 1176 Hyvnen J, Szab Z, Lazer D, Kaski K, Kertsz J, Barabsi A-L (2007), Structure and tie strengths in mobile communication networks. Proc.Natl.Acad. SciUSA104:7332.
[3] Hidalgo C, Rodriguez-Sickert C (2008) The dynamics of a mobile phone network. Phys A387:3017.
[4] Burt R. (2001b). Structural Holes Versus Network Closure as Social Capital. in Lin, N., Cook, K. and Burt, R.S.

Social Capital: Theory and Research. Sociology and Economics: Controversy and Integration series. New York: Aldine de Gruyter.
[5] Kossinets G., and Watts D. J. (2006). Empirical analysis of an evolving social network. Science 311, 5757.
[6] Podonly J., and Baron J. (1997). Resources and Relationships: Social Networks and Mobility in the Workplace. Americal Sociological Review 62, 673693.
[7] Miritello G, Lara R, Cebrián M, Moro E (2013) Limited communication capacity unveils strategies for human interaction, arXiv:1304.1979.
[8] Krings, G. Karsai, M., Bernharsson, S., Blondel, V.D., Saramäki, J. Effects of time window size and placement on the structure of aggregated networks, EPJ Data Science, 2012, Volume 1, Number 1, 4.
[9] Liben-Nowell D., and Kleinberg J. (2007). The link-prediction problem for social networks. In CIKM 03: Proceedings of the twelfth international conference on Information and knowledge management, New York, NY, USA, pp. 556-559.
[10] Raeder T., Lizardo O., Chawla N., and Hachen D. (2011). Predictors of short-term decay of cell phone contacts in a large scale communication network. Social Networks 33, 245-257.

Kernel-based approaches to large social networks

Till Hoffmann and Nick Jones
Department of Mathematics, Imperial College London

Mobile phone datasets such as CDRs provide rich data about user behaviour, mobility and call networks [4]. But the data are often difficult to interpret because of their size and complexity. Community detection algorithms attempt to shed light on the data by reducing their dimensionality [3], but usually give rise to a network of communities which presents us with the same difficulties–although at a smaller scale.

In contrast, we assume that the network can be explained by observable node attributes such as physical location or demographic data (as opposed to latent attributes such as community membership [14] or a latent embedding [5, 7]). We develop a model that connects network data to third-party information about node attributes. The model has a number of attractive properties: (1) it can be fitted to aggregated communication data protecting the privacy of users, (2) it is generative and can be used to synthesise (aggregated) network data that can be shared without privacy concerns, (3) fitting the model has relatively small computational cost, and (4) the model parameters are interpretable. We fit the model to synthetic communication data in the London metropolitan area to validate the inference algorithm.

I. INTRODUCTION

Social network data have become abundant in the digital age [4]. Seemingly private information such as sexual orientation, party affiliation, and ethnicity can be inferred from the behaviour of users of online social media [9]. How people use their mobile phones can yield information about their socioeconomic status [2]. Despite these successes, Golder and Macy [4] conclude that such data do not "provide the demographic background information needed by social scientists." By contrast, censuses and sample surveys provide detailed demographic information about respondents but are lacking data about their associates.

Fortunately, demographic and network data can be fused by leveraging the residential location of mobile phone users: the address of a user can be highly informative of their demographic background. For example, residents of areas with high property prices are likely to be more affluent than areas with cheaper accommodation. In this work, we assume that the residential area of mobile phone users is known and use aggregate communication volumes between different areas to constrain the parameters of a kernel which predicts the interaction rate between users conditional on their demographic background.

II. NETWORK MODEL

We assume that each node $i \in \{1, \dots, N\}$ of the call network is endowed with a q-dimensional coordinate vector z_i which captures node attributes. The undirected call volumes y_{ij} between two nodes i and j are modelled as conditionally-independent Poisson random variables with rate

$$\lambda_{ij} = \lambda\left(z_i, z_j, \theta\right)$$

where we have omitted the explicit dependence on the length-p parameter vector θ to avoid clutter. Unfortunately, despite the abundance and sensitivity of call data, they rarely provide individual-level demographic information [4]. Yet distributions for demographic attributes conditional on geographic areas are widely available from third parties such as the census or large surveys [11].

We assume that each node i has an associated residential area $g_i \in \{1, ..., K\}$. The aggregate call volume between two areas a and b is thus

$$m_{ab} = \sum_{i,j \in A, B: i<j} y_{ij},$$

where $A = \{i \in N : g_i = a\}$. Because the Poisson distribution is infinitely divisible [8], the aggregate call volume is also Poisson-distributed with aggregate rate parameter

$$\Lambda_{ab} = \sum_{i,j \in A, B: i<j} \lambda_{ij}.$$

The likelihood for the aggregated call volumes is thus

$$P\left(m|\theta, g\right) = \int dz\, P\left(z|g\right) \prod_{a<b} P\left(m_{ab}|\theta, z\right). \qquad (1)$$

As the integral over node attributes is non-trivial, we evaluate the expected aggregate call rate

$$\bar{\Lambda}_{ab} = \int dz\, \Lambda_{ab} P\left(z|g\right), \qquad (2)$$

and approximate the likelihood by

$$P\left(m|\theta, g\right) \approx \prod_{a<b} \frac{\bar{\Lambda}_{ab}^{m_{ab}} \exp\left(-\bar{\Lambda}_{ab}\right)}{m_{ab}!}. \qquad (3)$$

The approximation neglects the (co)variances of the aggregate call rates. But because they are U-statistics, the variances are known to be small for sufficiently large communities [10].

FIG. 1. Aggregated communication network amongst 32 local authority groups in the London metropolitan area.

III. SIMULATION AND INFERENCE

We use regional micro data from the 2011 census [**11**] in the United Kingdom to simulate a social network amongst 283,155 people over the age of 18 in the London metropolitan area with connectivity kernel

$$\lambda_{ij} = \exp\left(\sum_{k=1}^{p} \theta_k f_k\left(z_i, z_j\right)\right),$$

where $f : \mathbb{R}^q \times \mathbb{R}^q \to \mathbb{R}^p$ is a function that maps coordinates to features used to predict the interaction rate of two nodes. In particular, the first feature is one irrespective of the coordinates and accounts for the link density. The second and third feature are equal to one if i and j identify with a different religion and ethnicity, respectively. The fourth feature is the natural logarithm of the Euclidean distance between i and j. We let $\theta = (-9, -1, -1.4, -0.7)$ based on previous results from survey data [6, 13]. The synthetic network has an average degree of 27 and is shown in figure 1.

To evaluate the approximate likelihood in equation 3, we approximate the expected aggregate rate in equation 2 using a subset of the micro data, i.e. Monte Carlo integration. We use a Metropolis-Hastings sampler to draw samples from the posterior distribution. Kernel density estimates of the marginal posterior density for the regression coefficients is shown in figure 2. Although we only have access to aggregated communication volumes, the algorithm is able to recover the regression coefficients.

IV. DISCUSSION

We have developed an inference algorithm to fit a generative network model to aggregated communication volumes. Using only aggregated data can alleviate some of the privacy concerns associated with the sensitive nature of individual CDRs–even when they are anonymised [1].

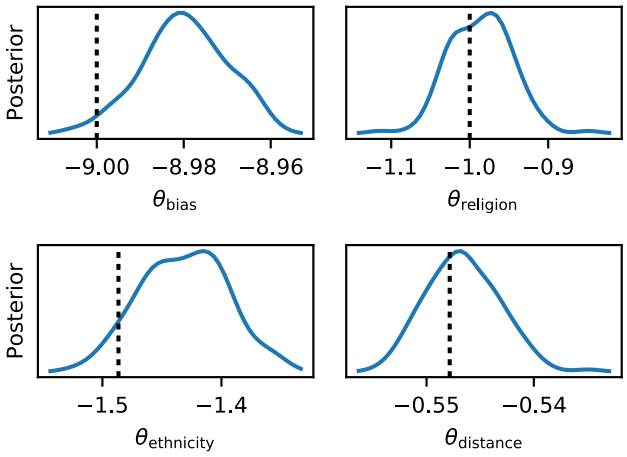


FIG. 2. Gaussian kernel density estimates of the marginal posterior density for regression coefficients inferred from simulated data.

Because our model is generative, the inferred rate kernel can be used to simulate networks with realistic attribute correlations. Synthetic networks can be shared without fear of deanonymisation [12] and allow us to study the effect of the connectivity kernel on network structure and dynamics unfolding upon the networks.

In contrast to latent embeddings of community detection algorithms, the kernel-based approach proposed here is easily interpretable. Each unit increase in the k^{th} feature changes the interaction rate between the corresponding individuals by a factor $\exp\theta_k$.

Finally, the algorithm has a relatively low computational cost even for large networks because the number of data points is of order K^2 rather than N^2.

In future work, we would like to extend the feature map f to a wider range of demographic attributes such as sex and age. More generally, we believe that building tools to derive meaningful information from aggregated data is an important step to prompt mobile phone providers and online social networks to make public data releases and allow a wider audience of researchers to access these rich data sets.

[1] Lars Backstrom, Cynthia Dwork, and Jon Kleinberg. Wherefore art thou r3579x?: Anonymized social networks, hidden patterns, and structural steganography. *Commun. ACM*, 54(12):133–141, 2011. doi: 10.1145/2043174.2043199.

[2] Joshua Blumenstock, Gabriel Cadamuro, and Robert On. Predicting poverty and wealth from mobile phone metadata. *Science*, 350(6264):1073–1076, 2015. doi: 10.1126/science.aac4420.

[3] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3–5):75–174, 2010. doi: 10.1016/j.physrep.2009.11.002.

[4] Scott A. Golder and Michael W. Macy. Digital footprints: Opportunities and challenges for online social research. *Annual Review of Sociology*, 40(1):129–152, 2014. doi: 10.1146/annurev-soc-071913-043145.

[5] Mark S. Handcock, Adrian E. Raftery, and Jeremy M. Tantrum. Model-based clustering for social networks. *Journal of the Royal Statistical Society A*, 170(2):301–354, 2007. doi:10.1111/j.1467-985X.2007.00471.x.

[6] John R. Hipp and Andrew J. Perrin. The simultaneous effect of social distance and physical distance on the formation of neighborhood ties. *City & Community*, 8(1): 5–25, 2009. doi:10.1111/j.1540-6040.2009.01267.x.

[7] Peter D Hoff, Adrian E Raftery, and Mark S Handcock. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460): 1090–1098, 2002. doi:10.1198/016214502388618906.

[8] Norman L. Johnson, Samuel Kotz, and Adrienne W. Kemp. *Univariate Discrete Distributions*. Wiley-Interscience, 1993.

[9] Michal Kosinski, David Stillwell, and Thore Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 2013. doi: 10.1073/pnas.1218772110.

[10] E.L. Lehmann. *Elements of Large-Sample Theory*. Springer, 2004.

[11] ONS. 2011 census microdata individual safeguarded sample (local authority): England and wales, 2015.

[12] JJ Pfeiffer, S Moreno, T La Fond, J Neville, and B Gallagher. Attributed graph models: Towards the sharing of relational network data. Technical report, Purdue University, 2014.

[13] Jeffrey A. Smith, Miller McPherson, and Lynn Smith-Lovin. Social distance in the united states: Sex, race, religion, age, and education homophily among confidants, 1985 to 2004. *American Sociological Review*, 79(3):432–456, 2014. doi:10.1177/0003122414531776.

[14] Yuchung J. Wang and George Y. Wong. Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association*, 82(397):8–19, 1987.

Familiar Strangers: the Collective Regularity in Human Behaviors

Yan Leng
MIT Media Lab
Cambridge, MA, USA
Email: yleng@mit.edu

Dominiquo Santistevan
Electrical Engineering & Computer Science
Cambridge, MA, USA
Email: niquo@mit.edu

Alex 'Sandy' Pentland
MIT Media Lab
Cambridge, MA, USA
Email: pentland@mit.edu

***Abstract*—The social phenomenon of familiar strangers was identified by Stanley Milgram in 1972 with a small-scale experiment. However, there has been limited research focusing on uncovering the phenomenon at a societal scale and simultaneously investigating the social relationships between familiar strangers. With the help of the large-scale mobile phone records, we empirically show the existence of the relationship in the country of Andorra. Built upon the temporal and spatial distributions, we investigate the mechanisms, especially collective temporal regularity and spatial structure that trigger this phenomenon. Moreover, we explore the relationship between social distances on communication network and the number of encounters and show that larger number of encounters indicates shorter social distances in social network. The understanding of the physical encounter network could have important implications to understand the phenomena such as epidemics spreading and information diffusion.**

We all have the experiences of encountering strangers during regular activities in our daily lives whom we are able to recognize [1], [2]. This is the interesting social phenomenon "Familiar Stranger" of urban environment, first identified by Stanley Milgram in 1972 by asking travelers to recognize "strangers" they met on a bus platform [1]. In 2013, Sun [3], for the first time, uncovered the encounter mechanisms of three million public transit users in Singapore to capture the time-resolved, in-vehicle encounter patterns and familiar strangers using transit smart-card data.

This hidden dynamic social network plays an unnoticeable but significant role in information diffusion, behavior synchronization and epidemics spreading process. Christakis [4] and Montanari [5] shed light on the diffusion of information, innovations and behaviors via social contagion driven by social interactions. Dong used a Markov jump process to capture the co-evolution of friendship and visitation patterns in a student dorms with monthly surveys and locations tracking through mobile phones [6]. Besides, a series of studies has focused on using large-scale or high-resolution data to empirically and computationally model the epidemics process. Danon showed that large-scale interaction data are needed to verify the assumptions of random transmission models and simple network structures in order to better understand and prediction the disease transmission [7]. Stopczynski utilized large-scale behavioral data for modeling epidemic prediction by physical proximity network [8]. Isella and Salathe tracked high-resolution proximity network to understand the transmission paths of diseases [9], [10].

I order to capture the physical proximity network in urban environment, we use countrywide mobile phone logs to explore the phenomenon and underlying mechanisms that trigger familiar strangers at a societal scale covering many aspects of social lives. This data simultaneously captures two layers of network: physical proximity and social networks. To our knowledge, we are the first to identify country-wide familiar strangers utilizing both mobility and social networks. We confirm the existence of the phenomenon of familiar stranger in urban environment. At a macro scale, we found that the collective regularities - temporal regularity and spatial structure - explain the phenomenon of familiar strangers. We also investigate the relationship between physical co-occurrences and the proximity in social networks and show that physical co-occurrences indicate shorter social distances via communication network. The understanding of collective regularity has implications for the prevention of epidemics and the facilitation of information spreading.

Data and settings

To capture both the social network and mobility network, we study the anonymized Call Detail Records of Andorra, a European country, for July, 2016. This dataset includes the caller, receiver, connected cell tower, start and end time of the connection. The spatial and social network encoded in Call Detail Records enable us to identify familiar strangers. We created a communication network where each user is a vertex and an edge exists if there exists direct contact between two users. In the month of July, we are looking at a total of 1,264,292 users. After filtering out users with more than 100 connections that may be hotels or vendors, we only consider the remaining 1,211,814 users. We identify the physical encounters by observing if two users called or text on the same tower within a one-hour window of each other. In our study, we define familiar strangers as individual pairs who physical co-locate at one cell tower within one-hour time window, but there exists no direct links on the communication network.

Spatial and temporal patterns of encounters

To understand the temporal distribution of physical encounters, we extract the time when an encounter happen as we see in Fig. 2. There existed prominent spikes between working hours (8 am - 11 am and 3 pm - 5 pm) on weekdays, and slight shifts for weekends (10 a.m. - 11 a.m. and 3 p.m. - 6 p.m.). Interesting, we see a small peak at 11 p.m. on Saturday night, which captures the encounters of Saturday night life.

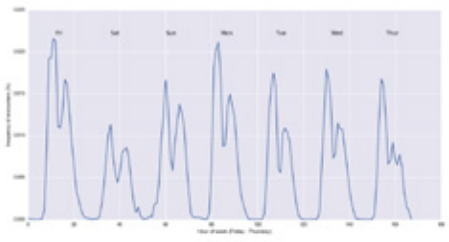


Fig. 1. Temporal distribution of physical encounters

We also analyze the spatial distribution of encounters and the number of encounters happens for each encounter pair. More specifically, each tower has different usage that is dependent on the popularity around that tower, but there is also a distribution over how many encounters happen between two users at a single tower. As shown in Fig. 2, there exists some towers that are not quite as popular with encounters, yet they have some of the highest number of encounters counts between pairs of users.

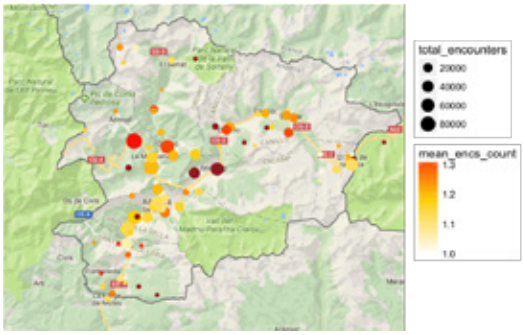


Fig. 2. Map of cell towers, where the size of the node is proportional to total number of encounters and the color is darker where a pair of users encounter more times on that tower.

COLLECTIVE REGULARITIES

Human mobility follows a high degree of temporal and spatial regularity [11] and physical encounter do not happen out of "coincidence". We believe that collective regularity of human behaviors triggers the physical co-occurrences. In particular, we focus on the temporal regularities and spatial structure of two consecutive encounters to uncover the underlying mechanisms that drive the encounter.

A. Temporal regularity

Human daily routines, such as commuting, follow certain temporal patterns. To explore the repeated encounters at a population scale, we created a encounter network based on mobility behaviors across a week and measure the inter-event time Δt between consecutive encounters of each familiar strange pair.

The left panel of Fig. 3 shows the distribution of inter-event time between two encounters. We observe noticeable peaks for every 24 hours and another lower peaks for $24 \cdot d \pm 6$ hour, where d is the d^{th} day after the first encounter. This indicates that people are very likely to encounter their familiar strangers on the same hour of day within the next d days. In addition, we analyze the time of the consecutive encounters. The right panel of Fig. 3 can be categorized into two types of encounters - weekday encounters and weekend encounters. Specifically, people who encounter during weekdays are less likely to encounter each other during weekends and vice versa. Besides, morning encounters are more likely to encounter again during morning, explaining the 24 hour peak as shown in the figure in the left panel. Both empirical exercises highlight the collective temporal regularity in people's daily routines partly explains the repeated physical encounters.

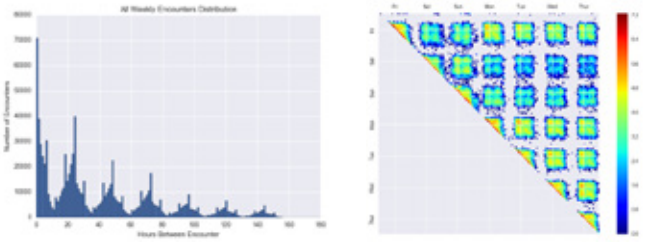


Fig. 3. Temporal regularity of collective mobility behaviors. The left panel shows the probability density function of inter-event time ($P_{\Delta t}$ between two consecutive encounters. The right panel shows the encounter time of two consecutive co-locations

B. Spatial structure

We also investigate the spatial structure of repeated encounters, specifically the popularity of the two locations, distances between the two locations and the Points of Interests.

1) Popularity of and distances between the consecutive encounter locations: To gain insights on the encounter and re-encounter pairs, we use the general gravity model to understand the relationships between spatial distances, attractions (popularity) of the two locations. The gravity model utilized in our study assumes the following functional form, as shown in (1). The fit of the model, and the relationships between encounter and re-encounter flow and the above-mentioned variables of interests are shown in Fig. 4.

$$T_{ij} = C\frac{N_i^{\alpha} N_j^{\beta}}{D_{ij}^{\gamma}} \tag{1}$$

where T_{ij} is the number of encounter and re-encounter flow between two geographical location i and j. D_{ij} is the distance between two geographical areas and N_i and N_j are the number of encounters of area i and j respectively. After applying a logarithmic transformation and fit parameters with a linear regression, we found $\alpha = 0.38, \beta = 0.407, \gamma = 0.823$.

FAMILIAR STRANGERS IN SOCIAL NETWORK

There has been research establishing the relationships between mobility behaviors and social ties. Crandall (2010) developed a framework to empirically and mathematically investigate the relationship between social ties and co-occurences [12]. Along the same line, Toole (2015) found that that the composition of a user's ego network in terms of the type of contacts they keep is correlated with mobility behavior[13]. Apart from the observed relationship between mobility behavior and the probability of the formation of a tie or the

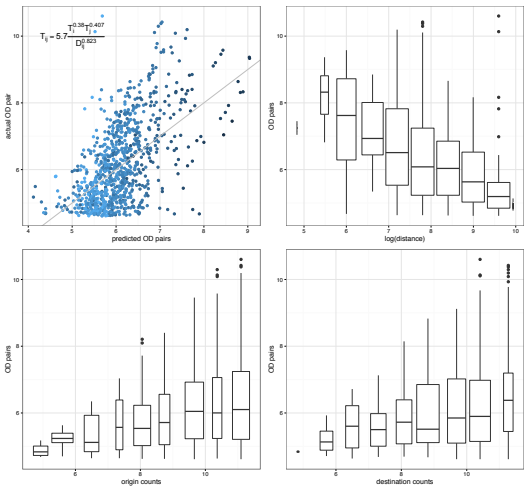


Fig. 4. Gravity law fit of encounter network. (A) Encounter and re-encounter flow obtained from data on the number of encounters the first and second locations, as a function of the distance between the encounter and re-encounter locations. (B-D) Encounter and re-encounter flow as a function of distance, encounter popularity of first and second locations.

strength of a ties, we are specifically interested in the social distances between 'strangers' who physically encounter one another multiple due to similar routines or interests. As shown in Figure 5, there exists a negative relationship between social distance and the number of encounters - the more time each familiar stranger pair encounter one another, the closer they are on the social networks.

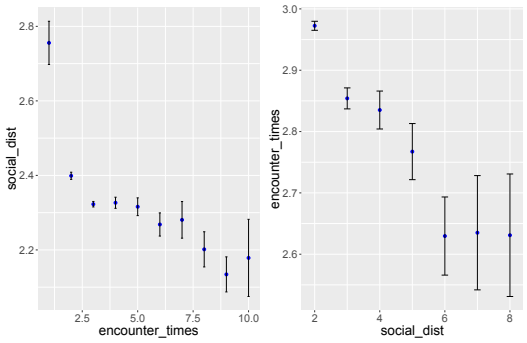


Fig. 5. Relationship between number of encounters and social distances in social networks of familiar stranger pairs

Discussions

In this study, we use the large-scale Call Detail Records of a European country Andorra to create the physical encounter network and phone communication network. We show the existence of the familiar stranger phenomenon in urban environment. By analyzing the temporal and spatial characteristics of the encounters, we uncover the underlying mechanisms, especially collective temporal regularity and spatial structure that trigger the phenomenon. In the end, we explore the relationship between social distances along social network and number of encounter in mobility network. We show that larger number of encounters indicates nearer social distances. The understanding of the physical encounter network could have important implications for epidemics preventions and information spreading facilitation.

Our study posits several interesting future studies. First of all, one natural future work is to investigate how the relationship of familiar strangers grows into actual friendship, and how their behaviors intersect and synchronize with each other. It would be even more interesting to understand the causal relationships between physical co-occurrences and the formation of social ties. Another interesting direction is to integrate the familiar stranger relationships into the modeling and simulation of epidemics spreading and information diffusion.

References

[1] S. Milgram, J. E. Sabini, and M. E. Silver, *The individual in a social world: Essays and experiments* . Mcgraw-Hill Book Company, 1992.

[2] E. Paulos and E. Goodman, "The familiar stranger: anxiety, comfort, and play in public places," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 223–230, ACM, 2004.

[3] L. Sun, K. W. Axhausen, D.-H. Lee, and X. Huang, "Understanding metropolitan patterns of daily encounters," *Proceedings of the National Academy of Sciences*, vol. 110, no. 34, pp. 13774–13779, 2013.

[4] N. A. Christakis and J. H. Fowler, "Social contagion theory: examining dynamic social networks and human behavior," *Statistics in medicine*, vol. 32, no. 4, pp. 556–577, 2013.

[5] A. Montanari and A. Saberi, "The spread of innovations in social networks," *Proceedings of the National Academy of Sciences*, vol. 107, no. 47, pp. 20196–20201, 2010.

[6] W. Dong, B. Lepri, and A. S. Pentland, "Modeling the co-evolution of behaviors and social relationships using mobile phone data," in *Proceedings of the 10th International Conference on Mobile and Ubiquitous Multimedia*, pp. 134–143, ACM, 2011.

[7] L. Danon, T. A. House, J. M. Read, and M. J. Keeling, "Social encounter networks: collective properties and disease transmission," *Journal of the Royal Society Interface*, p. rsif20120357, 2012.

[8] A. Stopczynski, A. S. Pentland, and S. Lehmann, "Physical proximity and spreading in dynamic social networks," *arXiv preprint arXiv:1509.06530*, 2015.

[9] L. Isella, J. Stehlé, A. Barrat, C. Cattuto, J.-F. Pinton, and W. Van den Broeck, "What's in a crowd? analysis of face-to-face behavioral networks," *Journal of theoretical biology*, vol. 271, no. 1, pp. 166–180, 2011.

[10] M. Salathé, M. Kazandjieva, J. W. Lee, P. Levis, M. W. Feldman, and J. H. Jones, "A high-resolution human contact network for infectious disease transmission," *Proceedings of the National Academy of Sciences*, vol. 107, no. 51, pp. 22020–22025, 2010.

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

[12] D. J. Crandall, L. Backstrom, D. Cosley, S. Suri, D. Huttenlocher, and J. Kleinberg, "Inferring social ties from geographic coincidences," *Proceedings of the National Academy of Sciences*, vol. 107, no. 52, pp. 22436–22441, 2010.

[13] J. L. Toole, C. Herrera-Yaqüe, C. M. Schneider, and M. C. González, "Coupling human mobility and social ties," *Journal of The Royal Society Interface*, vol. 12, no. 105, p. 20141128, 2015.

Impact of university admission on student's egocentric network

Sami Jouaber*, Yannick Leo*, Carlos Sarraute†, Eric Fleury* and Márton Karsai*
*Univ Lyon, ENS de Lyon, Inria, CNRS, UCB Lyon 1, LIP UMR 5668, IXXI, F-69342, Lyon, France
†Grandata Labs, 550 15th Street, San Francisco, CA, USA
sami.jouaber@ens-lyon.fr, yannick.leo@ens-lyon.fr, charles@grandata.com,
eric.fleury@inria.fr, marton.karsai@ens-lyon.fr

Introduction

Social relations are well-known to bring significant individual and collective benefits, however their maintenance requires efforts, which are limited by certain capacities. Examples are time [1], [2] or space [3] but constraints were also found on the cognitive level. As argued by the social brain hypothesis [4], there is a quantitative relationship between primate's brain capacities and the social group size they can effectively handle and maintain. Due to these reasons, individuals need to develop social strategies to maintain their immediate social network. These limitations and strategies not only determine the size of one's egocentric network but play also an important role in the maintenance dynamics and prioritizing of social ties. It has been recently shown that while the overall distribution of efforts that an ego commits to his/her friends is invariant in time [5], the turnover speed of creation and dissolution of social ties disclose individual social strategies on the long run [6]. These observations were possible due to the recent availability of large datasets collecting digital records of interactions of millions of individuals using mobile phone devices. Mobile phones became personal items of our everyday life and they allow to automatically capture temporal data of human interactions on the population level.

In this work we used mobile call communication data recorded in a developing country to study how some changes in the social environment affect the structure and the dynamics of one's egocentric networks. More precisely, we study the impact of university admission on the composition and evolution of the egocentric networks of 1675 freshmen students. The data was obtained via the combination of two anonymized datasets, one which records a sequence of anonymized call detailed records (CDR) of customers of a single telco operator over 2 years; and an anonymised bank dataset which was recorded over an inclusive period of 6 months and provided demographic informations (age, gender, postal code) and certain details about the evolution of the financial situation of students. This combined dataset gives us the opportunity to follow the egocentric network evolution of freshmen students, with observations starting 6 months before their university admission, and continuing over the following 18 months.

Social signature: persistency and network correlations

The first part of our study is based on the measurement of social signature, which was introduced recently by Saramäki and others [5]. Informally, social signature measures the ranked distribution of communication efforts (total number or duration of calls) by an ego toward his/her acquaintances over a period of time. In their original study, observing 24 students over 18 months, Saramäki et.al. found that although the acquaintances may change in one's egocentric network, the social signature remains rather invariant over time but highly specific to an individual. Using our considerably larger dataset first we confirm these findings by providing three important insights:

- The communication effort of people is not uniformly distributed (see Fig. 1a), as people focus their efforts on a small number of ties that may relate them to family or close friends. The shape of the social signature function is not well-balanced and reveal an unequal spread of efforts.
- Although acquaintances in the egocentric network of an individual are constantly changing, the overall shape of the social signature function stays rather invariant over time (as shown for three randomly selected students in Fig. 1a).
- By measuring the Jensen-Shannon Divergence (JSD) δ between the social signatures of the same ego observed in consecutive periods, its variance appears to be constant over time (see red bars in Fig. 1b),

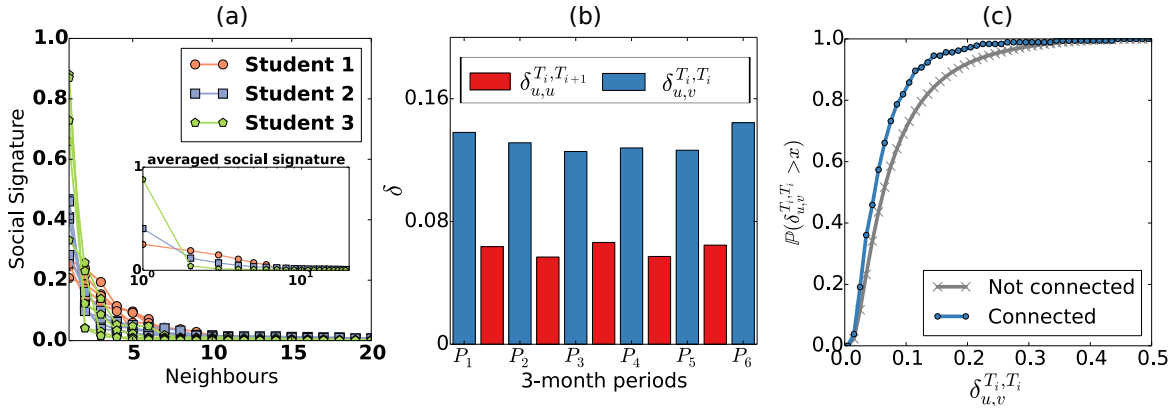


Fig. 1. **Measures of social signature.** In **(a)**, social signatures of three random selected students over 5 months (15 curves) are shown together with their average (see inset). **(b)** JSD δ measured between the social signature of different students (in blue) and between pairs of consecutive 3-month periods of the same student (in red). **(c)** Cumulative distribution of δ JSD of social signatures measured between connected pairs (in blue) and not connected pairs (in grey) of students.

and remains relatively smaller than the same measure between the average social signatures of randomly selected pairs of users (see blue bars in Fig. 1b).

Beyond these results we establish two new findings:

- The social signature of egos are more similar to the signature of their friends than to others. Fig. 1c shows the cumulative distribution of the JSD of social signatures of connected pairs of people (in blue) and unconnected pairs (in grey). It reveals that people who are connected through the communication network appear with a smaller difference as compared to unconnected others.
- People exhibiting similar social signatures can be clustered in groups, where similarities inside the group are significantly higher as compared to people from other groups (results not show here).

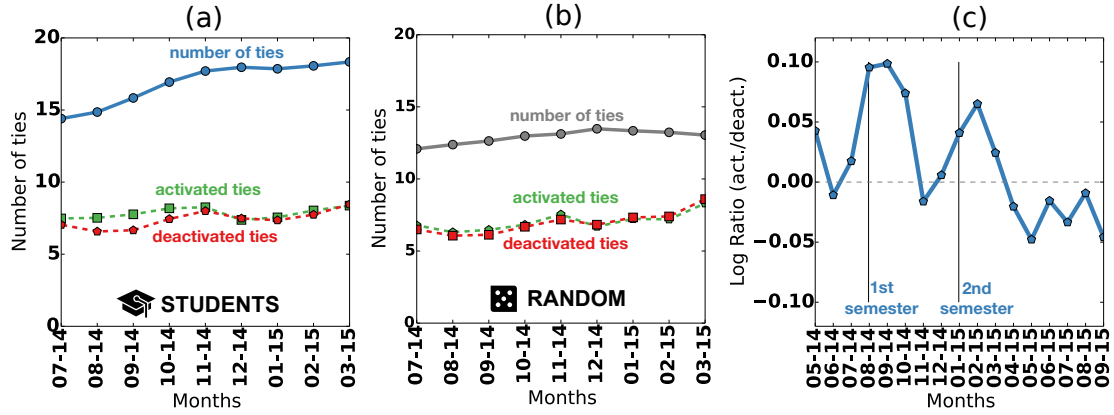


Fig. 2. **Formation and decay of communication ties. (a)** The average of the actual number of ties (in blue), the number of created (in green) and dissolved ties (in red) for 1675 students as a function of time (in month) **(b)** The same plot as (a) for a random sample of people measuring the actual number of ties (in grey), and the number of created (in green) and dissolved ties (in red). **(c)** The logarithm of the fraction of two average turnover functions measured for students and a random sample of people as a function of time. If this function is positive, there are larger turnover in the student egocentric network during two consecutive months as compared to random people.

Temporal formations and decays of communication ties

In the second part of our study we address the turnover of social ties in one's egocentric network, i.e. how new ties replace old relations as a function of time. To observe precisely evolution of egocentric networks we follow the methodology of Miritello et.al. [6], which allows us to carry out a continuous temporal analysis of decays and creations of ties by considering two reference periods and one experiment period.

Our observations confirm earlier results [6] that although the turnover of acquaintances is relatively high in one's egocentric networks, the actual number of active ties is rather constant as a function of time. This has been observed for a large control group of randomly selected people, who are not students, as shown in Fig. 2b. On the contrary, we found that students behave somewhat differently as during the year that they enter the university, the number of their active ties is increasing (from 15 to 18) suggesting social benefits and new social strategies developed during this period (see Fig. 2a). Moreover, the turnover of their egocentric network follows different patterns as other people. Here we measured the turnover as the fraction of newly created and deleted social ties of an ego during a given period. In Fig. 2c we show the logarithm of the ratio of average turnover functions measured for the set of students and in the control group. As we see most of new relations are created by students just after the entrance of the first and the second academic semester revealing that students choose different strategies to socialize during these periods to receive collective benefits.

REFERENCES

[1] G. Miritello, E. Moro, R. Lara, R. Martínez-López, J. Belchamber, S. G. Roberts, and R. I. Dunbar. Time as a limited resource: Communication strategy in mobile phone networks. *Social Networks*, 35(1):89–95, 2013.
[2] S. G. Roberts and R. I. Dunbar. The costs of family and friends: an 18-month longitudinal study of relationship maintenance and decay. *Evolution and Human Behavior*, 32(3):186–197, 2011.
[3] G. A. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2):81, 1956.
[4] R. Dunbar. The social brain hypothesis. *brain*, 9(10):178–190, 1998.
[5] J. Saramäki, E. A. Leicht, E. López, S. G. Roberts, F. Reed-Tsochas, and R. I. Dunbar. Persistence of social signatures in human communication. *Proceedings of the National Academy of Sciences*, 111(3):942–947, 2014.
[6] G. Miritello, R. Lara, M. Cebrian, and E. Moro. Limited communication capacity unveils strategies for human interaction. *Scientific reports*, 3, 2013.

SESSION 8
MOBILITY

More reliable and more accurate traffic models using mobile phone data

Klaas Friso, DAT.Mobility, Deventer, the Netherlands, kfriso@dat.nl

Jasper Keij, Mezuro, Weesp, the Netherlands, jasper.keij@mezuro.com

Abstract

Mobile phone data are a rich source to infer all kinds of mobility related information. Mezuro and DAT.Mobility are collaborating in developing innovative data-analysis techniques of transforming mobile phone data in information that can be used for traffic models. In this paper it is explained how the national origin-destination matrix (OD-matrix) is derived, based on the Call Detail Records (CDR) from Vodafone (one of the three mobile phone providers in the Netherlands), facilitating between 30 and 40% of the Dutch mobile phone usage. This OD-matrix is currently used in the update of the traffic model of the Rotterdam region. No other data source is known that gives travel information at a national scale at this high level. The developments of improving the quality of the OD-information are still going on. In this paper the next planned steps in our development will be highlighted.

Introduction

The aim of a traffic model is to provide insight into current and future traffic and transport flows as good as possible. Questions that can be answered with a traffic model are for example: How much traffic is driving on the road? At which location and at what time? Where are the locations of congestion? What is the effect of a change in the infrastructure or in traffic policy? What measures are needed to regulate the additional traffic of an event?

An important source for current traffic modelling in the Netherlands are the results of the national household travel surveys, called OViN (Onderzoek Verplaatsingen in Nederland) [1] conducted by Central Bureau of Statistics in the Netherlands. By conducting surveys (40,000 trip days annually) the movement behaviour of the Dutch population is described by place of origin and destination, time of travelling, means of transport used and the purpose of travelling. The results of this yearly survey will then be scaled to represent the entire Dutch population. Although this research is the gold standard in terms of available data sources, the sample size is very small and as a result it allows hardly spatial detail. On average OViN-data record the traffic movements of slightly more than 100 people on a particular day over the Netherlands. Therefore, it is not possible to make representative statements for specific regions (for example regions with a low number of inhabitants). Also the available information from OViN is not up-to-date (about 2 years old). Notice that information of trips made by trucks are not included in OViN and information of these trips is hard to gather by surveying.

Deriving trips from mobile phone data

The availability of this relatively new big data source is potentially an enormous enrichment for traffic models. Mezuro made it possible to define trips in the Netherlands based on CDR-data (Call Detail Records) of Vodafone users by developing several algorithms in the past few years. In this way it is possible to monitor the trip movements of more than 3.5 million persons daily. By using this data in traffic modelling the reliability and accuracy will increase because the analyses are not any longer based on a survey of 40,000 trips a year, but more than 1.25 billion trips per year.

At this moment it is not possible to replace all current data sources on which the traffic models are based by the mobile sensing data source of Mezuro. For example, there is no answer yet about the purposes of the travellers and all used modes. Furthermore, it is not possible yet to detect short distance trips. In order to answer such questions, the data will need to be linked with external databases.

Mezuro (specialists in processing big data) and DAT.Mobility (specialists in traffic consultancy/data) are collaborating in developing innovative data-analysis techniques of transforming mobile phone data in information that can be used for traffic models. A national origin-destination matrix (OD-matrix) for the Netherlands is available on a regular base. Therefore the Netherlands is divided in 1,261 traffic analysis zones where each zone represents a city or a town. The larger cities are split into separate city districts. The most detailed time granularity of the OD-matrix is an hour. In traffic modelling commonly the average working day and especially the peak periods are interesting for which the total mobility is needed. Because the mobile phone data gives information about 1/3 of the total amount of trips a scaling procedure is developed [2]. The scaling procedure depends on the number of inhabitants and penetration of mobile devices (per age group and area), the Vodafone market share per area and the change that a person will make a trip (for each day of the week). It is important to take into account the information per age group because there is an underrepresentation of mobile phone usage (in relation to trips made) amongst people younger than 12 years and older than 60 years.

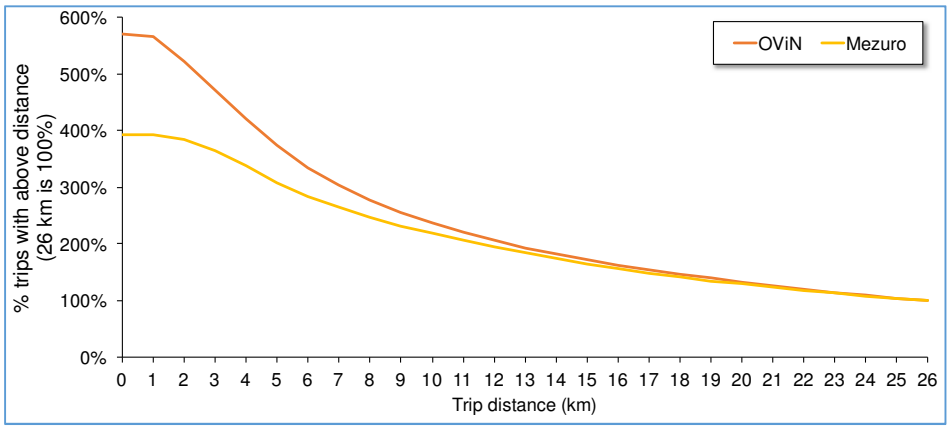


Figure 1: comparison between OViN and Mezuro-data of the relative number of trips per travel distance.

When compared to OViN a similar distribution of the number of trips on longer distances is seen, but diverging trip counts for trips with a travel distance below 10 kilometre (see Figure 1). We find there are about as many trips in both datasets over 10 kilometre, though below 10 kilometre OViN appears to record far more trips. Hence, we choose to only take trips over 10 kilometre into account as we have an underrepresentation below 10 kilometre. This is a bias that is hard to correct for provided it depends on the physical cell tower infrastructure. In areas with large reaching cell towers one might never be able to detect short distance trips, whereas in cities with smaller cells one might. An assumption which explains the underrepresentation of short distance trips is that people are more tended to leave their device at home, for example during a short walk with the dog or a trip to the bakery. So at this stage it is just possible to use trip information above 10 kilometre in traffic

modelling, but this is already very useful. The expectation is that in the near future this limit can be lowered to 5 kilometre because of technical improvements. In the meantime we will work on algorithms for an estimation procedure concerning the number of trips between 5 and 10 kilometre, because our goal is to make the national OD-matrix as complete as possible for all kinds of mobility applications.

Privacy

The privacy of the raw mobile phone data is assured by a rigorous protocol. Firstly the identifying information (phone number) is one way hashed. The one way hashing information is changed every month. In this way the movements of a single device can only be detected for one month. Secondly, all data on device level is stored in a 'black box' at the telecom provider. The data on device level cannot be seen and only algorithms tested beforehand can transform the raw data into movements. At last all results are aggregated. The export function to create results allows only 16 or more devices as output. OD-pairs having less than 16 devices are not within the dataset. In this way it is impossible to relate information to a single device of an individual.

National OD-matrix

The raw data is processed into basic location information removing data which is subsequently translated into OD-information based on the time ordered stay sequence. A trip between an origin and destination is defined when a mobile device is longer than 30 minutes at a certain location. A deviant location in between will result in an extra destination, which means two separate trips. In Figure 2 the OD-data of mobile phone data for a single day is presented showing a plausible spatial distribution. The figure shows that most trips are made in the Randstad area (Amsterdam, The Hague, Rotterdam and Utrecht).

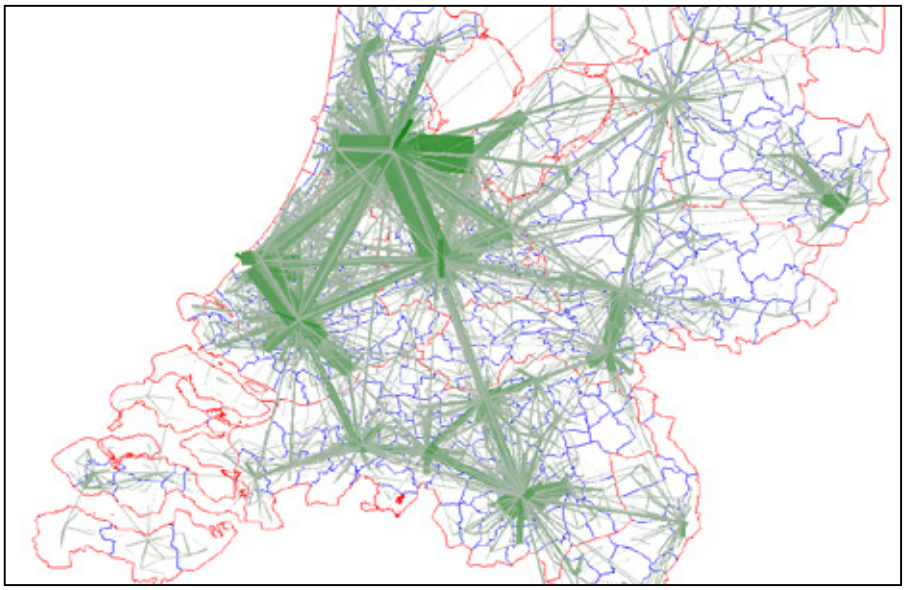


Figure 2. Spatial distribution of trips from mobile phones (data of one single day)

Derivation of trips by train and non-train

For a trip that is defined from the mobile phone data the mode by which this trip is made is not known directly from the data itself. For traffic modelling and also other applications it is very

interesting to determine the mode of a trip. At this stage we are able to distinguish between trips made by train and otherwise (non-train). For trips above 10 kilometres it means that the non-train trips are mostly made by car or truck. This is about 85%, the other 15% represent trips made by bus, tram, metro, scooters or cycling [1]. Figure 3 shows the share of trips made by train from the mobile phone data per origin zone per day together with the railroads and train stations. One of the next steps in our planned developments is therefore to split the non-train OD-matrix further into car and truck (and other).

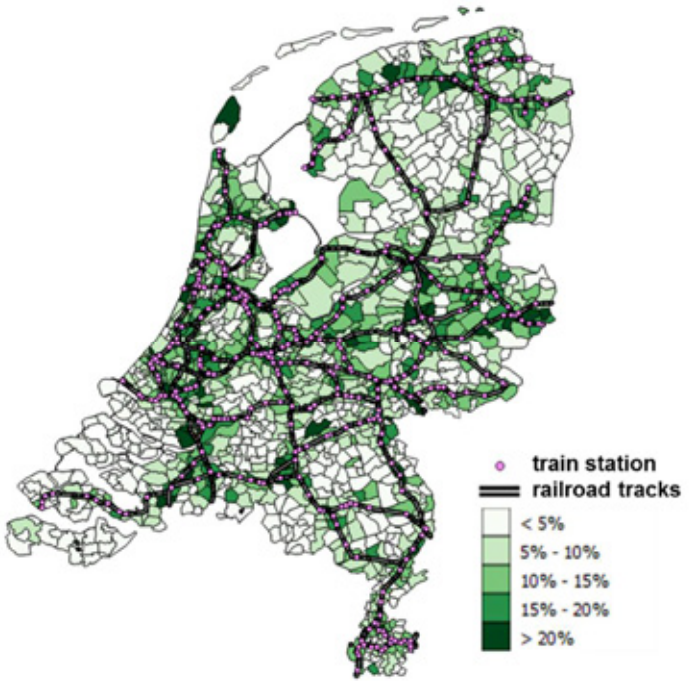


Figure 3: share of trips made by train (per origin zone)

Pilot studies

In 2015 we performed a successful pilot study in the Rotterdam region with enriching their synthetic a priori OD-matrix with the OD-matrix of mobile phone data [3]. Results show a better match of the assignment results of this enriched matrix with the counts indicating a better quality of the matrix. It also has been shown that specific (strong or weak (historic)) OD-relations which are known to be difficult to model become more plausible by the enriching procedure. In theory it is possible in traffic modelling to put special attention in known specific OD-relations. However, there exist unknown OD-relations that will not be modelled in a proper way. By making use of the measured OD-information from mobile phone data these relations will also improve in the model. In the update of the model of the metropole region Rotterdam-The Hague, which is currently in progress, the OD-matrix of mobile phone data will be actually applied in the matrix estimation for the base year.

In 2016 a pilot study has been performed with the National Model System of the Ministry of Transport in the Netherlands. In this pilot also an enrichment of the a priori model was performed. From both pilots we have learned that the average trip length as is measured in the mobile phone data is higher than of the synthetic models. Because the number of trips per zone was kept constant in the enrichment procedure this leads automatically to increased flows after assigning the OD-

matrix on the model network. In the Rotterdam pilot the choice was made to apply a correction factor on the number of trips per distance class with the result that the original trips length distribution was obtained again in the resulting OD-matrix. This was not part of the pilot with the National Model System.

It can be concluded that the enrichment procedure works well if the trip length distribution of the synthetic a priori model corresponds with the distribution of the mobile phone data, which is based on real measurements. It is possible, by means of usage of corrections factor to adjust for this in the model system. But, in fact, it means that the parameters of the a priori model should be reconsidered.

Goals for the near future

As mentioned before Mezuro and DAT.Mobility continue on the developments with regard to obtaining high quality OD-matrices out of mobile phone data which can be used already for a lot of applications with respect to transport analysis and as input for traffic models. At this stage we are able to deliver an OD-matrix at national level in the Netherlands for any given time period that is wanted for trips more than 10 kilometres. We are able to divide this matrix for the trips made by train and non-train. Our first goal is to work on algorithms to divide the non-train matrix further on in modes car, truck and other (bus, tram, metro, scooter and bike). Also we want to improve the usability of the data further on to work on methods that will obviate the underrepresentation of short distance trips (below 10 kilometre).

References

[1] OViN, Onderzoek Verplaatsingen in Nederland, www.cbs.nl/nl-NL/menu/informatie/deelnemersenquetes/personenhuishoudens/ovin/doel/default.htm

[2] J. Meppelink, J. van Langen, A. Siebes, M. Spruit, *Know Your Bias: Scaling Mobile Phone Data to Measure Traffic Intensities,* 2016

[3] L.J.J. Wismans, K. Friso, J. Rijsdijk, S.W. de Graaf, J. Keij, *Improving a-priori demand estimates transport models using mobile phone data. Rotterdam region case*. Mobile Tartu Conference, June 2016.

Understanding Drivers of Short Term Mobility

Sveta Milusheva[1]*, Elisabeth zu Erbach-Schoenberg[2,3]*, Linus Bengtsson[2,4], Erik Wetter[2,5], and Andy Tatem[2,3]
*Correspondence: svetoslava_milusheva@brown.edu, elisabeth.zu.erbach@flowminder.org
[1] Brown University, Providence, Rhode Island US [2] Flowminder Foundation, Roslagsgatan 17, 113 55 Stockholm, Sweden. [3]WorldPop, Geography and Environment, University of Southampton, University Road, Southampton SO17 1BJ, UK. [4]Dept. of Public Health Sciences, Karolinska Institute, Sweden. [5]Stockholm School of Economics, Stockholm, Sweden

I. Introduction

Populations are highly mobile, both in terms of long term movements of individuals relocating their place of residence as well as shorter term mobility such as commuting, work travel and recreational trips. Short term internal population movements have been shown to have an impact on spread of diseases [Oster 2012, Balcan et al 2009, Huang et al 2013, Wesolowski, Metcalf et al 2015, Wesolowski, Qureshi et al 2015], and increases in short term mobility will likely increase pollution and congestion and have potential implications for economic activity. Yet, the drivers of short term mobility have not been studied due to the lack of available data on fine spatial and temporal scales.

Call detail records (CDRs) provide a valuable data source for measuring population movements. By providing geo-location and timestamps for calls and text messages, they allow researchers to observe changes in location of individuals within a country and are commonly used to estimate short term mobility on the scale of whole countries. Long term movements are often measured as part of a census, available for many, but not all, countries. Previous studies have shown that short term movements measured using mobile phone data correlate with the long term migration variables available in census data. [Wesolowski et al 2013, Ruktanonchai et al 2016]. In this paper, we extend previous work in two ways. First, we include a time component to study how the relationship between long and short term movements changes over time. Second, we try to understand the drivers of this relationship at the individual level and the motivation behind an individual's choice to make a short or long term move based on previous movement patterns and the location of their social network. Working with call detail record data from Namibia and Senegal, we study the link between migration and short term movement and investigate how well the relationship found in one country translates to another country.

II. Data

For Senegal, we use CDR data provided by Sonatel and Orange in the context of the Data for Development Challenge in order to measure short term movement. The data consist of call and text data for Senegal between January 2013 and December 2013 for all of Sonatel's 9.5 million SIM cards, representing a large portion of the 14.2 million Senegalese population. In addition to timestamped locations for each communication, the data contains a hashed identifier for the second party involved in the call or text. For Namibia, CDR data was provided by MTC with the purpose of estimating internal mobility and its impact on malaria transmission. The data spans a period of 3.5 years, from October 2010 to May 2014 and includes all subscribers to MTC, corresponding to 72 billion entries for a user base of 4.5 million unique users.

Migration data for Senegal comes from the Agence Nationale de la Statistique et de la Demographie (ANSD) from a 10 percent sample of the 2013 census data. Migration data for Namibia was published by

the Namibia Statistics Agency [Namibia Statistics Agency 2015] using data collected in the context of the 2011 census.

III. Empirical Analysis

We study the relationship between long term migration and short term movements for both countries and then move on to explore two important drivers of short term movement--long term migration and social networks. For long term migration, we utilize the data for Namibia, which contains a longer time span of 3.5 years and use the Senegal data in order to explore the relationship between social networks and short term movements.

Long Term Migration as a Driver of Short Term Movement

We first examine the aggregate relationship between long term migration and short term movement for a full year and then incorporate the timing throughout the year into the analysis. Specifically, we can study how the correlation between short term and long term movement changes from day to day. Preliminary results for Senegal show positive and negative spikes in the coefficients throughout the year are associated with important holidays when people travel to visit family and friends. Usually we see a negative coefficient prior to the holiday, as individuals conduct short term moves in the opposite direction from where they migrated. After the holiday there are large positive coefficients, signaling individuals going back from their home location to their new place of residence. There is a consistent correlation between short term movement and long term migration patterns. Locations that have high levels of long term migration, also have high levels of short term movement.

We also explore the relationship between short and long term moves at an individual level, using the Namibia CDR data. We can explore the short term movement patterns of those individuals for which we observe a long term migration move in the CDR data using a fixed effects empirical strategy. This analysis helps us measure, for a district pair, whether long term migration between the two districts is driven by short term moves prior to the long term migration (the theory of someone gaining experience with a location prior to migration) or whether the relationship is the other way round, with short term moves following long term migration (individuals traveling to visit family and friends). And if both are present, this analysis can help us investigate how the magnitude of the effect varies.

Social Networks as a Driver of Short Term Movement

A large part of the relationship between short and long term movement can be explained by social networks. When individuals migrate, they often maintain social ties to individuals at the location of origin through calls and texts, and when possible will visit in person. Even for individuals that do not migrate, the locations visited during short term trips might be locations where they already have established contacts. The Senegalese CDR data is used to create a contact network of the locations where an individual has contacts that he or she has communicated with via mobile phone. We also have a locational network of the places the individual has physically visited, derived from an individual's location trajectory.

To study the relationship between the social network and short term movement, we regress the number of total short term moves on person contacts between two districts in Senegal. The relationship is very strong, with an R^2 of over .8. When analyzing the relationship between daily contacts and daily moves,

there is also a very high correlation and a weekly pattern emerges, with Sundays having the highest correlation. Conducting an analysis at the individual level as well, we determine the group of social contacts for each user. We use a fixed effects strategy and a logit model to estimate the effect of having a contact in a district on the probability of visiting that district. We also study the strength of the network and whether having a higher number of contacts increases the number of visits to that district.

IV. Conclusion

This paper aims to gain a better understanding of the drivers of short term movement. Using unique data for two different countries, we can exploit the longevity of the Namibia data and the detailed network structure of the Senegal data to study two important factors affecting short term movement. In addition, this is one of the first papers to combine mobile phone analyses from two different country contexts within sub-Saharan Africa. While it is necessary to study the factors of movement separately due to the different strengths of the two datasets, we are able to conduct similar analyses using census data for both country settings. As more and more work is conducted using mobile phone data in particular contexts, it will become important to examine how much of the analysis and results can be extrapolated to other country settings. This paper is a first step in doing this type of analysis, and it allows us to also consider how the results from the separate empirical analyses on the drivers of movement could translate to the other context and more broadly to other countries within sub-Saharan Africa.

References

Balcan, Duygu et al. (2009). “Multiscale mobility networks and the spatial spreading of infectious diseases”. In: Proceedings of the National Academy of Sciences 106.51 (2009), pp. 21484–21489.

Huang, Zhuojie, Andrew J Tatem, et al. (2013). “Global malaria connectivity through air travel”. In: Malar J 12.1 (2013), p. 269.

Namibia Statistics Agency (2015). “Namibia 2011 Census Migration Report”. Namibia Statistics Agency (2015).

Oster, Emily (2012). “Routes of Infection: Exports and HIV Incidence in Sub-Saharan Africa”. In: Journal of the European Economic Association 10.5 (2012), pp. 1025–1058.

Ruktanonchai, Nick W., et al. « Census-derived migration data as a tool for informing malaria elimination policy.” Malaria journal 15.1 (2016): 1.

Wesolowski, Amy, Caroline O. Buckee, Deepa K. Pindolia, Nathan Eagle, David L. Smith, Andres J. Garcia, and Andrew J. Tatem (2013). "The use of census migration data to approximate human movement patterns across temporal scales." PloS one 8, no. 1 (2013): e52971.

Wesolowski, Amy, CJE Metcalf, et al. (2015). “Quantifying seasonal population fluxes driving rubella transmission dynamics using mobile phone data”. In: Proceedings of the National Academy of Sciences 112.35 (2015), pp. 11114–11119.

Wesolowski, Amy, Taimur Qureshi, et al. (2015). “Impact of human mobility on the emergence of dengue epidemics in Pakistan”. In: Proceedings of the National Academy of Sciences 112.38 (2015), pp. 11887–11892.

Short-Term Traffic Prediction Using Visitor Location Registry Data

Sabra Ossen[1], Kulakshi Fernando[2], Rakshitha Godahewa[3], Fathima Dilhasha[4], Amal Shehan Perera[5], Malaka Walpola[6]
Department of Computer Science and Engineering
University of Moratuwa
Moratuwa, Sri Lanka

Abstract

Congestion due to road traffic is a major issue in urban areas. Currently, there are methods to predict short-term traffic using data based on Global Positioning System (GPS). This research aims at finding a cost-effective solution to congestion using Visitor Location Registry (VLR) data gathered through mobile cell towers. Prediction models were built using mobile network big data and results were validated using Closed Circuit Television (CCTV) records. Most suitable model for the given road segment is identfied by comparing multiple traffic prediction models. After extensive research on individual and ensemble models, we propose the majority-voting ensemble approach of Neural Network, Bayesian Combined Neural Network, and Kernel Regression method.
Keywords: Visitor Location Registry (VLR) Data; Traffic Prediction; Ensemble Models; Closed Circuit Television (CCTV) Data;

1. Introduction

Use of vehicles is the main transportation mechanism used worldwide in this century. Vehicles have made human life easier by reducing the travel times significantly compared to other transportation methods. However, the excessive usage of vehicles has resulted in many unexpected outcomes, but mainly traffic congestion.

At present, there are intelligent transportation systems which predict traffic with the help of road sensors. These types of methods are very expensive because of higher costs due to equipment, manpower and maintenance requirements. And with the increase of size and complexity of urban traffic networks the cost of road sensors can be seen to increase [5].

Visitor Location Registry (VLR) is a database in a mobile communications network and it has location details of mobile subscribers within the service area of a particular Mobile Switching Centre. A record in a VLR database has 4 attributes. Namely, International Mobile Subscriber Identity (IMSI) value of the user, Cell Global Identity (CGI) of the cell tower handling the mobile activity, date and time. A user is uniquely identified using the encrypted IMSI value.

Due to the availability of data, the possibility of using VLR data for traffic prediction has been analyzed in the paper by Singha and Kalita [7]. Alternative methods to build intelligent transportation systems using VLR data are analyzed in this research paper. With this motivation, we have proposed a traffic prediction model using VLR data based on machine learning as a low-cost solution.

2. Background

Traffic is identified as a dynamic, nonlinear process. In the literature, comparison on many parametric and nonparametric methods to modelling dynamic processes have been made and it can be seen that nonparametric methods perform better for dynamic processes [8]. But in general Neural Networks are known to handle such stochastic processes, similar to traffic, with better accuracy [1].

Literature highlights [2] that considering only a single link in the road network is not sufficient, and based on communication between the models an approach to model the overall network of junctions should be done. Similarly, it has been identified that individual neural networks exhibit good performance only on certain time periods. Therefore only a few of the individual methods are used for network wide prediction while most of the approaches suggest ensemble methods [9].

3. Methodology and Results

Traffic exhibits regular patterns over time due to the repetitive nature of human mobility [6]. Thus, we have introduced algorithms to identify patterns such as stationary users, pedestrians, and driving users.

A unique data record can be identified from a

distinct combination of IMSI, CGI, date and time values. Distinct data records were identified using Algorithm 1.

Algorithm 1 Duplicate Data Removal Algorithm

1: ***Categorize***: Categorize data set based on CGI and IMSI values.
2: ***Sort by Date***: For each subset categorized above , sort the records based on the date.
3: ***Sort by Time***: For each subset sorted above, sort the records by time. For each record considered in sorting if the time already exists discard that record.

If a user is stationary, he will take numbers of calls over a long time window within a day. We used this concept to identify the stationary users in Algorithm 2.

Algorithm 2 Stationary User Removal Algorithm

1: ***Get users with high frequency***: Identify the subset of records from the dataset containing IMSI values whose frequency is greater than five.
2: ***Get stationary users***: For each record in the subset, if the time gap between the first and last call is equal or greater than two hours, then that record is related to a stationary user.
3: ***Remove stationary users***: Identify and discard the IMSI values of stationary user records.

In Algorithm 3 as the mechanism of filtering Pedestrians, we used the number of unique CGI values that a particular person has gone through. It can be considered that the number of repetitions of CGIs is high in a slow movement.

Algorithm 3 Pedestrian Removal Algorithm

1: ***Get users with high frequency***: Identify the subset of records from dataset containing IMSI values whose frequency is greater than four.
2: ***Calculate CGI's traversed per user***: Calculate the unique number of CGIs for each user (each IMSI value) of the subset.
3: ***Get walking users***: For each record in the subset, if the time gap between the first and last call of a particular user is greater than or equal to two hours and during this time period he has passed only less than three CGIs, then classify those users as walking users.
4: ***Remove walking users***: Identify and discard the IMSI values of walking users.

With these filtering mechanisms, we were able to identify the users who were related to traffic. Each of the prediction models built below is validated using CCTV records where three traffic congestion levels "high", "medium", "low" were defined.

In the research conducted by Lee, Kim, Kim and Cho [3] comparison on models to predict short-term traffic has been made. Due to the dynamic nature of traffic, they have concluded that neural networks are more accurate when predicting the traffic. Based on the above, most of the models we have used include neural networks.

First is the Bayesian Combined Neural Network (BCNN) which is a combination of two neural networks using the Bayesian approach. Zheng and Lee have identified that existing traffic prediction models predict well only in a particular time period and they have proposed the above model [10]. Based on the above we used the Back Propagation Neural Network and Radial Basis Function Neural Networks. We have considered two roads of the roundabout in the junction in Sri Lanka as two segments that form the complete path of analysis.

The user count for the current time interval of the current road segment is calculated, based on previous and current time intervals of the previous road segment. For the testing data set the BCNN predicted values are calculated and then validated from a back propagation neural network using the normalized user count and the peak or non-peak quality of the time interval.

Next, we used two multi-layer feed forward neural networks with five hidden layers each. We divided the total time segment for ten minute time periods. For each of these periods, we found the normalized user count of that time period for a given road. An additional field to identify whether the time period is peak or non-peak was also used.

For the first neural network, the peak/non-peak bit for the current time period and normalized user count of the selected road was given as input. The user count for the next ten minutes of the road segment was considered as the output. Then, for the second neural network, we input the peak/non-peak bit for the current time period and the output we got from the first neural network (user count for the next ten minutes). Finally, the second neural network produced the traffic level for the next ten minutes for the selected road segment as the output.

Moving away from Neural Networks we next analyze the usage of regression models to forecast traffic [8]. Since nonparametric models do not assume an underlying distribution for data we test our data set with different nonparametric methods such as K-nearest neighbor regression, logistic regression, support vector machine regression and kernel regression.

Kernel regression tries to fit the local data

points into a distribution by using a kernel basis function for a specific kernel bandwidth (a selected window of data points). The kernel bandwidth can be a fixed bandwidth, an adaptive nearest-neighbor based bandwidth or a generalized nearest-neighbor based bandwidth which is specified using the method of Racine and Li [4].

Picking the simplest kernel which gives the best results decides the accuracy and applicability of kernel regression. Kernel regression method gave better accuracy when an Epanechnikov kernel, with bandwidth Type fixed and Kernel order of four, is used.

Finally, we used the majority-voting ensemble approach and combined the three prediction models described above.

CCTV data was used to validate the proposed models. For all days considered, we manually observed the amount of vehicles and assigned a traffic level (low, medium or high) for each road segment for each of the ten minute time periods. In model validation, we checked the predicted traffic level from a particular model against the actual traffic level which we identified through CCTV data. We measured the accuracy of the predictions of each model as follows.

$$Accuracy\% = \frac{No:ofCorrectlyClassifiedRecords \times 100}{TotalNumberofRecords}$$

Table 1 summarizes the accuracy levels for each of the prediction models described above.

Model	Accuracy Pct
NN Combination with User Count Prediction	80%
Bayesian Combined Neural Network	73%
Kernel Regression	78%
Ensemble Model (Majority Voting Approach)	85%

Table 1: Results of Different Prediction Models for Gnartha Pradeepa Mawatha SouthEast Road

4. Conclusions

Increasing road traffic is one of the major problems in the current world. Short-term traffic prediction can be considered as a necessary solution for the above congestion problem. Since existing methods of traffic prediction are very expensive, in this paper we investigated the approach of using VLR data as a cost-effective solution. With the raw VLR data obtained from telecom service providers, we were able to build the prediction models after applying custom filtering mechanisms. Based on the accuracy given by each prediction model we concluded that the majority voting ensemble approach of Bayesian Combined Neural Network, Neural Networks, and Kernel Regression is the most suitable approach. In the future, lightweight prediction models should be built to address the need of low computing power when predicting the traffic.

References

[1] K. Kumar, M. Parida, and V. Katiyar. Short term traffic flow prediction for a non urban highway using artificial neural network. *Procedia-Social and Behavioral Sciences*, 104:755–764, 2013.

[2] C. Ledoux. An urban traffic flow model integrating neural networks. *Transportation Research Part C: Emerging Technologies*, 5(5):287–300, 1997.

[3] S. Lee, D. Kim, J. Kim, and B. Cho. Comparison of models for predicting short-term travel speeds. In *5th World Congress on Intelligent Transp. Systems (CD-ROM)*, 1998.

[4] Q. Li and J. Racine. Predictor relevance and extramarital affairs. *Journal of Applied Econometrics*, 19(4):533–535, 2004.

[5] L. E. Y. Mimbela and L. A. Klein. Summary of vehicle detection and surveillance technologies used in intelligent transportation systems. 2000.

[6] K. Puntumapon and W. Pattara-Atikom. Classification of cellular phone mobility using naive bayes model. In *Vehicular Technology Conference, 2008. VTC Spring 2008. IEEE*, pages 3021–3025. IEEE, 2008.

[7] M. R. Singha and B. Kalita. Using mobile phone network for urban traffic management. *International Journal of Computer Applications*, 65(2), 2013.

[8] B. L. Smith, B. M. Williams, and R. K. Oswald. Comparison of parametric and nonparametric models for traffic flow forecasting. *Transportation Research Part C: Emerging Technologies*, 10(4):303–321, 2002.

[9] H. van Lint, A. Valkenberg, and A. van Binsbergen. Advanced traffic monitoring (atmo) for sustainable traffic management. In *Transitions Towards Sustainable Mobility*, pages 267–295. Springer, 2011.

[10] W. Zheng, D.-H. Lee, and Q. Shi. Short-term freeway traffic flow prediction: Bayesian combined neural network approach. *Journal of transportation engineering*, 132(2):114–121, 2006.

Understanding Multiday Activity Patterns Based on Mobile Internet Usage Behaviour

Yihong Wang, Gonçalo Correia, Bart van Arem
Department of Transport & Planning, Delft University of Technology
{Y.Wang-14, G.Correia, B.vanArem}@tudelft.nl

H.J.P. (Harry) Timmermans
Section of Urban Systems & Real estate, Department of the Built Environment, Eindhoven University of Technology
h.j.p.timmermans@tue.nl

Erik de Romph
Sustainable Urban Mobility & Safety, TNO
erik.deromph@tno.nl

I. Introduction

In traditional travel demand models, socio-demographic and socio-economic information is used to segment population and thus better explain the heterogeneity of travel behavior. However, such information is generally absent in anonymous mobile phone data due to privacy reasons. This gap should be bridged if we aim to replace traditional survey data with mobile phone data for travel demand modelling.

A possibly effective way is to infer one's personal attributes from calling behavior [1]. This approach has two drawbacks:

- Today, mobile phones are hardly used for calls, making it more irrelevant to study calling behavior, and people spend more time on services provided by mobile internet such as mobile apps [2]. Thus, mobile internet usage behavior seems to have a better potential to reflect individuals' traits [3].
- Supplementary surveys need to be conducted to collect ground truth data (i.e., the personal attributes of a small subset of mobile phone users) for training the inference model. However, such surveys are difficult to implement.

Therefore, the question we want to answer in this research is: is it possible to use persons' mobile internet usage behavior to better understand their travel behavior without knowing or estimating their personal attributes? Thanks to the data provided by the Shanghai Unicom WO+ Open Data Application Contest[1], we are allowed, in this study, to explore the possibility of finding latent factors behind mobile internet usage behavior and use them to explain heterogeneous travel behavior. Specifically, in this research, we are concerned with explaining persons' activity patterns across several days.

A by-product of this approach is the opportunity to find the relationship between travel and telecommunication usage in the mobile internet era. This issue has been extensively debated between two sometimes opposing views: there are those who think that this is a substitute relationship and those who think that it is a complementary relationship [4], [5]. However, the context of these debates is mostly outdated, and few studies have empirically examined this relationship considering mobile internet services as an emerging usage of phones.

In summary, the significance of this study is two-fold: (1) it proposes a possibility of overcoming an existing limitation of using mobile phone data to model travel demand; and (2) it addresses the timely question about the relationship between travel and telecommunication usage by providing empirical evidence.

[1] https://www.kesci.com/woplus/ *(retrieved date: January 5th, 2017)*

II. Methodology

A. Analysis of mobile internet usage behaviour

In this study, we consider an individual's mobile internet usage behavior in terms of the frequencies of using different types of mobile internet services (e.g., news and e-shopping) through mobile apps or websites over a period of several days. This information is provided in two main indicators: (1) the frequency of using all mobile internet services, which reflects an individual's usage intensity; and (2) the frequency proportion of using each type of mobile internet service, reflecting the different lifestyles and interests. The latter one contains a number of dimensions, corresponding to the types of mobile internet services, some of which are likely to be highly correlated with each other. Therefore, an exploratory factor analysis is applied to find the latent factors underlying those dimensions. Both the intensity and the latent factors underlying the frequency proportions are regarded as the proxy variables for one's personal characteristics. This assumption is supported by evidence obtained in existing research [3], [6].

B. Analysis of multiday activity programs

We use the term "activity program" to refer to daily choices of several activity types, their frequency, sequence, start time and duration [7]. Existing studies show that one's activity programs are strongly dependent on one's personal attributes [8], [9]. In our study, we especially focus on activity types and durations. For each individual, we calculate the share of time used in each type of activity in each day. Using the longitudinal mobile phone data, we can identify the time use patterns of all sampled users over several days, which can then be clustered to distinguish the heterogeneity of the users.

There are mainly three challenges in this step of the methodology. Firstly, stay points need to be distinguished from pass-by points among all mobile phone traces. Extensive literature has discussed this problem as well as its case-by-case solutions, and we mainly follow the approach of Alexander et al. [10], [11]. Secondly, missing traces would cause a deviation on the revealed activity programs; hence, in this study, we only sample the users leaving rather complete traces over the study period to prevent such errors. Thirdly, it is difficult to identify activity types in mobile phone data. One possible solution is to infer the specific types based on the ground truth (e.g., the characteristics of a certain activity type revealed in survey data [12]). Another solution is to only distinguish between different activity types without knowing which specific types they are [8], [13]. In our case, by applying the arbitrary thresholds (i.e., the least duration and the lowest frequencies of visiting a place over a certain period) to the spatial-temporal traces, we detect one's home and workplace locations, and we assume that the traces of staying at the detected home locations should be classified into the home activity. The same applies to classifying the work activity, and the remaining traces are labeled as the non-work activities.

C. *Classifying the patterns of multiday activity programs*

The first step of the methodology aims at calculating for each individual the mobile internet usage intensity as well as the latent factors reflecting the different lifestyles and interests based on the mobile internet usage behavior. The second step aims at labelling each individual's multiday activity pattern. In the final step, a multinomial logit model is estimated to explain the influence of the latent factors and the usage intensity on different multiday activity patterns.

III. Data

Having removed the noise in the data, we have a sample of 373,577 mobile users. Each anonymous user ID corresponds to not only a list of the page view counts for different types of mobile apps and websites, but also a list of spatial-temporal records within the city of Shanghai hour by hour from 27th of December, 2015 to 6th of January, 2016, in terms of longitude, latitude and timestamp. According to the mobile provider, the coordinates of the records are estimated with an accuracy of about 300 meters.

The data preprocessing steps are explained as follows. Firstly, we merged the page view counts of mobile apps and websites for the same type, thus producing a total of 14 types of mobile internet services: finance (excluding stock), stock, e-shopping, news, housing, tourism, sports, car, social network, entertainment, online education, job seeking, game and food. Secondly, we calculated the total page view counts for all mobile internet services for each user. For each of those who had ever used any mobile internet service, we constructed a vector containing the page view counts for the 14 types of mobile internet services. To calculate a fair indicator of the frequency proportion, the counts of each dimension were normalized over all users, and then the counts of each user were normalized over all dimensions. Thirdly, we extracted the stay points and labeled each of them with the activity type.

IV. Results

Based on the eigenvalue criterion, there are 4 latent factors that explain the mobile internet usage. The results of factor analysis using a varimax rotation are shown in Table 1. Factor loads > 0.30 are tagged with asterisks. Factor 1 seems to reflect the lifestyle of men with things such as stock and sports. Most of them prefer not to spend time on e-shopping and games. Factor 2 can best describe the lifestyle of young professionals who like travelling, social networking and online learning. Also, it appears to show that they care about finance and probably manage their money using online platforms. Factor 3 seems to represent the taste of more senior people who like car-related hobbies and reading news. Factor 4 is probably representing the relatively young feminine taste since the loads on e-shopping, online education and game are high, while the loads on finance and stock are negative.

We divided the mobile users into two groups: the detected commuters and the detected non-commuters, and we clustered their multiday activity patterns using the k-means clustering algorithm. The number of clusters can be determined by using the DB-index, which can indicate the compactness of a clustering solution [14]. The indices for both groups are shown in Fig. 1. Considering the trade-off between the DB-index and the interpretability (not too few and not too many clusters), we decided to have 4 clusters within each group. The average patterns of the clusters are presented in Fig. 2 and Fig. 3. Before interpreting the results, one should note that the first day during the study period is Sunday, and the sixth, seventh and eighth days are the public holiday for celebrating the western new year. Within the commuter group, people belonging to cluster 1 and cluster 3 seem to be those who seldom performed out-of-home activities except going to workplaces during the weekdays. The only difference between the two clusters is that the people from cluster 3 tend to have longer working times. People from cluster 2 were mostly active urban travelers who performed non-work activities from time to time. It is worth pointing out that the figure shows the average patterns of a cluster of people. Thus it does not necessarily mean that they performed non-work activities every day. The same applies to cluster 4: the people in this cluster did not necessarily work on every day of the weekend and the holiday. However, one who worked on any non-working day is more likely to be assigned to this cluster. Within the non-commuter group, people belonging to cluster 1 were different from the rest because they stayed at home for most of the time. People from cluster 2 generally spent even more time on out-of-home activities than the ones from cluster 3. Cluster 4 represents the non-commuters who usually stayed at home but hung out on the weekend and/or the holiday.

TABLE I. Factor Loads

	Factor 1	Factor 2	Factor 3	Factor 4
Finance (excl. stock)	0.58*	0.45*	0.01	-0.16
Stock	0.77*	0.01	-0.03	-0.18
E-shopping	-0.09	0.14	-0.07	0.40*
News	0.00	0.08	0.51*	-0.06
Housing	0.09	0.15	0.27	0.19
Tourism	0.02	0.41*	0.13	0.10
Sports	0.30*	0.01	0.14	0.10
Car	0.09	0.15	0.35*	0.05
Social network	0.04	0.48*	0.13	0.09
Entertainment	0.08	0.19	0.27	0.19
Online education	0.06	0.34*	0.25	0.40*
Job seeking	0.03	0.11	0.06	0.09
Game	-0.10	0.05	0.18	0.22
Food	0.01	0.10	0.03	0.05

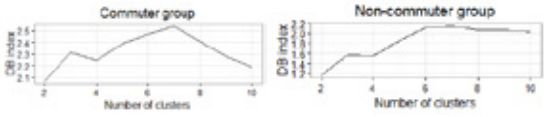


Fig. 1. DB indices for commuter group and non-commuter group

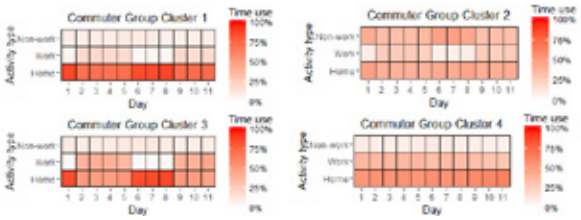


Fig. 2. Average activity patterns of 4 clusters for the commuters

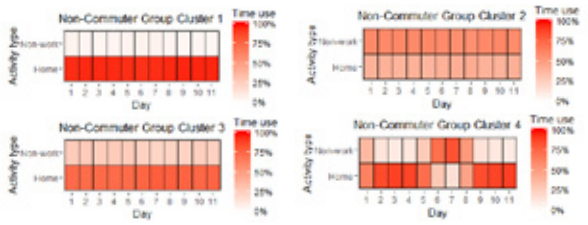


Fig. 3. Average activity patterns of 4 clusters for the non-commuters

For each group of users who had used any of the mobile internet service at least once during the study period, we estimated a multinomial logit model to classify their multiday activity patterns. The results are presented in Table 2. Firstly, it can be observed that the usage intensity is significantly related to the multiday activity patterns. People who used mobile internet services more intensely seem to be more engaged in non-work activities, and the non-commuters who preferred to always stay at home had remarkably less mobile internet usage. Secondly, the latent factors reflecting different lifestyles and interests are significantly related to some patterns. A higher value of factor 1 is more associated with the less interest in non-work activities. Comparing cluster 1 to cluster 3 within the commuter group, we find that people who worked longer are more likely to be younger, given their relatively higher values of factor 2 and factor 4. It is also interesting to see that the ones who like hanging out on non-working days are more likely to be characterized by a higher value of factor 4 representing relatively more young feminine taste. These results are mostly consistent with the prior knowledge about the relationship between one's activity patterns and the lifestyles represented by the corresponding factors.

TABLE II. RESULTS OF MULTINOMIAL LOGIT MODEL

Commuter group						
	Cluster 2		**Cluster 3**		**Cluster 4**	
	Param.	t-value	Param.	t-value	Param.	t-value
Intercept	-1.231	-18.9*	-0.811	-14.0*	-1.067	-17.3*
Intensity	0.109	4.7*	0.093	4.5*	0.083	3.8*
Factor 1	-0.104	-5.4*	0.006	0.36	-0.205	-10.4*
Factor 2	-0.144	-6.0*	0.149	7.3*	-0.338	-13.8*
Factor 3	0.122	5.0*	-0.087	-4.0*	0.071	2.94*
Factor 4	0.073	2.6*	0.194	7.9*	-0.028	-1.05
Non-commuter group						
	Cluster 2		**Cluster 3**		**Cluster 4**	
	Param.	t-value	Param.	t-value	Param.	t-value
Intercept	-3.540	-43.9*	-1.676	-38.3*	-3.499	-39.1*
Intensity	0.370	13.1*	0.201	12.8*	0.298	9.4*
Factor 1	-0.167	-6.0*	-0.031	-2.2*	-0.064	-2.1*
Factor 2	-0.346	-10.2*	-0.106	-5.8*	-0.023	-0.7
Factor 3	0.291	9.3*	0.191	10.8*	0.042	1.2
Factor 4	0.006	0.17	-0.015	-0.76	0.549	13.7*

V. CONCLUSION

Our approach was able to capture the latent factors behind the mobile internet usage behavior and cluster multiday activity patterns of users based on their spatial-temporal mobile phone traces. Moreover, we used a multinomial logit model to find the association between mobile internet usage behavior and multiday activity patterns. Two results stand out. Firstly, a complementary relationship was generally found between travel and mobile internet usage in our case because the usage intensity was observed to be significantly related to the preference for out-of-home activities. Secondly, the latent factors reflecting different lifestyles can explain some heterogeneity of activity patterns.

ACKNOWLEDGMENT

We would like to express our gratitude to the Shanghai Unicom WO+ Open Data Application Contest for making the data available for this research. Thanks go also to the TRAIL research school and the Dutch Organization for Scientific Research (NWO) for sponsoring the first author for his PhD study.

REFERENCES

[1] A. Arai, A. Witayangkurn, H. Kanasugi, T. Horanont, X. Shao, and R. Shibasaki, "Understanding User Attributes from Calling Behavior," in *Proceedings of the 12th International Conference on Advances in Mobile Computing and Multimedia - MoMM '14*, 2014, pp. 95–104.

[2] S. Richmond, "Smartphones hardly used for calls," 2012. [Online]. Available: http://www.telegraph.co.uk/technology/mobile-phones/9365085/Smartphones-hardly-used-for-calls.html. [Accessed: 03-Jan-2017].

[3] S. Seneviratne, A. Seneviratne, P. Mohapatra, and A. Mahanti, "Predicting user traits from a snapshot of apps installed on a smartphone," *ACM SIGMOBILE Mob. Comput. Commun. Rev.*, vol. 18, no. 2, pp. 1–8, Jun. 2014.

[4] P. L. Mokhtarian, "Telecommunications and Travel. The Case for Complementarity," *Journal of Industry Ecology*, vol. 6, no. 2. Wiley Online Library, pp. 43–57, 2003.

[5] I. Salomon, "Technological change and social forecasting: the case of telecommuting as a travel substitute," *Transp. Res. Part C Emerg. Technol.*, vol. 6, no. 1, pp. 17–45, 1998.

[6] G. Chittaranjan, J. Blom, and D. Gatica-Perez, "Who's Who with Big-Five: Analyzing and Classifying Personality Traits with Smartphones," in *2011 15th Annual International Symposium on Wearable Computers*, 2011, pp. 29–36.

[7] T. A. Arentze and H. J. . Timmermans, "A learning-based transportation oriented simulation system," *Transp. Res. Part B Methodol.*, vol. 38, no. 7, pp. 613–633, 2004.

[8] G. Goulet Langlois, H. N. Koutsopoulos, and J. Zhao, "Inferring patterns in the multi-week activity sequences of public transport users," *Transp. Res. Part C Emerg. Technol.*, vol. 64, pp. 1–16, 2016.

[9] K. Müller and K. Axhausen, "Using Survey Calibration and Statistical Matching to Reweight and Distribute Activity Schedules," *Transp. Res. Rec. J. Transp. Res. Board*, vol. 2429, pp. 157–167, Dec. 2014.

[10] L. Alexander, S. Jiang, M. Murga, and M. C. González, "Origin – destination trips by purpose and time of day inferred from mobile phone data," *Transp. Res. Part C*, vol. 58, pp. 240–250, 2015.

[11] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma, "Mining interesting locations and travel sequences from GPS trajectories," *Proceedings of the 18th international conference on World wide web - WWW '09*, no. 49. ACM, p. 791, 2009.

[12] F. Liu, D. Janssens, G. Wets, and M. Cools, "Annotating mobile phone location data with activity purposes using machine learning algorithms," *Expert Syst. Appl.*, vol. 40, no. 8, pp. 3299–3311, 2013.

[13] A. Pozdnukhov, "Demand Forecasting and Activity-based Mobility Modeling from Cell Phone Data," *UCCONNECT Final Reports*, 2016.

[14] D. L. Davies and D. W. Bouldin, "A Cluster Separation Measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979.

The Effect of Pokémon Go on The Pulse of the City

Eduardo Graells-Garrido*, Leo Ferres, and Loreto Bravo

Data Science Institute, Faculty of Engineering, Universidad del Desarrollo; Telefónica R&D, Santiago, Chile.

1 Introduction

Pokémon Go is a location-based augmented reality game that became an instant world-wide hit when it launched on August 3rd, 2016. At one point, people of all ages seemed to be caught in a frenzy of walking out of their way to find the next pocket monster. Due to the game's emphasis on the exploration of the physical world around the player, the general perception was that the game could go as far as making whole populations change their mobility patterns. This would imply a breakthrough, since changing mobility behavior at this scale is notoriously difficult to do, with well-known exceptions like catastrophes, some sports events, and the such. In this work we test the hypothesis whether the game had indeed a significant impact on the pulse of a city.

Our analysis relies on a set of mobile communications records (CDRs) from Telefónica Movistar, the largest telecommunications company in Chile, with a market share of 33% as of 2016. CDRs originally include logs of phone calls and SMS, while ours also include data-type network events (*e. g.*, Web browsing, application usage, etc.), aggregated by context-dependent amount of downloaded information [1]. Using these data we follow a natural experiment approach whereby we evaluate floating populations at two specific intervals of time: the seven days before and the seven days after the launch of Pokémon Go. We geographically circumscribe our data to Santiago, the capital of Chile and its most populated city. We select a specific number of devices to ensure that we analyze floating population patterns of active users that live in the city.

Our results show that Pokémon Go increased the number of people on the streets, after accounting for covariates like land use [4], availability of points of interest and daily patterns. However, this increment is not present during the entire day. There are two strong effects: one at lunch time and other at night. We discuss the relationship between mobility and these findings, as well as the implications for the city.

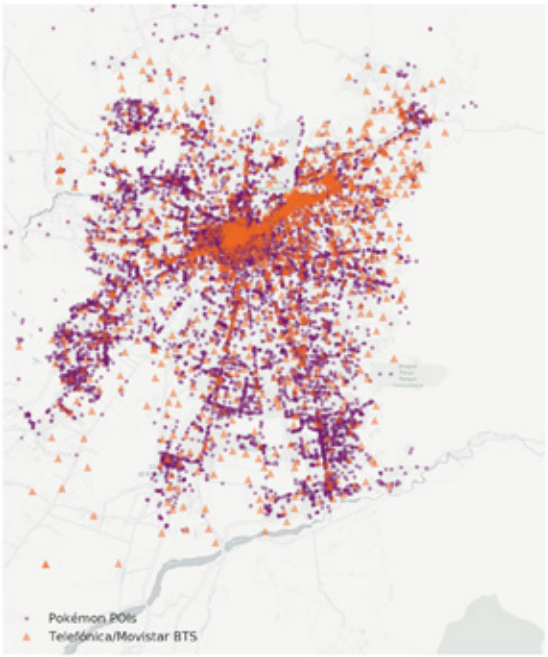


Figure 1: Telefónica cell tower stations and Pokémon points of interests in Santiago. Background tile data was provided by ©OpenStreetMap contributors and ©CartoDB.

2 Datasets

We study an anonymized CDR dataset of Internet network events from Telefónica Chile. The dataset contains records from the seven days prior to the launch of Pokémon Go (from July 27th to August 2rd) and the seven days after (from August 4th to

*egraells@udd.cl.

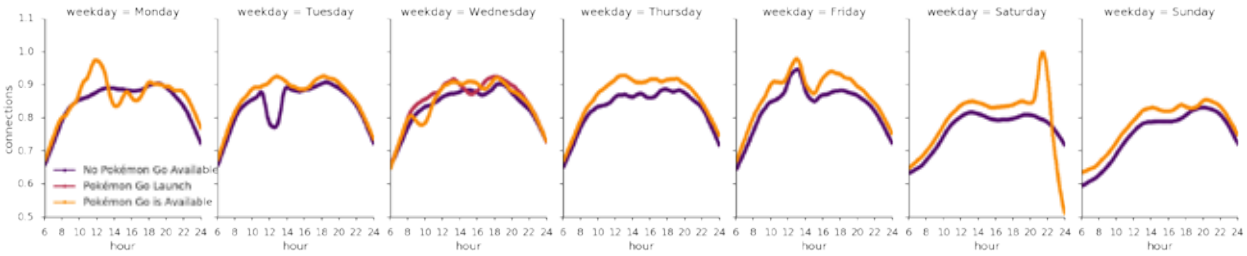


Figure 2: Amount of connected devices to the network for all days in the dataset. Note that the distributions are normalized by dividing the actual number of connections by the global maximum value. This is done to avoid publishing sensitive business information and for anonymization purposes.

August 10th). Since we focus on Santiago, we only analyze the records from the 1,464 cell towers within the urban area (see Figure 1). An assumption is that network events are triggered when devices are on the street. For instance, when comparing two different days at a specific time, an increase in the number of connected devices would mean that there are more people on the street. We do not analyze the records from the entire customer population in Santiago. We apply the following filtering procedure: first, we filter out those records that do not fall within the limits of the ODS explained above and also those with a timestamp outside the range between 6:00am and 11:59pm. In addition, to be considered, mobile devices must have been active every day under study. This is needed because a device that does not show regular events may belong to a tourist or someone who is not from the city. Finally, only devices that downloaded more than 2.5 MiB and less than 500 MiB per day are included, as that indicates either inactivity or an unusual activity for a human (*i. e.*, the device could be running an automated process). After these filters, the dataset comprises records from 142,988 devices. In other words, every measure is taken to ensure that events are triggered by humans.

Additionally, we use two external datasets: a travel survey from Santiago, held in the year 2012, and a crowdsourced list of *Pokéstops*, *i. e.*, points of interests where players could check-in and obtains items. As Figure 1 shows, these POIs are scattered around the city, having even more coverage than cell towers.

3 Approach

Our first step is to smooth the number of connected devices at each cell tower, according to several 1-minute snapshots of the tower network. A snapshot is the status of the cell phone network in a given time-window [6]. Figure 2 shows the city-level aggregated number of connected devices, having three categories of days: *before*, *during*, and *after* the launch of Pokémon Go. One can see that, although the curves tend to have similar shapes, after the launch of the game the number of connected devices is often greater. The patterns are stable across most days, with the means for connections generally higher when Pokémon Go was available. This means, intuitively, that there were more people connected to the network, presumably playing the game. Two rather surprising effects are found in Mondays between 10am and 12pm when Pokémon Go was available, and Tuesdays at about 12pm when it was not yet available. In the first case, we hypothesize that since it was the first Monday after the launch of the game, people were trying it out. In the second case, there does not seem to be any explanation for the sudden drop of connections before the launch of the game. This might be due to general network outages. The curve for when Pokémon Go was available in the same time period behaves as expected.

Then, we aggregate the device counts from each snapshot at the traffic analysis zone level, as defined in the travel survey. We do so by assigning tower events to the zones that contain the respective towers. These aggregated counts define a set of observations that we evaluate in a regression model, taking into account covariates that allows us to isolate and quantify the Pokémon Go effect. Thus, for every minute within 6am and midnight, we perform a Negative Binomial Regression using the following model:

$$\log E[X(t)] = \log a + \beta_0 + \beta_1 \text{PoGo} + \beta_2 \text{DayOfWeek} + \beta_3 \text{LandUse} + \beta_4 \text{PokéPoints},$$

where $E[X(t)]$ is the expected value of the number of active devices within a zone at time t. The PoGo factor represents the availability of the game; DayOfWeek and LandUse account for the fluctuations in population on different days according to land use; PokéPoints represents the number of Pokémon points

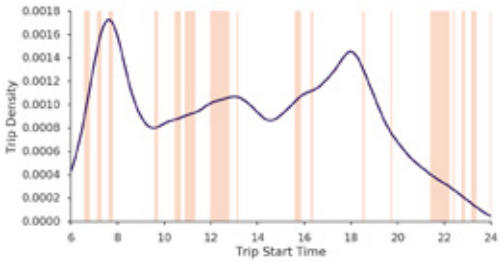


Figure 3: Trip start time from the Santiago travel survey. Each rectangle shows when the Pokémon Go factor is significant.

of interest within a given zone; and we specify the surface area a of each zone as exposure in the model.

The NB regression allows the following interpretation: the β coefficient assigned to a factor represents the difference of the logarithm of expected counts in a zone at time t, if all other factors were held equal. The exponential of this coefficient is defined as Incidence Rate Ratio, $IRR_\beta(t) = e^{\beta(t)}$. Then, we build a time-series of $IRR_\beta(t)$ values for each factor. By inspecting these time-series we determine when, in terms of time-windows within a day, there is a significant effect of each factor.

4 Results and Discussion

Figure 3 shows the time-windows within a day where the Pokémon Go factor is significant. There are two windows with prominent lengths: from 11:58 to 12:46 (with max $IRR = 1.138$), and from 21:24 to 22:12 (with max $IRR = 1.096$). The other time-windows exhibit shorter durations as well as smaller IRRs. This means that, having all other factors held equal, the availability of the game increased the amount of people connected to mobile towers in the city by 13.8% at lunch time and 9.6% at night. The figure also displays the trip start time distribution from the travel survey. One can see that before 6pm, which marks the end of laboral hours, the Pokémon Go effect tends to occur at moments where people starts to or is commuting, either because they are going to work or because they are taking a break. Instead of modifying their routines to play, people took advantage of commuting time and breaks during the day to play. As such, players tend to be nearby their work/study places, which are concentrated on downtown. Conversely, during weekends, where daily routines are not as strong and there is more flexibility, the relation with commuting is not present: the saturday night effect shown in Figure 2 is not present on the trip distribution. This implies that people played the game in places near their homes after having dinner.

These results show the potential of location-based augmented reality games to improve city life. For instance, the presence of pedestrians on the street at all times of the day is one of the four conditions needed for lively and safe cities, as proposed by the urbanist Jane Jacobs [5]. However, it is not clear how to motivate people to walk on the streets – just because there is a new park it does not mean that people will use it, specially on cities with biased perceptions of safety within the public space [2]. Pokémon Go showed that it is possible to motivate people to go out.

Even though the usage of mobile datasets to study urban theories is not new [3], such analysis is ex-post, and thus, it does not help understand causality. A more granular approach would be to perform natural experiments like ours. To the extent of our knowledge, this is the first natural experiment performed using mobile records of data-type. Our method makes possible to measure the effect not only of long-term interventions, but also short-term ones, opening a path to quantify how much specific actions help to improve quality of life in the city.

References

[1] Francesco Calabrese, Laura Ferrari, and Vincent D Blondel. "Urban sensing using mobile phone network data: a survey of research". In: *ACM Computing Surveys (CSUR)* 47.2 (2015), p. 25.

[2] Lucia Dammert and Mary Fran T Malone. "Fear of crime or fear of life? Public insecurities in Chile". In: *Bulletin of Latin American Research* 22.1 (2003), pp. 79–101.

[3] Marco De Nadai, Jacopo Staiano, Roberto Larcher, Nicu Sebe, Daniele Quercia, and Bruno Lepri. "The death and life of great Italian cities: a mobile phone data perspective". In: *Proceedings of the 25th International Conference on World Wide Web*. 2016, pp. 413–423.

[4] Eduardo Graells-Garrido, Oscar Peredo, and José García. "Sensing urban patterns with antenna mappings: the case of Santiago, Chile". In: *Sensors* 16.7 (2016), p. 1098.

[5] Jane Jacobs. *The death and life of great American cities*. Vintage, 1961.

[6] Diala Naboulsi, Razvan Stanica, and Marco Fiore. "Classifying call profiles in large-scale mobile traffic datasets". In: *IEEE INFOCOM 2014-IEEE Conference on Computer Communications*. IEEE. 2014, pp. 1806–1814.

Traffic Analysis of Railways using Call Detail Records

Hiroki Ishizuka
KDDI Research, Inc.
2-1-15 Ohara, Fujimino
Saitama 356-8502, Japan
hk-ishizuka@kddi-research.jp

Nao Kobayashi
KDDI Research, Inc.
2-1-15 Ohara, Fujimino
Saitama 356-8502, Japan
no-kobayashi@kddi-research.jp

Mori Kurokawa
KDDI Research, Inc.
2-1-15 Ohara, Fujimino
Saitama 356-8502, Japan
mo-kurokawa@kddi-research.jp

Chihiro Ono
KDDI Research, Inc.
2-1-15 Ohara, Fujimino
Saitama 356-8502, Japan
ci-ono@kddi-research.jp

Takahiro Hara
Osaka University
1-5 Yamadaoka, Suita Osaka
565-0871, Japan
hara@ist.osaka-u.ac.jp

ABSTRACT

Recently, increasing car traffic congestion especially at commuting time has degraded the efficiency of a transportation infrastructure. To overcome this situation, railway services that can carry a massive amount of people at once have been expected to play an important role. In order to reduce car congestion by improving public transports, we should optimize complex urban railway networks, and as this first step, we need to understand behaviors of railway commuters in practical period. For this aim, understanding human behaviors using call detail records (CDRs) that automatically recorded the location of the user and without the user's intentional operation have attracted attention. Although large-scale data are accumulated from CDRs through everyday use of mobile phones, the spatial and temporal resolution of CDRs is lower than that of existing positioning techs. To cope with this problem, conventional works have focused on the cellular handoff patterns as a substitute for the sparse location of CDRs. To handle the complex urban railway networks, the works does not work enough, because they have to learn the patterns for each route manually. To solve this issue, we propose the system that self-learns cellular handoff patterns for all urban rail lines using a huge amount of anonymous sparse location of CDRs. In this paper, We evaluate the three kinds of self-learning methods which include the voronoi diagram based approach, the mahalanobis distance based approach and the static range based approach using CDRs from a millions of subscribers. As the result, the accuracy of classifying the rail lines from the individual cellular handoff pattern achieved 81% against accurate 7600 GPS trajectories. In addition,the correlation about the traffic flow of urban railway networks between our output and the latest census data was 0.768. As the contribution of this paper, our proposed system indicates the possibility of understanding the traffic flow of the urban railway networks more often using only CDRs as substitute for the census.

Categories and Subject Descriptors

I.5.2 [**Pattern Recognition**]: Design Methodology—*classifier design and evaluation*

General Terms

Algorithms, Design, Experimentation

Keywords

GIS, CDRs, Transportation Estimation, Localization, Cellular Handoff Patterns, Self-learning

1. INTRODUCTION

Increasing traffic congestion at commute times has significantly degraded the efficiency of the transportation infrastructure and increased travel time, air pollution, and fuel consumption. As a result, urban areas are faced with the necessity of improving public transport to reduce car use. Urban areas (New York, London, Paris, Beijing, and Tokyo) have multimodal public transport systems consisting of railways and buses or light rail. In particular, railway services can carry massive amount of people at once are expected to be effective during commute times[16].

Tokyo is well known as one of the railway-oriented cities where the huge traffic demand generated from the megacity is well supported by a sophisticated urban railway system[7]. The results of the 2008 Person Trip Survey show that rail's modal share was 30%. Approximately 8 million commuting workers and students use railways on a daily basis in the Tokyo metropolitan area. Even now, railway networks surrounding Tokyo have grown from year to year[12]. In particular, the 2020 Summer Olympic and Paralympic Games in Tokyo have boosted massive investment in new railway infrastructures in Tokyo. To adjust and rebuild the huge complex railway networks, understanding the behavior of railway commuters in practical period is necessary.

From another point of view, the Tokyo metropolitan area experienced a shaker in the upper 5 during the Great East Japan Earthquake. As a result, almost all railways temporarily stopped their service. Since the earthquake occurred on a weekday afternoon, a large number of commuting workers and students were in their offices and schools in Tokyo. The stoppage of railways made it impossible for such people to return home[20][17]. During the Great East Japan Earthquake, there were 3.5 million as commuters stranded in the Tokyo metropolitan area. For the temporary evacuation of such stranded commuters, the government should

estimate a number of potential candidates for the stranded commuters by analyzing railway commuters in daily life.

In order to roughly understand the behavior of railway commuters, we can use the results of a metropolitan transportation census provided by Japan's Ministry of Land, Infrastructure, Transport, and Tourism's (MLIT) and local governments. However, the census has been conducted every five years. , e.g., the latest result was summarized in 2010. Because the environment of a transport infrastructure changes every year, understanding the behavior of railway commuters in detail is difficult using such old data. As another approach to find the behavior of railway commuters, we may use ticket sales from railway operators, however, railway operators have not provided such data to the public because of privacy issues.

On the other hand, mobile phones have become a key device for pervasive computing with users carrying them at almost all times. The ubiquity of these platforms has transformed mobile phones into one of the main surveyors of human behavior. However, conventional work using sensors on a mobile phone might not be suitable for large-scale, sustainable data collection in the analysis of human behavior. For example, participants (users) must install to analyze user activities. Moreover, the application wastes energy from the mobile phone battery with the unintended sensing operations. Though participatory sensing is one of the attractive solutions, collaborative users have been limited.

In order to cope with the problem, understanding human behaviors using call detail records (CDRs) that automatically recorded the location of the user and without the user's intentional operation have attracted attention. In fact, every time a subscriber makes or receives a phone call, an short message service(SMS), an multimedia messaging service(MMS), or data communication regarding the interaction and the location of the user (in the location of the base station used for the communication) is logged as CDR for billing purposes. CDRs have automatically and routinely been recorded without the user's intentional operation. Even if the user don't operate the mobile phone intentionally, some social networking applications on the mobile phone periodically are sending keep alive messages[4][1] in several minutes. Although large-scale data are accumulated from CDRs through everyday use of mobile phones, the spatial and temporal resolution of CDRs is lower than that of existing positioning technologies.

To overcome the weak points of CDRs, some existing works have focused on the cellular handoff patterns as a substitute for the sparse location of CDRs. Characteristics of the handoff pattern when people get on the same rail line are likely to be similar. Thus, the handoff pattern of base stations can be useful data to analyze railway commuters. To manage a large numbers of rail lines, however, the the conventional works are not suitable at the point of scalability, because the works have to learn the handoff patterns for each route manually. Moreover, the works don't adapt to be rebuilt the correct cellular handoff pattern including the modification when a new rail line or a new base station has been laid after once building the pattern.

To solve the issue, we propose the system that enable to maintain a large number of the cellular handoff patterns automatically. Our proposed system self-learns all cellular handoff patterns in urban complex railway networks using a huge amount of anonymous sparse location data of CDRs. In this paper, We evaluate the three kinds of self-learning methods which include the voronoi diagram based approach, the mahalanobis distance based approach and the static range based approach using CDRs from a millions of subscribers. As the result, the accuracy of classifying the rail lines from the individual cellular handoff pattern achieved 81% against accurate 7600 GPS trajectories. In addition,the correlation about the traffic flow of urban railway networks between our output and the latest census data was 0.768. As the contribution of this paper, our proposed system indicates the possibility of understanding the traffic flow of the urban railway networks more often using only CDRs as substitute for the census.

The system design of the proposed system is presented in Section 3. We describe the data cleaning process, the self-learning process and the classifying process in Section 4, 5 and 6, respectively. In Section 7, we represent the datasets that is both of CDRs and GPS data as the ground-truth data for our experiments. Then, we validate the accuracy of our system and the correlation with the latest census data in Section 8. Finally, we conclude our work in Section 9.

2. RELATED WORKS

Every mobile phone leaves location trajectories as CDRs while interacting with its infrastructure. Massive numbers of CDRs from mobile phone users have been used in a variety of urban-related applications, including understanding mobility patterns and models[6, 10, 9, 11] the use of urban spaces [21], travel demand during specific events [5], social network structure [19], and geographical distribution of mobile communications [13]. Thus, CDRs are now recognized as one of main data sources indicating user location trajectories.

Many conventional studies have tried to estimate the transportation mode as human movement behavior. As mentioned by Stenneth et al. [24], trying to classify public transportation from accelerometer data has limitations with regard to accuracy. Given that result and the easy availability of GPS on mobile phones, researchers can use location and GIS information. Hemminki et al. [8] discussed an approach using only accelerometers for distinguishing different modalities while testing the system across a few cities. In contrast to other work using accelerometers only, they obtained impressive results of around 80% recall and precision since they focused on periods of vehicle acceleration and deceleration. However, they need to keep sensing accelerometers while detecting the mode of transit without considering energy consumption. In addition, the method required the installation of a special sensing application on the user's mobile phone. Rahul et al.[22] used several sensors on a smartphone and GIS information. Their contributions were that they reduced the learning time of sensor data using General Transit Feed Specification (GTFS), which described the schedule for public transit. The accuracy of the proposed system was around 85%. However, they could not deploy the unit in other countries, including Japan, since opened GTFS was not so prevalent thourghout world.

In a parallel effort to our approach, Thiagarajan et al.[25] proposed CTrack, a system for trajectory mapping using base station fingerprints. CTrack was able to match the stream of new GSM fingerprints to road segments with a median accuracy of 75%. CTrack could also use information from an accelerometer and a compass to improve accuracy.

Becker et al. [3, 2] showed the practical capability of deriving the trajectories using handoff patterns of CDRs. The paper validated the cellular handoff patterns as relatively stable across different routes, speeds, directions, phone models, and weather conditions. Zhou et al. [28] used an accelerometer, audio, and the base station sequences to identify whether the user was on public transit or a car, which mostly worked for bus detection. The approach was power efficient as it used the base station sequences rather than GPS, however, training was required for all routes that the person took making deployment harder in new places or for new routes.

In contrast with conventional work, our approach has not required specific data gathering before creating cellular handoff patterns. We propose a self-learning algorithm of a cellular handoff pattern for a rail line using massive numbers of CDRs and the geo shape of the rail line.

3. SYSTEM OVERVIEW

Our proposed system enables the discovery of railway commuters throughout three steps. In the first step called the *Data Cleaning process*, we extract the moving data sequences from CDRs, because our system focuses on the commuters of railways. As the second step called the *Self-learning process*, we starts creating a database of self-learnt cellular handoff patterns for each rail line using the proposed three algorithms. At the final step called the *Classifying process*, we identify whether the sequences of the user's CDR were recorded when the user rode the specific line. the database of self-learnt cellular handoff patterns is used for the identification method of specific rail line that the user rode. Figure 1 shows the whole system design. We summarize the brief role of each process in the following.

1) ***Data cleaning Process***
Sequences of CDRs when a user is moving are only needed, since the aim of the proposed system is finding railway commuters. Thus, in this process, all CDRs are distinguished between *moving* and *staying*. We use only the CDRs of *moving* status in the following process.

2) ***Self-learning Process***
This process generates a training dataset using CDR sequences of *moving* status and the shape of the rail line. Using the training data set, a cellular handoff pattern for the rail line is built automatically into the proposed system. In this paper, we create cellular handoff patterns for 110 rail lines.

3) ***Classifying Process***
In this process, the rail line is discovered by collating all cellular handoff patterns. Similarities among the patterns of the user's CDRs and the patterns of all rail lines in a database are calculated using the Smith–Waterman algorithm.

4. DATA CLEANING PROCESS

Before starting the main processes, we should extract sequences of CDRs when a user is moving as the data cleaning process. Connection timestamps, base station IDs, and wireless connection information are recorded in the CDR. Therefore extracting user locations and associated timestamps from the CDR is viable. Our proposed method estimates user positions using CDRs when the user is driving

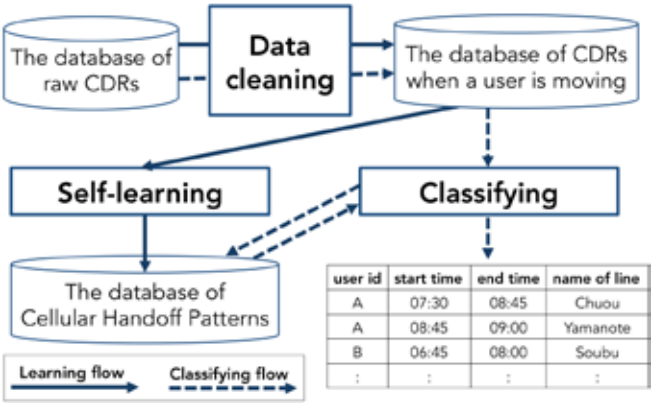


Figure 1: The overall architecture of our proposed system

or on a train, and as a first step, it is necessary to estimate the status of the user, namely, whether the user is currently moving or staying in one position.

Although multiple algorithms[14][27][26] for estimating user movement status and location and time information have been proposed, we propose a method in this paper that is better tailored for the use of CDRs. Our proposal consists the following three steps. First, we perform a temporal analysis in which, we record the timestamp and denote it as t_c for each record in the CDR. We then extract a set of locations whose associated timestamps fall within a predefined time window beginning from the specified time point. We denote the beginning time of this time window as t_0. Then in the spatial analysis, we calculate the distances between each location in the extracted location set and the location of the currently being processed record. We denote the movement status at t_c as s_{t_c} and we set s_{t_c} as staying if all the calculated distances fall below a predefined distance threshold, otherwise we set s_{t_c} as moving. Both the start time t_0 and the span of the time window are variable and depend on the estimated movement status from the precedent record. Specifically, if the estimated movement status is staying, we do not modify t_0 but we extend the span of the time window by $t_c - t_0$. If the estimated movement status is moving, we set t_0 to t_c and recover the span of the time window to its default value.

However, the ratio of moving and staying depends on the predefined distance threshold, which in turn can affect the precision and recall of our estimations. Specifically, setting a lower threshold can easily cause a rise in moving estimations which lowers estimation precision; on the other hand, a higher threshold causes a decreasing trend of moving estimations, which makes precision look better. Therefore, we define two distinct distance thresholds $(D1, D2 : D1 > D2)$ for movement status estimation, and the final decision is based on the aggregated judgment from these two thresholds.

Specifically, we extend our notation of s_{t_c} to $s_{t_c}^{d_i}$ to denote the movement status as determined by threshold d_i. As for the aggregated judgment, we compare pairs of $s_{t_c}^{d_1}$ and $s_{t_c}^{d_2}$ for each t_c, if they disagree with each other or are both *staying*, we set s_{t_c} to *staying*. Otherwise we set s_{t_c} to *moving*. Furthermore, if s_{t_c} is *moving*, we set s_{t_c}'s neigh-

bors $s_{t_c+/-1}$ to *moving* if $s^{d_2}_{t_c+/-1}$ is *moving*.

As is illustrated in figure 2, after the aggregated judgment, it is possible to estimate the ground-truth movement status more accurately.

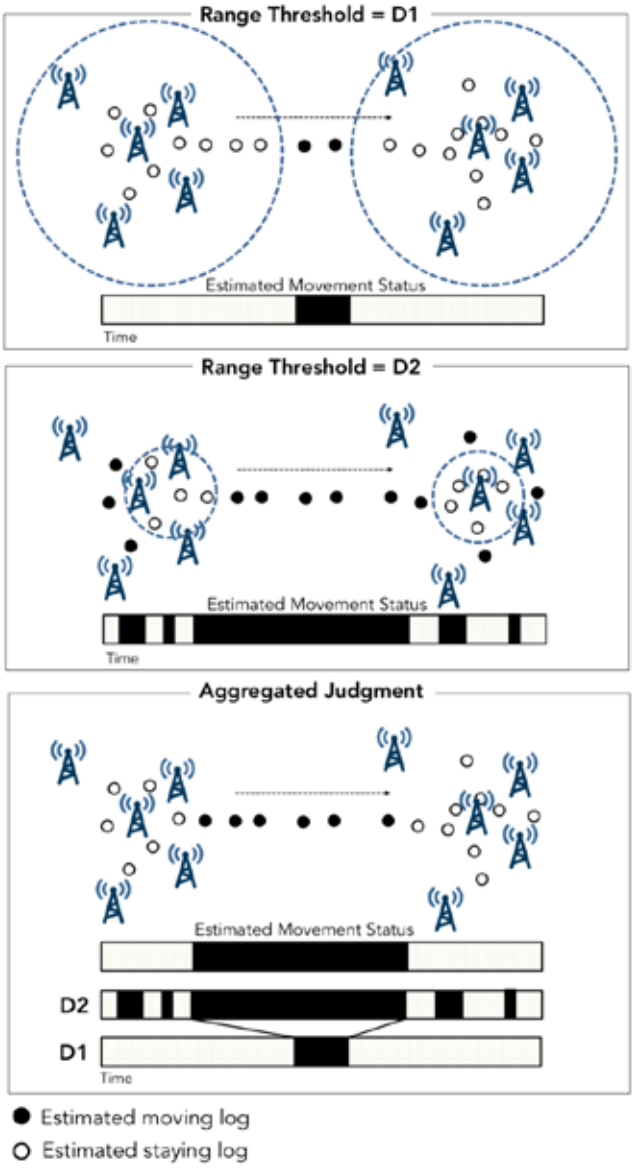


Figure 2: The process flow of the data cleaning

5. SELF-LEARNING PROCESS

We use a set(α) of location data$(p_1, p_2, ..., p_n) \in P_\alpha$ from a sequence of CDRs which were regarded as *moving* in the data cleaning process. Each p_i has three tuples like $(loc, time, tower\_id)$. A distance and a velocity between p_i and p_{i+1} are represented as $dist_i$ and $speed_i$, respectively. We also use a set of geographical points$(q_1, q_2, ..., q_n) \in Q_\beta$ of a rail line(β). $lineQ$ that is the shape of the rail line is a line that connected Q. A distance between p_i and $lineQ$ is shown as $dist\_lineQ_i$. Figure3 shows an election method of training data for a public transportation. A detail of the method is described in the next.

5.1 Creating Training Dataset

In order to learn the sequence of cellular handoffs for a specific rail line, we need a set of training CDR data that was recorded when the user rode on the line. To elect the training CDR data, our method processes a map-matching algorithm with the limitation of speed. The four conditions for electing training data are shown in the following. The first condition is that a data set for the *moving* state keeps recording at least $THLD_{cnt\_point}$ consecutive times, since such data are needed for learning *successive* cellular handoff patterns. Second, the average speed of movement using a vehicle should be over $THLD_{avg\_speed}$. As the third condition, we set the total distance of the movement($THLD_{total\_dist}$) to learn the length of a sequence of cellular handoffs. For the last condition, we set threshold $THLD_{dist\_lineQ}$ for $dist\_lineQ_i$, since we would like to elect only p_i points located along line Q. However, location data for CDRs are not accurate because they record the location using the result of triangulating by signals from just a few of cellular towers. A few $dist\_lineQ_i$ are over $THLD_{dist\_lineQ}$. Therefore, we set the ratio($THLD_{rate\_linematch}$) of p_i, which is located within $THLD_{dist\_lineQ}$ to all points of P.

1) $THLD_{cnt\_point} < \sum_{i=0}^{n} p_i$
2) $THLD_{avg\_speed} < \frac{\sum_{i=0}^{n-1} speed_i}{n-1}$
3) $THLD_{total\_dist} < \sum_{i=0}^{n-1} dist_i$
4) $THLD_{rate\_linematch} < \frac{\sum_{i=0}^{n} p_i | dist\_lineQ_i > THLD_{dist\_lineQ}}{\sum_{i=0}^{n} p_i}$

When the location data set α meets all conditions against the line β, α use as a training data for the line β. This process repeatedly executes against all location data sets regarded as *moving* to all 110 rail lines. After the process, all routes have sufficient training data to learn the sequence of cellular handoffs for each rail line.

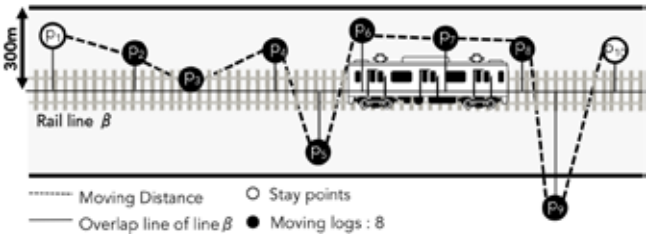


Figure 3: The example of creating the training dataset for rail line β

5.2 Building Cellular handoff patterns

We describe how to build a cellular handoff pattern using training datasets. The system assigns mobile base stations for each point composing a geographical polyline of a rail line using training datasets as mentioned in the previous section. In this paper, a cellular handoff pattern is defined as a geographical sequence of identified base stations representative of base stations assigned such points. We propose three kinds of assignment algorithms for the mobile base stations of each point. Assignment algorithms are expressed in detail in the following.

5.2.1 *Voronoi Diagram Based Assignment*

A p of a CDR has three tuples like $(loc, time, twoer\_id)$. A centroidC_i of the base station γ using locs of all p_j including the tower id of the base station γ as its $twoer\_id$ is calculated. Thus, the centroids of base stations covering all CDRs in the training dataset are calculated. We conduct Voronoi-based space partitioning using centroids for each base station. When a point composing a geographical polyline of a rail line is contained in a Voronoi region of a base station, the identification of the base station is learned as a piece of the cellular handoff pattern for the point. Note that the point does not have an assigned base station when any centroids from the base stations are not able to be found within 1000 m of it. An example of the Voronoi Diagram Based Assignment is shown in Figure 4(upper). Some conventional works[28] also elect this method.

5.2.2 *Mahalanobis Distance Based Assignment*

First, the method calculates the variance-covariance matrix using all centroid C calculated at creating the Voronoi region and locs of all p_j including the tower id of the base station γ as its $tower\_id$. The second step creates an ellipse region of equal probability using GDRs within the limit of the specific confidence interval(80%) in a Gaussian distribution of two variants derived by the variance-covariance matrix. When a point composing a geographical polyline of a rail line is contained in the ellipse region of a base station, the identification of the base station is learned as a piece of a cellular handoff pattern for the point. Though some ellipse regions overlap at the point, the method identifies the base station that has the nearest centroid to the point. In the same way as the method of the Voronoi diagram assignment, the point regards as it does not have an assigned base station when any centroids of base stations are not able to be found within 1000m around. Additionally, we set the maximum number of base stations for a point as three. An example result of Mahalanobis Distance Based Assignment is shown in Figure 4(middle).

5.2.3 *Static Range Based Assignment*

In the static range based assignment, we calculate the occupancies of base station IDs within the specified range (500 m in this paper) of the composition point of a geographical polyline for a specific transport. The base station id of the highest occupancy in a composition point adds to the base station alignment of the specific transport as the dominative base station. The following equation shows as extracting the base station IDs of the highest occupancy at composition point q_i.

$$tower\_id_{q_i} = argmax(\frac{\sum tower\_id_{q_i} | dist(p_i, q_i) < r}{\sum p_i})$$

As the first composition point in Figure 4(bottom), base station **A** is the dominative base station. For the second composition point, base station **B** is the dominative base station. Thus, we determine the cellular handoff alignment to choose the dominative base stations for each composition point. The figure 4(bottom) shows the example of the cellular handoff pattern in this method.

6. CLASSIFYING PROCESS

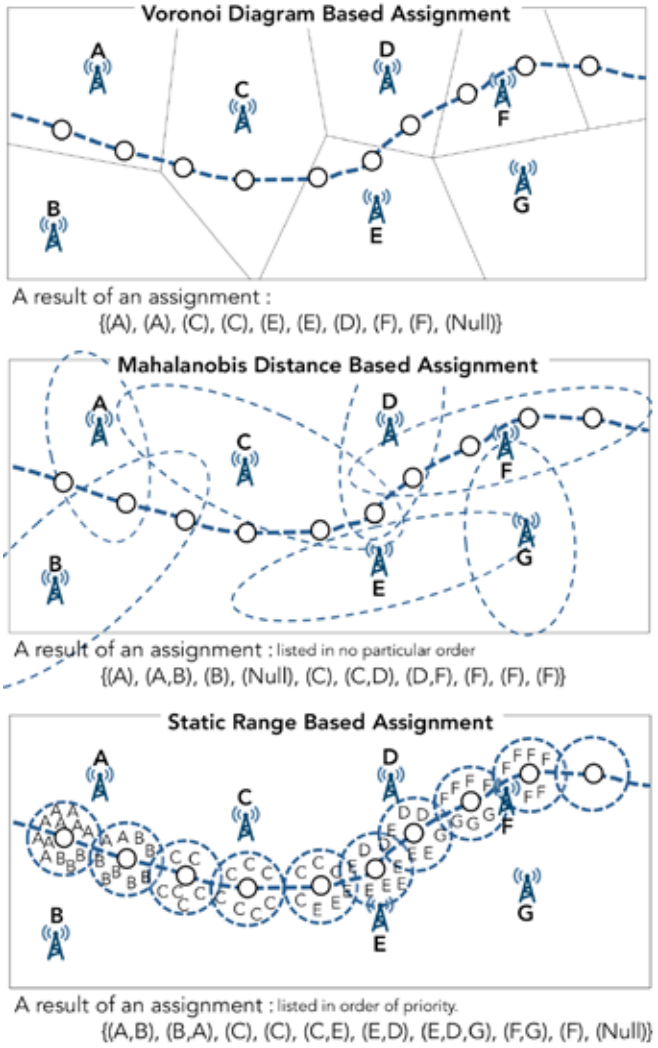


Figure 4: The cellular handoff patterns created by proposed three methods

By the classifying of a rail line that a user rode on, our system calculates the similarity between the transition series of base stations from the user's CDRs and learnt base station transition series of a rail line. By computing the similarity, the CDRs are associated with the nearest point of a rail line. The rail line with the highest similarity score is considered the rail line used by the rider. The calculation of the similarity is used a variant of the Smith-Waterman algorithm[23]. Smith-Waterman is a dynamic programming algorithm for local sequence alignment that has been widely used in bioscience (e.g., to determine similar regions between two nucleotide or protein sequences). In the scoring scheme of the Smith-Waterman algorithm, a positive value is assigned for a match and a negative or a lower value is assigned for mismatch. Once the scoring has started, it stops when a score is negative. Since, we want to compare the global sequences at both the base station sequence of the user's CDRs and a learnt base station sequence, the penalty is set as 0 when the compared two base stations are a mismatch. $\boldsymbol{U}(u_1, u_2, ..., u_m)$ and $\boldsymbol{L}(l_1, l_2, ..., l_n)$ are represented as a base station sequence of a user's CDRs and a learnt

base station sequence of a rail line, respectively. We formulate our scoring method in the following.

$$\boldsymbol{H}(i,j) = max \left\{ \begin{array}{l} 0 \\ \boldsymbol{H}(i-1,j-1) + s(u_i, l_j) \\ \boldsymbol{H}(i-1,j) \\ \boldsymbol{H}(i,j-1) \end{array} \right\}$$
$$1 \leq i \leq m, 0 \leq j \leq n$$
$$s(a,b) = \left\{ \begin{array}{l} 0, \ if \ b \in a \\ 1, \ otherwise \end{array} \right.$$

Using the above scoring matrix, the total score of the similarity between U and L is calculated in the following equation.

$$e(\boldsymbol{U}, \boldsymbol{L}) = \frac{max\boldsymbol{H}(i,j)}{min(m,n)}$$
$$0 \leq e(\boldsymbol{U}, \boldsymbol{L}) \leq 1$$

Since we want to calculate a top-k sequence matching against the user's CDRs, we use modified Smith-Waterman algorithm by [28]. The similarity between sequences is defined as the ratio of similar local sequences to all sequences of a user's CDRs.

7. DATA COLLECTION

For the sake of evaluations, we collect some kinds of data from subscribers of mobile communication services as provided from KDDI. Before using such data, we completely allow opt-in and obtain permission. In this section, we describe the details of the dataset for training and testing of our system.

7.1 Anonymous CDRs

We collected a large number of CDRs to create training data. For large-scale CDRs, our company as a mobile operator obtains clear opt-in permission for the use of CDRs from each individual. We want to learn the cellular handoff patterns for commuting rail lines from dormitory towns surrounding Tokyo to central Tokyo. Then, the area of data collection is the Tokyo metropolitan region with a horizontal width and vertical length of 230 km and 160 km, respectively. The number of users for the experiment is about 300,000. The period of data collection is two months (October and November 2014) and one month (April 2015). Each CDR has global location data that is calculated using the signal strengths of neighboring base stations using a simple triangulation algorithm. The average error of the location is around from 300 m to 500 m. All results of our proposed system are presented as aggregates. That is, no individual anonymous identifier was singled out for this study. By observing and reporting only on the aggregates, we protect the privacy of individuals.

On the other hand, we also need the geographical shapes of the rail lines to build cellular handoff patterns. We acquired all shape files of Japanese rail lines from the website[18] of the national land information division of the National Spatial Planning and Regional Policy Bureau in Japan. From the data on the rail lines, we selected 110 rail lines within the area matched to both the latest census and our data collection.

7.2 The Ground–truth Location Dataset

For evaluation of our classification results, we need the ground–truth location dataset against the location of CDRs. We invited data collection volunteers from our subscribers using an Android phone to allow us to use their CDRs and GPS data at the same time. The period of data collection is one month in April 2015. Acquiring location data from the GPS uses a Google Android application that can gather the data periodically. The area of data collection is the Tokyo metropolitan area as defined in the previous section. The total amount of data and the unique number of users are 10 million and 12175, respectively.

Moreover, we extract GPS data for riding on a railway to create accurate comparative data. The extraction method is similar to creating the training data set explained in the section on the self-learning process. Over 7600 cases of GPS location sequences moving along a rail line were discovered by the extracting process. To evaluate the accuracy of the extracted GPS location sequences, we carefully checked the randomly selected results visually, respectively. After the check process, we believed the datasets were sufficiently accurate for the comparative data as extracted GPS location sequences.

8. EVALUATION

To validate our proposed system, we conducted the two evaluations. First, we evaluated the accuracy of each self-learnt handoff pattern by comparing with our result and the ground-truth dataset. Second, we also investigated the correlation between the number of train commuters as a result of our system and the number of train commuters from the latest census data.

8.1 Rail Line Classification

In this paper, we proposed the three methods to create cellular handoff patterns. Then, we evaluated the accuracy of the rail line classification using the patterns which built by the three methods. If the ratio that two names of the line calculated from both the GPS sequence and our proposed method matched was higher, we consider that our handoff pattern is accurate enough. Note that the GPS and the CDRs were collected from the same user at the same time period. Table 1 shows the total accuracy for three methods. The number of trains represents the number of CDR sequences wherein a user rode on the rail line when the method was used. The number of TP (True-Positive) means the total number of accurate results. The results of the accuracy for each method indicated about from 80% to 81%. In addition, the result of using the Mahalanobis method was more accurate than using the other two methods.

Next, we compared the precision and recall of the results classifying rail lines among three methods. The precision of this paper is defined as the ratio of the number of results classified as a rail line δ using our proposed system to the actual number of GPS sequences of a user riding a rail line δ. And, FP (False-Positive) means the classified number as the rail line when the GPS sequence was not classified as the rail line. The precision and the recall were calculated using the following equations.

Table 1: Num of TPs and accuracy of three methods

	Voronoi	Mahalanobis	Static Range
Train users	461	810	315
True-Positive	369	653	258
Accuracy	80.0%	80.6%	81.9%

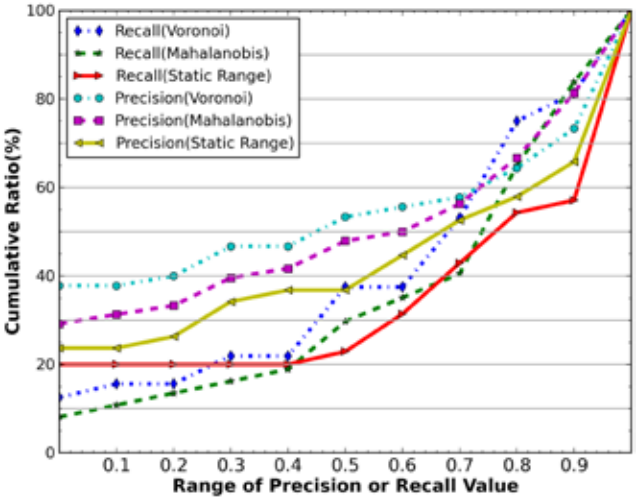


Figure 5: The cumulative frequency distributions of both the precision and the recall

$$Precision = \frac{TP}{(TP+FP)}$$

$$Recall = \frac{TP}{(TP+FN)}$$

Figure 5 shows the cumulative frequency distributions of both the precision and the recall. As the results, though there is no large difference among results from each three method, the precision values between 0.8 and 1.0 of the Mahalanobis method is better than the others. The precision values of the Voronoi and the Static Range method tend to indicate the higher ratio when a value is around 1.0. In conclusion, although the Mahalanobis method indicates better results than the others on the average, both the Voronoi and the Static Range methods have better classification results for the specific rail lines.

From the view of the results of the recall, the cumulative value of the Mahalanobis method is small when the values of recall are small. Furthermore, the cumulative value is large when the values of the recall are large. Thus, the Mahalanobis method is more accurate than the others. The Voronoi and the Static Range methods indicate a similar tendency for the result of the precision.

8.2 Correlation with the Latest Census Data

In the previous session, we mentioned that the accuracy of our classifying engine have been practical and feasible to extrac a line name that a individual user rode on. We also evaluate a macro performance by comparing a whole user's results of our system with the latest census. As the ground truth dataset, we utilize the survey of the number of

Figure 6: The mapping result of 110 rail lines selected for our experiment

metropolitan railway user of the 11th Metropolitan Travel Survey[15] that was conducted by Ministry of Land, Infrastructure, Transport and Tourism in 2010. 110 rail lines for this evaluation is elected as rail lines that are overlapped the dataset of the survey and the dataset of the geographical shapes of the rail lines mentioned at section 7.1. Notes that we excluded subways and super-express trains from evaluation dataset. The figure 6 shows a mapping result of 110 rail lines for this evaluation.

We create the mahalanobis distance based cellular handoff patterns for 110 lines using three hundred thousand user's CDRs during one month(Oct, 2014). The input data for classifying are also the same numbers of user's CDRs creating the pattern at Nov, 2014. The travel survey conducted from 16th to 18th Nov 2010. Therefore, we also use input data of same month to compare same condition. From the two points of view of the number of railway commuters at various time zones and the number of railway commuters at various rail lines, we compare results of our system with the latest census data.

8.2.1 *Hourly Railway commuters*

As the results of our system, we shows the average hourly number of railway commuters on weekdays in Figure 7. The figure is shown that our system can analyze a transition of the number of hourly train commuters on all rail lines. For verifying suitability of the number of hourly train commuters, we compare the results and the the latest census data. We set a comparing time as a core active time for people from am 6 to pm 11. Before the comparison, we normalize both our results and the latest census data, since the total number of user in two dataset is difference. the result of comparing our system results and the latest census data is shown in Figure 8. There are two peaks at am 8 and pm 6 in both results of the latest census data and our output. And trend of two transitions seem to be similar. The result of the latest census data is higher at the peaks. In contrast, the result of our system is higher in time between peaks.

8.2.2 *Railway commuters on various rail lines*

We analyze the correlation between our output and the

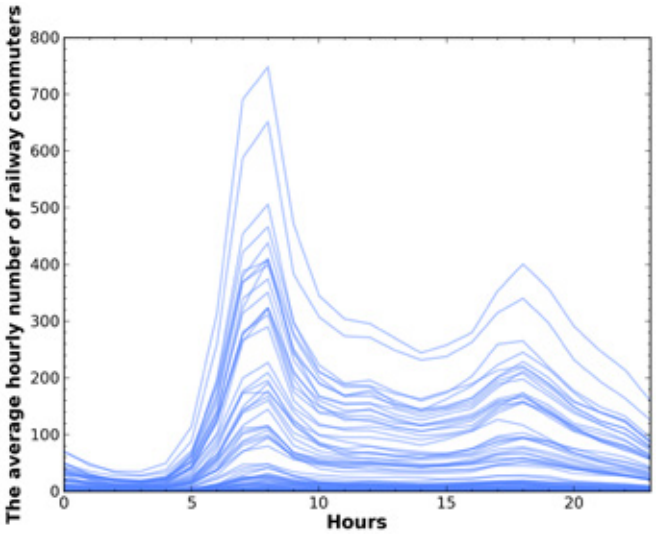


Figure 7: The average of an hourly number of railway commuters on weekdays in 110 rail lines

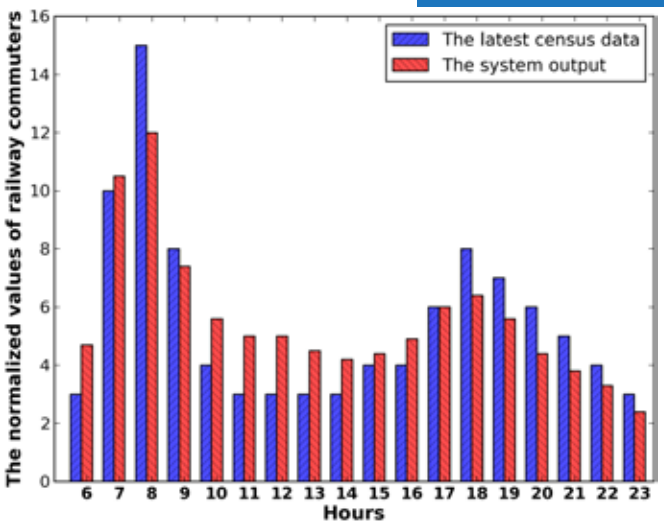


Figure 8: The comparing result of an hourly number of the railway commuters between our output and the latest census data

latest census data to compare the number of railway commuters for each rail line. The correlation is shown in Figure 9. In the figure 9, the x-axis, the y-axis means the number of railway commuters of the latest census data, the number of railway commuters of our output, respectively. The correlation coefficient is 0.768. Thus, there is the meaningful correlation between our output and the latest census data. Next, we analyze differences between our output and the latest census data to verify characteristics of our system. Since the absolute number of two data set is totally different, we rank each rail line according to the number of railway commuters. Then, the absolute value by differencing between ranks is calculated and sorted in descending order. We classify according to difference of the absolute value between our output and the latest census data under three groups(*diff_large*, *diff_middle*, *diff_small*). In other words, *diff_large*, *diff_middle* and *diff_small* are expressed rail lines of which a difference between our output and the latest census data is large, middle, and small, respectively. Figure 10 shows the sum and the average of the number of railway commuters in the latest census data including each group. As shown in the figure 10, we discovered that the difference between our output and the census data is large, when the number of railway commuters in the census data is small. Finally, though the result is reliable in a rail line that the number of users is large, a gap between our output and actual situation tend to be large in case of a rail line that the number of users is small.

9. CONCLUSIONS

The user's location trajectories with large-scale and long-term from CDRs have been attractive data source to understand human behavior. In particular, the cellular handoff pattern have been paid attention as the route or rail line classifying method that a user rode. To manage a large numbers of rail lines, however, the the conventional works are not suitable at the point of scalability, because the works have to learn the handoff patterns for each route manually.

To solve the issue, we propose the system that enable to maintain a large number of the cellular handoff patterns automatically. Our proposed system has process in three step which consist of the data cleaning, the self-learning, and the classifying to identify railway commuters. In the self-learning process, we propose the three kinds of self-learning methods which include the voronoi diagram based approach, the mahalanobis distance based approach and the static range based approach using a huge amount of anonymous sparse location of CDRs. As the result of evaluation, the accuracy of classifying the rail lines from the individual cellular handoff pattern achieved 81% against accurate 7600 GPS trajectories that regard as the ground truth dataset. We also evaluate the correlation about the traffic flow of urban railway networks between our output and the latest census data. the result of the correlation was 0.768. Figure 11 shows the mappting result of traffic flow from both our output and the latest census data. For visualizing the traffic flow as the color gradation, we normalized at the difference between the maximum value and the minimum value of the traffic flow. The two figure show the same trend at the point of the traffic flow. In conclusion, our proposed system indicates the possibility of understanding the traffic flow of the urban railway networks more often using only CDRs as substitute for the census. By using our system, we can measure the traffic flow of urban railway networks whenever you need.

10. 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.

11. REFERENCES

[1] A. Aucinas, N. Vallina-Rodriguez, Y. Grunenberger, V. Erramilli, K. Papagiannaki, J. Crowcroft, and D. Wetherall. Staying online while mobile: The hidden

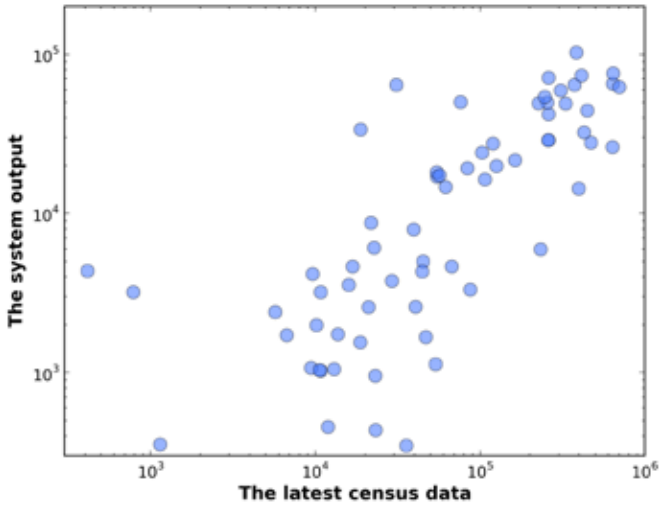


Figure 9: The correlation between our output and the latest census data

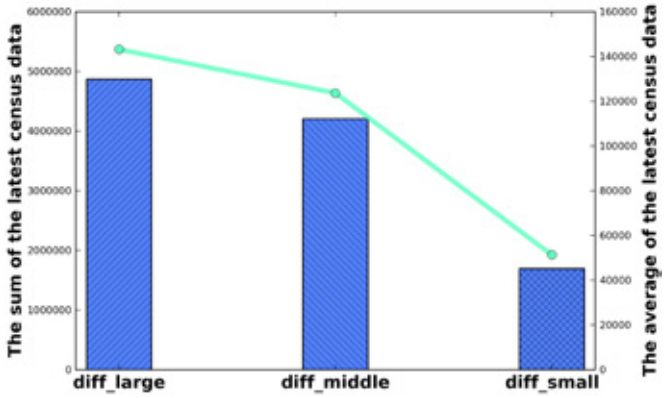


Figure 10: The sum and the average of a number of the railway commuters in the latest census data which be included in the defined three groups

costs. In *Proceedings of the Ninth ACM Conference on Emerging Networking Experiments and Technologies*, CoNEXT '13, pages 315–320, New York, NY, USA, 2013. ACM.

[2] R. Becker, R. Cáceres, K. Hanson, S. Isaacman, J. M. Loh, M. Martonosi, J. Rowland, S. Urbanek, A. Varshavsky, and C. Volinsky. Human mobility characterization from cellular network data. *Commun. ACM*, 56(1):74–82, Jan. 2013.

[3] R. A. Becker, R. Caceres, K. Hanson, J. M. Loh, S. Urbanek, A. Varshavsky, and C. Volinsky. Route classification using cellular handoff patterns. In *Proceedings of the 13th International Conference on Ubiquitous Computing*, UbiComp '11, pages 123–132, New York, NY, USA, 2011. ACM.

[4] M. Böhmer, B. Hecht, J. Schöning, A. Krüger, and G. Bauer. Falling asleep with angry birds, facebook and kindle: A large scale study on mobile application usage. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*, MobileHCI '11, pages 47–56, New York, NY, USA, 2011. ACM.

[5] F. Calabrese, F. C. Pereira, G. Di Lorenzo, L. Liu, and C. Ratti. The geography of taste: Analyzing cell-phone mobility and social events. In *Proceedings of the 8th International Conference on Pervasive Computing*, Pervasive'10, pages 22–37, Berlin, Heidelberg, 2010. Springer-Verlag.

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

[7] Y. Hayashi, X. Mai, and H. Kato. The role of rail transport for sustainable urban transport. In W. Rothengatter, Y. Hayashi, and W. Schade, editors, *Transport Moving to Climate Intelligence*, Transportation Research, Economics and Policy, pages 161–174. Springer New York, 2011.

[8] S. Hemminki, P. Nurmi, and S. Tarkoma. Accelerometer-based transportation mode detection on smartphones. In *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems*, SenSys '13, pages 13:1–13:14, New York, NY, USA, 2013. ACM.

[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 *Proceedings of the 9th International Conference on Pervasive Computing*, Pervasive'11, pages 133–151, Berlin, Heidelberg, 2011. Springer-Verlag.

[10] S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, J. Rowland, and A. Varshavsky. A tale of two cities. In *Proceedings of the Eleventh Workshop on Mobile Computing Systems & Applications*, HotMobile '10, pages 19–24, New York, NY, USA, 2010. ACM.

[11] S. Isaacman, R. Becker, R. Cáceres, M. Martonosi, J. Rowland, A. Varshavsky, and W. Willinger. Human mobility modeling at metropolitan scales. In *Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services*, MobiSys '12, pages 239–252, New York, NY, USA, 2012. ACM.

[12] H. Kato. Urban rail development in tokyo from 2000 to 2010. International Transport Forum Discussion Paper 2014-05, Paris, 2014.

[13] R. Lambiotte, V. D. Blondel, C. de Kerchove, E. Huens, C. Prieur, Z. Smoreda, and P. Van Dooren. Geographical dispersal of mobile communication networks. *Physica A Statistical Mechanics and its Applications*, 387:5317–5325, Sept. 2008.

[14] J. Liu, O. Wolfson, and H. Yin. Extracting semantic location from outdoor positioning systems. In *MDM*, page 73. IEEE Computer Society, 2006.

[15] Ministry of Land, Infrastructure, Transport and Tourism in Japan. 11th Metropolitan Transportation Census. `http://www.mlit.go.jp/sogoseisaku/transport/sosei_transport_tk_000007.html`, 2012.

[16] K. Nakamura and Y. Hayashi. Strategies and instruments for low-carbon urban transport: An international review on trends and effects. *Transport Policy*, 29(C):264–274, 2013.

[17] H. Nakanishi, K. Matsuo, and J. Black. Transportation planning methodologies for

(a) The result of the latest census data (b) The result of the output of the proposed system

Figure 11: The mapping results of the traffic flow from both the latest census data and our output

post-disaster recovery in regional communities: the east japan earthquake and tsunami 2011. *Journal of Transport Geography*, 31(Complete):181–191, 2013.

[18] National Spatial Planning and Regional Policy and Bureau in Japan. National Land Numerical Information Railway Data. `http://nlftp.mlit.go.jp/ksj-e/gml/datalist/KsjTmplt-N02.html`, 2012.

[19] J.-P. Onnela, J. SaramÃd'ki, J. HyvÃűnen, G. SzabÃş, D. Lazer, K. Kaski, J. KertÃl'sz, and A.-L. BarabÃąsi. Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences*, 104(18):7332–7336, 2007.

[20] T. Osaragi. Modeling a spatiotemporal distribution of stranded people returning home on foot in the aftermath of a large-scale earthquake. *Natural Hazards*, 68(3):1385–1398, 2013.

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

[22] R. C. Shah, C.-y. Wan, H. Lu, and L. Nachman. Classifying the mode of transportation on mobile phones using gis information. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '14, pages 225–229, New York, NY, USA, 2014. ACM.

[23] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197, March 1981.

[24] L. Stenneth, O. Wolfson, P. S. Yu, and B. Xu. Transportation mode detection using mobile phones and gis information. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '11, pages 54–63, New York, NY, USA, 2011. ACM.

[25] A. Thiagarajan, L. Ravindranath, H. Balakrishnan, S. Madden, and L. Girod. Accurate, low-energy trajectory mapping for mobile devices. In *Proceedings of the 8th USENIX Conference on Networked Systems Design and Implementation*, NSDI'11, pages 267–280, Berkeley, CA, USA, 2011. USENIX Association.

[26] Y. Zheng, Q. Li, Y. Chen, X. Xie, and W.-Y. Ma. Understanding mobility based on gps data. In *Proceedings of the 10th International Conference on Ubiquitous Computing*, UbiComp '08, pages 312–321, New York, NY, USA, 2008. ACM.

[27] C. Zhou, D. Frankowski, P. Ludford, S. Shekhar, and L. Terveen. Discovering personally meaningful places: An interactive clustering approach. *ACM Trans. Inf. Syst.*, 25(3), July 2007.

[28] P. Zhou, Y. Zheng, and M. Li. How long to wait?: Predicting bus arrival time with mobile phone based participatory sensing. In *Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services*, MobiSys '12, pages 379–392, New York, NY, USA, 2012. ACM.

SESSION 9
SOCIAL GOOD

A Framework for Evacuation Hotspot Detection after Large Scale Disasters using Mobile Phone Location Data

Takahiro Yabe, Yoshihide Sekimoto
University of Tokyo
yabe0505@iis.u-tokyo.ac.jp

Kota Tsubouchi
Yahoo Japan Corporation
ktsubouc@yahoo-corp.co.jp

Keywords
Disaster Management, Human Mobility, Urban Computing, Evacuation Hotspot Detection, Mobile Phone Location Data

1. INTRODUCTION
Large scale disasters such as Magnitude 7 earthquakes, tsunamis, and mega-typhoons damage urban infrastructure and cause severe social disorder. As a result of mass evacuation activities of the victims, it is often extremely difficult to grasp the locations of all the evacuation hotspots in the confusion after the disasters. This is made even complex especially because some people evacuate to locations that are not officially designated as evacuation shelters, such as parking areas of large shopping malls and local parks. Consequently, many evacuation hotspots couldn't be provided with food and supplies efficiently [2].

Therefore, there is an urgent need for an efficient framework for estimating evacuation hotspots right after a natural disaster. The framework needs to require less time and less work load for the decision makers compared to the conventional on-foot field search for evacuation hotspots.

Recently, GPS and call detail records (CDR) of mobile phones are being used for human mobility analysis [3-6], and are applied to various fields of study such as traffic management [7-9], urban planning [10], and pandemic simulations [11]. Some studies have analyzed the irregular human mobility after natural disasters such as Hurricane Sandy, Great East Japan Earthquake, and Haiti Earthquake [12-14], but none have proposed a framework for a real-time evacuation hotspot estimation.

In this paper, we propose a framework for estimating evacuation hotspots using location data from mobile phones that works efficiently in disasters of any scale. To demonstrate the accuracy and efficiency of the framework, we used Yahoo! Japan's GPS dataset of over 20,000 IDs for detecting and analyzing the evacuation hotspots after the Kumamoto earthquake. Through the demonstration on the Kumamoto earthquake, we show that our framework accurately detects evacuation hotspots, and also that this process can be completed at a significantly high speed and requires low effort compared to the conventional on-foot investigations.

Our key contributions of this paper are as follows:

- We propose a framework to detect evacuation hotspots following large scale natural disasters by using mobile phone location data.
- We validate the accuracy and efficiency of our framework by estimating the evacuation hotspots after the Kumamoto earthquake using actual GPS data provided by Yahoo Japan.

2. Mobile Phone GPS Dataset
The Yahoo Japan Disaster App collects the GPS data of mobile phones of individuals who have agreed to provide their location data to Yahoo! Japan when installing the app. Each GPS record contains an anonymized user ID, longitude, latitude, and timestamp. The GPS data are collected every day when the individuals move around, while the phone is turned on. In total, GPS data of around 1 million individuals (sample rate around 1% from all over Japan) have been collected. As shown in Table 1, for the experiment, we use a total of 22,124 users' 418,119 total GPS logs from a period of January 1st to May 16th of 2016 which were located in Kumamoto area. The data is dense to the extent that plotting all the GPS points taken on a single day, as shown in Figure 1, can draw the entire road network of the Kumamoto area.

3. PROPOSED EVACUATION HOTSPOT DETECTION FRAMEWORK
Our proposed framework calculates the congestion anomaly values (K) of each grid cell by comparing the nighttime congestion (M) after large disasters in each cell with the usual average (μ) and standard deviation (σ) of nighttime population in that cell, by the following equation.

$$K = {}^{M-\mu}/_{\sigma} \quad (3)$$

The framework is consisted of the manual parameter input process and the automated calculation process. The manual parameter input could be completed momentarily, since shapefiles are available online, and the parameters are easy to input. The automated calculation process is consisted of 4 parts; 1) location data collection, 2) aggregation and smoothing the collected GPS data, 3) calculating the anomaly value of each grid cell, and 4) visualization

Table 1. Number of unique IDs, number of logs of Yahoo Japan's GPS data taken from users' mobile phones

Period of data	Average daily number of IDs in Kumamoto area	Average daily total GPS logs in Kumamoto area
2016/01/01 ~ 2016/05/16	22,124 (1.2% sample rate)	418,119 (avg. 19 logs/user/day)

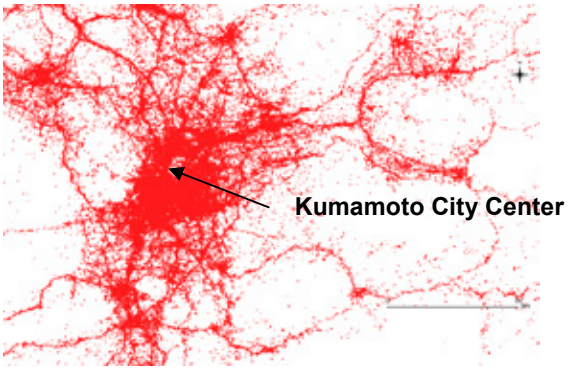


Figure 1. Collected GPS data points on April 1st 2016 of Kumamoto area. We can clearly observe the congested city center areas and also the road networks connecting the clusters, by plotting the dense GPS data of mobile phones.

Figure 2. Map of Kumamoto area; before, 2 days after, and 1 month after the earthquake. The grid cells with anomaly values K>3, 2<K<3, and 1<K<2 are colored red, orange, and yellow, respectively. We can observe the increase in irregularly congested grid cells right after the earthquake, and its decrease as time passes and start to get back to normal from the shock.

of estimated evacuation hotspots. This framework is performed iteratively after the occurrence of the disaster, to update the output locations of evacuation hotspots. Our framework is efficiently designed so that it could be quickly processed, and also has little burden for the framework users since most of the parts are automated. Users need to input the grid cell size for outputting the anomaly maps. The grid cell sizes can be set as either 100m, 250m, 500, or 1000m. This parameter needs to be selected according to the wideness of the disaster affected area. For detailed explanation on our method, please read Yabe et al. [1].

4. CASE STUDY: Kumamoto Earthquake

To demonstrate and to test our framework, we estimate the evacuation hotspots with real time GPS dataset after the Kumamoto earthquake. Furthermore, we verify the efficiency of our system by calculating the time taken for estimation.

4.1 Visualization of Evacuation Hotspots

Figure 2 shows the map of Kumamoto with grid cells with $K > 3$ colored in red, $2 < K < 3$ colored in orange, and $1 < K < 2$ in yellow. We can observe very few grid cells with $K > 3$ on April 1st before the earthquake, meaning that the majority of the grid cells have a population within usual range. However on the 18th, after the large earthquake, there is a significant increase in the number of grid cells with high anomaly values, especially near the city center and the southern part where many people evacuated. These grid cells indicate the "evacuation hotspots", where people evacuated at a significant rate compared to the usual population in that grid cell. It is also interesting to observe $K > 3$ grid cells located on roads near the coastline. The high anomaly values in these grid cells infer that many people stayed in their cars away from their houses to avoid being injured by building collapses. However, after a month from the earthquake, we can see a decrease in congested areas in Kumamoto area. This implies that many evacuees returned home (if their house was not completely damaged) or moved away to other areas of Japan for shelter, and many outsiders who were gathering in Kumamoto to volunteer work went back to their homes. Using these anomaly maps, we can detect locations that are significantly congested with people in real time, which can be utilized for rescue and supply distribution strategies. Using our framework, these maps will be easily obtained by the decision makers, by just entering three parameters.

4.2 Features of Evacuation Hotspots

In this subsection, we validate our estimation of evacuation hotspots by checking the facilities located in each grid cell using a map. Figure 3 shows a map of central Kumamoto area and the grid cells with $K > 3$ colored in red. The types of features located in each of the grid cells are written beside the grid cell. Many of the estimated hotspots contain facilities that have the capacity to contain large evacuation population, such as schools, city halls, convention centers and parks. From these results, we can conclude that our framework can accurately estimate the evacuation hotspots after the earthquake, and that our framework provides a convenient interface that shows the decision makers what features are congested with evacuees. We also have to note that out of the 1100 hotspots, around 100 were not designated as evacuation shelters before by the prefectural or the municipal government. This is an issue that made the detection of evacuation shelters difficult for the decision makers after the earthquake. As a result, people who evacuated to these non-designated facilities could not be satisfied with much supplies compared to people who evacuated to designated shelters.

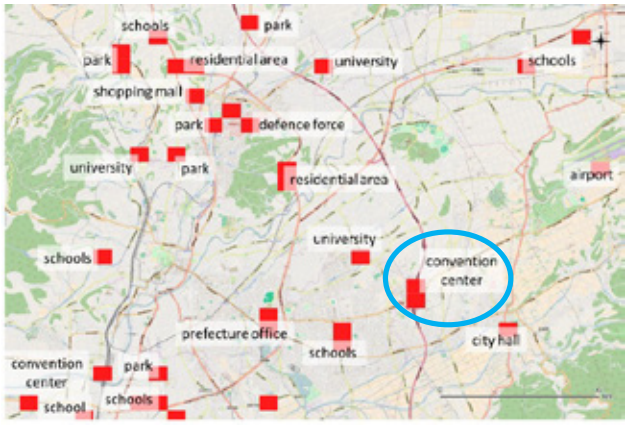


Figure 3. Map of central Kumamoto with K>3 grid cells colored in red. Close investigation on these grid cells revealed that most of these grid cells contain features that have a large capacity for evacuees such as schools, city halls, and convention centers. The blue circle shows Grand Messe Kumamoto (GMK).

4.3 Population Transition of Hotspots

To analyze the evacuation activities with more detail, we focus on an evacuation hotspot where $K > 3$ anomaly was detected, and plot the transition of the daily population in that selected hotspot. We focus on an area that includes "Grand Messe Kumamoto (GMK)" (circled in blue in Figure 3), a convention center in Kumamoto area, and analyze its daily transition of population near that facility. According to newspaper articles [15], many evacuees gathered in the parking area of GMK right after the earthquake, despite the fact

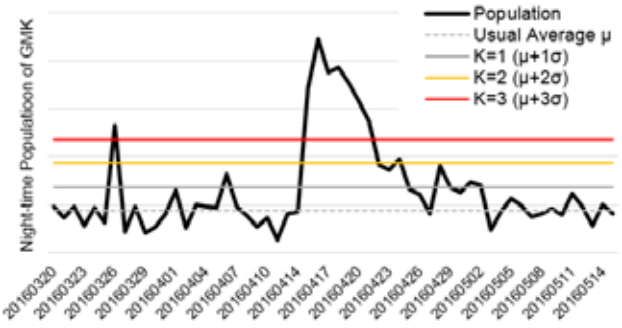


Figure 4. Transition of population density near "Grand Messe Kumamoto", a large shopping mall which was not designated as an evacuation place. We can observe high congestion of K>3 after the earthquake, on April 15th ~ 21st.

that this facility was not designated as an evacuation location. The broad black line in Figure 4 shows the daily transition of the night-time population in GMK from March 20th to 14th May 2016. The dotted gray line shows the average population μ of usual days near GMK, and the gray, orange, and red lines indicate the $K = 1$, $K = 2$, $K = 3$ lines for GMK, respectively. We can observe a rapid increase after the earthquakes on April 16th, and a significant anomaly in population density for more than a week over $K = 3$. We can also spot an instantaneous $K > 3$ on Saturday, March 26th. There was actually a large music festival held on this day at GMK, which is an example of an anomaly on a "usual" day. After April 18th, we can observe a quick decrease of population in GMK, and a gradual return to a usual level of population. By the beginning of May, the population in GMK has transferred back to the normal state. The evacuation hotspot in GMK was diminished after just a week after the earthquakes.

The increase of population on April 16th coincides with the information on the newspaper article [15]. We can conclude that the population in GMK, was accurately estimated from GPS dataset, and that the function as an evacuation shelter was finished after only one week from the earthquake.

5. DISCUSSION

Our framework for estimating evacuation hotspots using location data from mobile phones provides useful information to decision makers and shelter managers quicker and with less effort than conventional methods. Providing useful and accurate information can contribute to making efficient supply distribution and rescue operation plans after disasters.

By calculating the anomaly value of each grid cell in Kumamoto after the earthquake, we were able to estimate the distribution of irregularly congested grid cells, despite the sample rate (around 1%). We were able to quantify the significant increase in number of grid cells with high anomaly values right after the occurrence of the earthquake, and statistically show the irregularity of the situation compared to usual days. In addition, the gradual decrease of grid cells with anomaly values higher than 2, showed the settling down of irregularity after a few weeks from the occurrence of the earthquakes. Through the analysis and visualization of the anomaly with our framework, we were able to observe the impact of earthquakes on the people's evacuation activities.

Secondly, by checking the features located in each grid cell that were estimated to be extremely congested, we verified the accuracy of our estimation. We were able to detect locations where administrative organizations had not designated as evacuation shelters, such as Grand Messe Kumamoto.

6. CONCLUSION

We proposed a framework for estimating evacuation hotspots by calculating each grid cell's anomaly value after large disasters, using location data from mobile phones. To the best of our knowledge, this framework is the first to focus on estimating evacuation hotspots, and to actually demonstrate the framework using real GPS dataset. Our framework can function quicker and with less effort compared to conventional methods that involve on-foot searches for evacuation centers where people are gathering.

To validate our method, we analyzed the population density anomaly after the Kumamoto earthquake, and observed the sharp increase of high anomaly value grid cells in Kumamoto area caused by the evacuation activities of the victims. We then verified our estimation by observing the features included in each anomaly cell, and also newspaper articles that reported the population transition in one of the evacuation hotspots.

Through the validation in the case study of Kumamoto and the efficiency test, we have confirmed the high accuracy and practicality of our evacuation hotspot framework using location data from mobile phones.

7. REFERENCES

1. Yabe, T., Tsubouchi, K., Sudo, A., Sekimoto, Y., A Framework for Evacuation Hotspot Detection after Large Scale Disasters using Location Data from Smartphones: Case Study of Kumamoto Earthquake, Proc. 22nd ACM SIGSPATIAL Int'l Conference on Advances in Geographic Information Systems. ACM (2016).
2. West Japan Newspaper, Disparity between evacuation centers, April 20th (2016). (in Japanese)
3. Gonzalez, M. C., Hidalgo, C. A., & Barabasi, A. L. Understanding individual human mobility patterns. Nature, 453(7196), (2008).
4. Sevtsuk, A., & Ratti, C. Does urban mobility have a daily routine? Learning from the aggregate data of mobile networks. Journal of Urban Technology, 17(1), 41-60. (2010).
5. Ashbrook, D., & Starner, T. Using GPS to learn significant locations and predict movement across multiple users. Personal and Ubiquitous Computing, 7, 5, 275-286. (2003).
6. Calabrese, F., Di Lorenzo, G., Liu, L., & Ratti, C. Estimating origin-destination flows using mobile phone location data. IEEE Pervasive Computing, 10(4), (2011).
7. Demissie, M. G., de Almeida Correia, G. H., & Bento, C. Intelligent road traffic status detection system through cellular networks handover information, Transportation research part C: emerging technologies, 32, 76-88. (2013).
8. Wang, P., Hunter, T., Bayen, A. M., Schechtner, K. & González, M. C. Understanding road usage patterns in urban areas. Sci. Rep. 2, 1001 (2012).
9. Yang, Y., Gerstle, D., Widhalm, P., Bauer, D. & González, M. The potential of low-frequency data for the monitoring and control of bus performance. Transport. Res. Rec. J. Transport. Res. (2013).
10. Pentland, A. Society's nervous system: Building effective government, energy, and public health systems. IEEE Computer 45, 31–38 (2012).
11. Colizza, V., Barrat, A., Barthelemy, M., Valleron, A. J., & Vespignani, A.. Modeling the worldwide spread of pandemic influenza: baseline case and containment interventions. PLoS Med, 4(1) (2007).
12. Lu, X., Bengtsson, L., & Holme, P. Predictability of population displacement after the 2010 Haiti earthquake. Proc. National Academy of Sciences, 109(29), 11576-11581. (2012).
13. Wang, Q., & Taylor, J. E. Quantifying human mobility perturbation and resilience in Hurricane Sandy. PLoS one, 9(11), (2014).
14. Song, X., Zhang, Q., Sekimoto, Y., Horanont, T., Ueyama, S., & Shibasaki, R. Modeling and probabilistic reasoning of population evacuation during large-scale disaster. Proc. ACM SIGKDD (2013).
15. Kumamoto Daily Newspaper, Evacuation center at Grand Messe Kumamoto, April 24th (2016). (in Japanese)

Mapping poverty using mobile phone and satellite data

Jessica E Steele[1,2], Pål Roe Sundsøy[3], Carla Pezzulo[1], Victor Alegana[1], Tomas J Bird[1], Joshua Blumenstock[4], Johannes Bjelland[3], Yves-Alexandre de Montjoye[5], Kenth Engø-Monsen[3], Asif M Iqbal[6], Khandakar N Hadiuzzaman[6], Xin Lu[2,7], Erik Wetter[2,8], Linus Bengtsson[2,7], Andrew J Tatem[1,2,10]

[1] Geography and Environment, University of Southampton, Highfield Campus, Southampton SO17 1BJ, UK [2] Flowminder Foundation, Roslagsgatan 17 SE-11355, Stockholm, Sweden [3] Telenor Group Research, Snarøyveien 30, 1331 Fornebu, Norway [4] University of California, Berkeley, School of Information, USA [5] Imperial College London, Data Science Institute [6] Grameenphone, Dhaka, Bangladesh [7] Karolinska Institutet, Solnavägen 1, 171 77 Solna, Sweden [8] Stockholm School of Economics, Kungstensgatan 32, 113 57 Stockholm, Sweden [10] Fogerty International Center, National Institutes of Health, 16 Center Drive, Bethesda, MD 20892, USA

1 Introduction

Eradicating poverty in all its forms remains a major challenge and the first target of the Sustainable Development Goals (SDGs). To eradicate poverty, it is crucial that information is available on where affected people live. The geographic distribution of poverty in low and middle-income countries (LMICs) is often uncertain. Small area estimation forms the standard approach to produce poverty maps, but these techniques rely heavily census data, which in most LMICs are unavailable or out-of-date. Recently, there are promising signs that novel sources of high-resolution data can provide an accurate and up-to-date indication of living conditions. Recent analyses have shown the potential using of features derived from satellite and Geographic Information System data and mobile operator call detail records (CDRs). Here we use overlapping sources of satellite data, CDRs, and traditional survey-based data in Bangladesh to provide the first systematic evaluation of the extent to which different sources of input data can accurately estimate three different measures of poverty. We additionally compare our results with previous poverty estimates for Bangladesh at coarser and finer resolutions.

2 Data and Methods

Mobile phone CDR features were generated from 4 months of mobile phone metadata collected between November 2013 and March 2014. Grameenphone subscribers consented to the use of their data for the analysis. CDR features range from metrics such as basic phone usage, top-up patterns, and social network to metrics of user mobility and handset usage. They include various parameters of the corresponding distributions such as weekly or monthly median, mean and variance.

We further identified, assembled, and processed twenty-five raster and vector datasets into a set of satellite covariates for the whole of Bangladesh at a 1-km spatial resolution. These data were obtained from existing sources and produced ad hoc for this study to include environmental and physical metrics likely to be associated with human welfare such as vegetation indices, nighttime lights, climatic conditions, and distance to roads or major urban areas.

We employed a model selection stage using non-spatial generalised linear models, implemented via the R *glmulti* package to build every possible non-redundant model for every combination of covariates. Models were built on a randomly selected 80% of the data to guard against overfitting. Models were chosen using Akaike's information criterion (AIC), which ranks models based on goodness of fit and complexity, while penalizing deviance. This process was completed for national, urban, and rural strata, and using satellite data only, CDR data only, and combined satellite + CDR datasets (27 resulting models). Then, using the models selected by the previous step, we employed hierarchical Bayesian geostatistical models (BGMs) to predict the three poverty metrics at unsampled locations across the population.

All BGMs were implemented using integrated nested Laplace approximations (INLA), and the models fit using R-INLA, with the Besag model for spatial effects specified inside the function. The geostatistical models defined for the poverty data were applied to produce poverty predictions as a posterior

distribution with complete modelled uncertainty around estimates. The posterior mean and standard deviation were then used to generate prediction maps with associated uncertainty. Model performance was based on out-of-sample validation statistics calculated on a 20% test subset of data.

3 Results

We find models employing a combination of CDR and satellite data generally provide an advantage over models based on either data source alone. The fine spatial granularity of the resultant poverty estimates can be seen in figure 1, which shows the predicted distribution of poverty for the DHS wealth index. We also find that explicitly modelling the spatial covariance in the data was critically important. This resulted in improved predictions, lower error, and better measures of fit based on cross-validation and the deviance information criteria.

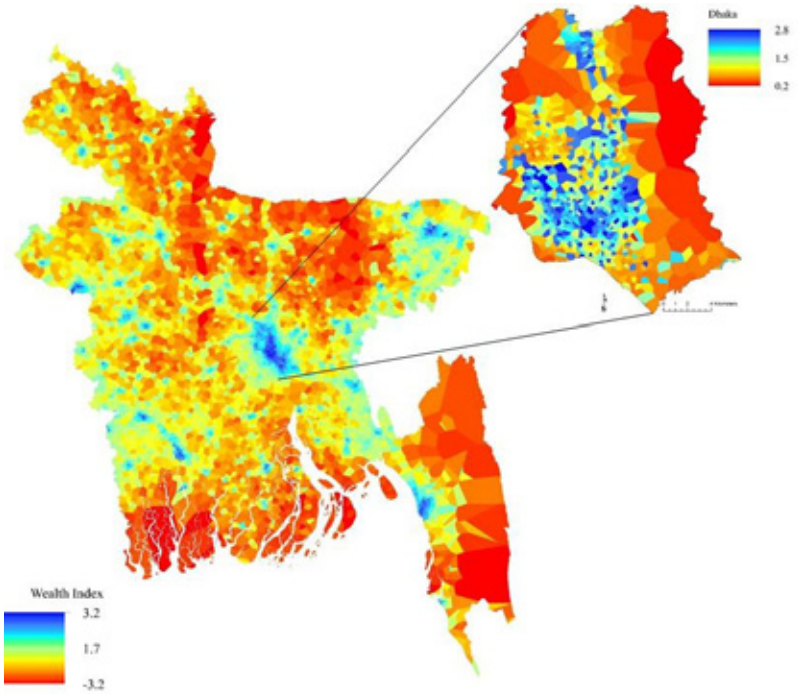


Figure 1. National level prediction maps for mean wealth index. The map was generated using CDRs + satellite data in Bayesian geostatistical models. Red indicates poorer areas.

Separating estimation by urban and rural regions highlights the importance of different data in different contexts. Nighttime lights, transport time to the closest urban settlement, and elevation were important nationally and in rural models; climate variables were also important in rural areas. Distance to roads and waterways were significant in urban and rural strata. Top-up features derived from recharge amounts and tower averages were significant in every model, affirming their importance in poverty work. People predicted to be poorer top up their phones more frequently in small amounts. Percent nocturnal calls, and count and duration of SMS traffic were significant nationally. Mobility and social network features were important in all three strata. In urban areas, SMS traffic was important, whereas multimedia messaging and video attributes were key in rural areas. The resulting predictions line up well with existing SAE estimates for Bangladesh, and with high resolution maps of slum areas in Dhaka.

4 Discussion

Models built with CDR data and models built with satellite data perform comparably in their ability to map poverty indicators, and integrating data from mobile phones and satellites provided improvement in predictive power and lower error. In urban areas, the CDR data were able to produce accurate, high-resolution estimates not possible using RS data alone. As such, CDRs potentially allow for estimation of wealth at much finer granularity – including the neighbourhood or even the household or individual – than the current generation of RS technologies. CDRs also provide significant advantages in temporal granularity: CDRs update in real-time versus RS data, which update far less frequently. Although in this study we have not utilised dynamic validation data, it is a clear future application for CDRs in real-time to better comprehend the dynamic nature of poverty.

5 References*

Pozzi, F., Robinson, T. & Nelson, A. Accessibility Mapping and Rural Poverty in the Horn of Africa. (Food and Agriculture Organization of the United Nations, 2009).

Rogers, D., Emwanu, T. & Robinson, T. Poverty Mapping in Uganda: An Analysis Using Remotely Sensed and Other Environmental Data. (Food and Agriculture Organization of the United Nations, 2006).

Tatem, A. J., Gething, P. W., Pezzulo, C., Weiss, D. & Bhatt, S. Development of High-Resolution Gridded Poverty Surfaces. (2014).

Blumenstock, J., Cadamuro, G. & On, R. Predicting poverty and wealth from mobile phone metadata. Science 350, 1073–1076 (2015).

Noor, A. M., Alegana, V. A., Gething, P. W., Tatem, A. J. & Snow, R. W. Using remotely sensed night-time light as a proxy for poverty in Africa. Popul. Health Metr. 6, 5 (2008).

Ghosh, T., Anderson, S. J., Elvidge, C. D. & Sutton, P. C. Using Nighttime Satellite Imagery as a Proxy Measure of Human Well-Being. Sustainability 5, 4988–5019 (2013).

Watmough, G. R., Atkinson, P. M., Saikia, A. & Hutton, C. W. Understanding the Evidence Base for Poverty–Environment Relationships using Remotely Sensed Satellite Data: An Example from Assam, India. World Dev. 78, 188–203 (2016).

Blangiardo, M., Cameletti, M., Baio, G. & Rue, H. Spatial and spatio-temporal models with R-INLA. Spat. Spatio-Temporal Epidemiol. 4, 33–49 (2013).

Blangiardo, M. & Cameletti, M. Spatial and Spatio-temporal Bayesian Models with R - INLA. (John Wiley & Sons, 2015).

Gruebner, O. et al. Mapping the Slums of Dhaka from 2006 to 2010, Mapping the Slums of Dhaka from 2006 to 2010. Dataset Pap. Sci. Dataset Pap. Sci. 2014, 2014, e172182 (2014).

glmulti: An R Package for Easy Automated Model Selection with (Generalized) Linear Models | Calcagno | Journal of Statistical Software. Available at: http://www.jstatsoft.org/article/view/v034i12. (Accessed: 21st January 2016)

Rue, H., Martino, S. & Chopin, N. Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations. J. R. Stat. Soc. Ser. B Stat. Methodol. 71, 319–392 (2009).

*Please see full details and references from this study published in the Journal of the Royal Society Interface

Estimating population behaviour to describe activity-based land-use in emerging economies using mobile phone event series

Gregor Engelmann
N-Lab, Horizon CDT
University of Nottingham

James Goulding
N-Lab
University of Nottingham

Gavin Smith
N-Lab
University of Nottingham

David Golightly
Human Factors Research Group
University of Nottingham

Understanding and monitoring the makeup of a city is integral to successful urban planning. Traditional ways of describing land-use involve the use of surveys, which are costly and quickly outdated. Unparalleled mass urbanisation make them particularly unsuited in emerging economies, where mobile phones are now nearly ubiquitous. This paper explores the potential of utilising automatically generated Call Detail Records (CDR) to understand population behaviour and by extension provide a summary of activity-based land-use. Using principal component analysis, non-negative matrix factorisation and unsupervised clustering, the paper shows the feasibility of describing activity-based land-use from CDR data in an exemplar East African city.

1. INTRODUCTION

The understanding and monitoring of activity-based land-use is intrinsic to successful urban planning. Land-use has traditionally been understood in terms of: land characteristics; ownership characteristics; and activity undertaken on the land itself [1]. Traditional techniques to monitor land-use have focused on manual surveys and more recently analyses of satellite imagery. These approaches are time consuming, expensive, infrequent, and quickly outdated. The rapidly increasing ubiquity of mobile phones, its role as a proxy for human movement, and the vast amounts of data generated by such devices provide much potential to supplement or even replace these traditional methods. This is of particular relevance in East Africa, where data gaps due to the expense and logistics required to employ sensor technologies exist. Closing such gaps is key in generating information to guide effective infrastructure investments and urban planning. In this paper we introduce a framework for identifying *activity-based land-use* via call detail record (CDR) data. The approach applies and compares principal component analysis (PCA) and non-negative matrix factorisation (NMF) to factorise underlying usage trends, prior to clustering regions into land-use classes via k-means clustering. Due to its dynamic nature, the use of CDR data to characterise actual behaviour rather than aggregate descriptors, is particular appropriate in emerging economies where mass urbanisation is resulting in rapidly changing environments.

A number of previous studies have utilized aggregate CDR data to analyse urban environments. One of the first was Reades et al [4], that found a correlation between calling activity and land use patterns in Rome. Subsequent studies have used either:
- Unsupervised [5, 6, 2], clustering according to aggregate statistics related to network behaviour such as average number of connected mobile phone subscribers, or call volumes;
- Semi-structured [3], seeding clustering algorithms with a small set of known points of interest to calibrate results or;
- Supervised [7], using ground truth or existing land use data for calibration techniques to identify land use classes at the base transceiver station (BTS) level
approaches. Accurate and fine grained land-use data necessary for the use of semi-structured and supervised approaches is extremely hard to find in East Africa. As we are restricted by the lack of accurate zoning for validation we focus improving unsupervised learning approaches instead. Our work goes further than previous unsupervised approaches, however, in that we: 1. first convert raw CDR event series to time series at different levels of temporal granularity to perform sparsity analysis to identify BTS outliers with anomalous behaviour that would skew the observed behaviour patterns; and 2. we then also apply and compare principal component analysis (PCA) with non-negative matrix factorisation (NMF) to extract latent features within population behaviour, reflecting daily patterns of life. Using factorisation techniques allows us not only to generate a vocabulary to describe underlying behaviours (such as nightlife, commuting, or industrial patterns), but then also allows us to go on to produce interpretable clusters in this compressed behavioural space. The resulting land-use archetypes can be understood and characterised in terms of how much they express each of these building block behaviours.

2. METHODOLOGY

The data used as part of this study covers a total of 433.6 million network events covering calls and SMS for 415k mobile phone subscribers taking place across the Dar es Salaam region of Tanzania over a period of 122 days in the autumn of 2014[1]. A raw CDR record is automatically created for each

[1]Due to both individual and commercial privacy, the anonymised data used as part of this study is not publicly available, and was provided to us through a partnership with a major private sector network operator in Tanzania

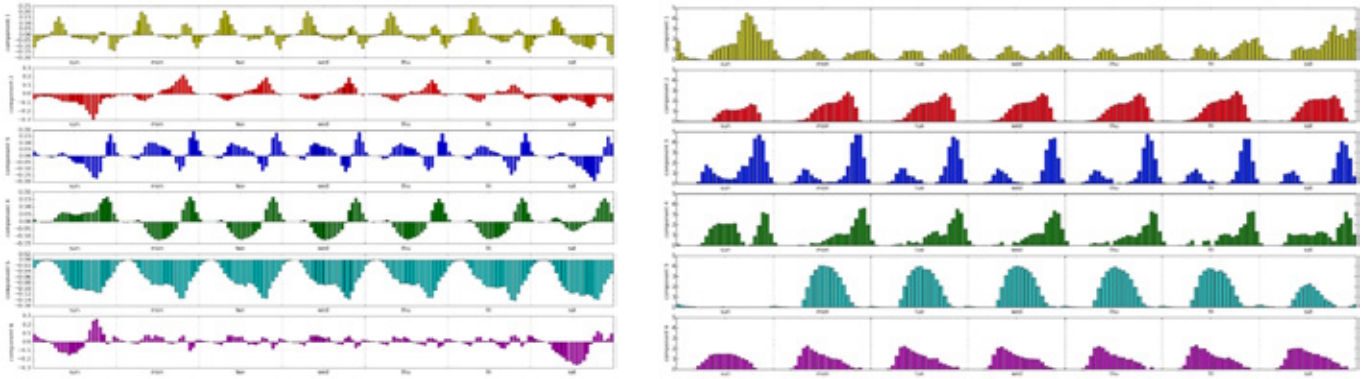


Figure 1: Six trends extracted via PCA (left) and NMF (right) from the CDR data. Each describes a different underlying population behaviour, which form the building blocks for our activity based land-use clustering approach. With the exception of component 5, the NMF components are far easier and more intuitive to interpret.

network event and includes a range of attributes including: *timestamp*; an anonymised mobile phone *subscriber ID*; *call duration*; and a *Base Transceiver Station (BTS) identifier*.

Our approach for estimating land-use patterns took this mass data set and followed a multi-step process.

First an event series of network events was extracted for each BTS. These event series were then themselves split into hour, day and week periods, and binned to produce time series of hourly event counts. The resulting time series instances were used to identify BTS with high levels of sparsity indicating potential issues with the operational status of a BTS over the study period. Any BTS showing activity for less than 55% of the bins was removed as noise from subsequent analysis. We then used the time series generated from the week periods for further analysis. This data set contained approximately 9,500 weekly time series instances, each reflecting 168 hourly intervals (24×7). Some BTS showed a uniform distribution of activity during day time, while others showed a higher network activity pattern either at evening times only, or at both morning and evening times.

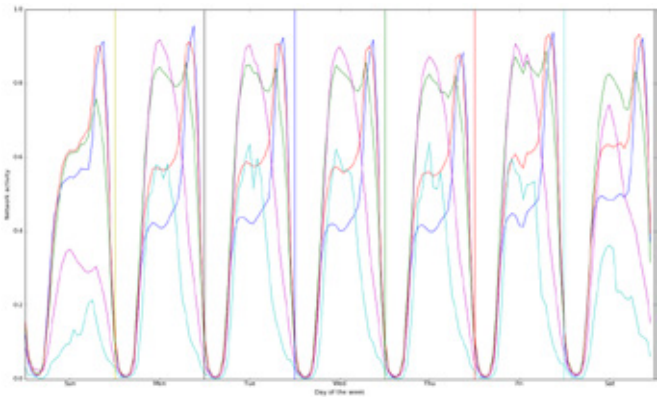


Figure 2: Observed behaviour clusters over an average week in Dar Es Salaam, Tanzania. Each cluster corresponds to a different land-use class

The total amount of network events differed quite significantly across the study area, making direct comparison and clustering difficult. In order to compensate for differing network event counts, feature scaling was applied to each time series for standardisation. Standardising by z-scores was not possible, as NMF does not allow for the input matrix to contain any negative values. We applied both PCA and NMF to these time series to identify latent features occurring in weekly usage behaviour. This factorisation process reduced both noise and reduced our feature space from 168 to 7 dimensions, eliminating outliers and addressing the curse of dimensionality. Moreover, manual analysis of the resulting latent features allowed us to interpret subsequent cluster archetypes. We found that due to allowance of negative loadings, the factors resulting from PCA are far harder to interpret (see Figure 1).

Following construction of our latent feature space, k-means clustering was applied in order to identify k different land-use areas based on our interpretable activity trends.

3. RESULTS

The factors resulting from our analysis reflecting underlying activity patterns are illustrated in Figure 2. These are informative of population behaviours, with common weekly trends being revealed. Component 2, for example, reveals general underpinning network activity patterns (and is very similar to the average weekly time series for all towers) with a gradual increase from 7am until 10am, plateauing out before an early evening spike in network events. Component 3, however, reflects a predominant residential activity pattern, with population leaving an area in the middle of the day, and returning after work. In contrast, Component 5 reflects a workplace behaviour, with high daytime activity, zero nighttime events, and a significant weekend drop. Any particular area may be composed of a combination of land-uses (for example, half residential and half industrial), and so may express each of these building block behaviours to a different amount. These components provide us with a vocabulary through which we can discuss those combinations, without need for imagery or demographic data.

With these building blocks in hand and each time series projected into the lower dimensional space they represent, k-means clustering technique was applied. The choice of k remains an arbitrary one dependent on the task in hand. As such the number of clusters n was varied from 2 to 15 clus-

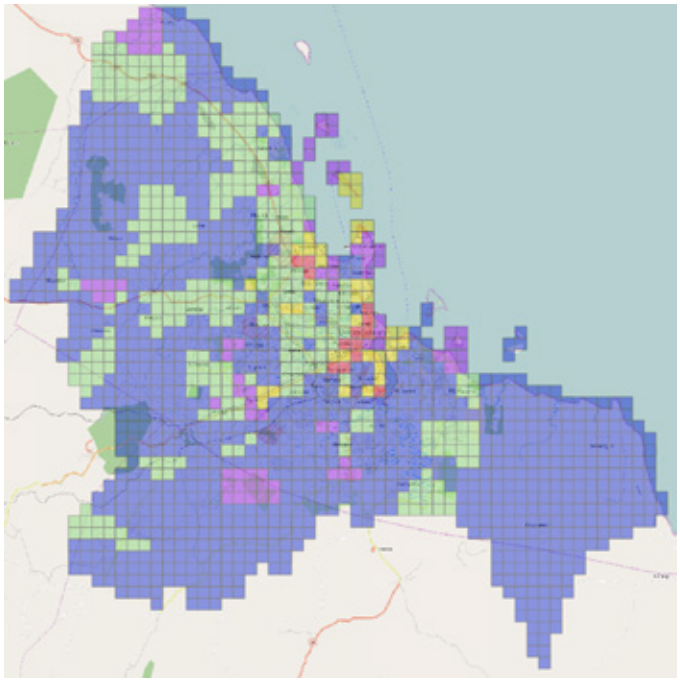


Figure 3: Spatial distribution of activity-based land-use areas in the central Dar Es Salaam region. Purple is 'Affluent-Commercial', Green is 'Slum', Yellow is 'Formal-Night-Active', Red is 'Industrial' and Blue is 'Residential-Commuting'

ters. The most interpretable results were obtained at $n = 5$. The centroids for each of the resulting clusters are illustrated in Figure 2[2]. Annotation of each class was performed by examining the extent it expressed each underlying behaviour building block in combination with external knowledge:

Cluster 1 - Affluent-Commercial: consistent activity throughout the day (zones that bring people in due to tourism, job opportunities and amenities).

Cluster 2 - Slum: characteristic of a poor demographic with lower daytime activity, low morning activity and significant peak in the early evening (perhaps due to lack of mobility).

Cluster 3 - Residential-Commuting: this profile expresses a far higher expression of component 3 (the residential activity pattern) than other behaviours, suggesting a commuting pattern.

Cluster 4 - Industrial: high expression of component 5 (non-residential). Some commuting, but a highly significant lack of mobility activity in the mornings, evenings and weekend.

Cluster 5 - Formal-Night-Active: average activity over the course of the day, but with significant spikes in the evening and night.

[2]A lower number of clusters mixed different behaviours, while a larger number of clusters produced some with insufficiently distinct behaviours.

The map in Figure 3 shows a plot of the spatial distribution of these clusters for the centre region of Dar es Salaam. In order to protect commercial interests of the network operator that provided us with the data, we interpolate BTS catchment areas into a 500m x 500m grid-cell representation.

4. CONCLUSION

In this study we have shown how factorisation techniques and k-means clustering might be used to identify interpretable activity-based land-use clusters from CDR data. While not always feasible, as data may have been pre-aggregated, hourly aggregates seem to be the best temporal scale for performing sparsity analysis for outlier detection. We also found that NMF is far superior to PCA in describing underlying behaviour trends due to the absence of negative loadings.

5. REFERENCES

[1] J. R. Anderson, E. E. Hardy, J. T. Roach, and R. E. Witmer. A land use and land cover classification system for use with remote sensor data. Technical Report 964, USGS, 1976.

[2] D. M. R. S. Kaushalya Madhawa, Sriganesh Lokanathan. Using mobile network big data for land use classification. Technical report, LIRNEasia, July 2015.

[3] T. Pei, S. Sobolevsky, C. Ratti, and S. L. S. C. Zhou. A new insight into land use classification based on aggregated mobile phone data. *International Journal of Geographical Information Science*, 28:1988–2007, 2014.

[4] 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.

[5] V. Soto and E. Frias-Martinez. Automated land use identification using cell-phone records. In *HotPlanet '11 Proceedings of the 3rd ACM International Workshop on MobiArch*, pages 17–22. ACM, 2011.

[6] V. Soto and E. Frias-Martinez. Robust land use identification using cell-phone records. In *Proceedings of the 1st Workshop in Pervasive Urban Applications, in conjunction with 9th Int Conf Pervasive Computing*, 2011.

[7] J. L. Toole, M. Ulm, M. C. Gonzalez, and D. Bauer. Inferring land use from mobile phone activity. In *Proceedings of the ACM SIGKDD International Workshop on Urban Computing - URBComp*. ACM, 2012.

Targeted disease containment strategies based on spatial network analysis

Stefania Rubrichi [a], Mirco Musolesi [b], and Zbigniew Smoreda [a]

a – Sociology and Economics of Networks and Services dept., Orange Labs, Paris
b - Department of Geography, University College London

Abstract

Epidemic outbreaks are an important healthcare challenge, especially in developing countries where they represent one of the major causes of mortality. Approaches that can rapidly target subpopulations for surveillance and control are critical for enhancing containment processes during epidemics. The identification of optimal candidates for targeting preventive and containment measures entails both theoretical and computational issues. Moreover, spatial heterogeneity of risk distribution introduces new challenges in the way we describe epidemic dynamics. In particular, in the case of infectious diseases where transmission takes places because of individuals' proximity, an effective study of epidemic spreading requires to take into account both the social and the spatial dimensions of the problem, explicitly including people's mobility.

Using a real-world dataset from Ivory Coast, this work presents a method for interventions based on the socio-geographical heterogeneity of disease transmission dynamics extracted from cellular data. More specifically, exploiting structural properties of a spatial network derived from mobile phone Call Detail Records (CDRs), we investigate different target strategies based on the identification of the highest risk geographical areas or the highest risk individuals as the best candidates for isolation in order to slow down an epidemic.

The approaches presented provide in this communication further evidence that mobile phone data can effectively be exploited to model disease spreading at different geographic levels. In particular, we show that combining meta-population models with explicit spatial properties of mobility networks from CDR data is a promising approach for developing containment strategies for supporting decision-making during country-level pandemics.

*

The empirical evaluation of this work is based on mobile phone and epidemiological data. We analyzed an anonymized set of mobile phone data, Call Details Records (CDR) collected by Orange Cote d'Ivoire. It consists of billing information of about 8 million mobile phone users (i.e., 35% of the country population), collected between February and October 2014 in Ivory Coast, for a total of about 4.5 billion records.

Concerning the epidemiological data, to model realistic disease spread dynamics, we considering a scenario based on values of the parameters estimated from the Ebola outbreak in Sierra Leone in 2014.

In order to describe the countrywide-scale infectious disease spread, where individuals change location over the time, we use a meta-population model. This framework has

traditionally provided an attractive approach to epidemics modeling. It permits to include a more realistic contact structure, and to reflect the spatial separation of the sub-populations, in which case the contact rate might vary with spatial separation.

To model the process of disease transmission we consider the SEIR epidemiological model. Figure 1 shows a schematic representation of the model.

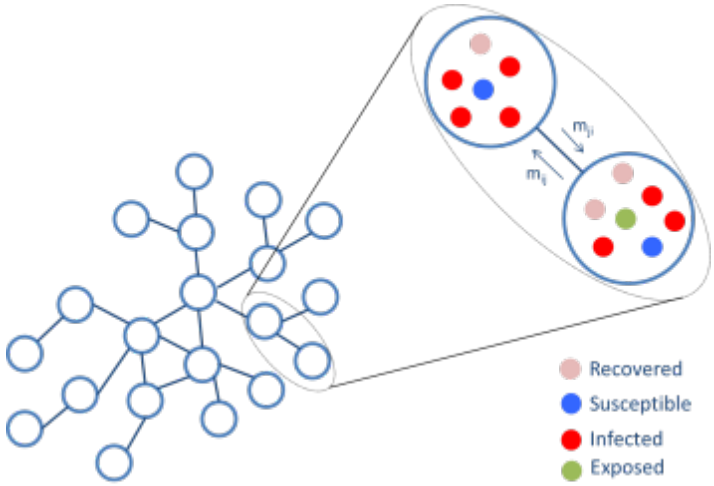


Figure 1. Schematic illustration of a meta-population epidemiological model with SEIR dynamics. On the left side, the subpopulations network consists of patches connected among them according to individuals' moving. The individuals in the patches are either susceptible (blue filled circles), infected (red filled circles), exposed (green filled circles) or recovered (pink filled circles). In each patch, the disease transmission and recovery occur according to the SEIR compartmental model. Then individuals move among patches according to specific mobility pattern (i.e., m_{ij}, m_{ji}).

Further, we investigate the chance of success when developing control strategies coverage, at both spatial and individual basis. For this purpose, we propose a method for identifying the higher-risk sub-populations as best candidate for containment strategies. In particular, we tried to estimate a measure of "spreading power" associated with a geographic area as well as with a single person, based on spatial and behavioral indicators, respectively, inferred from mobile phone data.

First, we consider the spatial targeting. We approached this problem as the identification of the influential spreader nodes within a complex spatial network. We consider, among the others, the accessibility, which have been shown able to provide valuable insights into the relationship between structure and spreading dynamics (Viana et al. 2012). We approached accessibility issue by considering a recently proposed measure for valuing the spatial interactions, namely the "place rank" (El-Geneidy & Levinson 2011).

As curbing the spread of a disease in an entire geographical region might be restrictive and somewhat unlikely, to improve the targeting process, we consider the spreading power of single person based on their mobility profiles. We investigate the effect of specific spatial behavioral indexes, linked to users' mobility, on the identification of individuals at highest risk: radius of gyration, entropy of visited sub-prefectures, and percentage of time spent at home.

We focused on the case of person-to-person transmitted diseases, where social and environmental factors (e.g., crowded setting) are primary determinants of transmission, and therefore must be considered when implementing control strategies. However, although central in describing the course of epidemics, these factors bring an intrinsic spatial variation, whose incorporation in epidemiological models remains a key theoretical challenge.

We have therefore considered the problem of the spatial localization of interactions and tried to capture and characterize the social-geographical heterogeneity of transmission. This task was carried out using two techniques which a focus on the influence of the structural features of the network on the epidemic process. In the first approach, we took into consideration geographic diversity, seeking at identifying geographic areas with the higher opportunity of contact (i.e., where the majority of exchange is likely to originate). By taking a fundamentally novel method in defining space accessibility and attractiveness, we measured the "spreading power" of the nodes in a spatial network to be isolated.

In the second approach, using a spatial-range-based mobility measure, we quantified the "spatial behavior" of single individuals as a measure of spreading risk of the nodes. Based on this, we selected a subpopulation that is expected to become infected and simultaneously infectious earlier and with higher probability than average population because of his/her mobility profile.

The results show the importance and relevant effects of the spatial dimensions on the spreading of infectious disease. While space influence has frequently reported anecdotally and there have been relatively little systematic investigations, our work tries to bridge this gap through a twofold scheme: first, it seeks to quantify and capture the associated complex diversity, then it utilizes such a diversity as predictive correlates, to enhance our ability to understanding the dynamics of the epidemiological process and defining effective targeted control interventions.

Moreover, the proposed method for identifying the most influential elements is, in our knowledge, fundamentally different than that taken by most typically used measures. On the one hand, using a flow-based measure, we sought to incorporate the dynamic of the spreading that unfolds along links into the measure itself. On the other hand, removing the direct dependence of its definition on the location of cellular towers, we could rely on a mobility measure which prevent from bias due to variations in tower density.

References:

Viana M P, Batista J L, Costa L da F, "Effective number of accessed nodes in complex networks." Phys. Rev. E: Stat Nonlin Soft Matter Phys., 2012.

El-Geneidy A & Levinson D, "Place rank: Valuing spatial interactions." Netw. Spat. Econ., 11(4), 643–659, 2011

What does mobile metadata measure? Insights from a pilot study during a sudden emergency

Carolina Mattsson
mattsson.c@husky.neu.edu
Northeastern University

Drew Margolin
dm658@cornell.edu
Cornell University

Stefan Wojcik
stefan.wojcik@colorado.edu
Northeastern University

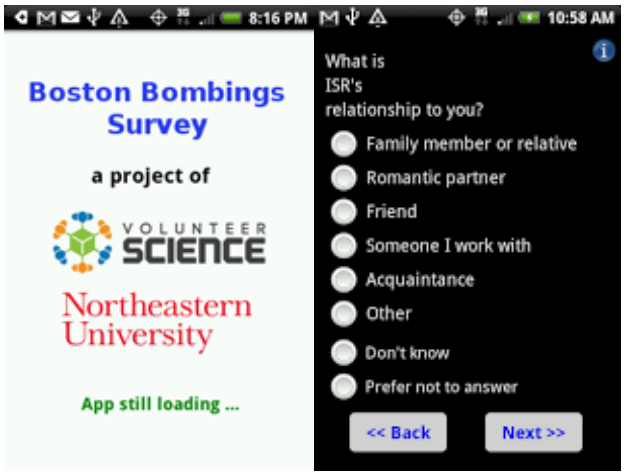
David Lazer
d.lazer@neu.edu
Northeastern University
Harvard Kennedy School

1. Introduction

Mobile data is a frontier in understanding society, and has helped create the discipline of computational social science.[1] Behavioral data from mobile phones have been used to study the societal reaction to emergencies[2,3] and many other phenomena. A smaller group of studies have sought to tie this big data revolution back to the heart of the social sciences.[4,5] Continuing to do this is crucial, since big data approaches struggle when separated from subject matter expertise.[6] There is a pressing need for calibration of our most powerful tools.

In a pilot study following the Boston Marathon bombing[7] we collected mobile records, asked traditional social network survey questions, and elicited bombing-relevant information through a smartphone app. Using this data we first attempt to classify ties into social categories based on mobile metadata, uncovering issues that illustrate the need for contextualized measurement. Tragedy provides us with an unwelcome but profound source of context. By measuring social ties when they are at their most poignant, we probe the meaning behind mobile metadata measurement. We also point out likely biases introduced by mobile metadata in studying emergency communication.

2. Data Collection



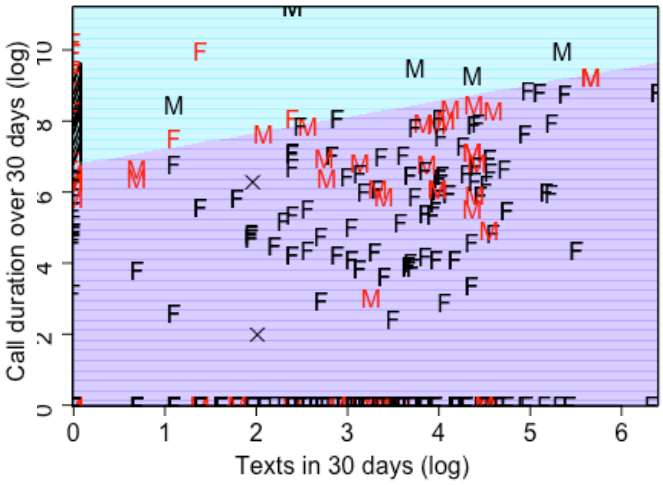
The authors created a smartphone app that accessed the call and text logs on the respondent's phone and administered a survey based on those logs. The survey first asked respondents about demographics, phone use, family composition, and details of their situation at the time of the bombings. The app then introduced into the survey up to eight mobile contacts with whom the respondent communicated after the bombing – although with a timing error of one hour. Respondents were asked each contact's gender and location at the time of the bombing, their relationship to that contact, and whether they exchanged help or information on the day of the bombing. Lastly, the app collected aggregated behavioral data over the 30 days before from the phone itself. If given additional permission, the app kept the full call and text logs.

Data was collected after the Boston Marathon bombing from 152 voluntary respondents. 87 of the respondents agreed to donate their fully detailed call and text logs. To motivate participation by those most affected by the bombing, the researchers made a

$3 donation per respondent to The One Fund Boston, the official charity for the victims. The smartphone app was available for download on Android phones for several months. These efforts created a rich and focused dataset, with 48 respondents and 97 top mobile contacts known to have been within a mile of the bombings. 29 and 56 of these, respectively, are covered by fully detailed log data

3. Results

Using the 30 days of aggregated communication history available for most respondents we explored the feasibility of classifying the two most common relationship categories: mother and friend. We find that some variables do fairly well – ex. dyads with no texts and high call volumes are overwhelmingly mothers, and vice versa. But basic classifiers only reach an out-of-sample AUC of 84, which remains below the base rate of friends (85.5%, N=531). As a classification task, social ties are a formidable challenge.



LDA classifier with apparent error 0.115
Correctly classified in black, incorrect in red.
Note neater sorting along the axes.

But is this primarily a technical issue? Wherever the highlighted partition is drawn, most of the misclassifications would be between mothers and those friends who text and talk especially much – 'best friends' perhaps. What constitutes friendship, and to a lesser extent maternal ties, is a social construct that may not map 1:1 onto measurable actions and behaviors that we care to study with mobile metadata.[4]

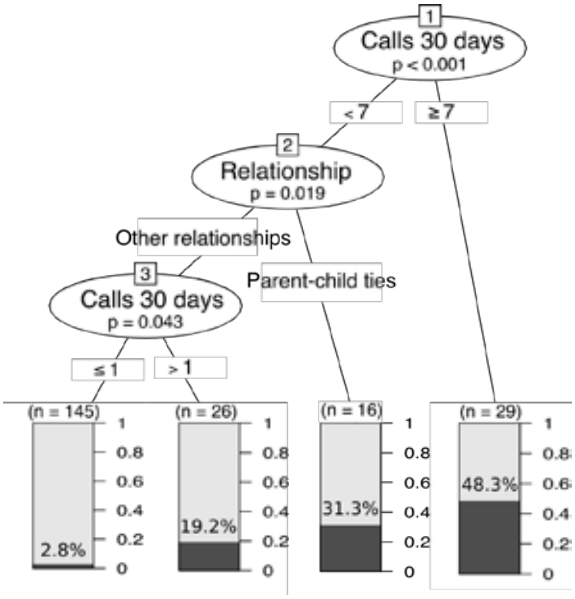
Our app collected several measurable actions and behaviors in the context of a disaster. We use classification trees where the splits are based on statistical significance to compare the predictive power of mobile metadata to traditional relationship categories. We limit our sample to the 216 dyads where the closest person was within a mile of the bombing, and where we have both mobile history and day-of mobile records.

Providing help: Our survey asked whether you gave or received help from specific contacts, thus allowing for a dyad measure of providing help. We find that traditional relationship categories do not significantly distinguish between dyads where help was provided and not. Interestingly, the *fraction of communication days* is a predictor.

Crisis communication: From the mobile records for the day of we determine whether the dyad communicated in the first 30 minutes after the bombing. We find that the most significant predictor is based on mobile history: log number of calls in the past 30 days. The second predictor is based on relationship category: parent-child ties. These are powerful predictors – both dyads who had 7 or more calls in the past month, and dyads between parents and children who had fewer calls were an order of magnitude more likely to share a call in the immediate aftermath of the bombing than other surveyed ties. Both 'mothers' and 'best friends' are likely to communicate early in this case.

We find that mobile metadata are useful in predicting mobile communication and help provision in a sudden emergency. They appear to pick up on social categories – 'strong' and proximate ties – that become poignant in this context. However, metadata does not pick up on *all* such ties, missing parent-child ties that do not call routinely.

Understanding what is and what is not captured by mobile metadata in contexts we care about is a promising avenue for further research. Our sample is small and comes from one very specific source – the Boston Marathon bombing - so specific results must be taken with a large grain of salt. Larger studies employing this context-rich methodology across different scenarios have the potential to determine more generally what mobile metadata is measuring.



Did the dyad communicate via mobile in the first 30 minutes after the bombing?
Classification tree with statistically significant splits. Bars show the fraction of dyads who did communicate early in each group.

4. Additional findings

The methodology presented here is also useful for illuminating biases introduced when using mobile metadata to study particular social processes. For example, in our pilot study we find that many, but not all, of those most affected used their mobile phones in the immediate aftermath of the bombing. Those who do use their phones appear to be oversampled from those away from their family and friends at the time of the bombing.

5. Conclusions

Methodologies such as those used here can provide much-needed context for evaluating the power of mobile metadata as a measuring tool for social ties. Results of a pilot study show cause for cautious optimism – despite feasibility issues classifying social ties, mobile metadata does pick up on real relationships that are called upon in real emergency situations. The pilot study also uncovers several sources of systematic bias that future purely 'big data' studies must contend with.

---

[1] Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A.-L., Brewer, D., … Van Alstyne, M. (2009). SOCIAL

[2] Bagrow, J. P., Wang, D., & Barabasi, A.-L. (2011). Collective response of human populations to large-scale emergencies. *PloS One*, *6*(3), e17680.

[3] Sundsoy, P. R., Bjelland, J., Canright, G., Engo-Monsen, K., & Ling, R. (2012). The Activation of Core Social Networks in the Wake of the 22 July Oslo Bombing (pp. 586–590). IEEE.

[4] Eagle, N., Pentland, A. S., & Lazer, D. (2009). Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, *106*(36), 15274–15278.

[5] Boase, J., & Ling, R. (2013). Measuring Mobile Phone Use: Self-Report Versus Log Data. *Journal of Computer-Mediated Communication*, *18*(4), 508–519.

[6] Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google flu: traps in big data analysis. *Science*, *343*(6176), 1203–1205.

[7] Lazer D ,Kennedy R, &Margolin D. (2013). Communication in the aftermath of the Boston Marathon bombing.

SESSION 10
TELCO APPLICATIONS

Predict Cellular Network Traffic with Markov Logic

Marco Lippi, Marco Mamei, Franco Zambonellli
Dipartimento di Scienze e Metodi dell'Ingegneria
University of Modena and Reggio Emilia, Italy
{marco.lippi,marco.mamei,franco.zambonelli}@unimore.it

ABSTRACT

We propose a Markov logic approach to cellular network traffic forecasting, that exploits spatial relationships between close cells in the network grid. Experimental results show the advantages of using such information.

1. INTRODUCTION

In this work we present a novel approach based on Markov Logic [2] to predict cellular network traffic across the city. As this traffic is a widely used proxy for people presence, our approach can be applied to predict crowd distribution. We focus on *aggregated* CDR data (i.e., data measuring the number number of CDRs generated from a region, without any reference to the IDs of the people being involved). This kind of data presents a number of ready-to-market applications as privacy concerns are much reduced (in contrast with CDRs with anonymized IDs). In fact, all major telecommunication companies already have services for the analysis and commercial exploitation of this data. While there are several works analyzing properties of aggregated CDR data and predicting user movements from individual CDRs [3], the task of predicting future density of CDRs from aggregated data is relatively under-explored.

2. METHODOLOGY

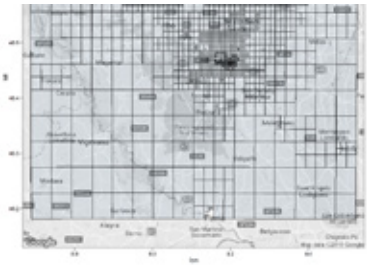
We focus the analysis on aggregated CDR data that count SMSs, calls and Internet traffic over specific areas of the city and at time intervals. Specifically, the geographic area under analysis is tessellated with an irregularly shaped grid, similar to a Voronoi tessellation. Thus, the more cell network antennas present, the denser the grid (see Figure 1a).

The resulting dataset (illustrated in Table 1b) is a set of counters estimating, for each cell of the grid and 15-minutes-interval, the number of SMSs, calls and Internet traffic. Counters can also be fractional to take into account CDR interactions originating in a cell and ending in another one. The original data comprises about 12 million records like the ones depicted. For each cell and CDR type, the typical plot resembles the one in Figure 1c (top) where it is possible to observe daily and weekly patterns in city dynamics.

To better highlight variability in our data, for each cell, we computed the mean (μ_t) of the time series in that 15 minutes interval (i.e., the mean among all days at that time) and obtain a "standardized" score by computing $\hat{x}_t = (x_t - \mu_t)/\mu_t$. The resulting time series shows the deviation from the mean of that cell at that time, in mean units (e.g., $\hat{x}_t = 1$ means that there is twice – 100% more – as many people than usual), see Figure 1c (bottom). Finally, we discretized $\hat{x}_t$ into a set of classes. Working with discrete values notably simplify computations, without compromising the actual significance and interpretability of the results.

To predict cellular network traffic, we apply Markov logic, a statistical relational learning method, to perform *collective* classification on a grid of cells. While traditional machine learning classifiers typically treat the examples as independent and identically distributed, statistical reltional learning approaches are capable of taking into account relations and inter-dependencies between the examples to be classified, so that a joint classification spanning multiple examples can be performed. In particular, we aim to exploit the spatial relationships between cells, as the nature of CDR data is inherently relational along this dimension: at time t, the traffic at two cells c_1 and c_2 spatially close in the network will typically be strongly inter-related. Inspired by the work in [1] for road traffic flow forecasting, we modeled our domain with a set of logic predicates that describe the dynamics of CDR traffic data during time and across different cells. Supposing to discretize the amount of traffic in C classes, predicates `Class0(cell,time)`, ..., `ClassC(cell,time)` can be used to indicate the fact that, at a certain cell and at a given time, the traffic quantity falls in one of such classes. A simple predicate `Neighbors(cell,cell)` indicates that two cells are spatially close in the grid. Time is modeled with predicate `Next(time,time)`. Additional information about the day of the week and the part of the day can be also easily modeled with logic predicates. Given a domain described in terms of logic facts, a Markov Logic Network (MLN) consists in a set of weighted rules that

a)



b)

Timestamp	Cell	Type	Count
1/3/2015 13:00	3943	SMS In	4.34
1/3/2015 13:00	3943	Call In	2.34
2/3/2015 13:15	3943	Network	9.34
...	...	...	...
30/4/2015 20:45	3943	Call Out	3.45

c)

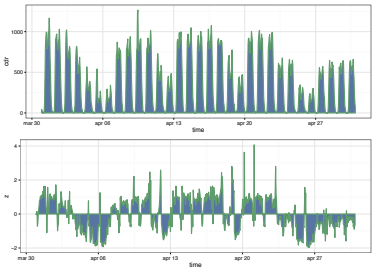


Figure 1: a) Irregular grid tessellating the area b) Example of aggregated CDR data c) Example of data in a grid cell: (top) Original behaviour extracted from mobile phone data. (bottom) "standardized" score showing the deviation from the mean of that cell at that time, in mean units.

model relationships among the existing predicates. For example, a simple predictor that forecasts, for cell c and time t, the same traffic class observed at cell c at time $t-1$, is obtained with the rules:

```
Class0(c,t1) ∧ Next(t1,t2) => Class0(c,t2)
                    ...
ClassC(c,t1) ∧ Next(t1,t2) => ClassC(c,t2)
```

Clearly, such rules are not *always* true, but they are true with a certain probability. Given a collection of observations of past events, the Markov logic framework allows to *learn* the weights of such rules directly from data. The higher the weight, the higher is a probability that the rule will be true. Once the weights of the MLN have been learned, one can use the model to compute the truth value of some query predicates. In the case of this work, the aim is to forecast the dynamics of CDR traffic in the future. Given the current state of the CDR traffic network, by performing inference over the MLN it is possible to retrieve the cell configuration that maximizes the probability of the rules in the model. In order to exploit spatial relationships, rules like the following one can be used:

```
ClassC(c1,t1) ∧ Next(t1,t2) ∧ Neighbors(c1,c2)
                => ¬ Class0(c2,t2)
```

Such rule means that, if there is a high traffic (class C) in a certain cell $c1$ at time $t1$, then at the next time step $t2$ it is unlikely that a neighbor cell $c2$ will have very low traffic (class 0). Complex relationships and dependencies can by modeled with such rules.

3. EXPERIMENTS

We focused the analysis on the province of Milan and we used aggregated CDR data collected over a period of two months: from March 1, 2015 to April 30, 2015 and including calls and SMS sent and received and network traffic. In our experiments we sum together all SMS and calls sent/received, while we do not consider network traffic (as it is expressed in a different format in our data). For each cell, at a given time t this sum is our x_t. The area under analysis is tessellated in 1,419 cells. Cell areas can range from 0.04 Km^2 in the city center to 40 Km^2 in the suburbs. Temporal resolution is 15 minutes. For each cell, we compute the standardized score $\hat{x}_t = (x_t - \mu_t)/\mu_t$ and we discretized those values in 5 classes associated to intervals: $[-\infty, 0.25]$, $[0, 25, 0.50]$, $[0.50, 0.75]$, $[0.75, 1]$, $[1, \infty]$ (i.e., the class $[0.75, 1]$ contains those values having from 75% to 100% more traffic than the mean at that time). Overall, 79% of data fall in the first class, 11% in the second one, 4% both in the third and fifth one, 2% in the fourth one.

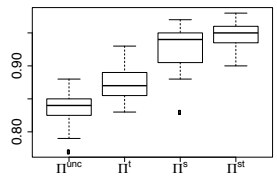
Following the work in [3], we try to establish upper bounds for the predictability of aggregate (discretized) CDR behavior across cells. We compute different entropy measures for each cell: *(i)* The random entropy $s^{rand} = log_2 N = 2.3$ (N is the number of values exhibited by the cell – all cells have 5 values). *(ii)* The uncorrelated entropy $s^{unc} = -\sum_{j=1}^{N} p_j * log_2 p_j$. *(iii)* The time (Markov) correlated entropy $s^t = -\sum_{\hat{x}_t} \sum_{\hat{x}_{t-1}} p(\hat{x}_t, \hat{x}_{t-1}) log_2 p(\hat{x}_t | \hat{x}_{t-1})$. *(iv)* The spatial correlated entropy $s^s = -\sum_{\hat{x}_t} p(\hat{x}_t, \hat{x}_t^1 ... \hat{x}_t^k) log_2 p(\hat{x}_t | \hat{x}_t^1 ... \hat{x}_t^k)$. where $\hat{x}_t^1 ... \hat{x}_t^k$ are values in neighbor cells. *(v)* Finally, the spatio-temporal correlated entropy $s^{st} = -\sum_{\hat{x}_t} \sum_{\hat{x}_{t-1}} p(\hat{x}_t, \hat{x}_{t-1}, \hat{x}_t^1 ... \hat{x}_t^k) log_2 p(\hat{x}_t | \hat{x}_{t-1}, \hat{x}_t^1 ... \hat{x}_t^k)$.

Naturally, for each cell, we will have $s^{rand} \geq s^{unc} \geq s^t, s^s \geq s^{st}$. We then compute the predictability Π associated to each entropy according to Fano's inequality [3]. This is an upper bound for any algorithm predicting $\hat{x}_t$. $\Pi^{rand} = 20\%$ for all cells. Results for other predictabilities are in Figure 2a. For example, from these results, we can infer that the upper bound for a classifier using only temporal information (1 step – 15 minutes back) is about 85% (median value). Despite these encouraging predictability results, it is worth noting that the class distribution is highly skewed (e.g., 79% of all the $\hat{x}_t$ are in one class), and thus a simple majority classifier would get very good results in terms of accuracy, according to Π^{unc}. Therefore, more careful analysis is needed.

We run experiments to test the Markov Logic predictor focusing on a subset of 23 cells[1], and we employed

[1]We chose the cells whose id contains the prefix 3943.

a)



b)

Predictor	Accuracy	Avg F_1
Random	65.7	19.8
Majority	80.3	16.1
RW	74.3	38.8
MLN	77.4	40.6

c)

	0	1	2	3	4
0	48,015	3,738	931	437	976
1	4,257	1,894	381	110	122
2	951	418	568	103	145
3	470	134	164	111	191
4	1,163	163	194	208	1,546

Figure 2: a) Predictability according to Fano's inequality associated with uncorrelated, time, spatial and spatiotemporal correlated entropies. b) Comparison of the classifiers employed in the experimental study. We report accuracy and average F_1 over the five classes. c) Confusion matrix of the MLN predictor. Rows: true labels, columns: predictions.

the first half of the data (March) for training our system, while the remaining part (April) was used as test set. For Markov logic, we used the Alchemy software,[2] training our model for 1,000 epochs with the voted perceptron algorithm. All the other software parameters were left to their default values. We compared four different predictors (see Table 2b). As a first baseline, we measured the performance of a classifier that randomly predicts one of the four classes, by drawing from a probability distribution that knows the true proportions between the classes (called Random in Table 2b). As a second baseline, we employ a classifier that simply always predicts the most frequent class (named Majority in Table 2b), that is class 0 in our case, corresponding to low traffic. As a third predictor, we use a Random Walk (RW in Table 2b) that produces as a forecast at time t the same class that was observed at time $t-1$ (for each cell independently). Finally, we employ an MLN exploiting spatial relationships between the cells. The task is to predict the status of the grid 15 minutes ahead in the future. We report both the accuracy and average F_1[3] over the five classes. As already introduced when discussing Π^{unc} predictability, the accuracy is clearly dominated by the most frequent class, that is present 80.3% of the times in the test set (which is, in fact, the accuracy of the Majority predictor), while the average F_1 gives the same importance to each of the five classes, and it is thus more significant in this setting. For example, the Majority predictor achieves the best accuracy, but actually it is a completely useless system, as it never predicts something different from the low-level traffic. It is interesting to see that the RW predictor already achieves a significant improvement over Random, thus proving to be a very strong competitor. This behavior suggests that CDR traffic has a dynamic which changes smoothly through time, and 15 minutes ahead in the future is a short horizon to observe big changes in the network configuration. Nevertheless, the MLN approach achieves better results than RW, both in terms of accuracy, and of average F_1. Table 2c reports the confusion matrix for the MLN model: rows/columns represent the true/predicted values, respectively (position i, j in the matrix indicates the number of examples of class i that are predicted to belong to class j).

[2]http://alchemy.cs.washington.edu
[3]The F_1 is the harmonic mean between precision and recall.

Figure 3: CDR traffic at time t (left) and $t+1$ (right). Traffic at cells A and B increases, following the trend in the neighboring cells.

Figure 3 shows a case study in which spatial relationships help to improve the accuracy of the predictions. The traffic classes for the cells in the network are represented for some timestamp t (left), and for the subsequent timestamp $t+1$ (right). In this scenario, the traffic in cells A and B increases (from green to yellow), which is a case where a Random Walk predictor would fail. The MLN model, on the other hand, correctly forecasts the traffic classes for cells A and B by exploiting spatial relationships, as most of the neighbors at time t belong to a high (yellow, orange or red) traffic class.

Acknowledgment

Source of the Dataset: TIM Big Data Challenge 2015, *www.telecomitalia.com/bigdatachallenge*

4. REFERENCES

[1] M. Lippi, M. Bertini, and P. Frasconi. Collective traffic forecasting. In *ECML/PKDD Proceedings, Barcelona, Spain*, 2010.
[2] M. Richardson and P. Domingos. Markov logic networks. *Machine Learning*, 62(1):107–136, 2006.
[3] C. Song, Z. Qu, N. Blumm, and A. Barabási. Limits of predictability in human mobility. *Science*, 327(5968), 2010.

Towards a data science model for device upgrade

Roberto Caporicci, Elia Polo, Dario Patane, Francesco Calabrese
Vodafone Research
Email: {Roberto.Caporicci, Elia.Polo, Dario.Patane, Francesco.Calabrese}@vodafone.com

I. Introduction

Over the last decade, especially due to the introduction of Smartphones, the flexibility of their operating systems and Social Networks, the basic phone user data requirements are increasing to an exponential rate requiring faster, more responsive and efficient network to rely on. Following, the telecommunication industry is designing ever evolving infrastructures with continuous improvements and mobile providers are adapting to the changing interest of their customer base.

It is of principal interest in the whole mobile communication industry to guarantee an extraordinary user experience by investing and give strategic importance to the latest wireless technology standards. Mobile users, in order to adopt a new major broadband update (commonly labeled generations) need a specific data plan from the provider and a compatible phone device which has the necessary baseband capabilities, leading to, in most cases, a purchase of a new mobile device from the customer. Improving the estimation of the upgrade rate on which consumers are spontaneously switching to a new wireless generation is of fundamental importance for telecommunication providers. Therefore knowing which are the crucial underlying factors to motivate a customer to switch to a more recent device is of increasing relevance.

Among the most pertinent aspects for device switchers we can certainly recognise demographics. Younger people are often more suitable to upgrade devices and are affected in a larger portion by recent trends. Country, province and wealth may represent crucial indicator for capacity and willingness to spend on handsets. Communication habits are also playing an important role in this decision making, thus amount and trends of calls data traffic might be an early indicator of future data demanding customers

Additionally, wear status, manufacturer, hardware and software characteristics of the current phone might be determinant elements to motivate a user to upgrade to a more modern one and, going beyond, the possibility to acquire a device from the same manufacturer. Classical examples are brands such Apple and Samsung which rely on a great customer loyalty for expensive phone tiers.

The main objective of this work is to understand the characteristics of users currently making use of previous wireless architectures (such as 2G and 3G) and identify groups of them with a particular tendency to switch to a new phone with 4G connectivity.

Previous work has partially addressed the problem by focusing on the operating change switch [4, 1]. In [2], a model for phone replacement based on user traffic data (as a proxy for user's economic condition) was developed. Such model however was not used for actual prediction. A more comprehensive approach was proposed in [3], which however highlights the importance of identifying a larger set of features for the prediction task.

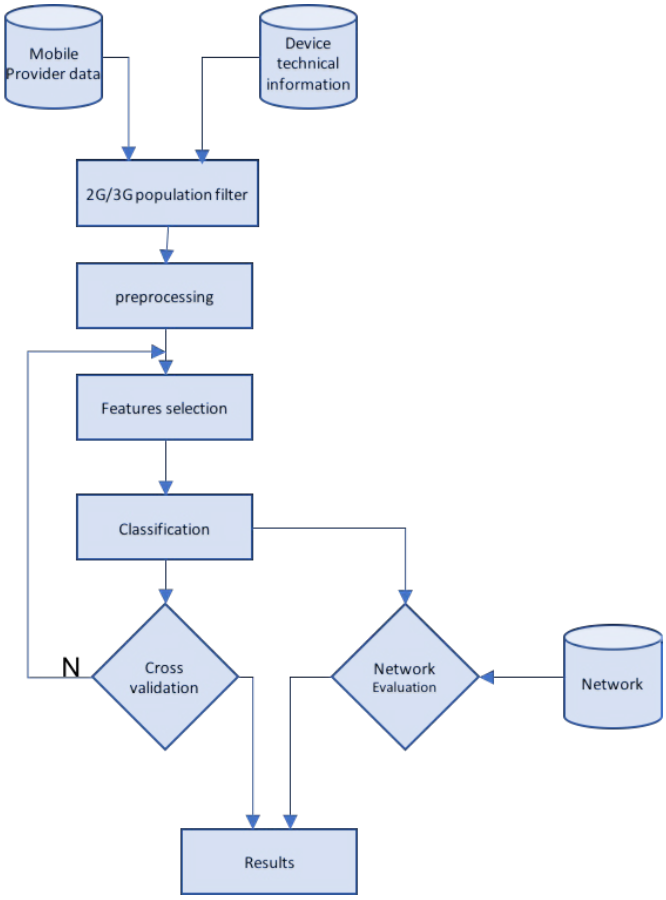


Fig. 1. Proposed data science flow for device and recommender models.

We dispose of data on million customers regarding demographics: age, gender, location, on traffic habits: trends and average values of data traffic, inbound and outbound calls and SMS, on data plan: monthly data and calls allowance, provider costs and information on the current device identified by the TAC (Type Allocation Code). We could enrich device information with characteristics such as hardware performance, software version, wireless connection capacity, price, release date and popularity. This set of features extends what was proposed in previous work, by combining both device information, device usage, and customer demographics and billing info.

The idea is to introduce a flexible and scalable solution to identify with a certain likelihood which users in a specific

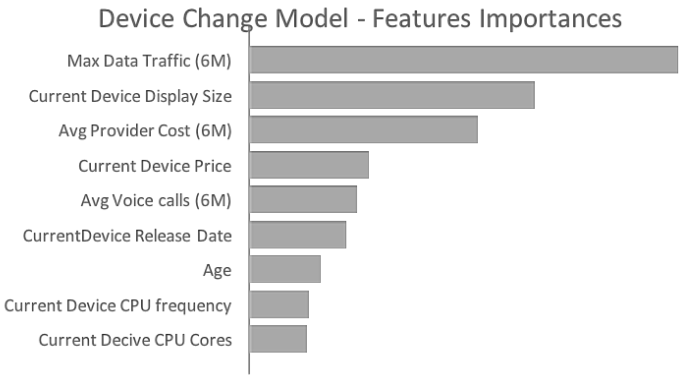


Fig. 2. Important features for the Device change model.

customer base are more inclined to switch to a latest wireless technology device in the immediate future. We thought to create two Machine Learning models in order to provide an estimation of the device switch probability (Device Change Model) and, additionally, which would be the most relevant device family of interest for each user (Device Recommender Model). Figure 1 showcases the proposed data science flow for the device change models.

II. Device Change Model

The Device Change Model has been shaped as a binary classification problem by considering attributes of a certain month (some of which have been built analyzing trends of previous months). The positive cases are determined by cases of device switch from a 2G/3G wireless device to a 4G one in the subsequent N months, where N is a forecasting window parameter. Such parameter can be tuned based on the specific business application of the model. By construction, the nature of the label is very unbalanced causing a series of problematics that have to be addressed properly.

The model has been implemented as a pipeline of data cleaning, preprocessing and standard attributes selection and it is making use of a Random Forest algorithm for classification to determine the probability of belonging to each binary class with a great capacity to scale to very large customer bases. The most important selected attributes are related to traffic behavior and current device characteristics: data usage and trend, voice calls and SMS, device price, display size and device release age have the higher roles followed by customer age, see Figure 2.

For what regards the Random Forest parameter search and evaluation we made use of a cross-validation technique and using as scoring metric the area under the ROC (Receiver operating characteristics) curve (AUC) because of its ability to deal with unbalanced class problems obtaining an AUC value of 0.80 and looking at the 10% top scored population an uplift value of 3.5.

III. Device Recommendation Model

In order to introduce the Device Recommendation Model a prior step has to be described: Device Clustering. Given

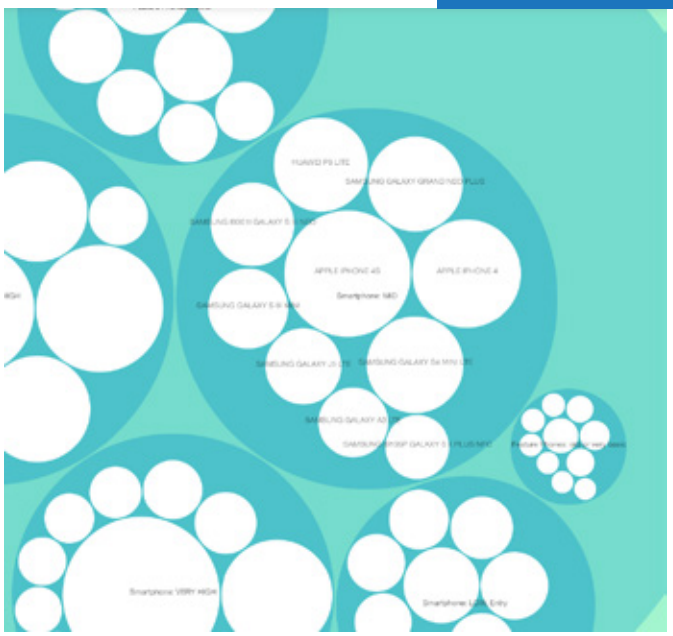


Fig. 3. A graphical representation of a subset of devices clustered based on the defined features. The circle size is related to the number of customers owning such device.

the very large number of different manufacturers and device models, we decided to cluster together very similar phones from hardware, software and capabilities prospective. Clustering attributes are including: Wireless Technology, release date, CPU/RAM, operating system, popularity, price and camera features. The clustering method used was K-Means with a few number of clusters in order to give also a commercial meaning to them (e.g. Feature Phone, Entry-level Smartphone, Mid-level smartphone, High, etc.). A graphical representation of devices clustered based on the defined features is shown in Figure 3.

Differently to the Device Change Model the Device Recommender Model has been implemented by several binary classification problems (one for each cluster) using as positive label the event of switching to a device belonging to such cluster and on input the same attributes of the Device Change with the addition of the cluster identifier of the current device. The considered population has been filtered to a subset for which a phone switch has been observed, reducing substantially the data points for training and evaluation. The reason for this choice is to train the model only with relevant information just about positive change cases. The most important features for the device recommender models are shown in Figure 2.

Equivalently to the Device Change Model the evaluation and parameter tuning of each binary model have been determined by cross-validation and an AUC value which scored with a weighted average on cluster popularity around 0.80.

As an additional evaluation step the Device Change Model has been trained with the most recent data and we checked the model quality on Network data by observing the real devices switch at the present time by looking to the complete customer base who switched device from 2G/3G to 4G and putting into relation to the positive event the likelihood previously

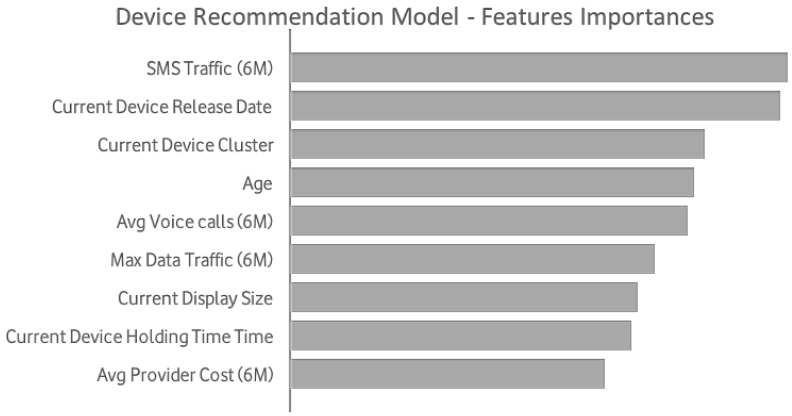


Fig. 4. Important features for the Device recommendation model.

computed by the model.

By looking at the top 13% of scored population by the model a lift of 2.73 has been observed looking to a future time window of one month. This initial result shows that a supervised learning approach to build a forecasting model to predict the population willing to adopt a new wireless technology device might be really effective if we consider together demographics, phone traffic and current device attributes.

The proposed data science models have applicability in different scenarios. From one side, they can be used to perform targeted marketing campaigns toward customers with high propensity for device change. Moreover, they could be used to identify the right level of device orders in different stores based on the predicted device preferences and replacement rates. These aspects will be investigated in our future work.

REFERENCES

[1] S. Böhm, F. Adam, and W. C. Farrell. Impact of the mobile operating system on smartphone buying decisions: A conjoint-based empirical analysis. In *International Conference on Mobile Web and Information Systems*, pages 198–210. Springer, 2015.

[2] J. Liu, Z. Lei, L. Chen, and Y. Zhou. Understanding how users change their mobile phones by massive data analysis. In *2015 7th International Conference on Intelligent Human-Machine Systems and Cybernetics*, volume 1, pages 232–237, Aug 2015.

[3] Q. Ma, S. Zhang, W. Zhou, S. Yu, and C. Wang. When will you have a new mobile phone? an empirical answer from big data. *IEEE Access*, 4:10147–10157, 2016.

[4] Y. Wang, H. Zang, P. Devineni, M. Faloutsos, K. Janakiraman, and S. Motahari. Which phone will you get next: Observing trends and predicting the choice. In *2014 IEEE Network Operations and Management Symposium (NOMS)*, pages 1–7, May 2014.

On Added Value Of Feature Engineering for Churn Prediction

Sandra Mitrović
KU Leuven, Belgium
sandra.mitrovic@kuleuven.be

Bart Baesens
KU Leuven, Belgium
bart.baesens@kuleuven.be

Wilfried Lemahieu
KU Leuven, Belgium
wilfried.lemahieu@kuleuven.be

Jochen De Weerdt
KU Leuven, Belgium
jochen.deweerdt@kuleuven.be

I. Introduction

Churn prediction (CP) in telecommunications has been a topic of active research for more than a decade, with more and more studies experimenting with different explanatory variables (features), especially recently, as (social) network analytics becomes more popular and call detail records (CDRs) are being used to construct customer networks and extract network features [1], [2], [3], [4]. Reported results show that network and traditional features combined perform better than traditional (also called local or individual) ones alone. Additionally, some studies combine traditional, network and historical information and claim such approach to be even more beneficial [2]. Obviously, adding more features leads to better performance, but distinguishing between feature types is important for various reasons. First, some studies claim that network features alone do not express enough predictive power [1], while others claim the opposite [3], [4]. Second, since previous benchmarking studies have shown a rather flat maximum effect in terms of performance of analytical techniques for CP [5], the best way to improve the performance of a CP model is to creatively think about new feature types as showcased in this research. Third, due to differences in the computational burden needed to operationalize and calculate each of the different types of features (e.g. network features are a lot more resource intensive to obtain than local ones), setting off the performance difference provides strategic input to firms taking their first steps in CP by indicating what feature types to invest in. To the best of our knowledge, previous studies have not performed as thorough analysis on the importance of each particular feature type. Hence, in this work, we introduce a more fine-grained categorization of features and perform an experiment of stacking different feature types (adding one more at every step) with the aim to investigate the influence that each particular feature type has on the resulting performance and retained feature set. Our results show that, in general, certain types of network features (indirect ones) and trend features do not improve performance in the setting at hand. Additionally, local and network features are of different importance for prepaid and postpaid datasets.

II. Datasets

We use one prepaid and one postpaid dataset, both consisting of CDRs and monthly snapshots. CDRs contain only calling information, while snapshots mostly contain information about customer handset characteristics, the date of the first usage and tariff plan information, scarce demographic data (zip code for both datasets and the date of birth for postpaid dataset). Additionally, for prepaid customers, we also have information about last recharge and monthly amount spent on voice and SMS services. For postpaid customers, we have an indicator whether the customer has a fixed-time contract and if so, the number of days left till its expiration.

III. Methodology

Our fine-grained classification consists of ten different feature types, whose classification can be seen in Table I, with the hierarchy represented in Table II (abbreviations can be used to match the two tables; labels 'engineered'('E')/'non-engineered'('NE') serve to distinguish features requiring/not-requiring additional computational effort). As depicted in Table II, the first division is made on observational (the most recent month with customer base for which we make prediction) and historical (the previous month) features, due to the fact that historical data is not always available. Next, as in many related studies, we distinguish between local and network features. Network features can be further categorized into direct and indirect, an idea from the fraud detection domain [6]. Finally, as the same features are calculated for two consecutive months, a non-trend and trend classification imposes as a rational follow-up. Definitions provided in Table I, can be applied to any type of dataset (even beyond the Telco domain). On the contrary, some examples, especially if related to L-E/L-NE classification, are dataset dependent (e.g. age, can be considered as L-NE/L-E when provided in the dataset/calculated from the date of birth).

For the prepaid dataset, we construct the ten above mentioned feature types, ending up with 112 features in total. For the postpaid dataset, we omit the HLT-E feature type, since our dataset has no **continuous** local non-engineered features. The total number of features for postpaid is 94. Hence, we perform ten (for prepaid) and nine (for postpaid) iterations, where on each step we extend the feature space for one more feature type set, in the same order as presented in Table II.

Due to the fact that our datasets do not contain customers' churn date, we define churn as usage inactivity in consecutive calendar month (similar to approaches in [1], [4]). Additionally, we opt for predicting churners in month $M+2$ (with M denoting the observational month) as we aim at detecting

TABLE I
FEATURE TYPE DEFINITION

Abbrev.	Feature Type Name	Definition	Example
L-NE	Local Non-Engineered	Customer based features which do not require any further preprocessing (model-dependent pre-processing is not considered)	gender; handset characteristics (e.g. POLYPHONIC); number of reloads
L-E	Local Engineered	Customer based features for which some kind of model-independent preprocessing is necessary	number of days being a customer derived from customer activation date
ND-E	Network Direct Engineered	Features calculated from the customer ego-network (1^{st}-level neighbourhood) across different dimensions/granularities	number of calls/outgoing calls/outgoing calls toward home network in month M
NI-E	Network Indirect Engineered	Network features which cannot be calculated from customer ego-network only	Page Rank score; 2^{nd} degree of a node in month M
HLNT-NE	Historical Local Non-Trend Non-Engineered	The same type of features as L-NE, except that they refer to the one month before the observed month (month $M-1$)	handset characteristics; number of reloads in month $M-1$
HLT-E	Historical Local Trend Engineered	Trend features calculated based on local non-engineered (L-NE) and historical local non-engineered (HLNT-NE) variables	recharge amount trend between months M and $M-1$
HNNTD-E	Historical Network Non-trend Direct Engineered	The same type of features as ND-E, except that they refer to the one month before the observed month (month $M-1$)	number of calls/outgoing calls/outgoing calls toward home network in month $M-1$
HNNTI-E	Historical Network Indirect Engineered	Historical (one month before i.e. month $M-1$) versions of the NI-E variables	Page Rank score; 2^{nd} degree of a node in month $M-1$
HNTD-E	Historical Network Trend Direct Engineered	Trend features calculated based on direct network features corresponding to month M (ND-E) and month $M-1$ (HNNTD-E)	trend in number of incoming calls between months M and $M-1$
HNTI-E	Historical Network Trend Indirect Engineered	Trend features calculated based on indirect network current (NI-E) and indirect network historical (HNI-E) features	trend in 2^{nd} degree of a node between months M and $M-1$

TABLE II
HIERARCHY OF FEATURE TYPES

Hierarchy				Abbrev.
Observational	Local	Non-Engineered		L-NE
		Engineered		L-E
	Network	Direct		ND-E
		Indirect		NI-E
Historical	Local	Non-Trend		HLNT-NE
		Trend		HLT-E
	Network	Non-Trend	Direct	HNNTD-E
			Indirect	HNNTI-E
		Trend	Direct	HNTD-E
			Indirect	HNTI-E

churn in a timely way by focusing on early warning indicators (we address this as the "one-month gap" approach), instead of just predicting churners in the next month ($M+1$), in which case our models would also detect those churners who already made up their mind and are essentially lost causes not worth targeting with a churn prevention campaign.

We convey our experiments using logistic regression (LR) with stepwise feature selection. Our motivation for choosing LR as the churn classifier is twofold. First, previous benchmarking studies have empirically confirmed its performance [7]. Furthermore, since LR is a white-box classifier, it also facilitates interpretation which is key to the development of churn prevention campaigns. For the stepwise selection method, we start from the whole set of features and alternately exclude and include one feature at a time. The stopping criterion is based on the AUC score and a predefined threshold $t(=1\%)$: we attempt excluding/including features as long as the AUC score either improves or does not get relatively lower than 1% both from the initial AUC score (with complete feature set) and the AUC score of the previous iteration. This procedure enforces exploration of various feature combinations and obtaining the least number of features while still retaining an AUC score closed to the initial one.

IV. RESULTS

The distribution of the retained features across different iterations can be seen in Fig. 1 and Fig. 2, for prepaid and postpaid datasets, respectively. It is important to emphasize that each iteration involves all preceding feature types as well, e.g. the iteration denoted with NI-E involves all features belonging to the L-NE, L-E and ND-E feature types as well. Each feature is uniquely represented by a symbol, determined by its colour and shape. Each colour corresponds to a particular feature type and has the same meaning both for prepaid and postpaid (e.g. blue for ND-E and red for HNNTD-E). Different shapes correspond to different features (and these, unlike colours, have different meaning for prepaid and postpaid). The presence of the same shape in different colour indicates that the same feature was present (and retained) in different iterations (typically both observational and historical), e.g. for prepaid, the recency of the outgoing calls toward the customers of the home operator (denoted with left triangle) is retained both in the observational (for NI-E and HLT-E) and historical version (NHTI-E).

It can be observed that for both datasets, only features of five feature types are always retained after feature selection: L-NE, L-E, ND-E, HLNT-NE, HNNTD-E.

In case of prepaid, network features (both observational: HLNT-NE and historical: HNNTD-E) appear to have the most important role: once introduced, some of the features of this type will always remain in the selected set of features. We can also observe a limited effect of the local feature types (both L-NE and L-E), which fade away with inclusion of other feature types and eventually disappear when historical features are introduced. Local historical features have even more restricted influence since they disappear as soon as historical network features are introduced, even though one of them (the date of first monthly reload), reappears in the last iteration.

In case of postpaid, L-NE features are much more dominating and at least one of them is retained throughout all the iterations. Although this might look surprising at the first sight, the most frequent L-NE feature retained in our case is the number of days remaining till fixed-time contract expiration (denoted by a '+' sign), which is actually consistent with reality: unlike prepaid customers which have no restrictions and can freely behave based on their interactions influence, the behaviour of postpaid customers is heavily dependent and steered by their individual contracts as well. We can also observe that for postpaid, historical features (both HLNT-NE and HNNTD-E) are always retained, once introduced. The effect of ND-E features is still present, although they disappear when historical network features are added.

Obtained results indicate a clear difference between postpaid and prepaid datasets, which is not surprising.

Due to differences in AUC performances for different iterations (see Fig. 3), we aim at verifying if these differences are significant. For this, we follow the approach from [8] and apply bootstraping (resampling 1000 times). Next, we apply a non-parametric Friedman test. The resulting p-value of the Friedman's test for the postpaid dataset is $0,5469\,(>0,05)$, which is not significant. The obtained p-value for the prepaid dataset is $0,0002\,(<0,05)$, which indicates that the differences between feature sets in this case are significant. To discover ex-

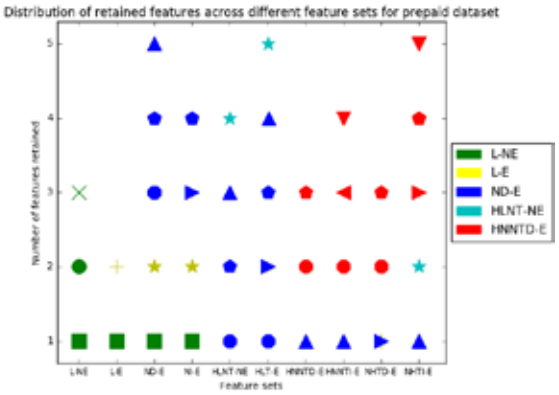


Fig. 1. Distribution of retained features across iterations for prepaid dataset

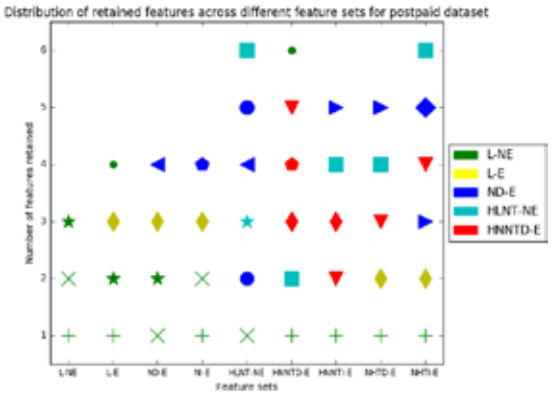


Fig. 2. Distribution of retained features across iterations for postpaid dataset

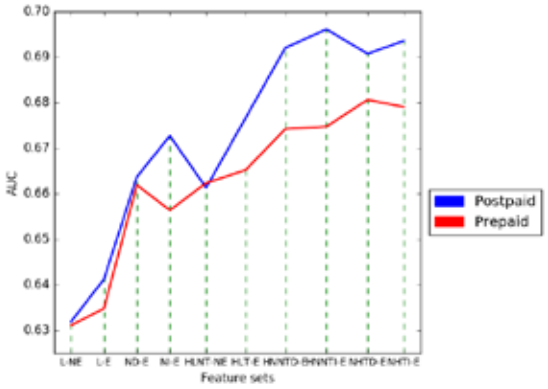


Fig. 3. AUC changes over different iterations for prepaid and postpaid datasets

actly which pairs of feature sets differ, we apply the Nemenyi post-hoc test. Fig. 4 shows that there are significant differences between several feature sets for the prepaid dataset. It is worth noting that feature set ND-E is significantly different from all other feature sets except NHTD-E and NHTI-E at significance level of 95%.

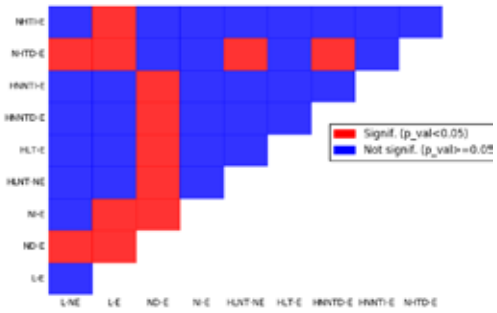


Fig. 4. The results of post-hoc Nemenyi test for prepaid dataset

V. Conclusion and Future Work

In this work, we simulate the usual process of adding different feature types (but on a more fine-grained level) and detect which feature types' feature(s) remain among the most important ones in the final model. Obtained results show that local and direct network features both in present and historical version dominate in the collection of retained features, although local features are more dominant for postpaid, while network ones are more significant for prepaid. Indirect network and trend features seem not to be of much importance.

We are aware that the one-month gap approach lead to lower AUC than usually reported. For future work, we plan to experiment with different ways of stacking features (for example, from local to historical and then to network) to check whether it would lead to the same (type of) retained features.

References

[1] P. D. Kusuma, D. Radosavljevik, F. W. Takes, and P. van der Putten, "Combining customer attribute and social network mining for prepaid mobile churn prediction," in *Proc. the 23rd Annual Belgian Dutch Conference on Machine Learning (BENELEARN)*, 2013, pp. 50–58.

[2] X. Zhang, J. Zhu, S. Xu, and Y. Wan, "Predicting customer churn through interpersonal influence," *Knowledge-Based Systems*, vol. 28, pp. 97–104, 2012.

[3] K. Kim, C.-H. Jun, and J. Lee, "Improved churn prediction in telecommunication industry by analyzing a large network," *Expert Systems with Applications*, vol. 41, no. 15, pp. 6575–6584, 2014.

[4] A. Backiel, B. Baesens, and G. Claeskens, "Mining telecommunication networks to enhance customer lifetime predictions," in *International Conference on Artificial Intelligence and Soft Computing*. Springer, 2014, pp. 15–26.

[5] W. Verbeke, K. Dejaeger, D. Martens, and B. Baesens, "Customer churn prediction: does technique matter?" in *Proceedings of the Joint Statistical Meeting, JSM2010, Vancouver, Canada*.

[6] B. Baesens, V. V. Vlasselaer, and W. Verbeke, "Social network analysis for fraud detection," *Fraud Analytics Using Descriptive, Predictive, and Social Network Techniques: A Guide to Data Science for Fraud Detection*, pp. 207–278, 2015.

[7] B. Baesens, T. Van Gestel, S. Viaene, M. Stepanova, J. Suykens, and J. Vanthienen, "Benchmarking state-of-the-art classification algorithms for credit scoring," *Journal of the operational research society*, vol. 54, no. 6, pp. 627–635, 2003.

[8] G. Loterman, M. Debruyne, K. V. Branden, T. Van Gestel, and C. Mues, "A proposed framework for backtesting loss given default models," *The Journal of Risk Model Validation*, vol. 8, no. 1, p. 69, 2014.

Layers of social acquaintance for telecommunication churn prediction

Davide Cellai
Idiro Analytics, Clarendon House, 39 Clarendon Street,
Dublin 2, Ireland
davide.cellai@gmail.com

James P. Gleeson
MACSI, Department of Mathematics & Statistics,
University of Limerick, Co. Limerick, Ireland

January 15, 2017

Telecommunication churn is the phenomenon where subscribers of a phone operator leave the service. Predicting churn is a very important topic in the telecommunication industry, especially in mature markets where the number of subscriptions exceeds, sometimes quite substantially, the country population [2].

In this work, we propose a new predictive model of telecommunication churn based on social network approaches, called m-*exposure model* [1]. We use a data set of anonymized mobile phone calls from an operator with a large segment of the considered market. The records report all outgoing and incoming calls, text and data of about 1.4 million subscribers, and the data set spans over 12 weeks.

Our model is based on two ideas: a multilayer description of different social acquaintances to identify the layer where social influence is largest, and a social threshold model of churn.

First, we give a score to each edge in the network that differentiates ties based on the time of the day communication occur. In this way, we identify distinct layers of communications in the network and show the existence of a layer (called *mixed layer*) that carries most of the socially relevant information. Then, we use this information to develop a threshold model based on the network proximity of churning events. We also introduce a parameter that encapsulates the delay between the observation data and the prediction time interval.

We apply the model to port out churn (churners that move to another operator) and calculate the performance of the model in terms of churn probability and lift. The churn probability increases remarkably on the mixed layer, meaning that our method correctly individuates links that are more likely to lie on a predictive churn path (Fig. 1).

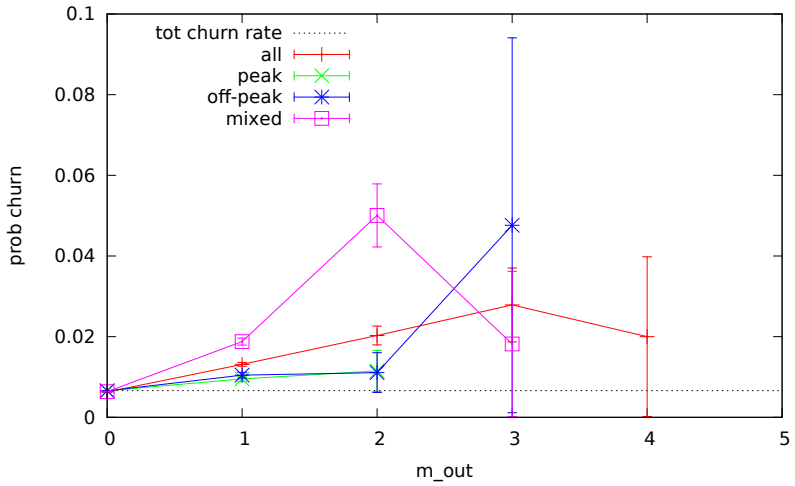


Figure 1: Churn probability as a function of the number of recent churning neighbours in a voice-based network. The graph aggregation time interval is 8 weeks. The plots obtained by restricting contacts to one of the three layers are also reported. It is easy to see that the subscribers exposed to churn on the mixed layer have a higher churn probability than on the other layers.

We also define lift as the ratio between the fraction of correctly predicted churners and the fraction of churners predicted with a random method:

$$L = \frac{c_m}{p_{ch} n_m}, \tag{1}$$

where p_{ch} is the total fraction of churners in the considered time interval and c_m is the number of churners in the considered subset of subscribers (of size n_m).

We show that our m-exposure model performs significantly better on the mixed layer, increasing the lift from 2.34 (as in the aggregated network) to 3.57, without reducing too much the set of captured churners. We also discuss the effect of other network-related features such as the exposure to traffic to other operators and temporal clustering of subscribers' actions.

We compare the model with EDIX, another social network-based model of churn, and find similar performance ($L = 3.2$ for m-exposure and $L = 3.5$ for EDIX), but, remarkably, a small overlap in the predicted churners, meaning that the m-exposure model is able to capture high risk subscribers that were not captured by other social network models.

Finally, we argue that this mechanism to identify social relationships in a mobile phone network can be also applied to expiry churn, with a model based on the inactivity time of a subscriber's nearest neighbors, and other types of users' behaviours.

References

[1] D. Cellai and J. P. Gleeson. Layers of social acquaintance for telecommunication churn prediction. (to appear), 2017.

[2] Derek Doran and Veena Mendiratta. Propagation models and analysis for mobile phone data analytics. In Dariusz Król, Damien Fay, and Bogdan Gabryś, editors, *Propagation Phenomena in Real World Networks*, volume 85 of *Intelligent Systems Reference Library*, pages 257–292. Springer International Publishing, 2015.

LEARNING ABOUT NEW TECHNOLOGIES: EVIDENCE FROM PHONE PLANS IN RWANDA

DANIEL BJÖRKEGREN*
BROWN UNIVERSITY

Although the spread of new technologies is vital for economic development, it is difficult to study with traditional sources of data. The mobile phone represents a new technology which automatically records every potential learning experience, and nearly every remote interaction with peers who could share their own learning experiences. In 2006, a Rwandan mobile phone operator introduced a new plan that represented substantial savings for over 85% of subscribers; however, it took years to diffuse. This project uses operator data to investigate how individuals learned about this new plan.

I exploit several features of the data. First, subscribers learn about call charges by requesting their balances between transactions; I observe these balance inquiries and thus the generation of information. Second, I exploit the fine timing of the data to isolate actions that follow the generation of information. Finally, I use data on the network of subscribers to identify peer effects. Balance-conscious peers helped diffuse information about the plan, and it appears they communicated rich information about the relative prices.

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

1. Introduction

Learning about new technologies is vital for economic growth. However, it is difficult to study: seldom is it possible to gather data rich enough to both describe behavior and differentiate between channels of learning. The mobile phone is an economically important technology that has seen widespread adoption. It has the relatively unique feature that operators must maintain a log of all actions taken on the network in order to provide service. These records provide a realtime window into how individual subscribers learn to use the network. This project uses rich data from a mobile phone network to track how subscribers learn about a new cheaper phone plan, to determine how profitable technologies diffuse through society.

This setting has three features beneficial for studying learning over networks.

First, mobile phone networks provide a rich, passively collected source of data on both actions and social networks. The data represents nearly the entire network of mobile phone subscribers. The temporal resolution of the data makes it possible to use micro-level event studies for identification: it is possible to determine whether actions preceded or followed the provision of information.

Second, subscribers are making economically significant decisions. Telephony represents 5% of expenditure in subscribing households.[1] Plan choice itself is not a trivial decision: on average, subscribers saved almost half of their spending on domestic calls by switching to the newly offered plan; they could have saved approximately 10% more by switching earlier. The operator introduced the plan as part of a sequence of billing changes that made the network more attractive to poor consumers. Although the introduction of the plan also affects operator revenues, this paper focuses on consumers, for which the new plan is analogous to a new technology with large but heterogeneous benefits.

Third, the context is information-poor: the mobile phone network in this country represents the vast majority of remote communication apart from radio. This feature makes it easier to distinguish between confounding sources of information relative to settings with more complex communication networks.

2. A New Plan

Almost all subscribers have prepaid plans, with a balance that is depleted with use. This balance can be refilled by purchasing an airtime scratch card from any of the operator's agents, who are located throughout the country. The party that initiates a call pays for the call; receiving a call is free.

Initially, the operator offered one prepaid plan ("PerMin"), which was billed by the first minute and then every 30 seconds thereafter. As the network expanded to reach poorer consumers, the operator made a sequence of changes in billing policy that reduced prices. The most notable of these was the introduction of a plan billed by the second ("PerSec"). The pricing of the new plan was set so that a 45 second call cost the same under either plan; shorter calls were cheaper under PerSec. The top panel of Figure 1 compares pricing between the two plans.

Most subscribers make very short calls (the median call length is 22 seconds), and as a result PerSec was cheaper for most subscribers. The bottom panel of Figure 1 shows the distribution of call durations. Over all domestic transactions, PerSec was cheaper for 97% of subscribers. Even the charges incurred while subscribed to PerMin would have been cheaper under PerSec for 85% of subscribers. The savings associated with switching were substantial: switchers

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

Revision December 28, 2016. Preliminary and incomplete. I am grateful to Michael Kremer, Greg Lewis, and Ariel Pakes for guidance and encouragement. Thank you to Nathan Eagle for providing access to the data, computing facilities, and helpful conversations. In Rwanda, I thank the staff of my telecom partner and government agencies. This work was supported by the Stanford Institute for Economic Policy Research through the Shultz Fellowship in Economic Policy.

[1] According to the government's household survey, EICV 2010-2011.

FIGURE 1. Price schedule and density of observed calls by plan type, 1.2005-1.2008

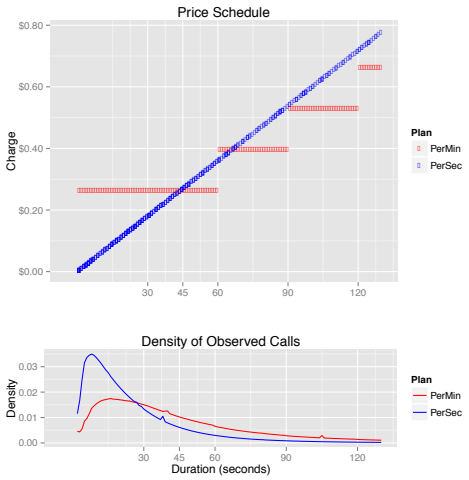
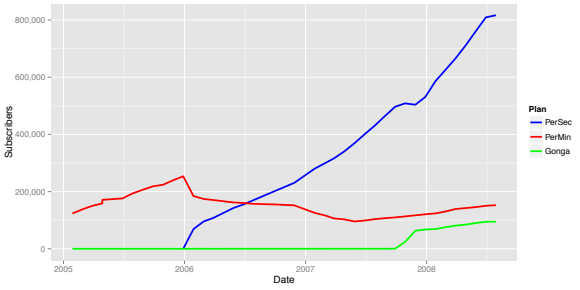


FIGURE 2. Subscribers by plan and date



saved on the order of half of their domestic calling expenditure.

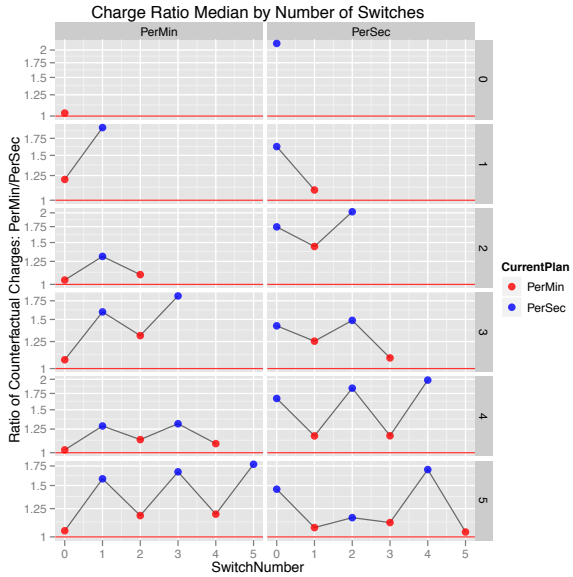
The process of switching was simple and did not incur a fee: to switch plans, a subscriber would enter a short code into the phone, not dissimilar from the codes entered when topping up airtime. Thus in order to switch, a subscriber must know both that a different plan is available and the short code to enter into the phone.

Despite the large savings associated with switching to per second billing, it took years for the plan to diffuse through society, as shown by the evolution of subscriber counts in Figure 2.

3. ANALYSIS

Agents matter. Roughly half of switches occur immediately following a top up, suggesting the operators' agents are involved in the switching process, either by providing information or by lowering the switching cost.

FIGURE 3. Switch Patterns: Median Charge Ratio by Current Plan



A separate graph is drawn for subscribers of different initial plans and total number of switches. Subscribers change behavior upon switching, but PerSec is still optimal for the median.

Subscribers settle on the optimal plan, which for most is PerSec. Plans can be switched multiple times by entering a short code, at the user's convenience. We might expect subscribers to switch between plans multiple times for two reasons: subscribers may switch to learn about which plan is best for them, or to reduce costs on calls of a certain type (e.g., one might switch to PerMin before making a long call). A not insignificant number of subscribers switch back and forth between plans; those who do mostly settle on PerSec. Subscribers adjust their behavior in tandem with switching: they make shorter calls when subscribed to PerSec and longer calls when subscribed to PerMin.

Figure 3 shows the median charge ratio by the current plan for subscribers with different sequences of plans. Subscribers make calls that are cheaper under PerSec when subscribed to PerSec (higher ratio), and calls that are cheaper under PerMin when subscribed to PerMin (lower ratio). However, the most striking feature of the graph is that all of the median ratios lie above 1, denoted by the red line: the median subscriber would be better off simply sticking with PerSec rather than engaging in complex switching behavior. In fact nearly the entire distribution of subscribers would be better off sticking to PerSec than switching back and forth.

Peers matter. To analyze the peer effects, I estimate the hazard rate of switching with regressions.

I find that an increase in the fraction of my social network neighbors who have switched from 0% to

100% is associated with between 38-67 percentage point increase in my hazard rate, including controls.

Subscribers who would have saved more by switching based on past usage are more likely to switch. I am more likely to switch when I have more neighbors who have switched; however, this could be due to learning from peers or due to confounding homophily or correlated shocks.

We can control for the most obvious form of homophily: I am likely to be clustered with others who make similar types of calls, and thus would obtain similar benefits from switching. Indeed I find I am more likely to switch when my neighbors would have saved more by switching (based on usage while they are subscribed to PerMin). However, when I include the direct peer effect as well as a control for neighbors' savings, the peer effect barely moves.

Peers provide information. Information about plans is generated either by the operator (and shared through agents or marketing material), or through direct experience on the network. Since accounts are prepaid, subscribers receive no monthly bill; instead, the precision of the information an individual gets from experience on the network depends on how much feedback he requests from the network. If an individual submits a balance inquiry after every transaction, he can learn from every transaction; if an individual seldom requests balance information, he will have much less precise feedback about the costs of his usage. The model would thus predict that the more precise information that an individual's neighbors generate, the better his decision. This also suggests a placebo test: when the old plan is sufficiently well understood, only the precision of neighbors who have switched to the new plan should matter.

I build on the previous regressions by controlling for how much information is generated by neighbors who have switched, the ratio of balance inquiries to calls. This information control absorbs some of the variation that had been explained by the fraction of neighbors switched, suggesting that at least part of the peer effect could operate through transmission of information learned by experience. But if I add the corresponding control for neighbors who have not switched (who cannot generate any information about the new plan from their experience) I find it to be near zero.

Switchers reoptimize immediately for the new pricing schedule. The pattern of calls sent adjusts immediately following a switch. Figure 4 shows that the median call duration drops immediately following a switch, consistent with subscribers reoptimizing behavior for the new plan (since shorter calls are cheaper). This suggests that subscribers are aware of the pricing policy under PerSec at the point of switching. It suggests that the information being shared is not simply that the new plan is cheaper, but a richer description of the billing difference.

FIGURE 4. Subscribers immediately adjust their calling to take advantage of the different prices under PerSec

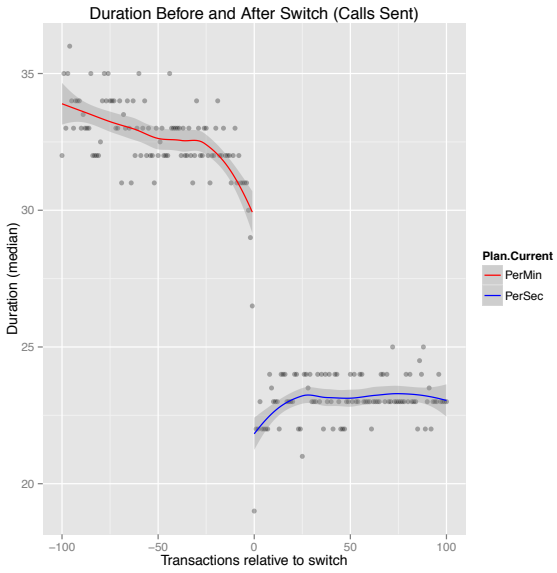
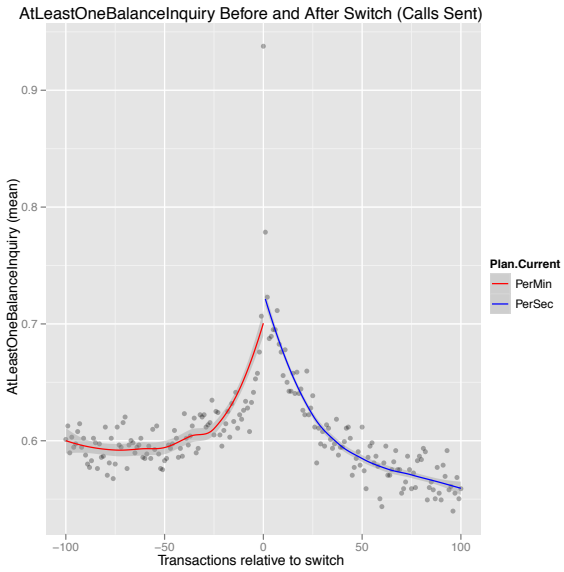


FIGURE 5. Subscribers check their balance frequently as they are switching plans



A further check is the pattern in balance inquiries. If subscribers are learning about prices under the new plan, we would expect to see a jump in balance inquiries immediately following a switch. In Figure 5, looking at the fraction of calls that were preceded by at least one balance inquiry, do we see a jump, but the increase starts before the switch.

For more details, find the full paper at http://dan.bjorkegren.com