NetMob 2010

Workshop on the
Analysis of Mobile Phone Networks

MIT, Cambridge, USA
May 11, 2010

Program and Book of Abstracts

Edited by: Vincent Blondel and Gautier Krings

Organizing Committee:

Chair: Vincent Blondel	UCLouvain (Belgium)
Francesco Calabrese	Senseable City Lab, MIT
Gautier Krings	UCLouvain (Belgium)
Benjamin Waber	Media Lab, MIT

Scientific Committee:

Chair: Vincent Blondel	UCLouvain (Belgium)
Laszlo Barabási	Northeastern University
Rob Claxton	British Telecom (UK)
Vittoria Colizza	ISI Foundation (Italy)
Massimo Colonna	Telecom Italia (Italy)
Nathan Eagle	Santa Fe Institute
Alexandre Gerber	AT&T Research
Marta Gonzales	MIT
Cesar Hidalgo	Harvard University
János Kertész	Budapest University of Technology (Hungary)
Renaud Lambiotte	Imperial College (UK)
David Lazer	Northeastern University
Jure Leskovec	Stanford University
Nuria Oliver	Telefonica Research (Spain)
Jukka-Pekka Onnela	Harvard University
Asu Ozdaglar	LIDS, MIT
Alex (Sandy) Pentland	Media Lab, MIT
Mason Porter	University of Oxford (UK)
Carlo Ratti	Senseable City Lab, MIT
Jari Saramäki	Helsinki University of Technology (Finland)
Leonardo Soto	AirSage
Zbigniew Smoreda	Orange Labs (France)
John Tsitsiklis	LIDS, MIT
Paul Van Dooren	UCLouvain (Belgium)

NetMob 2010

Workshop on the

Analysis of Mobile Phone Networks

PROGRAM

8:45-9:15 Registration

Session A (9:15-10:45)

9:15-9:30 Workshop overview: trends in mobile phone network analysis
Vincent Blondel, UCLouvain

9:30-9:45 Age, Gender and Communication Networks
A. Stoica (1,2), Z. Smoreda (1), C. Prieur (2), J.-L. Guillaume (3)
(1) SENSe/Orange Labs
(2) LIAFA/Université Paris 7
(3) LIP6/Université Paris 6
p.11

9:45-10:00 Geography of Social Groups
S. Arbesman, J.-P. Onnela, N.A. Christakis
Harvard Medical School
p.15

10:00-10:15 From Wireless Contacts to Community Structures
T. Hossmann, T. Spyropoulos, F. Legendre
Computer Engineering and Networks Laboratory, ETH Zurich
p.16

10:15-10:30 Social Relationship and Behavior Analysis in Mobile Social Networks
H. Zhang, R. Dantu
Dept. of Computer Science and Engineering, University of North Texas
p.19

10:30-11:00 Coffee break

Session B (11:00-12:15)

11:00-11:15 Timescales in evolving mobile networks
G.M. Krings (1), M. Karsai (2), J. Saramäki (2), V.D. Blondel (1)
(1) UCLouvain
(2) BECS, School of Science and Technology, Aalto University
p.27

11:15-11:30 Towards an investigation of the structure and temporal dynamics in a large scale telecoms dataset
F. Reid, N. Hurley
Clique Research Cluster, University College Dublin
p.28

11:30-11:45 Dynamics and temporal correlations in mobile phone based social networks
M. Karsai (1), L. Kovanen (1), M. Kivelä (1), R. K. Pan (1), J. Saramäki (1), J. Kertész (2), A.-L. Barabási (3,4), K. Kaski (1)
(1) BECS, School of Science and Technology, Aalto University
(2) Institute of Physics, Budapest University of Technology and Economics
(3) CCNR, Northeastern University
(4) CCSB, Dana-Farber Cancer Institute
p.32

11:45-12:00 Characterize Mobile Communication Network Using Communication Motifs
Q. Zhao, N. Oliver
Telefonica Research and Development
p.33

12:00-12:15 Mobility, Data Mining and Privacy: The GeoPKDD Paradigm
F. Giannotti, F. Pinelli, S. Rinzivillo, R. Trasarti
KDD Lab – ISTI – CNR
p.36

12:15-14:00 Lunch break

Session C (14:00-15:45)

14:00-14:15 Epidemics on Mobile Phone Networks
P. Wang (1), M.C. González (2), R. Menezes (3), A.-L. Barabási (1)
(1) Center for Complex Network Research, Northeastern University
(2) Dept. of Civil and Environmental Engineering, MIT
(3) Dept. of Computer Sciences, Florida Institute of Technology
p.43

14:15-14:30 Human movements and the spread of infectious diseases
V. Belik (1), T. Geisel (1), D. Brockmann (2)
(1) Max-Planck-Institute for Dynamics and Self-Organization
(2) Northwestern University
p.44

14:30-14:45 A Tale of Two Cities
S. Isaacman (1), R. Becker (2), R. Cáceres (2), S. Kobourov (3), J. Rowland (2), A. Varshavsky (2)
(1) Dept. of Electrical Engineering, Princeton University
(2) AT&T Labs – Research
(3) Dept. of Computer Science, University of Arizona
p.50

14:45-15:00 Predicting travelling activity space based on phoning activity space in social networks
D. Janssens, S. Moerdijk
Hasselt University
p.53

15:00-15:15 Preliminary findings on application of mobile phone data analysis to urban studies
F. Manfredini, P. Pucci, P. Tagliolato, P. Dilda
Dipartimento di Architettura e Pianificazione, Politecnico di Milano
p.56

15:15-15:30 Analyzing cell-phone mobility and social events
F. Calabrese, G. Di Lorenzo, F. Pereira, L. Liu, C. Ratti
Senseable City Laboratory, Massachusetts Institute of Technology
p.59

15:30-15:45 Paris by Night
C. Cariou (1), C. Ziemlicki (2), Z. Smoreda (2)
(1) Everydatalab
(2) SENSe/Orange Labs
p.62

15:45-16:15 Coffee break

Session D (16:15-17:00)

16:15-16:30 The “Friends and Family” Mobile Phone Study: Overview and Initial Report
N. Aharony, C. Ip, W. Pan, A. Pentland
Media Lab, Massachusetts Institute of Technology

p.69

16:30-16:45 Efficient Collaborative Application Monitoring Scheme for Mobile Networks
Y. Altshuler (1), S. Dolev (2), Y. Elovici (1), N. Aharony (3)
(1) Deutsche Telekom Labs, Ben Gurion University
(2) Computer Science Dept. Ben Gurion University
(3) Media Lab, Massachusetts Institute of Technology

p.72

16:45-17:00 Beyond San Francisco Cabs: building a *-lity Mining Dataset
M.-O. Killijian (1), M. Roy (1), G. Trédan (2)
(1) CNRS, LAAS, Université Toulouse
(2) TU Berlin

p.75

Concluding Talk

17:00-17:30 Mining Large Scale Cell Phone Data
J. Bolot
Sprint

p.79

Session A

Age, Gender and Communication Networks

Alina Stoica[a,b], Zbigniew Smoreda[a], Christophe Prieur[b], Jean-Loup Guillaume[c]
a. SENSe/Orange Labs (*alinamihaela.stoica, zbigniew.smoreda@orange-ftgroup.com*)
b. LIAFA/Université Paris 7 (*prieur@liafa.jussieu.fr*)
c. LIP6/Université Paris 6 (*jean-loup.guillaume@lip6.fr*)

***Abstract*—In this paper, we address some sociological and topological issues associated with mobile phone communication. Based on a dataset of a few million users, we use customers' age and gender information to study relation between these parameters and the average behavior of users in terms of number of calls, number of SMS and calls duration. We also study the dataset from a networking point of view: we define different profiles based on the topological properties of the personal network of each individual and study the relations between these profiles and the age of customers.**

Keywords: mobile phone, age, gender, network structure

I. INTRODUCTION

The ICTs landscape has been entirely changed by the cell phone diffusion. This individual and ubiquitous device, offering voice and text communication features, has transformed the frequency and the geography of communication as compared to older fixed phone practices. We are now virtually always accessible to others wherever we are. Moreover, the mobile phone gives us a direct access to a person: the phone line is no more filtered by household's or bureau's "switchboard" [4]. This offers a useful insight into individual behavior and personal and social network analysis.

The recent possibility to analyze large datasets of behavioral data coming from telecommunication operators gives us the opportunity to revisit some older research on telephone usages. It offers also a new prospect to work on nearly complete interpersonal communication networks. Among an increasing amount of behavioral traces collected by technical systems (internet, mail, IM, SNS…), the interpersonal communication data seem to be the best proxy of social interactions [1,2,3]. Indeed, we usually talk to people with whom we also have many other links, which is not always true in the case of communication with email contacts or SNS "friends". Such datasets therefore open the door to the analysis of close social relationship.

II. DATASET

The raw data analyzed in this contribution – the CDRs (Call Detail Records) – contain all mobile phone exchanges observed in 2006/2007 over a six-month period between Belgian customers of a local mobile operator. After data cleaning the dataset contains 3.3 million users that exchanged over 6 billion calls and short messages (SMS). In addition to communication details (date, hour, duration of call), this anonymous dataset also includes customers' age and gender.

We compared the age and gender distribution of the mobile phone customers in our dataset to the general national population and concluded that there is no systematic bias in operator's customer as regarding these two characteristics (except for people aged over 60 who are underrepresented amongst cellular users).

III. RESULTS

A. Voice–text usages: a generational transformation?

The mobile phone diffusion started in the mid-1990. Classically, it first touched the young and wealthy part of the developed countries population before the rapid, massive and nearly universal adoption [5]. From the usage point of view, it means that nowadays only the youngest groups of the population were entered in their "communication age" directly with a cell phone at hand. Hence, it seems interesting to look at some basic indicators of the mobile phone usages by age. Figure 1 shows the average number of calls and SMS, and the mean call duration by age. We observe that the differences in voice call frequency or duration between ages are relatively minor. The main distinction concerns SMS usage: younger users send more SMS than older ones. In the age group 18 to 25 this tendency is really impressive: the SMS is used 4 times more frequently than a conversational exchange!

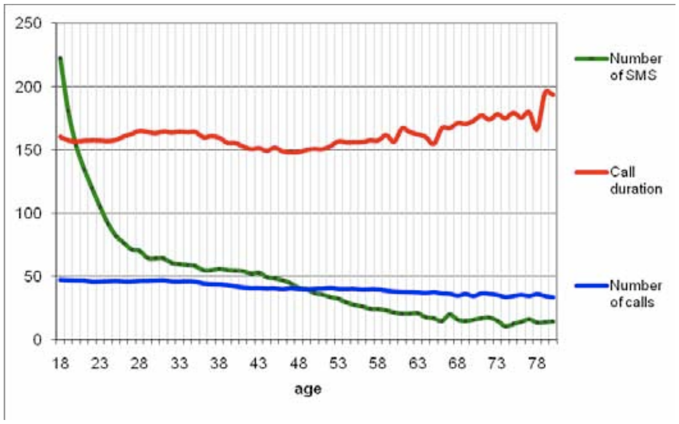


Figure 1. Average monthly number of calls, number of SMS and call duration as a function of phone user's age.

These data show that today interpersonal mobile communication is clearly distributed between voice and text exchanges.

B. Age groups and individual network profiles

To develop these observations, we decided to test whether there is a relationship between the age of a person and the way she is linked to others, or not. We thus analyzed the connections between individuals from the perspective of the communication network they belong to.

We represented the interpersonal communications by an undirected simple graph, where the vertices are the mobile phone customers. Two vertices are connected by an undirected edge if there had been at least one communication in each direction between the corresponding users during each month of the observation. This means that we only consider relatively strong interpersonal links. This gives us a social network with approximately 3 million vertices and 7 million edges that we use in order to characterize customers. Rather than a global approach, we decided to analyze the local structure of the graph around each vertex. More precisely we studied the personal (or ego-centered) network of each vertex (*ego*), i.e., the graph whose vertices are ego's neighbors and whose edges are the edges between the neighbors (note that ego is not included in its ego-centered network).

For every ego we computed several parameters of its ego-centered network: the number of vertices (ego's degree), the number of edges (the edges between ego's neighbors), the number of isolated vertices (the neighbors that are connected only to ego and not to any other neighbor of ego), the number of triangles (a group of 3 interconnected neighbors) and the number of “stars” (a group of 4 neighbors where one of them is connected to the other 3 that are unconnected between them). We use these simple network motives to identify specific individual profiles.

Note that the degree of vertices must be taken into account as the values of the different parameters are biased by it, we chose to compute profiles separately for each degree and distributed the vertices of each degree into 6 profiles defined as follows:

- profile 1: densely connected networks: the number of edges is high and the number of isolated vertices is low;
- profile 2: sparsely connected networks (the opposite situation): the number of edges is low and the number of isolated vertices is high;
- profile 3: mixed situation where there is a densely connected group of neighbors (many triangles) and a sparsely connected one (many isolated vertices);
- profile 4: medially dense networks but with many triangles: these networks do not belong to the first 3 profiles but have a high number of triangles;
- profile 5: medially dense networks but with many stars: these networks do not belong to the first 4 profiles but have a high number of stars;
- profile 6: medially dense networks with no special structure: unclassified vertices.

The question is to determine whether there is a connection between the different profiles and the age of a person or not. To answer this question, we computed, for each age from 18 to 60[1], the probability that an individual of that age belongs to a given profile (see: figure 2). The range of probabilities is different for different profiles which is due to the over-representation of profiles 2 and 6 caused by the heterogeneous distribution of parameters, with many small values. However, there are important differences between these probabilities for different ages. We observe that middle age people (30 to 45) are generally involved in sparser structures when younger and older groups are more densely connected. However, the oldest keep a densely connected group even if they have isolated contacts, while the youngest belong to some cliques (profile 4) or have one or more correspondents who are the “stars” of their ego-centered networks (profile 5).

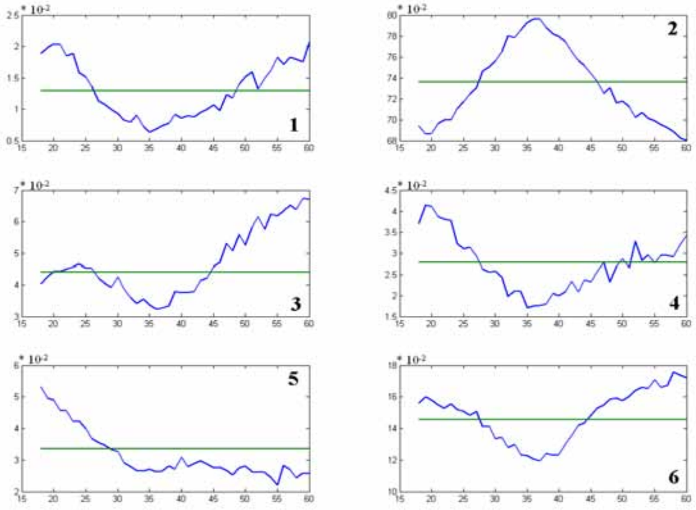


Figure 2. The probability of belonging to the 6 profiles by age (the green line is the average profile probability)

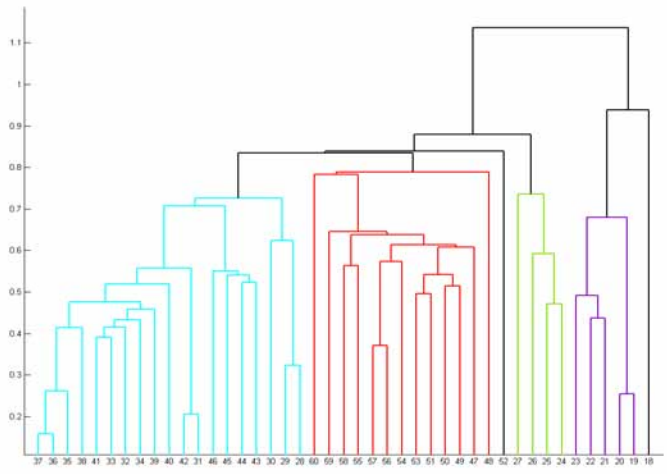


Figure 3. Hierarchical clustering of ages on probabilities in the 6 connectvity profiles

Let us now group together the ages that have similar probabilities for the 6 profiles. A hierarchical clustering was performed on profiles probabilities (see: figure 3). We observe

[1] The persons older than 60 are underrepresented amongst mobile phone users.

that there are 4 main homogeneous age groups very similar to life stages categories [6]: 18-23 (students), 24-27 (young people starting their active life), 28-46 (in couple, usually with children), and 47-60 (at the final stage/end of professional life, children are adult or living apart). Interestingly, here the classification is based exclusively on structural characteristics of their local communication network (where network size effect was neutralized).

C. Gender effect in mobile communication networks

The second personal characteristic of people in our network is their gender. Some years ago, in a French study on the residential use of the (fixed) telephone, the communication of several hundreds of adult men and women has been followed for 4 months using telephone billing records [5]. The study focused on the correlation between the observed duration of phone calls and the gender of callers and receivers. Data have shown that the duration of calls are correlated with the gender of the call receiver and is on average longer when a woman is called. Therefore the reasons why women speak more on the phone [6] seem more related to the gender homophily in telephone networks than to "their intrinsic tendency to talk." An in-depth Conversation Analysis work on a recorded telephone talk's dataset [7] has suggested that politeness rules governing the telephone call can explain in part why it is the gender of receiver that has the biggest effect on how the call is managed and on its overall duration. The conversations involving women tended to go through longer introductive sequences, to be more multi-thematic and digressive in nature with a corresponding lengthening and multiplication of closure sequences; and the conversations with men had a tendency to be linear and monothematic. In summary, the callers seem to adjust their interaction style to the gender of the receiver.

Ten years after, the mobile usages still fit this gender communication pattern. As we can see in the figure 4, mobile phone calls towards a woman are, in average, longer than calls to a man, whatever caller gender is.

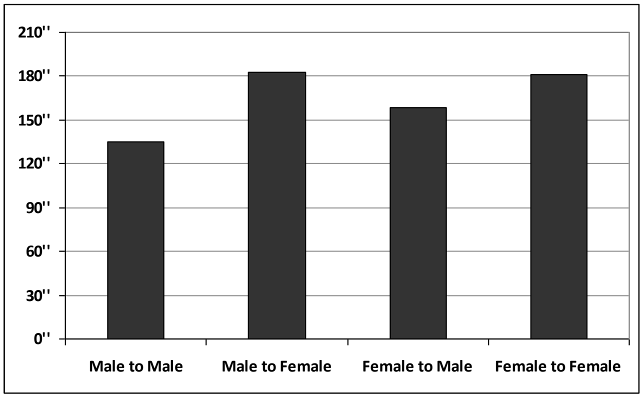


Figure 4. Mean call duration (in seconds) according to call initiator and receiver gender

To go further, we isolated all mixed-gender two-way communication pairs in our network and calculated average durations of call between them. We obtained 171 seconds when a male calls a female and 162 seconds when a female calls a male. As a consequence, it seems that we do not face a technological artifact but a more general social interaction pattern.

The personal networks composition was also scrutinized looking at the fraction of men in ego communication network (see figure 5). We observed that the overall gender homophily in communication networks evolves with age for men and, less, in women's networks. The life cycle transitions modify sociability patterns—from external to the household contacts for young people to more and more family-oriented links for older individuals [10]—and influences the shape of gender relations. In fact, as domestic and familial spheres are still associated with the role of woman, with age the male's network starts to be populated by females. At the end of lifecycle, there are more women than men in the man's mobile phone directory.

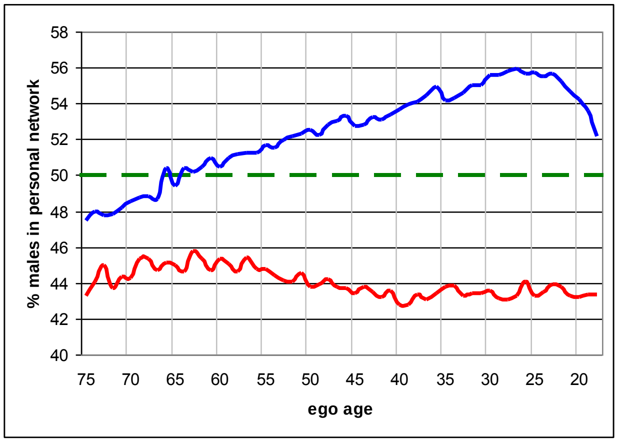


Figure 5. Fraction of males in the communication network of females (red) and males (blue) by age

Going back to communication practices, we can speculate about their hypothetical transformation. Of course, we do not have long-term time series, thus our interpretation remains tentative. However, some tendencies (see: figure 6) indicate differentiation of gendered mobile usages by age. The SMS usages seem to be more "feminine" in general (fig. 6-a): for younger part of population (aged 18-25), we also notice that the between-gender "texting" is particularly popular. And, in fact, it develops at the expense of voice calls (fig. 6-c); the number of voice communication is going down in young adults. The duration of calls varies less, but for the young adults it diminishes sharply for same gender calls. The mixed-gender conversation length remains at the level of other age groups.

We can consider that in younger generations the mobile phone appropriation was deeper as it includes both communication functions offered by cell phones: text and voice. Some authors indicate that heavy SMS use in youngster's relation with other gender is related to seduction tactics where a direct voice contact can be more "risky" for interlocutors [11]. Anthropologists also emphasized the propensity of girls to write personal diaries, letters, etc., as well as women responsibility in familial correspondence [12]. The changing balance between voice and text in general and in between-gender communications can be in part a reflection of new form of accommodation to a new communication channel offered by a popular technology.

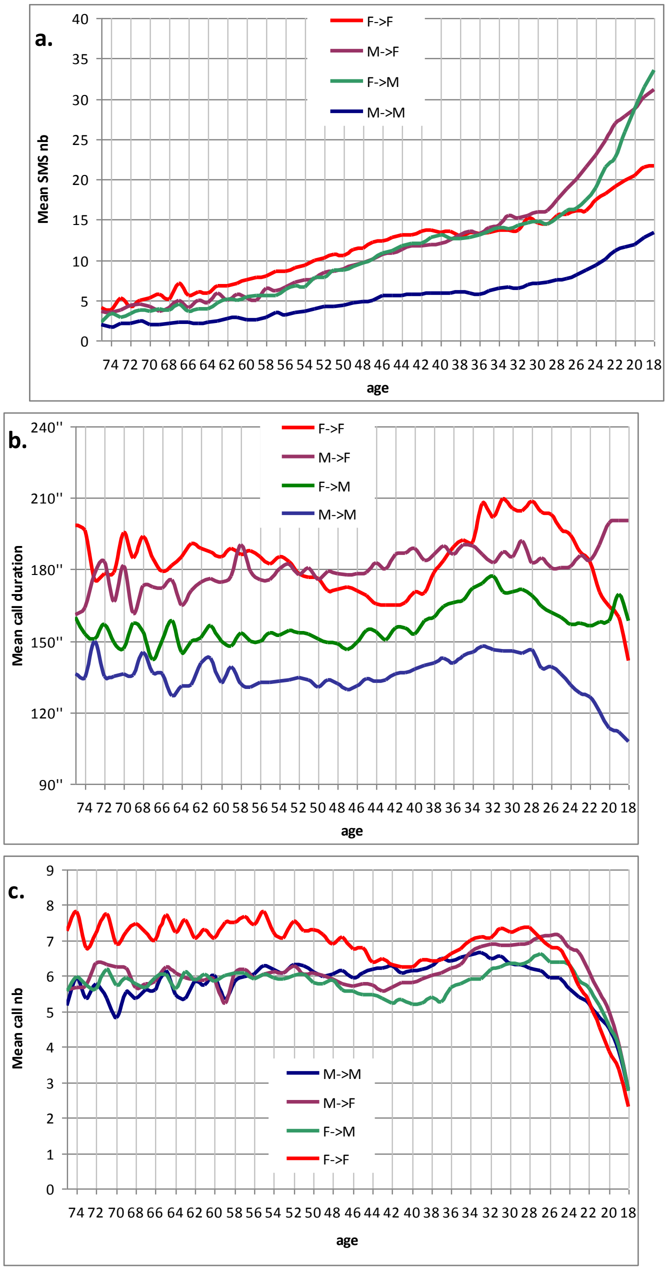


Figure 6. Mobile communication practices by age and gender: (a) mean monthly SMS number, (b) average call duration, (c) mean monthly number of calls.

IV. CONCLUSION

We presented here a stage report of an ongoing research. Based on CDR of a mobile phone operator, we studied the relations between phone usages and networks properties of customer's and their age and gender. We have shown that some already known sociological results are still valid in the context of mobile phone. In particular, there are very different behavior depending on the age of the users and their gender: the gender of the receiver of phone calls is strongly correlated with the duration of the call, while the gender of the caller is less important.

We also built ego-centered networks for each customer and tried to find some correlations between the topological structure of their neighborhood and their behavior. Automatic profiling based only on simple topological properties yielded groups of users which correspond to life stages categories.

Much work remains to be done in this context, in particular more complex network properties and their correlations with age or gender could be studied. We could also propose models taking such behaviors into account.

V. REFERENCES

[1] J-P.Onnela et al., "Analysis of a large-scale weighted network of one-to-one human communication", New Journal of Physics, vol. 9, 179, 2007.

[2] C.A.Hidalgo, C.Rodriguez-Sickert, "The dynamics of a mobile phone network", Physica A, vol. 387, pp. 3017–3024, 2008.

[3] R.Lambiotte et al., "Geographical dispersal of mobile communication networks", Physica A, vol. 387, pp. 5317-5325, 2008.

[4] R.Ling, The Mobile Connection: The Cell Phone's Impact on Society, San Francisco, Morgan Kaufman Publ., 2004.

[5] See: http://www.itu.int/ITU-D/ict/statistics/index.html

[6] N.Shulman, "Life-cycle variations in patterns of close relationship", Journal of Marriage and Family, vol. 37, pp. 813–821, 1975.

[7] Z.Smoreda, C.Licoppe, "Gender-Specific Use of the Domestic Telephone", Social Psychology Quarterly, vol. 63, pp. 238-252, 2000.

[8] L.R.Rakow, Gender on the Line. Women, the Telephone, and Community Life, University of Illinois Press, Chicago, 1992.

[9] R.Akers-Porrini, "Efficacité féminine, courtoisie masculine: la durée inégale des appels téléphoniques", Réseaux, vol. 18(103), pp. 143-182, 2000.

[10] A.Degenne, M.Forsé, Introducing Social Networks, London Sage, 1999.

[11] R.Ling, B.Yttri, 2003. "Kontroll, frigjøring og status: Mobiltelefon og maktforhold i familier og ungdomsgrupper." in F.Engelstad, G.Ødegård (eds.) *På terskelen: makt, mening og motstand blant unge*, Oslo: Gyldendal Akademisk (english version: http://www.richardling.com/papers/2004_Control_Emancipation_and_status.pdf)

[12] D.Fabre (ed.), Ecritures ordinaires, Paris, P.O.L, 1993.

Abstract for NetMob

Authors:
Samuel Arbesman, Harvard Medical School (arbesman@hcp.med.harvard.edu)
Jukka-Pekka Onnela, Harvard Medical School (onnela@hcp.med.harvard.edu)
Nicholas A. Christakis, Harvard Medical School (christak@hcp.med.harvard.edu)

Presenter: Samuel Arbesman

Geography of Social Groups

The relationship between geographical position and network position within social networks is a complex one. Previous research has shown that connectedness is inversely proportional to distance, and that tie strength is related to geographical position as well. Here we focus on groups of individuals within social networks. By examining a mobile phone network from a European country, we create a metric of the geographic span of collections of people, in order to determine the localities of groups of individuals. Through the development of a suitable null model, we verify that locality is indeed a very important component of locality for group structure. In addition, we find that this span is related to the size of the group in a clearly quantifiable manner.

From Wireless Contacts to Community Structures

Theus Hossmann, Thrasyvoulos Spyropoulos, and Franck Legendre
Computer Engineering and Networks Laboratory
ETH Zurich, Switzerland
lastname@tik.ee.ethz.ch

***Abstract*—Human mobility and resulting contacts are driven by *intention*, *co-location*, and *social* relations between people. Based on wireless contact traces (Bluetooth, Wifi), we aim at characterizing the structure in human contacts. Instead of investigating the microscopic properties of contacts (e.g., duration and occurrence distributions), we are more interested in a macroscopic view of mobility that can more easily capture the range of human inter-relations. We hence turn to community detection. However, since these algorithms require one-dimensional tie strength metrics, we present a method to map contacts features (evolving with time) to a scalar feature value. We then analyze the outcome of the community detection by looking at inter- and intra-community ties. This provides interesting insights on the diversity of human inter-relations, which have applications to diffusion processes, for example.**

I. Introduction

The rapid proliferation of smartphones with wireless networking capabilities (Bluetooth, Wifi) creates amble opportunity for opportunistic networks where devices connect to other devices in proximity (when within radio range), “on the fly”, to exchange or spread information. This is a novel networking paradigm that is envisioned to co-exist with (and often complement) existing broadband wireless technologies (e.g. cellular, WiFi, etc.). Since actions of interest can only occur during a wireless contact, *contacts* and their statistical properties become of key importance in the design and performance evaluation of such opportunistic networks. To this end, a number of efforts have been made recently to collect relevant mobility data and analyze contact patterns; this is done either implicitly, by looking at the access points and base stations users are associated with over time in WiFi or cellular networks [1], or explicitly with experiments designed to log peer contacts (e.g. via Bluetooth) [2], [3], [4]. The majority of these traces reveal a considerable heterogeneity in contact patterns, but also significant structure and (statistical) predictability of these patterns e.g. due to time-of-day periodicity, location preference, etc. Nevertheless, the vast majority of trace analysis research in networking has focused on the *inter-contact* and *contact duration* statistics [5], [6], which are important for network performance analysis but limits mobility analysis to a microscopic view.

Recently, researcher have been looking at mobility at large-scale [1], its predictability [7], and spatial connectivity properties [8]. Human mobility and resulting contacts are actually driven by *intention*, *co-location*, and *social* relations between nodes (e.g. friends, colleagues). The latter influences someone to decide the destination (and often time) of a mobility trip; *location* on the other hand dictates the path, as well as (unknown) nodes encountered regularly at preferred/home locations (“familiar strangers”) or occasionally (“random encounters”). This creates a rather intricate contact structure that is not readily observable or usable at contact and inter-contact pattern levels. To this end, a more abstract, *macroscopic* view of mobility is needed that can more easily capture the range of node inter-relations.

In this abstract, we present a detailed study and comparison of the community structure of 4 mobility traces, namely the Haggle trace [2], the MIT Reality Mining trace [3], and the ETH trace [4]. We apply a state of the art community detection algorithm [9] to study the nature of inter-community links (e.g. bridging links vs. bridging nodes vs. community overlap, etc.), and the inter- and intra-community weight distributions in order to highlight the diversity of human relations. To our best knowledge, this is the first in depth comparative study of these properties.

In our context, nodes and contacts can be represented on a *contact graph*, where a link between two nodes indicates a measured “strong” relationship between nodes (e.g. frequent meetings, or a recent meeting [10]) through its existence (binary graph) or an edge weight (weighted graph). A variety of metrics and algorithms could then be used to characterize *node importance* on this graph, such as degree centrality, pagerank, etc., as well as to identify *similar* nodes through (implicit or explicit) *community detection*. Yet, the actual “social properties” of mobility traces, such as the modularity of communities and the distribution of inter- and intra-community weights, have not received the same amount of attention. These properties are particularly important for two reasons: *first*, they allow us to better understand the underlying structure governing human mobility and facilitate the design of improved mobility models. *second*, they give hints on the impact of the social structure on the dynamics of diffusion processes e.g., in terms of delays but also in terms of capacity (or conductance).

The outline of this abstract is the following. In Section II, we describe the contact data used for our analysis and how we pre-process them by mapping and aggregating pair-wise contacts (i.e., different characteristics evolving over time) to a scalar value suited for community detection algorithms with weighted edges. In Section III, we analyze the outcome of the community detection algorithm. Eventually, we conclude by discussing ongoing work in Section IV.

II. Data Description

We start by describing the data used for our analysis in Section II-A. We then describe a metric of tie strength based on the principal component of contact frequency and duration (Section II-B).

TABLE I
CONTACT TRACES CHARACTERISTICS.

	MIT	INFO	ETH
Scale and context	97 campus students and staff	41 conference participants	20 lab students and staff
Period	9 months	3 days	5 days
Periodicity	300s (Bluetooth)	120s (Bluetooth)	0.5s (WiFi)
# Contacts			
Total	$100'000$	$22'459$	$23'000$
Per dev.	$1'030$	547	$1'150$

A. *Contact Traces*

We define a *contact* as the period of time during which two devices are within radio transmission range of each other. A contact contains of the information about the two nodes involved, a starting time and a duration. In a opportunistic network, such a contact is an opportunity to exchange or spread information.

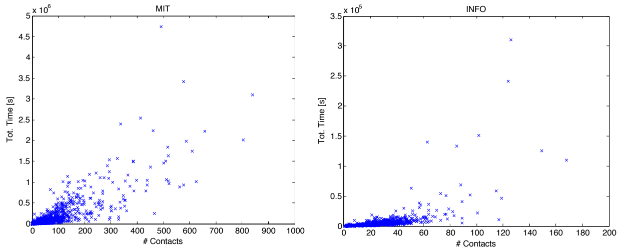
In order to cover a broad range of mobility scenarios with our analysis, we use different measured contact data: the MIT *Reality Mining* [3] (**MIT**), the iMotes Infocom 2005 (**INFO**) [2] and the ETH [4] (**ETH**). Their characteristics are summarized in Table I. Note that in the MIT trace, despite its long duration, a lot of short contacts were supposedly not logged due to its time granularity of 5 minutes. For our evaluation we cut the trace at both ends and used $100'000$ contacts reported between September 2004 and March 2005. Note that this time period contains holidays and semester breaks and thus still captures varying user behavior. The ETH trace contains more than $23'000$ reported contacts and is unique in terms of time granularity and reliability. Although its measurement period spans a considerably shorter time than MIT, we have on average more than 1000 reported contacts per device. This is roughly the number of contacts per device we also have for the MIT trace.

B. *Tie Strength*

To assess the strength of the tie between two nodes in a contact graph different metrics such as the age of last contact [11], contact frequency [12], [13] or aggregate contact duration [13] have been used (i.e., in DTN routing protocols). Here we consider two features: contact *frequency*[1] and aggregate contact *duration*. In a first step, we assign each pair of nodes $\{i, j\}$ a two-dimensional feature vector $\mathbf{z_{i,j}} = (f_{i,j}, d_{i,j})$, where $f_{i,j}$ is the number of contacts in the trace between nodes i and j, and $d_{i,j}$ is the sum of the durations of all contacts between the two nodes – both dimensions centered (zero empirical mean) and normalized to their respective standard deviation.

Figure 1 shows the scatter plots of the number of contacts vs. the total contact duration (pair-wise) for the MIT and INFO traces. They clearly show a high correlation between both features.

Since state-of-the-art community detection requires one-dimensional tie strength metrics, we transform the two-dimensional feature vector to a scalar feature value: We use the *principal component* (e.g., [14]), i.e., the direction in which the data vector Z has the largest variance the direction of the Eigenvector $\mathbf{e_1}$ with the largest corresponding Eigenvalue. We define the tie strength between i and j as the projection of $\mathbf{z_{i,j}}$ on the principal component, $w_{i,j} = \mathbf{e_1}^T \mathbf{z_{i,j}} + w_{\text{min}}$, where we add w_{min} – the smallest tie strength of all node pairs – in order to have positive tie strengths. With this metric we are able to combine the frequency and duration in a scalar value that naturally represents the heterogeneity of node pairs. We can now define the weight matrix $\mathbf{W}$ with the respective $w_{i,j}$.



(a) MIT (b) INFO

Fig. 1. Scatter-plots of number of contacts vs. total contact duration over the whole duration of the traces.

Note that with this aggregation of the contact data, we loose the timing information about contacts. We are not so much interested in the actual timing of the contacts, but rather try to capture the underlying structures that govern mobility.

The number of communities and the resulting modularity is given for each contact trace in Table II.

III. COMMUNITY STRUCTURE ANALYSIS

We will now focus on the community structure of human contacts contacts. Using the Louvain as well as Spectral community detection algorithm and the Newman modularity metric, we will first (Section III-A) assess how strongly modular contacts are. In a second step (Section III-B), we will focus on the the conductance *between* the communities, and assess how strongly communities are connected to other communities and how the conductance between them is distributed (i.e., bridging links, bridging nodes or hierarchical overlap).

A. *Intra-Community Ties*

In order to assess the *modularity* of a given partition of nodes to communities we compute the widely used Q function as introduced by Newman [15]. The Q function

$$Q = \frac{1}{2m} \sum_{ij} \left(w_{i,j} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j),$$

where $k_i = \sum_j w_{i,j}$ is the strength of node i and $m = \frac{1}{2} \sum_j k_j$ is the total weight in the network. c_i denotes the community of node i thus, the Kronecker delta function $\delta(c_i, c_j)$ is one if nodes i and j share the community and zero otherwise. $Q = 0$ is the expected quality of a random community assignment and [15] reports modularities of above $Q = 0.3$ for different networks (social, biological, technical, etc.) for state-of-the-art community detection algorithms[2].

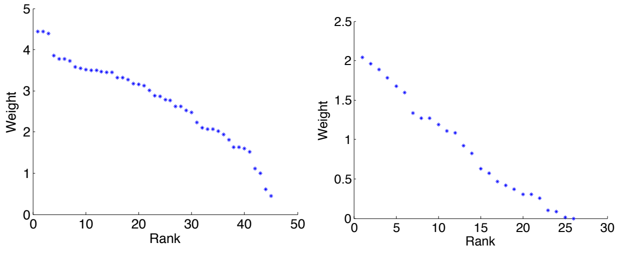
In Table II we present some statistics of the trace networks' community structure as found by Louvain. A first thing to note is that the two clustering algorithms find different communities but with similar modularity. In general the modularity of the

[1] Note that contact age – assuming a stationary contact process – can be considered an approximation of contact frequency, therefor we do not consider it here explicitly.

[2] Note that the quality of a community assignment is a function of (i) the network, since it can be more structured or less, and (ii), the community detection algorithm, since it can find a good community assignment or not.

Trace/Model	# Comm.	Q
MIT	5	0.49
ETH	2	0.23
INFO	6	0.12

TABLE II
NUMBER OF COMMUNITIES AND MODULARITY (Q) OF CONTACT TRACES USING THE LOUVAIN ALGORITHM.



(a) ETH, Community 1

(b) MIT, Node 41

Fig. 2. Ranked Community Internal Weights (per Community and per Node).

Louvain communities is slightly higher communities than the Spectral. In the rest of the paper, we will present all results for the Louvain algorithm, though, the results hold also for the Spectral clustering. Second, the modularity varies broadly among the traces. We observe a strongly modular MIT trace, lower modularity in the ETH case and very low modularity in the INFO case. Similar values for other community detection algorithms (K-Clique and Newman), different traces and other strength metric (total contact duration) have already been reported in [13], thus we confirm these findings as a first result.

To find out more about the insides of communities we look at the distribution of intra-community weight. Figure 2 shows some typical representatives of community-internal tie strengths, ranked over all node pairs of a community, as well as per node. We observe that the weights are strongly skewed. A community can thus not be thought of as a homogeneous group of strongly connected nodes (like a mesh). Instead, there is strong heterogeneity even within a community. This observation is consistent throughout all traces and all communities (only few are shown in Figure 2 due to space limitations).

B. Inter-Community Ties

Comm. Index	1	2	3	4	5
1	**20.5%**	1.0%	0.5%	0.1%	0.01%
2	1.0%	**31.8%**	4.2%	2.9%	0.2%
3	0.5%	4.2%	**13.4%**	2.7%	0.2%
4	0.1%	2.9%	2.7%	**8.9%**	0.1%
5	0.01%	0.2%	0.2%	0.1%	**2.1%**

TABLE III
PERCENTAGES OF TOTAL WEIGHT WITHIN AND BETWEEN COMMUNITIES (MIT TRACE). ALL WEIGHTS SUM TO 100% AND INTER-COMMUNITY WEIGHTS ARE HALVES BETWEEN TIED COMMUNITIES.

We now change our focus on the interface *between* the communities. Table III shows an example matrix for the MIT trace of how the total weight in the network is distributed within the communities and between the communities. In the matrix we see that the inter-connections of communities are weak in many cases. For instance, communities 1 and 2 together contain more than 50% of the weights and 50% of the nodes. However, between them there is only 1% of the weight.

IV. DISCUSSION AND CONCLUSIONS

The results presented herein are preliminary investigations of using community detection algorithms to highlight the community structure of contact traces. Actually, it does not only matter how much of the weight falls between two communities, but also how this weight is distributed. Thus, we are currently aiming at identifying the type of interface as either (i) bridging links (people linked to one specific person in another community), (ii) bridging nodes (people part of two communities i.e., overlap), or (iii) hierarchical communities. We characterize these three types in the following.

Note that certain community detection algorithms inherently identify some of these interfaces. For instance the K-Clique algorithm [16] allows nodes to be in more than one community and thus identifies bridging nodes. Similarly, a class of algorithms such as Newman Girvan [17] is based on a hierarchical tree (dendrogram) and thus inherently identifies hierarchies. However, neither of them provides a distinction between all the three types of inter-connection. We are hence currently combining the peculiar features of existing algorithms at once.

REFERENCES

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

[2] P. Hui, A. Chaintreau, J. Scott, R. Gass, J. Crowcroft, and C. Diot, “Pocket switched networks and human mobility in conference environments,” in *WDTN*, 2005.

[3] N. Eagle and A. Pentland, “Reality mining: sensing complex social systems,” *Personal and Ubiquitous Computing*, May 2006.

[4] V. Lenders, J. Wagner, and M. May, “Measurements from an 802.11b mobile ad hoc network,” in *EXPONWIRELESS*, 2006.

[5] V. Conan, J. Leguay, and T. Friedman, “Characterizing pairwise inter-contact patterns in delay tolerant networks,” in *ACM Autonomics*, October 2007.

[6] T. Karagiannis, J.-Y. Le Boudec, and M. Vojnovic, “Power law and exponential decay of inter contact times between mobile devices,” in *MobiCom ’07: Proceedings of the 13th annual ACM international conference on Mobile computing and networking*. New York, NY, USA: ACM, 2007, pp. 183–194.

[7] C. Song, Z. Qu, N. Blumm, and A.-L. Barabasi, “Limits of predictability in human mobility,” *Science*, vol. 327, no. 5968, pp. 1018–1021, February 2010.

[8] P. Wang and M. C. Gonzales, “Understanding spatial connectivity of individuals with non-uniform population density,” *Philosophical Transactions of The Royal Society A*, vol. 367, pp. 3321–3329, August 2009.

[9] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, “Fast unfolding of communities in large networks,” *J.STAT.MECH.*, 2008.

[10] T. Hossmann, T. Spyropoulos, and F. Legendre, “Know thy neighbor: Towards optimal mapping of contacts to social graphs for dtn routing,” in *Infocom 2010*. IEEE, March 2010.

[11] H. Dubois-Ferriere, M. Grossglauser, and M. Vetterli, “Age matters: efficient route discovery in mobile ad hoc networks using encounter ages,” in *MobiHoc*, 2003.

[12] A. Lindgren, A. Doria, and O. Schelén, “Probabilistic routing in intermittently connected networks,” *SIGMOBILE Mobile Computer Communications Revue*, July 2003.

[13] P. Hui, J. Crowcroft, and E. Yoneki, “Bubble rap: Social-based forwarding in delay tolerant networks,” in *MobiHoc*, May 2008.

[14] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification (2nd Edition)*. Wiley-Interscience, November 2000.

[15] M. E. J. Newman, “Modularity and community structure in networks,” *Proceedings of the National Academy of Sciences*, June 2006.

[16] G. Palla, I. Derenyi, I. Farkas, and T. Vicsek, “Uncovering the overlapping community structure of complex networks in nature and society,” *Nature*, June 2005.

[17] M. E. J. Newman, “Analysis of weighted networks,” 2004.

Social Relationship and Behavior Analysis in Mobile Social Networks

Huiqi Zhang, Ram Dantu
Department of Computer Science and Engineering
University of North Texas
Denton, TX 76203 USA
huiqizhang@my.unt.edu, rdantu@unt.edu

I. Introduction

The modern telecommunication, such as mobile communications, unites people around the world into a Wide Area Social Network (WASN). In WASN, people form groups or clusters based on interests, goals, etc. Since the mobile phones have become an important tool of modern human daily life, telecommunication patterns may reflect different human relationships and behaviors, and changes in telecommunication patterns may expose signs of social relationships and behavior changes. For example, the calling patterns of a person with his/her friends differ from those with spammers.

A social network dynamically changes since the social relationships (social ties) change over time. The evolution of a social network mainly depends on the evolution of the social relationships. The social-tie strengths of person-to-person are different one another even though they are in the same groups.

Almost all existing social network research has focused on overall social-network structures and properties. These research efforts lack analysis for one-to-one or one-to-many relationships and behaviors in the detail-necessary when interested in special groups or clusters of people. These detailed features of human relationships are more important for detecting terrorists, spam and user preferences. Because of human social-behavior's diversities and complexities, applying one technique will not detect the many different features of human social behaviors. Therefore, we use multiple probability and statistical methods, integrating them for social-network and human-behavior analysis. We propose an integrated platform for analyzing the properties of social structures and human behavior; for quantifying and measuring interpersonal relations in groups; for predicting social ties; for detecting change points, unusual consumption events, opt-in bursts; and for identifying willingness levels of users to communicate each other based on human telecommunication patterns.

The integrated platform consists of several components including zoom, scale, and analysis tools, which are used for analyzing network structures, for discovering social groups and events, for quantifying relationships and so on. The integrated platform is extensible; new tools can be added as needed. By zoom-in we may use multiple scales to analyze social-group member behavior up to one-to-one. By zoom-out we may analyze general social-network structures and properties.

Pentland uses the mobile phones programmed, electronic badges and microphones as a Socioscope to sense and capture human behavioral data (location, proximity, body motion) [9], [10]. These behavioral data are then used to analyze the characterization of group distribution and variability, conditional probability relationships between individual behaviors and focuses on human relationship analysis based on physical distance proximity. Eagle extends the approach in [9], [10] to study a variety of human cultures as a culture lens [11]. Eagle *et al.* present a method for measuring human behavior, based on contextualized proximity and mobile phone data, to study the dyadic data using the nonparametric multiple regression quadratic assignment procedure (MRQAP) [2]. Our approach focus on quantifying human behaviors, interpersonal relationships, changes of relationship by studying human calling patterns based on mobile phone call detail records.

II. Overview of the Proposed Approach

As presented in Figure 1, the integrated platform consists of a number of components, including data extraction and transformation; network visualization; and zoom, scale, and several analysis tools used for analyzing network structures, discovering and quantifying social groups, predicting social ties and detecting events, quantifying relationships, etc. The integrated platform is extensible. New tools can be added as we identify additional features. The model is composed of the three layers briefly described below.

Data Processing: This layer consists of two components: *Data Extraction* and *Data Transformation.* In data extraction, related information is extracted from raw datasets and then transformed in data transformation into required data format for visualization and analysis.

Visualization and Zooming: An open-source visualization tool is used for drawing the social networks. Using zooming-in levels we may use multiple scales to analyze social-group member behavior up to one-to-one. Using zooming-out levels we may analyze general social-network structures and properties.

Behavior Analysis and Detection Tools: This layer, the core of the model, consists of four components: *Quantifying Social Groups, Reciprocity and Predicting Social Ties, Change Point Detection* and *Unusual Consumption Detection, Opt-in Pattern Detection* and *Willingness Level inference.*

We describe these components and solutions in details next.

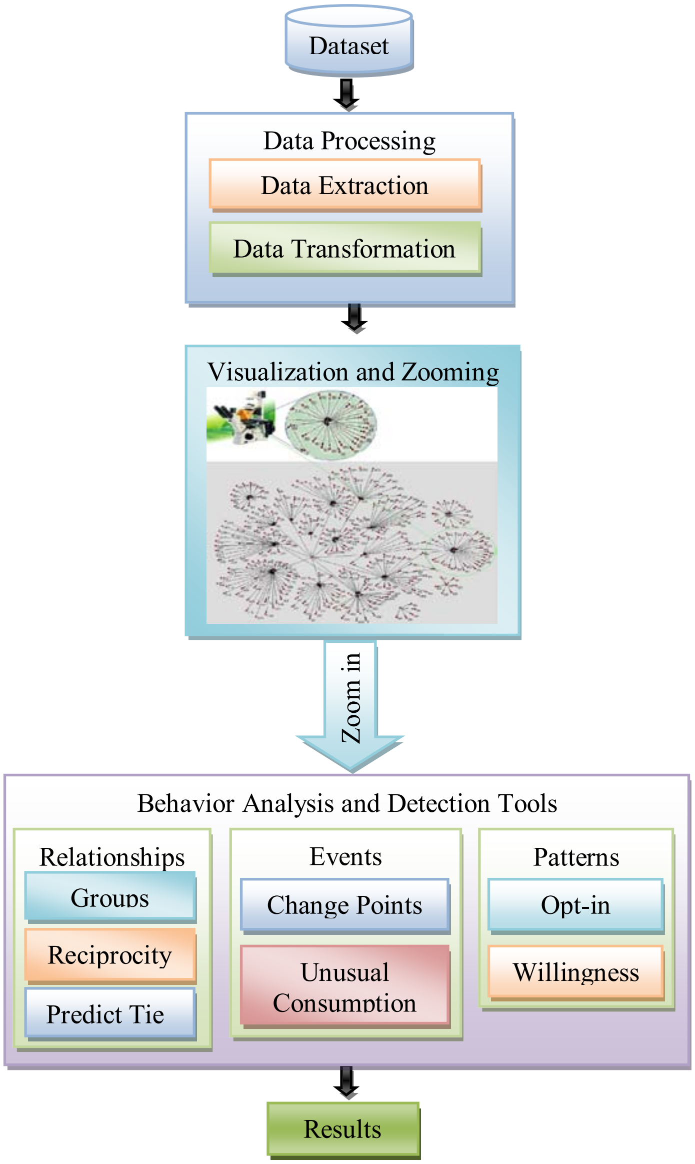


Fig.1 The architecture of the proposed approach.

III. Quantifying Groups and Predicting Ties

The approach proposed here for the social-group identification relies first on a computation of a reciprocity index. We then use the index to compute the affinity between two users. Finally we use this affinty to define the socially related groups. These steps are presented in the next Sections.

A. *Dyads and Reciprocity Index*:

In social networks, one of the important relationships between people is reciprocity. Reciprocity can be defined as the action of returning similar acts [1], [5]. In this study, our interest is to investigate how people use technology to construct social relationships.

In [5] the authors propose the index of mutuality, ρ_{kp}. This index focuses on the probability of a mutual choice between two actors i and j in a graph.

Our approach for reciprocity differs from the above work. We observed that the structure and transactions in reciprocity are different when compared with face-to-face interactions. *Existing approaches measure the tendency of mutual choices for actors (nodes) in a graph. These approaches do not deal with frequency and duration of real-time electronic communications between two actors. In real life, the frequency of communication plays an important role in the relationship between persons. To the best of our knowledge no similar work has been reported. We propose a new reciprocity index based on mobile-phone call-detail records.*

In a mobile-phone social networks, actor i and actor j may call each other multiple times. Reciprocity reflects their relationship in a period of time. The mutual index in [5] and other existing mutual indices cannot measure this kind of relationship. The existing mutual (reciprocity) indices measure the tendency of mutual choices for actors (nodes) in a graph. They do not deal with communication frequency. We propose a reciprocity index, $\rho_{a\leftrightarrow b}$ which does measure the tendency of reciprocity for actors *a* and *b* in a group.

Suppose that the number of phone call arrivals is a Poisson process. Then the probability of no arrivals in the interval [0, t] is given by

$$P(\tau > t) = e^{-\lambda t}$$

where λ is the arrival rate and τ is interarrival time. The occurrence of at least one arrival between 0 and t is given by

$$P(\tau \le t) = 1 - e^{-\lambda t}$$

Considering actor *a* calls actor *b* at time t_i with rate $\lambda_a t$, the probability of actor *b* calling actor *a* back (reciprocity) at a time t_j with rate $\lambda_b t$ can be computed by

$$P(a \to b \,\&\, b \to a) = P(a \to b)P(b \to a \mid a \to b)$$
$$= P(a \to b)[P(b \to a) + \rho_{a\leftrightarrow b} P(b \xrightarrow{not} a)]$$
$$= (1 - e^{-\lambda_a t_i})[(1 - e^{-\lambda_b (t_j - t_i)}) + \rho_{a\leftrightarrow b} e^{-\lambda_b (t_j - t_i)}]$$

The expected value, $E(R \mid \rho_{a\leftrightarrow b})$, of number of reciprocity from *b* to *a* is the total number of calls, S, from *a* to *b* times this probability, i. e.

$$E(R \mid \rho_{a\leftrightarrow b}) = S(1 - e^{-\lambda_a t_i})[(1 - e^{-\lambda_b (t_j - t_i)}) + \rho_{a\leftrightarrow b} e^{-\lambda_b (t_j - t_i)}]$$

After rearranging the terms, we have

$$\rho_{a\leftrightarrow b} = [R - S(1 - e^{-\lambda_a t_i})(1 - e^{-\lambda_b (t_j - t_i)})] / S(1 - e^{-\lambda_a t_i}) e^{-\lambda_b (t_j - t_i)} \quad (1)$$

where R is observed number of reciprocity.

The $\rho_{a\leftrightarrow b}$ is 0 if there is no tendency toward reciprocity and 1 if there is a maximal tendency toward reciprocity.

B. *Social Group Identification*

Groups correspond to data clusters. Cluster analysis concerns a set of multivariate methods for grouping data variables into clusters of similar elements. In our work we use probabilistic models for classification of variables by their *affinity* [3].

Affinity measures the similarity between probability distributions. Because our problem belongs to discrete events, we only consider finite event spaces. Let

$$S_N = \{P = (p_1, p_2, \ldots p_N) \mid p_i \ge 0, \sum_{i=1}^{N} p_i = 1\}$$

be the set of all complete finite discrete probability distributions and $P, Q \in S_N$. The Hellinger distance between P and Q is defined as

$$d_H^2(P,Q) = \frac{1}{2}\sum_{i=1}^{N}(\sqrt{p_i}-\sqrt{q_i})^2$$

$d_H^2(P,Q) \in [0,1]$, $d_H^2(P,Q) = 0$ if P = Q and $d_H^2(P,Q) = 1$ if P and Q are disjoint [3].

The affinity between probability measures P and Q is defined as

$$A(P,Q) = 1 - d_H^2(P,Q) = \sum_{i=1}^{N}\sqrt{p_i q_i}$$

$A(P,Q) \in [0,1]$, $A(P,Q) = 1$ if P = Q and $A(P,Q) = 0$ if P and Q are disjoint [3].

We clasify our social network members into three categories: socially close members, socially near members and socially far members.

In this paper, we use three attributes incoming (*in*), outgoing (*out*) and reciprocity (*reci*) of calls and messages.

Let m_i, n_i be the number of calls, where $i \in \{in, out, reci\}$. $P = (p_{in}, p_{out}, p_{reci})$ is a vector of normalized frequencies over the training period. $Q = (q_{in}, q_{out}, q_{reci})$ is a vector of normalized frequencies of the same attributes observed over the testing period. Then $p_i = m_i \big/ \sum_i m_i$ where $i \in \{in, out, reci\}$ and $q_i = n_i \big/ \sum_i n_i$ where $i \in \{in, out, reci\}$.

The reciprocity part is computed by Eq. (1).

We compute affinity between P and Q is as follows:

$$A(P,Q) = \sum_i \sqrt{p_i q_i} \quad \text{where } i \in \{in, out, reci\} \qquad (2)$$

C. Predicting Social Ties

We map call-log data into time series (social-tie strengths) by our affinity model and apply Seasonal Auto Regressive Integrated Moving Average *(SARIMA)* models for predicting the future values.

Seasonal Auto Regressive Integrated Moving Average (*SARIMA*) models integrate Seasonal (periodic), Auto Regressive (*AR*), Integrated (I), and Moving Average (*MA*) into a general comprehensive time series model [12].

The seasonal $ARIMA(p,d,q)$ model with period *s* is given as

$$\phi_p(B)\Phi_P(B^s)\Delta^d\Delta_s^D Z_t = \theta_q(B)\Theta_Q(B^s)e_t$$

or

$$\phi_p(B)\Phi_P(B^s)(1-B)^d(1-B^s)^D Z_t = \theta_q(B)\Theta_Q(B^s)e_t$$

which are denoted by $SARIMA(p,d,q)\times(P,D,Q)_s$.

The Box-Jenkins method [12] uses an iterative approach of identifying a possible model from a general class model. The chosen model is then checked against the historical data to see whether it accurately describes the series. The model fits well if the residuals are generally small.

IV. Event Detection

Another important element that we can extract from call records is the occurrence of events. This capability could be used for detecting network attacks. To identify events in the call records we first use a wavelet de-noising method to process the data and then we apply the modified method described in [6] for detecting change points based on number of calls and call durations. These steps are described next.

Social network structures and relationships dynamically change over time. Still, change point and event detection methods can be used to discover human relationship and behavior changes based on human communication pattern changes.

A. Wavelet Denoising

Generally, for the denoising, the wavelet scaling function should have properties similar to the original signal. The general wavelet denoising procedure follows 3 steps: wavelet selection, threshold selection and inverse wavelet transform.

In a Discrete Wavelet Transform (DWT) the scale factors between levels are usually chosen to be powers of 2. For DWT, the mother wavelet is defined as:

$$\psi_{j,k}(t) = 2^{-j/2}\psi(2^{-j}t-k)$$

where *j, k* є *Z*.

The DWT is given by

$$W_{j.k} = 2^{-j/2}\int_{-\infty}^{\infty} x(t)\psi(2^{-j}t-k)dt$$

where $W_{j,k}$ is wavelet coefficients, x(t) is the signal to be transformed, and ψ(t) is the mother wavelet or basis function.

The inverse transform of DWT used to compute original data is given as:

$$x(t) = \sum_{j=1}^{\infty}\sum_{k\in Z} W_{j,k}\psi_{j,k}(t) + \sum_{j=1}^{\infty} c_{j,k}\phi_{j,k}(t)$$

where $\phi_{j,k}(t)$ denotes the scaling function and $c_{j,k}$ denotes scaling coefficients, which are defined by

$$c_{j,k} = 2^{-j/2}\int_{-\infty}^{\infty} x(t)\phi_{j,k}(2^{-j}t-k)dt$$

The inverse wavelet transform (IWT) reconstructs the signal from its coefficients.

B. Change Point Detection Procedures:

A change-point analysis method attempts to find a point along a distribution or trend of values where the characteristics of the values before and after the point differ.

Let $X = (X_1, X_2, ..., X_\theta)$ be a process.

The multiple change points for the process is defined as [6]

$$X_1 = (x_1, ..., x_{\tau_1}) \sim f_1$$
$$X_2 = (x_{\tau_1+1}, ..., x_{\tau_2}) \sim f_2$$
$$................................$$
$$X_\theta = (x_{\tau_{\theta-1}+1}, ..., x_{\tau_\theta}) \sim f_\theta$$

where $f_1, f_2, ..., f_\theta$ are either known or unknown probability density functions or trends; $\tau_1, \tau_2, ..., \tau_{\theta-1}$ are change points.

We first use a wavelet-denoising method to pre-process the data, and then we use a sequential estimation scheme in [6] for detecting multiple change points which chooses increasing subsamples and finds one change point at a time until all change points are found.

The general procedure for change point detection is as follows.

a) Sequential Detection:

A widely used change-point-detection method based on Page's cumulative sum (cusum) rule is defined by

$$T(a) = \inf\{n : S_n \geq a\}$$

where

$$S_n = \max_{1 \le k \le n} \sum\nolimits_{i=k+1}^{n} \log \frac{g(x_i)}{f(x_i)}$$

is maximum likelihood ratio based cumulative sum, a is a threshold, and f and g are pre- and post-change density functions. When density functions are unknown, the best estimates for each value k of a change point are used. The cusum is computed by

$$\hat{S}_n = \max_{1 \le k \le n} \sum\nolimits_{i=k+1}^{n} \log \frac{\hat{g}(x_i)}{\hat{f}(x_i)}$$

and the stopping rule described in [39], defined by

$$\hat{T}(a) = \inf\{n : \hat{S}_n \ge a\}$$

We use a leased-squares method for nonlinear trends. Whether a trend is linear or exponential is decided by the lower sum of squares for each segment. In this way, the maximum likelihood ratio is replaced by a minimum weighted sum of squared residuals

$$\hat{S}_n = \min_k \{\sum\nolimits_{i=1}^{k} \sqrt{\hat{f}_i}(x_i - \hat{f}_i) + \sum\nolimits_{i=k+1}^{n} \sqrt{\hat{g}_i}(x_i - \hat{g}_i)\}$$

and the stopping rule is defined as

$$\hat{T}(\alpha) = \inf\{n : p_n \le \alpha\}$$

where p_n is a p-value testing significant based on $\hat{S}_n$ of a change point at k.

b) Post-estimation:

The detected change point must be estimated after a stopping rule has detected it. The change point estimator we use is based on the cusum stopping rule $\hat{T}$ and the minimum p-value

$$\hat{\tau} = \arg\min_{1 \le k < \hat{T}} p(k, \hat{T}, X)$$

where $p(k, \hat{T}, X)$ is the p-value of the likelihood ratio test comparing $X_1 = (x_{1,\dots,} x_k)$ and $X_2 = (x_{k+1,\dots,} x_{\hat{T}})$.

c) Significance Tests:

To eliminate false change points, we use ANOVA F-type tests. If the test is significant, we repeat step 1-3 to search for next change point, else, the test is a false change point and the search continues based on an initial sequence after the last significant change point.

For fitting linear trend $E(x_i) = a + bt_i$, we use a standard least-square estimate

$$\hat{b} = \frac{\sum_i (x_i - \bar{x})(t_i - \bar{t})}{\sum_i (t_i - \bar{t})^2} \quad \text{and } \hat{a} = \bar{x} - \hat{b}\bar{t}$$

and for fitting exponential trend $E(x_i) = \exp(a + bt_i) - 1$, we use

$$\hat{b} = \frac{\sum_i \log(1 + x_i)(t_i - \bar{t})}{\sum_i (t_i - \bar{t})} \quad \text{and } \hat{a} = \overline{\log(1 + x)} - \hat{b}\bar{t}$$

for initial approximation.

When the preliminary estimator of a change point is obtained, we perform a refinement of this estimator by least square fitting from the segment in the neighborhood of the preliminary estimator. If the change points are not significant for the chosen level α, they are removed and the corresponding segments merged.

After the iterations end, all the change points are significant at the chosen level α.

We select a Coiflets5 wavelet and the Minimax threshold method to denoise the data by the principles described above and, then, apply the sequential change point detection method in [6].

C. Unusual Consumption Event Detection

Change-point-detection methods do not deal with bursts of short width in time series. Bursts in time series are related to events such as attacks in networks. We propose the inhomogeneous Poisson model for detecting these bursts, which we term as unusual consumption events.

We assume that the number of calls follows an inhomogeneous Poisson process and call duration follows an inhomogeneous exponential distribution.

The maximum likelihood estimates are used to estimate average number of calls and call duration. Next we consider the maximum average number of calls and call duration obtained for all weekday/weekend and week by week. Suppose that the m week data is used to compute the rates of number of calls and call duration for user p. Let $\hat{\lambda}^p_{d1}, \hat{\lambda}^p_{d2}, \dots, \hat{\lambda}^p_{d7}$ be the rate of number of calls obtained for all weekday/weekend and $\hat{\lambda}^p_{w1}, \hat{\lambda}^p_{w2}, \dots, \hat{\lambda}^p_{wm}$ be the rate of call duration obtained week by week for m weeks of user p respectively. Let $\hat{\mu}^p_{d1}, \hat{\mu}^p_{d2}, \dots, \hat{\mu}^p_{d7}$ be the mean of call duration obtained for all weekday/weekend and $\hat{\mu}^p_{w1}, \hat{\mu}^p_{w2}, \dots, \hat{\mu}^p_{wm}$ be the mean of call duration obtained week by week for m weeks of user p respectively.

Then the maximum means of the number of calls and call duration are respectively computed by:

$$\hat{\lambda}^p_{\max} = \max(\hat{\lambda}^p_{d1}, \hat{\lambda}^p_{d2}, \dots, \hat{\lambda}^p_{d7}, \hat{\lambda}^p_{w1}, \hat{\lambda}^p_{w2}, \dots, \hat{\lambda}^p_{wm})$$
$$\hat{\mu}^p_{\max} = \max(\hat{\mu}^p_{d1}, \hat{\mu}^p_{d2}, \dots, \hat{\mu}^p_{d7}, \hat{\mu}^p_{w1}, \hat{\mu}^p_{w2}, \dots, \hat{\mu}^p_{wm})$$

where $\hat{\lambda}^p_{\max}$ and $\hat{\mu}^p_{\max}$ are the maximum likelihood estimates of number of calls and the call duration for user p over the number of days specified respectively. The thresholds define the limits for all weekday/weekend and week by week. Our assumption is that the calling pattern could be different. Each caller has his/her own thresholds, and if the number of calls or call duration is greater than their usual thresholds for some day, we define that some event has occured in that day.

To calculate the threshold of number of calls for user p, N^p_{thres}, we define

$$N^p_{thres} = \hat{\lambda}^p_{\max} + \hat{\sigma}^p_{\max}$$

where $\hat{\lambda}^p_{\max}$ and $\hat{\sigma}^p_{\max}$ are the maximum rate of number of calls and correspondent standard error with positive $\hat{\sigma}^p_{\max}$.

To calculate the threshold of call duration for user p, D^p_{thres}, we define

$$D^p_{thres} = \hat{\mu}^p_{\max} + \hat{\delta}^p_{\max}$$

where $\hat{\mu}^p_{\max}$ and $\hat{\delta}^p_{\max}$ are the maximum mean of call duration and correspondent standard error with positive $\hat{\delta}^p_{\max}$.

Definition of an unusual consumption event

A collection of call-log data can be represented as $C =< (t_1, a_1, d_1, l_1), (t_2, a_2, d_2, l_2), \dots (t_n, a_n, d_n, l_n) >$, where t_i is a time point, d_i is a call duration, l_i is a location and a_i is a pair of actors, caller-callee $< s_i$, $r_i >$ where s_i is an actor who initiates a call at time t_i and r_i is an actor who receives the call. An unusual consumption event is defined as a subset E $\subset$C of a tuple

$E = \{(t_1,a_1,d_1,l_1),(t_2,a_2,d_2,l_2),...(t_m,a_m,d_m,l_m)\}$

such that either $\sum_{i=1}^{m} d_i > D_{thres}$ or $count(d_i) > N_{thres}$ defined as the above in the time period $\Delta t = t_m - t_1$.

V. Pattern Recognition

Opt-in is an approach to e-mail or phone marketing in which customers must explicitly request to be included in an e-mail or phone call campaign or newsletter. In addition, customers can easily choose to be removed from a mailing or phone list if the advertisements are unwanted, thus eliminating unsolicited emails or phone calls. People may be interested in some advertisements for a period of time, but will not want receive those advertisements later. Ultimately, the customer comes to consider this traffic as spam and decides to opt-out. We believe that current spam filters have great difficulty detecting this type of traffic. Note that several kinds of opt-ins exist. We consider opt-ins whom customers show lot interest for a short period of time and later have no interest but still keep getting unwanted emails or calls as *opt-in* bursts.

Another instance where researchers and developers can find pattern recognition useful occurs with presence awareness. We propose use of a Bayesian inference model to compute the willingness level of people's communications with one another at a given time. Another example of willingness level of people's communications is a computer and telecommunication presence. Emergency of presence-aware communications allows people to quickly connect with others via the best choice of communication mean, whether on the road, in meetings, or working from remote locations. Presence awareness also lets users know when others in their contact list are online. For those interested in studying presence awareness, presence information can include more user details, such as availability, location, activity, device capability and other communication preferences. Researchers and developers can use presence to detect and convey willingness and ability to talk on the phone. Presence-enabled telephony services can reduce telephone traffic, as well as tag and improve customer satisfaction. The existing approaches provide presence for “online,” “busy,” “away,'' “offline,” etc.

A. Opt-in Detection

Opt-in burst detection is related to burst detection on data streams and to time series which are continuous data. However, the Opt-in behavior resembles accumulated impulses and is not continuous. The existing approaches to detect bursts are used for text, novel and unusual data points or segments in time-series that either have contents or are traffic data. *However, none of the previous work focuses on the specific problem we study here, opt-in bursts by studying the calling pattern based on call detail records to detect opt-in bursts that reflect human activity.*

We define the opt-in bursts as dense sequences of accumulated impulses with an interval of length *w*.

Let B={b$_1$, … b$_k$} be a subsequence of bursts contained in a sequence S ={s$_1$, …, s$_n$}. The *i*th burst value is defined as

$$b_i(t) = \sum_{j=1}^{w_i} s_t \delta(t - t_j^i)$$

where w$_i$ is total number of impulses of the *i*th burst, i.e. the *i*th burst width, t_j^i is the occurrence point of the *j*th impulse of the *i*th burst and δ(t) is a delta function denoting the occurrence of a impulse at point t= t_j^i.

The *i*th burst amplitude A$_i$ can be calculated as

$$A_i = \frac{1}{w_i}\sum_{j=1}^{w_i} s_t \delta(t - t_j^i)$$

where s$_t$ is the value of an pulse at point t.

To detect the bursts, we define the sliding window SW$_k$ as

$$SW_k(t) = A_k rect(\frac{t - t_m}{\tau_k})$$

where A$_k$ is the amplitude of a sliding window k, *rect((t-t$_m$)/τ$_k$)* is rectangle function denoting the occurrence point of a burst at time t=t$_m$ and τ$_k$ is the width of a sliding window k.

Definition of an opt-in burst

A collection of call log data can be represented as

$C =< (t_1,a_1,d_1,l_1),(t_2,a_2,d_2,l_2),...(t_n,a_n,d_n,l_n) >$,

where t_i is a time point, d_i is a call duration, l_i is a location and a_i is a pair of actors, caller-callee $< s_i$, $r_i >$ where s_i is an actor who initiates a call at time t_i and r_i is an actor who receives a call. An opt-in burst is defined as a subset E $\subset$C of a tuple E = {(t_1, a_1, d_1, l_1), (t_2, a_2, d_2, l_2), …, (t_m, a_m, d_m, l_m) } such that $0 < count(d_i) < N_{thres}$ in the time period $\Delta t = t_m - t_1$, where N_{thres} is a threshold which can be estimated from the historical data.

We process the sequence S by Exponentially Weighted Moving Average (EWMA), and then apply the dynamic-size sliding windows to detect opt-in bursts. The EWMA places more emphasis on the most recent data. Therefore, EWMA would be more useful in dynamic systems.

Let S ={s$_1$, …, s$_n$} be a sequence. Then the moving average (MA) is given by

$$\bar{s}_k = \frac{1}{M}\sum_{i=k-M+1}^{k} s_i$$

where $\bar{s}_k$ is moving average of the k's instance and M is the number of latest values. The EWMA can be derived from MA as

$$\bar{s}_k = (1-\alpha)s_k + \alpha(1-\alpha)s_{k-1} + \alpha^2(1-\alpha)s_{k-2} + \alpha^3\bar{s}_{k-3}$$

where 0 ≤ α < 1 is a constant. This is a recursion function.

B. Willingness Level Inference

When callers want to make a call, they would like to know if the callee is in a mood to receive a call. In other words, callers would like to know when it is a good time to call the particular callees. We estimate the chance that a callee will accept a call based on the time of the day, call duration and the location.

We propose a Bayesian inference model to compute a receiver's willingness level in a given time.

Let X and Y be two events. We have Bayes' theorem [8]:

$$P(X \mid Y) = \frac{P(Y \mid X)P(X)}{P(Y)}$$

P(X | Y) is called posterior probability, P(Y | X) is referred to as likelihood and P(X) is prior probability.

Let X = (incoming call = accept, incoming call = missed)

Let Y = (T_i, D_j, Loc_l), where T_i is time interval, i = 0, 1, 2, …23, (*e.g.* : 0 – 1 O'clock), D_j is a day, j=1,2, …, 7, *i.e.*, D_1=Sunday, D_2=Monday, … D_7= Saturday, Loc_l= location name, l=1, 2, …n

Then, by Bayes theorem, the willingness level to accept a call is

$$P(incoming = accept \mid T_i, D_{j,} Loc_l)$$
$$= \frac{P(T_i, D_j, Loc_l \mid incoming = accept)P(incoming = accept)}{P(T_i, D_j, Loc_l)}$$

VI. Experimental Results and discussions

A. Real-life Data Sets and Parameters

***Real-life traffic profile*:** In this paper, actual call logs are used for analysis. These call logs of 81 users were collected for a period of 8 months at MIT [4] by the Reality Mining Project group. Additionally, the call logs of 20 users were collected for a period of 6 months by the Network Security team at UNT.

The Reality Mining Project group collected data on mobile phone usage of 81 users, including user ID (unique number representing a mobile phone user), time of call, call direction (incoming or outgoing), incoming call description (missed or accepted), talk time, and tower ID (location of phone user). These 81 phone users were students, professors, and staff members. The collection of the call logs was followed by a survey to gather feedback from participating phone users about behavior patterns such as favorite hangout places; service providers; talk-time minutes; and phone users' friends, relatives and parents. More information about the Reality Mining Project can be found in [4].

(Because of the limited space we will show the experimental results in the workshop)

The experimental results show that our approach is effective.

VII. Conclusion

In this paper we propose an integrated platform for social relationship and human behavior analysis based on mobile phone call detail records. Because of the diversities and complexities of human social behavior, one technique cannot detect all the different features of human social behaviors. Thus, we used multiple probability and statistical methods.

We propose a new *reciprocity index* for measuring the levels of reciprocity between users and their communication partners, propose *affinity* model for quantifying social groups, map call-log data into time series and apply *SARIMA* model for predicting social ties, combine *wavelet denoising* and *sequential detection* for detecting change points, propose *inhomogeneous Poisson* and *inhomogeneous exponential model* for detecting unusual consumption events, combine *dynamic slide window* and *exponentially weighted moving average* for detecting opt-in bursts and propose *Bayesian* inference model for quantifying willingness levels.

We may quantify relationships for a short-term period, say a month, or a long-term period, say, a year or more, using our model by adjusting the parameters. Errors appear when the number of calls is small. However, these kinds of cases seldom occurred in our experiments.

This work is useful for homeland security and for detecting unwanted calls, *e.g.*, spam, telecommunication presence, and marketing. The experimental results show that our model achieves high accuracy. In our future work we plan to analyze and study social-network dynamics and evolution.

Acknowledgment

We would like to thank Nathan Eagle and Massachusetts Institute of Technology for providing the call logs of the Reality Mining dataset.

References

[1] S. Wasserman and K. Faust. Social Network Analysis: Methods and Applications. Cambridge University Press, 1994.

[2] N. Eagle, A. Pentland, and D. Lazer. "Inferring friendship network structure by using mobile phone data," in *Proceedings of the National Academy of Sciences*. vol. 106, no. 36, pp. 15274-15278, 2009.

[3] M. Fannes and P. Spincemaile. "The mutual affinity of random measures". *Periodica Mathematica Hungarica,* vol. 47 (1–2), pp. 51–71, 2003.

[4] Massachusetts Institute of Technology: Reality Mining. http://reality.media.mit.edu/ 2009.

[5] L. Katz and J. Powell. "Measurement of the tendency toward reciprocation of choice," *Sociometry*, vol. 18, pp. 659-665, 1955.

[6] J. W. Cangussu, M. Baron. "Automatic identification of change points for the system testing process," in *Proceedings of the 30th Annual IEEE International Computer Software and Applications Conference (COMPSAC 2006),* Chicago, IL, Sept. 18-21, 2006.

[7] J. W. Harris and H. Stocker, *Handbook of Mathematics and Computational Science,* New York: Springer-Verlag, 1998, §21.10.4, p. 824.

[8] N. Nilsson. *Artificial Intelligence, a new synthesis, first edition*, San Fransisco: Morgan Kaufmann Publishers, 1998.

[9] A. Pentland, "Collective intelligence," *IEEE Computational Intelligence Magazine*, vol. 1, pp 9-12, 2006.

[10] A. Pentland, "Automatic mapping and modeling of human networks," *Physica A: Statistical Mechanics and its Applications,* Elsevier, vol. 378, pp. 59-67, 2007.

[11] N. Eagle. "Behavioral inference across cultures: using telephones as a cultural lens," *IEEE Intelligent Systems*, vol. 23, no. 4, pp. 62-64, 2008.

[12] G. E. P. Box and G. M. Jenkins and G. C. Reinsel. *Time series analysis: Forecasting and control*, 3rd ed., Prentice-Hall, Englewood Cliffs, NJ, 1994.

Session B

Timescales in evolving mobile networks

Gautier M. Krings
Université catholique de Louvain

Márton Karsai
Aalto University

Jari Saramäki
Aalto University

Vincent D. Blondel
Université catholique de Louvain

Mobile phone networks have raised a wide interest in the past years, thanks to the recent availability of large sets of mobile phone call records (CDRs). Those datasets consist of calls made by users from mobile phone providers during time periods ranging from a couple of weeks to several years. They provide interesting advantages compared to other datasets as fixed phone call records or emails. First, a mobile phone is a personal object (in opposition to fixed phones that are usually shared between groups of people), this allows to derive a social network directly from the call data records. Secondly, as a call can only involve two persons at the time, the calls contained in the dataset represent real interactions between individuals. This is not always the case, as for email networks, where one email can be sent to a large number of recipients in a small time.

Moreover, those datasets possess a strong temporal component, as the edges of a mobile phone network are originated by calls between users. The timestamp of a call is particularly interesting as it allows to quantify the strength and the regularity of a relationship. This is a significant advantage compared to datasets provided by online social networks, where friendships are usually defined in a static way.

The availability of such datasets have led to various analyses of the structure of mobile phone networks, which followed all a systematic procedure : from the dataset, a digraph is constructed by representing customers as nodes, and drawing an edge between two nodes if they have made at least one call over the whole period. The edges are then usually weighted, either with the number of calls, or with the total calling time between the two users. In some cases, some thresholding is performed to remove outliers from the dataset.

When one uses this approach, the strength of an edge represents the frequency of interaction between two users as a fixed value, with the assumption that the frequency of calls is constant over time. When one thinks about it, the frequency of interaction between two users should logically be varying over time, but the representation of the network by a static graph hides completely the dynamics of the edges.

To avoid this, one needs to represent the network as an evolving graph.

As said earlier, a dataset representing a mobile phone network is a collection of events $((i, j), t)$, meaning that customer i has called customer j at time t.

When the time resolution is high, only a few events happen at a single timestep and the graph is composed of sparse edges changing at a high frequency. It is then hard to make a clear analysis of the dynamical behavior of such a graph.

The classical approach to eliminate the high-frequency noise and extract global trends from an evolving graph is the segmentation of the dataset into time windows. The time interval is sliced into several continuous time windows, and for each window, one static graph that contains all the events happening within the window is built.

This approach is similar to a moving average technique in time series analysis that are used for denoising purposes. It presents also the same issues : if a window is too small, high frequency noise is not eliminated and hides the general trends. If it is too large, the result is a static network, without dynamical information.

Therefore, a good choice for the length of time windows is crucial, but usually left behind by researchers and motivated by intuition.

In this work, we propose a method to analyze the influence of the length of a time window on the graph it generates and to characterize the growth of an evolving graph with time.

We consider the evolving graph as a stochastic process, each possible edge being a variable that appears with a non-stationnar probability. We estimate the quantity of information of a growing set of events and track the redundancy of the set of events. We use the redundancy to quantify the graph growth and as stopping criterion for the graph aggregation.

We apply our method to a mobile phone network of 3 million users making 360 million calls over 6 months, and show that after 3 weeks, the network growth reaches a steady state that can be expressed matematically. We also analyze similarity measures on the sequence of graphs created by windows of different sizes to validate 3 weeks as an adequate window length.

Towards an investigation of the structure and temporal dynamics in a large scale telecoms dataset

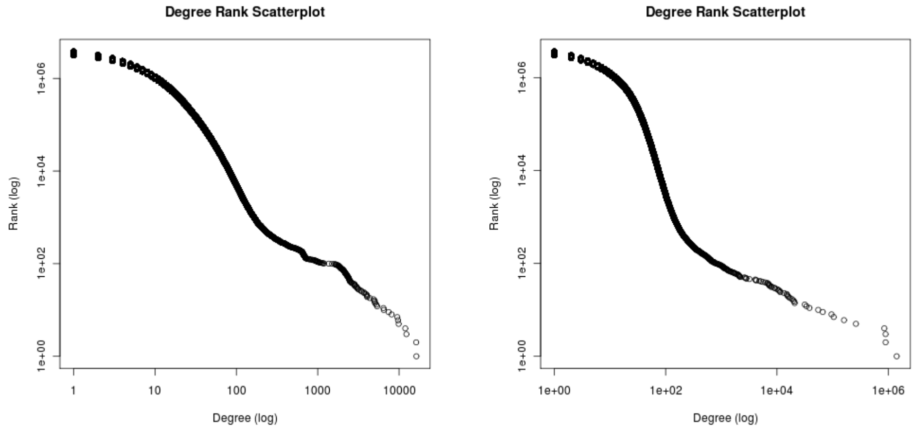
Fergal Reid
Clique Research Cluster
University College Dublin
Dublin 4, Ireland
fergal.reid@gmail.com

Neil Hurley
Clique Research Cluster
University College Dublin
Dublin 4, Ireland
neil.hurley@ucd.ie

1 Motivation

Clique (`clique.ucd.ie`) is a research project started in January 2009, focussing on complex network analysis and visualization. In conjunction with three industrial partners, Clique is addressing four core research challenges, namely: the identification of coherent communities, the identification of nodes that have pivotal roles, the identification of network structure that is remarkable or anomalous and the question of how to model and analyze information flow. Our research concerns datasets that are massive, multi-attribute and dynamic and is driven by the industrial contexts provided by our industrial partners.

One of our research domains is very large scale mobile telecoms datasets. These datasets provide a rich source of human social interaction data and provide an excellent opportunity to investigate large scale complex networks. The particular data provided by our industry partner [1] yields call graphs where a single week worth of data contains many millions of nodes, with the sparse edge structure typical of such networks. Having access to week by week snapshots of this rich data, over a period of time spanning many months, we are conducting research to find structure and regularity in this data, with a view to investigating the underlying social structure and attempting to discover the channels through which influence propagates in complex and social graphs. We are specifically considering the problems of finding interesting graph structures - such as the clusterings of nodes that form communities - and diffusion modeling of user influences in areas such as network churn and handset adoption.

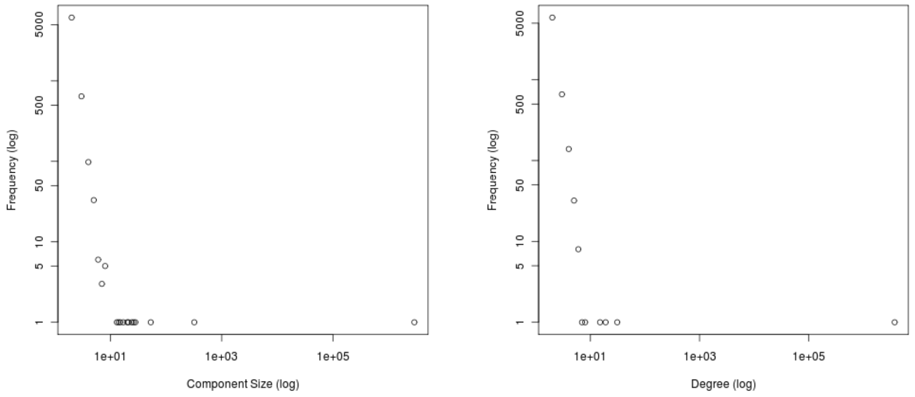


(a)

(b)

Figure 1: The degree distribution of (a) one months worth of raw call data and (b) one months worth of SMS data. No filtering of high degree nodes, call centers etc has been performed on this data.

[1] We would like to acknowledge Idiro Technologies for providing us with access to this data.



(a) (b)

Figure 2: The connected component size distribution of (a) one weeks raw SMS data (b) one months raw call data. Huge components clearly visible.

2 Completed Work

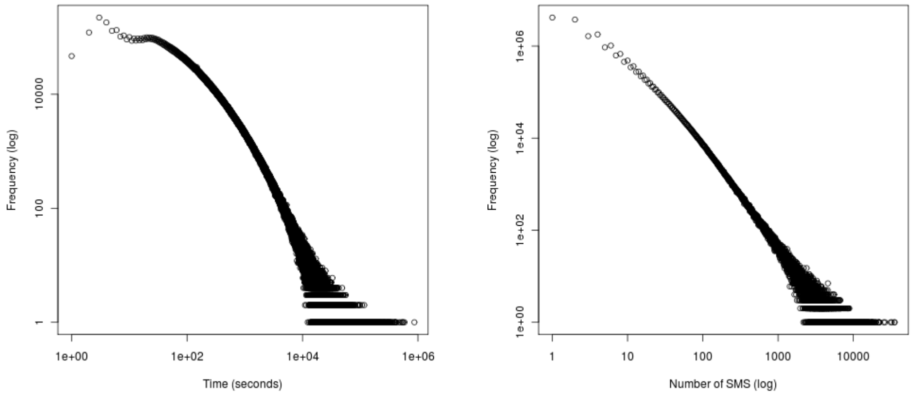
As a prerequisite for developing techniques to predict influence diffusion within a social graph, it is first necessary to investigate the structure and dynamics of the graph. Much work has been done on community assignment in graph datasets - thoroughly summarized in the review of [Fortunato, 2010]. However, after the analysis of [Leskovec et al., 2008] some question exists over whether the existing community assignment solutions will successfully operate within the densely overlapping cores of large social graphs, and effectively extract useful community structure. It is against this backdrop that we attempt to find structure in our data.

As a first step, we attempt to characterize our data in terms of several metrics that have previously proven useful in the study of complex networks. The degree distribution of static graphs extracted from a month long time period is shown in Figure 1, where we separately examine call and SMS data. Note that the degree distribution of the both networks, when viewed as an undirected, unweighted graph, exhibits the heavy tail properties previously described in the work of [Onnela et al., 2007] and [Dasgupta et al., 2008]. In Figure 2, the distribution of connected component sizes of a weeks worth of SMS data, and a months worth of call data is shown. Again, in agreement with the findings of [Onnela et al., 2007],[Dasgupta et al., 2008] the vast majority of nodes are part of a huge connected component. This result holds true in SMS and call graphs, and even when we consider graphs covering time periods as short as one week.

In reality, the edges in the call graph have multiple attributes, including call type, total call duration and total number of calls, accumulated over the extracted time-period. In Figure 3, the distribution of edges weighs is plotted, where the edges are weighted according to firstly the total number of seconds two nodes spend calling each other, and secondly the total number of SMSs two nodes send to each other.

Having conducted some simple characterization, we have attempted to replicate the experiments of [Onnela et al., 2007], which investigate the 'weak tie hypothesis' of [Granovetter, 1973]. Our initial results indicate that our dataset, whether we view a weeks worth of data at a time, or consider longer time periods, such as month by month, also displays the property that the ties that are most important for network connectivity are those with smaller dyadic edge weight. Our work indicates that this analysis holds regardless of whether we weight edges according to the number of calls sent and received, or by call duration, or by number of SMS messages exchanged. These initial results are part of an on-going analysis and we expect that structural characteristics such as this will have an important impact on the development of a good model of influence diffusion.

Investigating correlations between different graph statistics is also important in order to obtain a useful characterization of the data. Currently, we are investigating whether relationships exist between the degree of a node and the summed weight of its edges, whether there is assortivity [Newman, 2002] between high degree nodes in our dataset, and to what extent the SMS and call graphs are similar to, or different from, each other. A further area of research is an attempt to classify nodes by the proportion of communications that they emit and receive, and to study the



(a) (b)

Figure 3: The edge weight distribution of (a) one months worth of raw SMS data (b) one months worth of raw call data.

dynamics of changing calling patterns across nodes with different local structural properties.

Edge significance State-of-the-art research has attempted to filter the more ‘casual’ edges from a mobile telecoms network in order to eliminate noise from the social graph, and also decrease the computational complexity of processing such large networks. For instance, reciprocated communications are used in [Lambiotte et al., 2008][Dasgupta et al., 2008] as evidence of more meaningful social connection. With similar motivations, and believing network dynamics an interesting area of investigation, we are attempt to characterize the dynamic properties of edges, such as the distribution of the length of time an edge that appears in the graph persists for, and correlating this with other edge properties such as call duration, bidirectionality, and number of calls.

Influence diffusion In addition to anonymised summary call data records, we also have access to data describing which network users *churned* from the network at certain times. We also have information on the handset currently used by each user. One of our ongoing goals is to investigate whether we can find a significant social dimension to the spread of changes in handset adoption, and network churn, after [Dasgupta et al., 2008], as a tool of investigating the social properties of the spread of influence.

3 Outlook

We intend to proceed along this line of investigation, attempting to find structure in the the social network we approximate with mobile call graphs. We are investigating whether well defined structure can be found in such large networks. We are also investigating the impact of social effects on influence change, specifically on churn and handset uptake. We are scientifically interested in whether we can develop a generic way of investigating influence spread that can extend across mobile datasets and to other complex network domains, and whether we can leverage the dynamic nature of the graph across time to make better models and find richer structure than we could with a single time slice.

We hope to present some results on this ongoing work by the time of the workshop.

References

K. Dasgupta, R. Singh, B. Viswanathan, D. Chakraborty, S. Mukherjea, A.A. Nanavati, and A. Joshi. Social ties and their relevance to churn in mobile telecom networks. In *Proceedings of the 11th international conference on Extending database technology: Advances in database technology*, pages 668–677. ACM, 2008.

Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75 – 174, 2010. ISSN 0370-1573. doi: DOI:10.1016/j.physrep.2009.11.002.

M.S. Granovetter. The strength of weak ties. *American journal of sociology*, 78(6):1360, 1973.

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(21):5317–5325, 2008.

J. Leskovec, K.J. Lang, A. Dasgupta, and M.W. Mahoney. Statistical properties of community structure in large social and information networks. 2008.

MEJ Newman. Assortative mixing in networks. *Physical Review Letters*, 89(20):208701, 2002.

J.P. Onnela, J. Saramaki, J. Hyvonen, 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*, 104(18):7332, 2007.

Dynamics and temporal correlations in mobile phone based social networks

Márton Karsai*,[1] Lauri Kovanen†,[1] Mikko Kivelä‡,[1] Raj Kumar Pan§,[1] Jari Saramäki♯,[1] János Kertész**,[2, 1] Albert-László Barabási††,[3, 4] and Kimmo Kaski‡‡[1]

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

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

[3]CCNR, Northeastern University, Boston, MA, USA

[4]CCSB, Dana-Farber Cancer Institute, Boston, MA , USA

Complex networks are rarely static, but involve dynamics on several time scales. Often, the networks studied in the literature represent either static "snapshots" of dynamic entities or aggregates over some period of time. In the first case, all dynamics are entirely lost, while in the second case some trace of dynamics can be retained e.g. by representing edge activation frequencies as edge weights. However, in this case, all information on short-term temporal correlations between dynamics of edges is lost.

Here, we study the microdynamics of a large social network reconstructed from mobile communication billing records with time stamps for each communication event (calls and text messages). It is natural to assume that such events are temporally correlated i.e. incoming calls trigger outgoing calls as some information is relayed, and call patterns of social groups contain temporal correlations as messages are exchanged between group members in a conversation.

We begin with the smallest scale interactions, showing that there is clear evidence of correlations where incoming calls trigger outgoing calls, and discuss the associated time scales. Then we move on to multipoint correlations by showing that "temporal motifs" exist within short time windows in significantly greater numbers than in a temporally uncorrelated reference system. Having shown the existence of correlations, we assess their significance in the function of a social system in terms of information transmission. We show that unexpectedly, temporal correlations appear to slow down the transmission dynamics compared to a null reference.

emails:

∗ mkarsai@lce.hut.fi
† lkovanen@lce.hut.fi
‡ mtkivela@cc.hut.fi
§ rajkp@lce.hut.fi

♯ jsaramak@cc.hut.fi
∗∗ kertesz@phy.bme.hu
†† alb@neu.edu
‡‡ Kimmo.Kaski@hut.fi

Communication Motifs: A Novel Approach to Characterize Mobile Communications

Qiankun Zhao Nuria Oliver
Telefonica Research and Development, Spain
zhao@tid.es nuria@tid.es

Abstract: Social networks mediate not only the relations between entities, but also the patterns of information propagation among them and their communication behavior. In this paper, we extensively study the temporal annotations (*e.g.*, time stamps and duration) of historical communications in mobile phone networks and propose two novel tools for – qualitative and quantitative – characterizations of the patterns of information propagation in these networks. Specifically, we define *communication motifs* and *maximum-flow communication motifs* in a mobile network as structures that satisfy certain communication and topological constraints. We propose two motif discovery algorithms and apply them to the mobile phone network. Using the discovered motifs to characterize the communication behavior and information propagation patterns in the mobile networks, we verify the following hypothesis: 1) the functional behavioral patterns of information propagation within mobile networks are stable over time; 2) the patterns of information propagation in mobile networks seem to be sensitive to the cost of communication; and 3) the speed and the amount of information that is propagated through a mobile network are correlated and dependent on individual profiles.

Summary: Social networks represent the links between a set of entities connected to each other with different types of relationships. For example, papers are linked by citations in a citation network, bloggers are linked by comments or blogrolls in a blog network, while cell phones are connected via phone calls in a cell phone network [8, 11, 1]. In the literature, social networks have been extensively studied from a graph theory perspective (*e.g.*, power laws, small worlds phenomenon, coverage, etc.) [3, 6, 2]. Properties of different types of complex networks have been compared [9, 11, 3]. Recently, research studies on social networks from a behavioral perspective have received a lot of attention. These works, dealing with problems such as community identification, spam detection, or modeling information flows [8, 4] have a lot of applications in recommender systems, social search, economics, and advertising [12, 5, 7].

In this paper, we study the communication characteristics (from a behavioral perspective) within mobile phone networks. In particular, we aim at identifying topological subgraph structures with frequency and temporal constraints, which we refer to as *motifs*, to characterize communications in mobile networks.

caller	callee	timestamp	duration
2	*3*	*9:00* am, 11-Aug-08	*2* minutes
3	*4*	*9:10* am, 11-Aug-08	*5* minutes
4	*3*	*9:41* am, 11-Aug-08	*3* minutes
3	*7*	*9:51* am, 11-Aug-08	*2* minutes
3	*8*	*10:36* am, 15-Aug-08	*6* minutes
7	*4*	*10:08* am, 11-Aug-08	*6* minutes
7	*2*	*10:43* am, 13-Aug-08	*6* minutes
8	*4*	*9:36* pm, 15-Aug-08	*6* minutes

Table 1: CDR(Call Detail Records) data

A fundamental issue in analyzing information flow or propagation patterns within communication oriented social networks is how to represent the communication data in such a way that it captures every piece of useful information. Consider the example in Table 1, which shows eight entries from a phone call detail record (CDR) dataset, where calls between users and their associated durations and timestamps are recorded, with one entry per call. A number of alternative representations of the CDR data in Table 1 are shown in Figure 1. Here G_1 is constructed by taking each user as a node and each call between any two users as an edge between the two corresponding nodes, while G_2 extends G_1 by adding weights to the edges proportional to the frequency of calls between any two nodes) and G_3 is obtained by removing the direction of the edges in G_1.

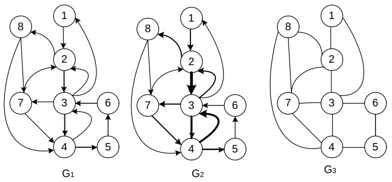


Figure 1: Graph Representations of CDR data.

These representations are meaningful and valid in certain social networks such as friends or citation networks, where the nature of the relationship is embedded in – or may be easily derived from – the records. However, in the case of social networks derived from communication logs, it is difficult to properly infer the nature of the relationships due to the multiplicity of reasons for making a call (*e.g.*, business, personal, service, etc.) and the importance of the temporal context. In other words, once one paper cited another, the *cited* relationship between both papers always holds true. However, phone calls are made for different reasons and hence the nature of a relationship between two nodes in the network may also depend on the temporal context of calls, *i.e.*, a call made during working hours is probably of different nature than a call made at night. The same applies to other temporal attributes such as duration and frequency of the interaction, or temporal distance between two calls (inter-call time delay). As a result, the representations that are used in existing information propagation studies are not valid under the context of our study in this paper. Furthermore, many studies on information propagation assume that consecutive interactions transmit the same piece of information within the inferred networks, and we shall argue that this assumption is not always valid.

In [6], Kleinberg *et al.* notice the importance of the temporal annotation in instances of communications. They find an information pathway where users are updated with the latest information at the quickest speed based on the temporal distance between the communications. However, their work does not consider the case of different pieces of information being propagated in the same network. Moreover, by using part of the temporal information only, they are not able to generalize the local behavior patterns of individuals or small groups to the entire network.

In this paper, we propose to study information propagation and behavior patterns in communication oriented social networks, taking a mobile phone network as an example, by leveraging the **temporal annotations** of these communications together with the topological structure of the network. The proposed approach is based on the following observations:

- Calls or interactions between the same two users in the network may have different purposes and thus transmit different information, causing different effects in terms of information propagation. The interactions between any two users may range from being intense and frequent to being isolated events without further impact on the network. As a result, we propose to differentiate the calls by incorporating the temporal information of the calls into the social networks.
- The semantics associated with each interaction (*e.g.*, topics discussed, purpose of a call) are hard to infer [9, 11, 3]. It is possible that two adjacent interactions (*i.e,* interactions that share at least one common user) have no causal relationship between them. However, the temporal attributes associated with the interactions may shed some light on the amount of information propagated. For example, it seems reasonable to assume that the closer in time two adjacent interactions take place, the more likely it is that they are about the same topic.
- The amount of information passed from one node to another in the social network may be quantified in different ways. For example, in a CDR dataset the amount of information can be quantified by the duration of the call, whereas in a Facebook dataset the amount of information can be quantified by the length of the text typed on a user's wall. In both cases, we assume that the longer an interaction is, the larger amount of information is propagated via this interaction.

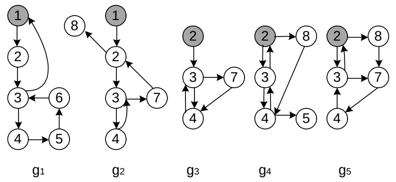


Figure 2: Representation with temporal constraint.

Based on the above observations, we aim to find novel and meaningful ways to model the users' behaviors and information flow patterns. For example, based on the CDR data in Table 1, five different calling sequences are constructed by taking the call timestamps into account as shown in Figure 2. In these graphs, two edges are connected if and only if their corresponding timestamps are within 30 minutes of each other. Compared to the graphs in Figure 1, the five sequences in Figure 2 take into account the temporal information of the calls.

In order to model how information is propagated within a mobile phone network, we propose the concepts of *communication motifs* and *maximum-flow motifs*. The proposed motif concepts are an extension of the network motifs in biological networks [10], which refer to patterns that recur within a network much more often than at random. We propose two efficient algorithms to automatically identify communication motifs in a network, carry out extensive experiments with a large mobile phone network dataset and discuss the properties exhibited by the discovered motifs. Finally, we characterize the information propagation behavior in the mobile phone network using these motifs as a measurement, which leads us to the following findings:

- The functional behavioral patterns of information propagation are stable over time;
- The patterns of information propagation are sensitive to the cost of communication;
- The amount of information being propagated and its speed are correlated and depend on user profiles in the mobile phone network.

As supported by our research results, we shall claim that the temporal attributes associated with communication data are critical to model information propagation and characterize the nature of the relationship between any two nodes in the network. In future work, we plan to continue this line of research.

1. REFERENCES

[1] A. Anagnostopoulos, R. Kumar, and M. Mahdian. Influence and correlation in social networks. In *ACM SIGKDD*, pages 7–15, 2008.

[2] W. Chen, Y. Wang, and S. Yang. Efficient influence maximization in social networks. In *ACM SIGKDD*, pages 199–208, 2009.

[3] N. Du, C. Faloutsos, B. Wang, and L. Akoglu. Large human communication networks: patterns and a utility-driven generator. In *KDD '09*, 2009.

[4] L. Guo, E. Tan, S. Chen, X. Zhang, and Y. E. Zhao. Analyzing patterns of user content generation in online social networks. In *ACM SIGKDD*, pages 369–378, 2009.

[5] A. Gürsel and S. Sen. Producing timely recommendations from social networks through targeted search. In *AAMAS*, 2009.

[6] G. Kossinets, J. M. Kleinberg, and D. J. Watts. The structure of information pathways in a social communication network. In *KDD*, 2008.

[7] T. Lappas, K. Liu, and E. Terzi. Finding a team of experts in social networks. In *ACM SIGKDD*, pages 467–476, 2009.

[8] Y.-R. Lin, Y. Chi, S. Zhu, H. Sundaram, and B. L. Tseng. Facetnet: a framework for analyzing communities and their evolutions in dynamic networks. In *WWW*, 2008.

[9] R. MIlo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon. Superfamilies of evolved and designed networks. *Science*, 303(5):1538–1542, March 2003.

[10] R. MIlo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: Simple building blocks of complex networks. *Science*, 298(25):824–827, October 2002.

[11] J.-P. Onnela, J. Sarama, J. Hyvonen, G. Szabo, D. Lazer, K. Kaski, J. Kertesz, and A.-L. Barabasi. Structure and tie strengths in mobile communication networks. *PNAS*, 104(18):7332–7336, May 2007.

[12] P. Raghavan. The changing face of web search: algorithms, auctions and advertising. In *STOC*, 2006.

Mobility, Data Mining and Privacy: The GeoPKDD Paradigm

Fosca Giannotti
KDD Lab - ISTI - CNR
Pisa, Italy

Fabio Pinelli
KDD Lab - ISTI - CNR
Pisa, Italy

Salvatore Rinzivillo
KDD Lab - ISTI - CNR
Pisa, Italy

Roberto Trasarti
KDD Lab - ISTI - CNR
Pisa, Italy

March 4, 2010

Abstract

The technologies of mobile communications and ubiquitous computing pervade our society, and wireless networks sense the movement of people and vehicles, generating large volumes of mobility data. Miniaturization, wearability, pervasiveness are producing traces of our mobile activity, with increasing positioning accuracy and semantic richness: location data from mobile phones (GSM cell positions), GPS tracks from mobile devices receiving geo-positions from satellites, etc. This paper gives a short overview of the analytical tools developed within the European Project GeoPKDD (Geographic Privacy-aware Knowledge Discovery and Delivery), a project funded by European Commission under the FET program of the IST FP6 framework.

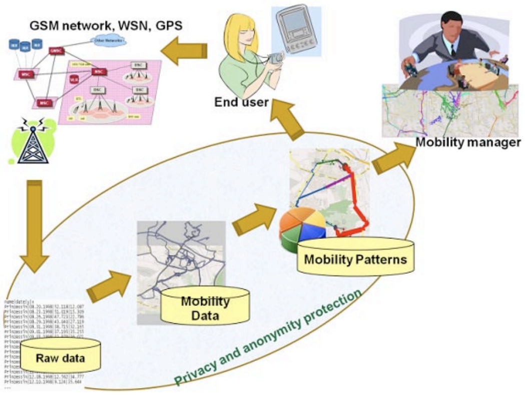


Figure 1: The GeoPKDD process

1 Introduction

Research on moving-object data analysis has been recently fostered by the widespread diffusion of new techniques and systems for monitoring, collecting and storing location aware data, generated by a wealth of technological infrastructures, such as GPS positioning and wireless networks. These have made available massive repositories of spatio-temporal data recording human mobile activities, that call for suitable analytical methods, capable of enabling the development of innovative, location-aware applications. The GeoPKDD project [1], since 2005, investigates how to discover useful knowledge about human movement behavior from mobility data, while preserving the privacy of the people under observation. GeoPKDD aims at improving decision-making in many mobility-related tasks, especially in metropolitan areas.

The GeoPKDD system, originally presented in [5], allows to handle the whole knowledge discovery process from mobility data, in particular it provides tools for reconstructing a trajectory from raw logs, storing and querying trajectory data, classifying trajectories according to means of transportation (pedestrian, private vehicle, public transportation vehicle, extracting spatio-temporal pattern and models as useful abstractions of mobility data, find an optimal tradeoff between privacy protection and quality of the analysis.

2 Experimenting on real GSM data

To demonstrate the power of our framework we have tested it on different real scenarios and different data sources. Here we present a set of experiments on a dataset of real GSM data logs. The observations are collected by the *Telecom Italia Lab*, the research laboratory of the main telecommunication company in Italy, using an estimation of the position of the devices by means of triangulation. The dataset contains the points recorded along a whole day (in particular the 21st May, 2009). The first step was the importing of the observations in the data management system which is based on Oracle 11g database. The trajectories are built starting from this raw data using the trajectory reconstruction algorithm, and cleaning them from errors and outliers obtaining a set of useful data (Fig.2).

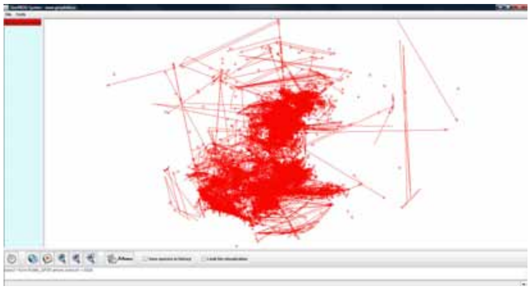


Figure 2: The reconstructed moving points dataset

The resulting set is composed by more than four million trajectories in the urban area of Rome.

2.1 Statistical Analysis. A set of statistical analysis can be easily computed having the data storage system integrated with a set of spatio-temporal primitives, that allow to efficiently compute spatial and temporal measures, like temporal gaps and spatial distances between consecutive points.

2.1.1 Distribution of movements during the day. For this analysis we partition the day in hours (0-24) and we will intersect the trajectory dataset with this periods counting the presence of the trajectories in them. The result is shown in Fig.3.

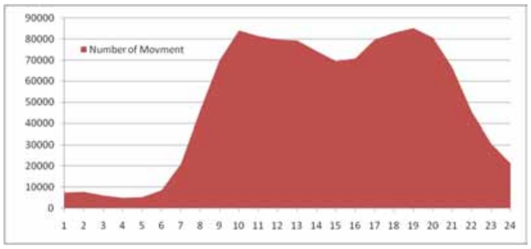


Figure 3: The Time distribution of movements

The analysis shows two major peeks during the day corresponding to the periods of time where the people are going or coming from work. Another aspect of the mobility is highlighted: the period between the peeks has a very high number of movements which gives to the mobility agent a clue about the sustainability of the traffic during such hours.

2.1.2 Density of movements in space. The distribution of movements can be analyzed not only in time but also in space. For example, by dividing the territory in a grid of 50×50 cells, we can compute the density within each cell. In this case we can take advantage from the previous analysis obtaining a spatio-temporal density distribution which can be navigated in both dimensions. The global result is shown in Fig.4.

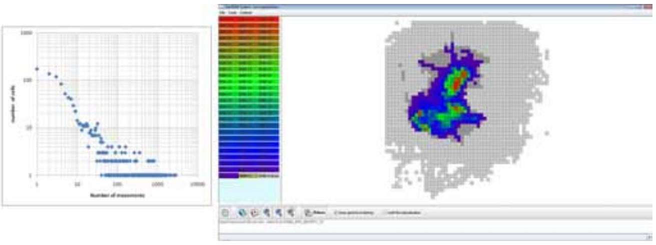


Figure 4: The density distribution plot (Left) and the density distribution on the grid (Right)

Using the temporal dimension we can focus the view only in a specific period, say from 6 am to 12 am, obtaining the Fig.5. As we can see in the morning the mobility is greater between two dense points in the south part of the city giving to the mobility manager the idea of where the peek of detected in the first analysis is focused.

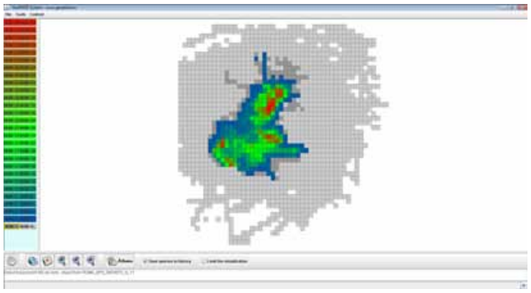


Figure 5: The density distribution in the morningof movements

To better understand the traffic flows of the traffic we proceed with the next statistical analysis called O-D Matrix.

2.1.3 O-D Matrix. Usually this analysis is obtained by the mobility agencies using a survey obtained from direct or by telephonic interview to the citizens. The quality of the data obtained in this way is very poor and has very high costs. Instead thanks to our system this can be done with a very low cost and high precision avoiding incomplete and incorrect data. For this example we introduce a bigger grid 20×20 which simulates the districts of the city. Having more information like real districts or regions of interest given by the mobility agency we can use them to perform this analysis. Joining the matrix with the used grid we can browse it by selecting a region as origin or destination. In Fig.6 we show how the people move from and to a cells.

The next section will show a further step: the statistical analysis using data mining algorithms such as *T-Clustering*

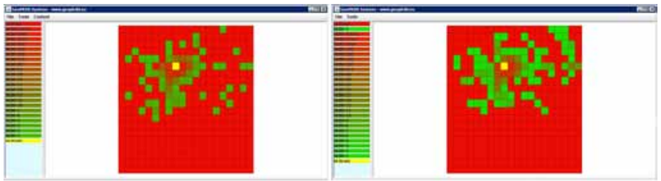


Figure 6: The OD Matrix focused on the yellow cell. The destinations (Left) and the origins (Right)

[3] and *T-Pattern* [2] and how they can interact with the previous analysis thanks to the unifying system.

2.2 Data mining analysis. The understanding of the mobility in a city is a very complex process, here we present an example of how the presented framework helps the analyst in the discovery process. An example of analysis is *looking for common behavior of people who move toward a common destination*. To perform this task the first step is to find this communities of people: the Fig.7 shows two clusters obtained applying the *T-Clustering* using the *common destination* distance function. Once the analyst identifies the clusters of intersect, it is possible to refine the analysis by investigating other regularities in their behavior, for example by applying the *T-Pattern* algorithm on the trajectories of one of the selected cluster (say, for example, the red cluster in Fig.7). The resulting T-Patterns are shown in Fig.8.

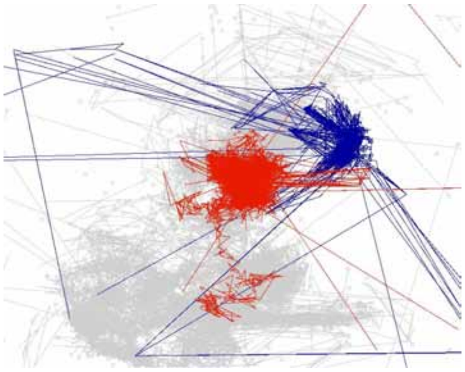


Figure 7: Two Clusters discovered using common destination distance function

To give semantic to the *T-patterns* extracted we can intersect the regions of the *T-Patterns* with a set of interesting places (specified by the mobility agency), and discover that the T-Pattern in Fig.8(b) represents the behavior of people coming in a common area. This simple process show the capabilities of the system allowing the user to perform iterative querying mixing together different data mining algorithms obtaining a deep understanding of the data.

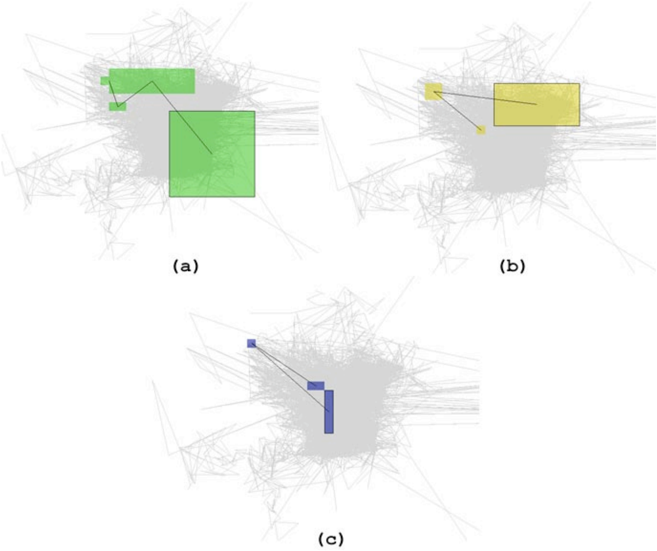


Figure 8: The T-Patterns discovered on the trajectories in a cluster

The approach of combining the extracted patterns with the information available in the system can be also exploited to learn another type of model: the *Location Prediction* [4]. The location prediction algorithm aggregate the local patterns found in the previous step to produce a global model for the considered dataset: the model gives a high level description of the mobility allowing also to predict the possible destinations of an individual by observing his/her movements in the recent past.

3 Conclusions

The analysis capabilities of our system have been applied onto a massive real life GSM dataset and we demonstrated how the various methods and systems developed in the project support the creation of novel analytical services for mobility management, such as: i)Analysis of the movements in space and time ii) the automated construction of origin/destination matrices from mobility data in a timely, reliable and objective manner. It allows to analyze users's flows between geographical areas, overcoming the limitations of the current survey-based approach; iii) discovery of mobility patterns with different data mining tools which can be combined to go deep on the data understanding iv) the detailed analysis and discovery of systematic movement behaviors, i.e., the movements that repeat periodically during the week, with particular emphasis to commuting patterns like home-to-work and work-to-home.

References

[1] Geographic Privacy-aware Knowledge Discovery and Delivery Project. GeoPKDD, http://www.geopkdd.eu/

[2] F. Giannotti, M. Nanni, F. Pinelli, and D. Pedreschi. Trajectory pattern mining. In KDD, pages 330-339, 2007.

[3] G.Andrienko, N. Andrienko, S. Rinzivillo, M. Nanni and Dino Pedreschi. A Visual Analytics Toolkit for Cluster-Based Classification of Mobility Data. SSTD, pag. 432-435(2009)

[4] Anna Monreale, Fabio Pinelli, Roberto Trasarti, Fosca Giannotti: WhereNext: a Location Predictor on Trajectory Pattern Mining.KDD 2009. 15th ACM SIGKDD Conference on Knoledge Discovery and Data Mining

[5] Riccardo Ortale, E. Ritacco, Nikos Pelekis, Roberto Trasarti, Gianni Costa, Fosca Giannotti, Giuseppe Manco, Chiara Renso, Yannis Theodoridis: The DAEDALUS Framework: Progressive Querying and Mining of Movement Data. ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM GIS 2008).

Session C

Authors: Pu Wang, Marta C. González, Ronaldo Menezes and Albert-László Barabási

Affiliations:

P. Wang is with the Center for Complex Network Research, Northeastern University, Boston, Massachusetts, USA and the Department of Physics, University of Notre Dame, Notre Dame, Indiana, USA

M. González is with the Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

R. Menezes is with the Department of Computer Sciences, Florida Institute of Technology, Melbourne, Florida, USA

A.-L. Barabási is with the Center for Complex Network Research, Northeastern University, Boston, Massachusetts, USA and the Department of Medicine, Harvard Medical School, Harvard University, Cambridge, Massachusetts, USA

Emails:

pwang2@nd.edu

marta.gonzalez.v@gmail.com

rmenezes@cs.fit.edu

barabasi@gmail.com

Abstract:

The history of technological viruses is intrinsically linked to the history of computational devices. Since the inception of the Internet, programmers begun writing self-replicating executables such as Creeper, the first known instance of a computer virus. From there on, the field of computer security grew together with the ability of programmers to write increasingly more sophisticated viruses. In recent years mobile phone devices have become the new frontier for self-replicating programs. The availability of these devices together with its constant online presence makes them an ideal breeding environment for technological viruses. Yet, society has not been affected by major mobile virus outbreaks. A natural question then becomes: Why are desktop and laptop computers so vulnerable to a problem that mobile phones do not seem to be? This paper shows that the spread of mobile phone viruses is limited by the type of viral behavior (topological or scanning), the existence of a market share providing sufficient homogeneity in the world of devices, and the mobility of individuals. We find that hybrid mobile viruses that include some level of random scanning have a higher chance to bring havoc to the mobile community than their standard topological counterparts even under adverse conditions to the virus, such as protection mechanisms used by phone providers.

Human movements and the spread of infectious diseases

Vitaly Belik[1], *Theo Geisel*[1] *and Dirk Brockmann*[2]

[1]Max-Planck-Institute for Dynamics and Self-Organization, Göttingen, Germany

[2]Northwestern University, Evanston, IL, USA

Abstract

Modern technologies allow collection and processing of an abundance of information on human activity, in particular, its mobility patterns. This opens unprecedented opportunities for development of computational models of many human-mediated phenomena such as the spread of infectious diseases. The main modeling approach for spatial epidemics explicitly incorporating traveling behavior of the host is the reaction-diffusion model. It assumes random, markovian movements of the host. This approach seems to be inappropriate for humans that typically return to their home locations after traveling. We investigate a model explicitly incorporating bidirectional human movements on star-like network topologies consisting of central home nodes and distant locations. We show for various topologies that dependent on parameters both models exhibit strong differences as well as similarities. An important result is the attenuation of epidemics in bidirectional systems as compared to generic reaction-diffusion systems. Global outbreak of an epidemic is determined by a threshold value of the characteristic time spent by an individual on distant locations. Our results provide a framework for incorporating an abundant data on human mobility available today.

1 Introduction

Infectious diseases remains a pressing challenge for humankind. Understanding the dynamics of epidemics would significantly contribute to the battle against them. In the past mathematical epidemiology approached this problem relying on rare empirical data and posteriorly predictions. Nowadays availability of large amount of data on human activity, in particular its mobility patterns, allows us to approach an ultimate goal of forecasting the progress of an epidemic and planning response measures on the fly. We already experienced such kind of forecasting during the recent outbreak of H1N1 influenza pandemic [1, 2]. However, the detailed computational platforms simulating epidemic spread could be used in non-ambiguous way only if one fully understands the driving factors and essential constituents of a model. This requires a thorough theoretical investigation of the underlying epidemiological model. To this end in the present contribution we address the role of human mobility on the spread of infectious diseases.

Recent empirical studies on human mobility revealed complex although universal dynamical features of human movements. The laws governing human mobility include anomalous diffusion [3] and high degree of predictability [4, 5]. An important property of human mobility is their tendency to return to a few most preferred locations such as home and workplace. This empirical evidence lead us to devising an epidemiological model, explicitly incorporating the last property.

A major theoretical framework for modeling the spread of infectious diseases explicitly taking into account host movements is the reaction-diffusion approach. This approach assumes that hosts are indistinguishable and move randomly between all available locations like chemical particles. However recent empirical findings suggest a high degree of regularity and predictability in human movements and thus the reaction-diffusion approach is not appropriate for description of human epidemics.

In what follows we introduce a model explicitly incorporating human travel into epidemiological framework. We show for various topologies that there can be a significant difference in the behaviour of reaction-diffusion and bidirectional models. We discovered a novel kind of a threshold behaviour of an epidemic outbreak depending on the dwelling time of individuals on distant locations. We show that even in the parameter range where the global outbreak occurs, there is still a pronounced differences in the speed of propagation of both kinds of epidemics.

2 Model

We consider an epidemiological model explicitly incorporating movements of hosts distinguishable according to their location of origin [6]. If X_n^m denotes the number of individuals originating from location m and sojourning in location n. Then the traveling between different locations is described by the following scheme

$$X_n^m \underset{\omega_{mk}^n}{\overset{\omega_{km}^n}{\rightleftharpoons}} X_n^k, \tag{1}$$

where ω_{km}^n and ω_{mk}^n are forward and backward travel rates. Note, that considering all individuals indistinguishable we get rid of n-dependence of travel rates $\omega_{km}^n \equiv \omega_{km}\,\forall\, n$ and recover random walk or diffusive travel.

We chose an SIR epidemiological model [7] to model the local infectious dynamics, which subdivide individuals into classes according to their infectious status. These include infecteds I_n^m, susceptibles S_n^m which can catch a disease, and recovereds R_n^m denoting either immune or dead individuals depending on which disease is modeled. Then the local epidemic is described by the following scheme

$$\begin{aligned} I_n^k + S_m^k &\overset{\alpha}{\rightarrow} I_n^k + I_m^k \\ I_n^k &\overset{\beta}{\rightarrow} R_n^k, \end{aligned} \tag{2}$$

where α and β denotes infection and recovery rates. Combined, (1) and (2) define our model. We assume that individuals in locations are well mixed and one can apply standard chemical kinetics rules. Note, that up to now we did not specify any conditions on travel rates ω_{km}^n, which define how and along which links in a mobility network individuals travel. In the following we restrict our-self to the case of traveling on overlapping star like topologies [8, 9], where nodes represent a center of a star and links to the next neighbors in a network of locations represent rays of the star. Individuals of one particular location can travel only on the corresponding star. If they visit one neighbor location, they need to return home, before visiting another neighbor location. Mathematically we can describe this by choosing travel rates as $\omega_{km}^n = \omega_{nm}^n \delta_{kn} + \omega_{kn}^n \delta_{mn}$. We will call this model bidirectional because individuals travel forth and back between locations always returning home before traveling further in contrast to erratic usually not returning random walk travel of reaction-diffusion model. This setup is a good approximation to describe movements of humans, in particular, commuting behaviour.

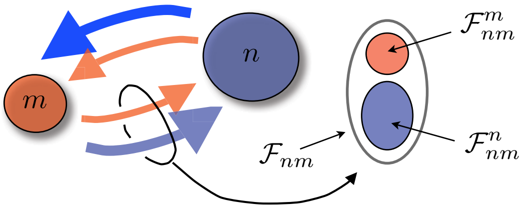


Fig. 1: Illustration of the approach used to parametrize the bidirectional model. Large circles represent two locations populated by different species (red and blue). Color arrows correspond to fluxes of different species. On the right a transection of the total flux $\mathscr{F}_{nm}$ between locations is shown consisting of sub-fluxes $\mathscr{F}_{nm}^n$ (blue) and $\mathscr{F}_{nm}^m$ (red).

2.1 Parametrization

In order to investigate bidirectional model and compare it with widely used reaction-diffusion model, we need to use conventions concerning the parameters of the model and underlying mobility network. As an artificial mobility network we will use a regular one-dimensional lattice and uncorrelated scale-free network [10, 11]. We also considered other random networks (Erdős-Rényi, Watts-Strogatz), which we do not discuss, but the major results are independent of particular topology and are present in all of them.

After numerically generating a particular topology, we assign to every node a nominal population size N_n, chosen randomly from a uniform distribution. In order to compare epidemiological models explicitly incorporating travel, i.e. bidirectional model and reaction-diffusion model, we calibrate both system in a way, that the total fluxes of travelers between locations are kept the same in both systems. Being aware only of the total fluxes between locations it is impossible to distinguish between bidirectional and reaction-diffusion traveling mechanisms. Further convention is a constant return travel rate $\omega^n_{nm} \equiv \omega \quad$ corresponding to equal dwelling times on distant locations independent from the location of origin. However the knowledge of only total fluxes between locations is not enough to extract travel rates of bidirectional model as it becomes clear from Fig. 1. You see there, that the total flux $\mathscr{F}_{nm}$ from location m to location n consists of partial flux $\mathscr{F}^m_{nm}$ of individuals originating from location m and traveling to location n as well as partial flux $\mathscr{F}^m_{nm}$ of individuals originating in location n, having dwelled in location m and returning to location n. We solve this dilemma, by proposing a vein model which states that partial fluxes are proportional to the sizes of different local populations:. $\mathscr{F}^n_{nm} = \mathscr{F}_{nm} N_n/(N_n + N_m)$. Using the last expression, conservation of individuals originating from a particular location, and detailed balance conditions on travel rates we derived a system of linear equations which uniquely maps total fluxes onto travel rates

$$\omega_{nm} + \frac{\mathscr{F}_{nm}}{N_n + N_m} \sum_k A_{km} \frac{\omega_{km}}{\omega} = \quad \frac{\mathscr{F}_{nm}}{N_n + N_m}.$$

Numerical solutions of the last equation reveals a lower boundary for the return rate. We could explain the existence of such a boundary by considering a simple one dimensional homogeneous lattice topology, where we can easily obtain relation between forward travel rate ω^+ and random walk traveling rate ω being the total flux normalized by the number of individuals pro location

$$\omega^+ = \frac{\omega \quad \omega}{2(\omega \qquad \omega)},$$

where from follows $\omega \quad > \omega$ which is a constrain of constant total flux rate ω. To assign total fluxes in an inhomogeneously populated complex network of location we use an ansatz $\mathscr{F}_{nm} \quad N_n N_m$ reminiscent of gravity law [12].

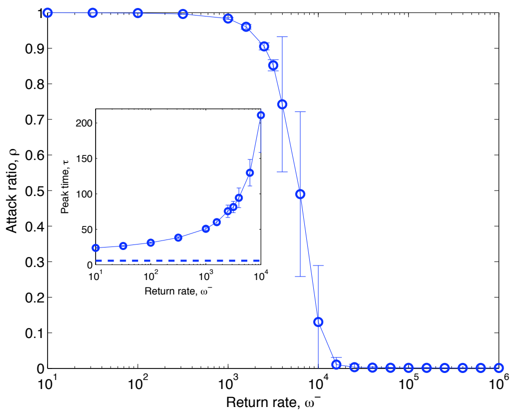


Fig. 2: Attack ratio $\quad$ in dependence on return rate $\omega \quad$ for a fixed value of total averaged flux rate $\quad \omega \quad = 1$ for a SIR epidemic on a scale-free network with power-law exponent $\quad = 1.5$ and with 1000 nodes populated uniformly with $\quad N \quad = 250$ individuals each. Inset shows the dependence of peak time $\quad$ on return rate $\omega \quad$ for bidirectional epidemic (circles) and reaction-diffusion model (dashed line). Results were averaged over 50 stochastic realisations. Infection rate $\alpha = 1$, recovery rate $\beta = 0.1$.

3 Results

Most important questions which could be answered by an epidemiological modeling are under which conditions an epidemic outbreak occurs and how fast it spreads. The conditions under which an epidemic outbreak occurs are usually given by threshold parameters. The most important threshold parameter in epidemiology is a basic reproduction number $\mathscr{R}_0$, giving a number of newly infected individuals in a fully susceptible population caused by one infected [7]. For SIR model $\mathscr{R}_0 = \alpha/\beta$. If $\mathscr{R}_0 > 1$ there is an epidemic outbreak, otherwise — no outbreak. In the context of reaction-diffusion SIR epidemic model, the concept of global invasion threshold was recently introduced [13]. This global invasion threshold is determined by a minimal flux between locations in order for epidemic to affect the majority of locations. We discovered a new invasion threshold in a bidirectional SIR model which is determined by the return rate ω or by the time an individual spends on distant locations. The existence of this threshold is evident from Fig. 2 where the dependence of the attack ratio on the return travel rates ω for a bidirectional SIR epidemic on a uncorrelated scale-free network [11] is presented. Attack ratio gives a fraction of the overall population affected during an epidemic. We fixed total flux at a value sufficient for a global outbreak in a reaction-diffusion system. As one observes for low return rates the attack ratio is close to unity as it is expected for given parameters α and β thus witnessing a global outbreak. However, with growing values of the return rate, the attack ratio drops almost to zero reflecting the absence of the global outbreak. The regime of high return rates corresponds to small dwelling time on distant locations and thus an infected has not enough time to transfer an infection to susceptibles in unaffected locations before returning home. Thus we can define a threshold value of the return rate determining with a high probability the existence of the global outbreak (in Fig. 2 $\omega_{\text{threshold}}$ 10^3 10^4). If we consider the dynamics under the threshold, where the global outbreak occurs, there are still significant differences in the behaviour of bidirectional and reaction-diffusion models. In the inset of Fig. 2 we present the dependence of a peak time $= \int I(t)tdt / \int I(t)dt$ or a time from the start of an epidemic to the the average moment when the number of infecteds reaches its peak. We see that by varying return rate we can obtain ten-fold increase in the peak time in the bidirectional model as compared to the reaction-diffusion model, which of course do not depend on return travel rate and only on the total flux. Elsewhere [6] we showed analytically that attenuation of the bidirectional epidemic on a lattice as compared to reaction-diffusion model is a generic feature of a bidirectional model.

To assess the mutual impact of major travel parameters, i.e. forward and backward (return) travel rates as well as total flux on the dynamics of bidirectional SIR epidemics we calculated the attack ratio for various travel parameter values in a lattice of homogeneous locations. The results are presented in Fig. 3. One observes that the global invasion threshold in terms of the total flux rate ω at low values of travel rates ω^+, ω still exists in the bidirectional system. This confirms again the similarity of both reaction-diffusion and bidirectional framework at low travel rates, what we showed elsewhere. However for increasing total flux (going away from the beginning of coordinates), we observe that for small value of return rates there is a global outbreak, but for increasing return rate we arrive in the region of no global outbreak, even for the same value of a total flux rate ω. This effect could not be observed in the reaction-diffusion system, which is fully determined by a total flux, and is specific for the bidirectional model. Except upper left region of no epidemic outbreak one can also observe an another smaller outbreak-free region in the low right corner. This region corresponds to high values of the forward travel rates and is unrealistic, because then individuals would spent most of the time on distant location compromising the notion of a home location. However the effect is very similar to the case of high return rates, because infected individuals spent so little time at home that they could not infect significant number of people there.

4 Conclusions

Summarizing, we have investigated a bidirectional epidemiological model, explicitly incorporating the highly regular movement patterns of human. The model reflect the tendency of humans to travel among a few most preferred locations bidirectionally. We systematically compared bidirectional model to the standard, also explicitly incorporating travel behavior of the host, reaction-diffusion epidemic model on regular topology and complex networks. We showed that as compared to an equivalent reaction-diffusion system, bidirectional travel with the same total fluxes between locations, leads to significant attenuation of epidemic spread. The global invasion threshold in terms of total flux

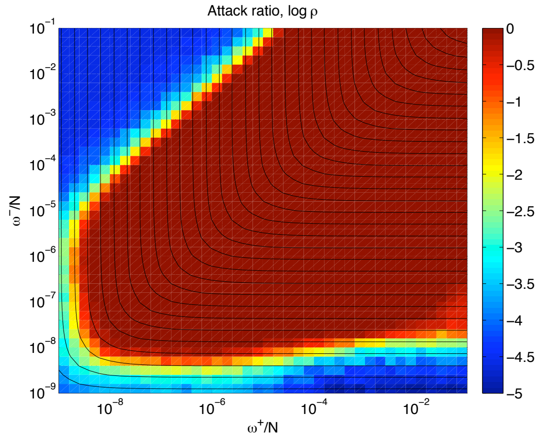


Fig. 3: Attack ratio in dependence on forward travel rate ω^{+} and backward (return) travel rate ω for a bidirectional SIR epidemic in a homogeneous one-dimensional lattice of locations. Black lines corresponds to constant values of the logarithm of the total flux rate ω. Dark red regions corresponds to an infection outbreak, dark blue — to the absence of epidemic. Infection rates $\alpha = 1$, recovery rate $\beta = 0.1$. Lattice consists of 100 nodes with 1000 individuals each. Results were averaged over 50 stochastic realisations.

between locations known in the reaction-diffusion framework (SIR epidemic) still exists in bidirectional model at low travel rates. This indicates the equivalence of reaction-diffusion and bidirectional approaches in this regime. However bidirectional epidemiological model exhibits a new global invasion thresholds in terms of return travel rates. Even in the regime of global outbreak in terms of the total flux of individuals, high enough return rates would lead to the extinction of epidemic. In the light of new findings we convinced that bidirectional model provides a framework for incorporation of the abundance of data on human mobility available today into comprehensive computational epidemiological models.

References

[1] Donald G McNeil Jr. Predicting the Flu with the aid of Washington (Well, George). *The New York Times*, May 4th:1, 2009.

[2] Duygu Balcan, Hao Hu, Bruno Goncalves, Paolo Bajardi, Chiara Poletto, Jose Ramasco, Daniela Paolotti, Nicola Perra, Michele Tizzoni, Wouter Broeck, Vittoria Colizza, and Alessandro Vespignani. Seasonal transmission potential and activity peaks of the new influenza A(H1N1): a Monte Carlo likelihood analysis based on human mobility. *BMC Medicine*, 7(1):45, 2009.

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

[4] Marta C González, César A Hidalgo, and Albert-László Barabási. Understanding individual human mobility patterns. *Nature (London)*, 453:779, 2008.

[5] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of predictability in human mobility. *Science*, 327:1018, 2010.

[6] V V Belik, T Geisel, and D Brockmann. The impact of human mobility on spatial disease dynamics. In *Proceedings of 2009 International Conference on Computational Science and Engineering*, volume 4, pages 932–935. IEEE Computer Society, 2009.

[7] R M Anderson and R M May. *Infectious diseases of humans*. Oxford university press, 1991.

[8] Lisa Sattenspiel and Klaus Dietz. A structured epidemic model incorporating geographic mobility among regions. *Mathematical Biosciences*, 128:71, 1995.

[9] Matt J. Keeling and Pejman Rohani. Estimating spatial coupling in epidemiological systems: a mechanistic approach. *Ecology Letters*, 5:20–29, 2005.

[10] R Albert and A L Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74:47, 2002.

[11] Michele Catanzaro, Marián Boguñá, and Romualdo Pastor-Satorras. Generation of uncorrelated random scale-free networks. *Phys. Rev. E*, 71(2):027103, Feb 2005.

[12] G K Zipf. The P1 P2/D Hypothesis: On the Intercity Movement of Persons. *American Sociological Review*, 11:677, 1946.

[13] Vittoria Colizza and Alessandro Vespignani. Invasion threshold in heterogeneous metapopulation networks. *Phys. Rev. Lett.*, 99:148701, 2007.

A Tale of Two Cities

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

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

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

Characterizing human mobility patterns is critical to a deeper understanding of the effects of human movement. For example, the impact of human travel on the environment cannot be understood without such a characterization. Similarly, understanding and modeling the ways in which disease spreads hinges on a clear picture of the ways that humans themselves spread [2]. Other examples abound in fields like urban planning, where knowing how people come and go can help determine where to deploy infrastructure [1].

Human mobility researchers have traditionally relied on surveys and observations of relatively small numbers of people to get a glimpse into the way that humans move about, for instance by studying airline flight paths [4]. These methods often result in small sample sizes and may introduce inaccuracies due to intentional or unintentional behaviors on the part of those being observed. However, with the advent of cellular wireless communication, ubiquitous networks are now in place that must know the location of the millions of active cell phones in their coverage areas in order to provide the phones with voice and data services. Given the almost constant physical proximity of cell phones to their owners, location data from these networks has the potential to revolutionize the study of human mobility.

In this work, we explore the use of location information from a cellular network to characterize human mobility in two major cities in the United States: Los Angeles (LA) and New York (NY). More specifically, we analyze anonymous records of approximate cell phone locations at discrete times when the phones are in active use. Our data set spans two months of activity for hundreds of thousands of phones, yielding hundreds of millions of location events. We then compile aggregate statistics of how far humans travel daily. We introduce the concept of a *daily range*, that is, the maximal distance that a phone, and by assumption its owner, has been seen to travel in one day. Finally, we make various observations about these daily ranges in the two populations of interest.

Our results show significant differences in mobility patterns between Angelenos and New Yorkers, and bring out unexpected aspects of human behavior. Our main observations are as follows:

Fridays are Weekend Days: When considering daily ranges, Fridays are more similar to Saturdays and Sundays and therefore we treat them as weekend days.

Weekends are Varied: Although more daily range maxima occur on weekends, this does not necessarily correlate to greater distances traveled on weekends. Our results show that weekends tend to be more variable than weekdays. Some people travel less on weekends and some travel more, when compared to their typical weekday behavior.

Angelenos Commute Farther: There is a nontrivial differences between the weekday travel patterns of Angelenos and New Yorkers. Specifically, the median for weekday daily range in LA is 34% larger than in NY, whereas the 25^{th} percentile weekday LA ranges are 53% larger.

This trend of Angelenos traveling farther than New Yorkers continues when examining maximum daily ranges. Our results show that people living in the LA area travel about 20% farther than those from the NY area, regardless of the percentile considered.

Commuting Estimates: The median daily range of people in the greater NY and LA regions are 3.8 and 5.0 miles, respectively. Moreover, as shown in Figure 1, cell phone users in downtown LA have median daily ranges that are nearly double those of their Manhattan counterparts.

Interestingly, even though the data released by the US Census Bureau [6] indicates that people in NY have the longest commutes in the nation by *time*, our data suggests that people in NY have significantly shorter commutes than people in LA by *distance*. If not necessarily contradictory, our data indicates that commuting is done differently in NY and LA. It is possible that generally slower forms of transportation, such as public transport or walking, are responsible for the long commute times reported in NY.

City of Neighborhoods: We found that variations in mobility are striking even between subareas of the same city. Within LA, variations span from 1.3x (at the median) to 3x (at the 98^{th} percentile). The differences within LA itself are thus equal to, or perhaps even a bit greater than, differences between LA and NY.

Differences within NY are even greater — variations span from 1.8x (at the 75^{th} percentile) to 4.2x (at the 98^{th} percentile). The map overlays in Figure 2 also show that LA is more self-similar than NY.

Manhattanites Travel Very Far: Examining maximum daily ranges only for residents of the city centers reveals an interesting reversal of the general pattern of Angelenos traveling farther than New Yorkers. Here the medians are at 69 and 29 miles for Manhattan and downtown LA, respectively. For the 75^{th} percentiles the corresponding numbers are 735 and 129 miles. These numbers show that when Manhattanites travel far, they travel very far and farther than Angelenos.

US vs Unnamed European Country: It is possible to compare some of our statistics to those computed by González et al. for an Unnamed European Country (UEC) [3]. Our maxima show that in the greater LA area, 50% of people traveled more than 36 miles on at least one day, and that in the NY area 50% traveled more than 27 miles. This is in sharp contrast to González et al.'s findings that nearly 50% of all the people in their study remained within a 6-mile range over a 6-month period. The LA maxima are more than 5x larger than those in UEC and the NY maxima are more than 4x larger. While it is not surprising that the numbers in the US are larger, as the US is more car-oriented, the magnitude of the difference is unexpected.

Overall, we conclude that the study of operational records from cellular networks holds great promise for the large-scale characterization of human mobility patterns without compromising individual privacy. For more details on our data set, our analysis methodology, and our results, we refer the reader to our full paper [5].

1. REFERENCES

[1] The journey to work: Relation between employment and residence. Technical Report No. 26, American Society of Planning Officials, May 1951.

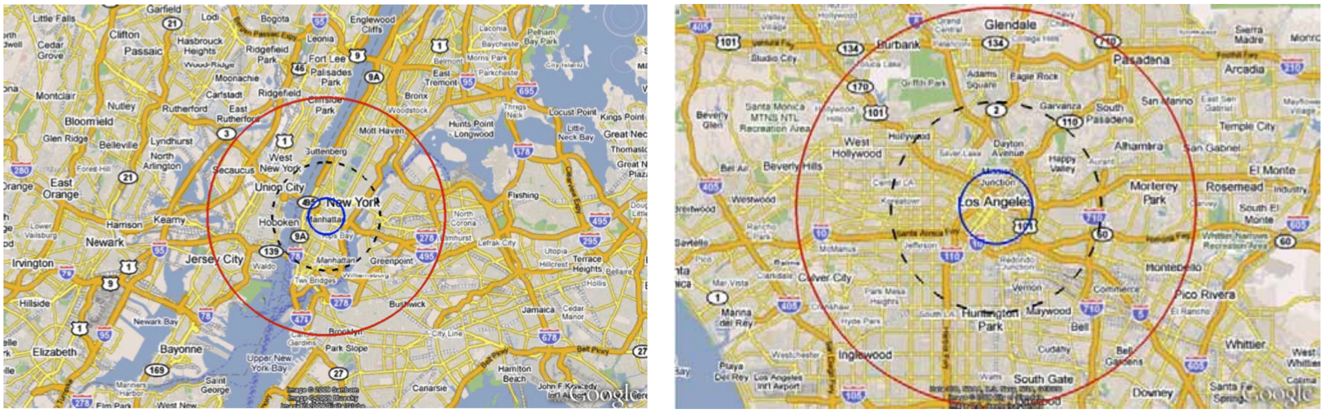
[2] D. Brockmann, V. David, and A. M. Gallardo. Human mobility and spatial disease dynamics. *Proc. of the Workshop on Social Computing with Mobile Phones and Sensors: Modeling, Sensing and Sharing*, Aug. 2009.

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

[4] R. Guimera and L. Amaral. Modeling the world-wide airport network. *Eur Phys J B*, 38, Jan. 2004.

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

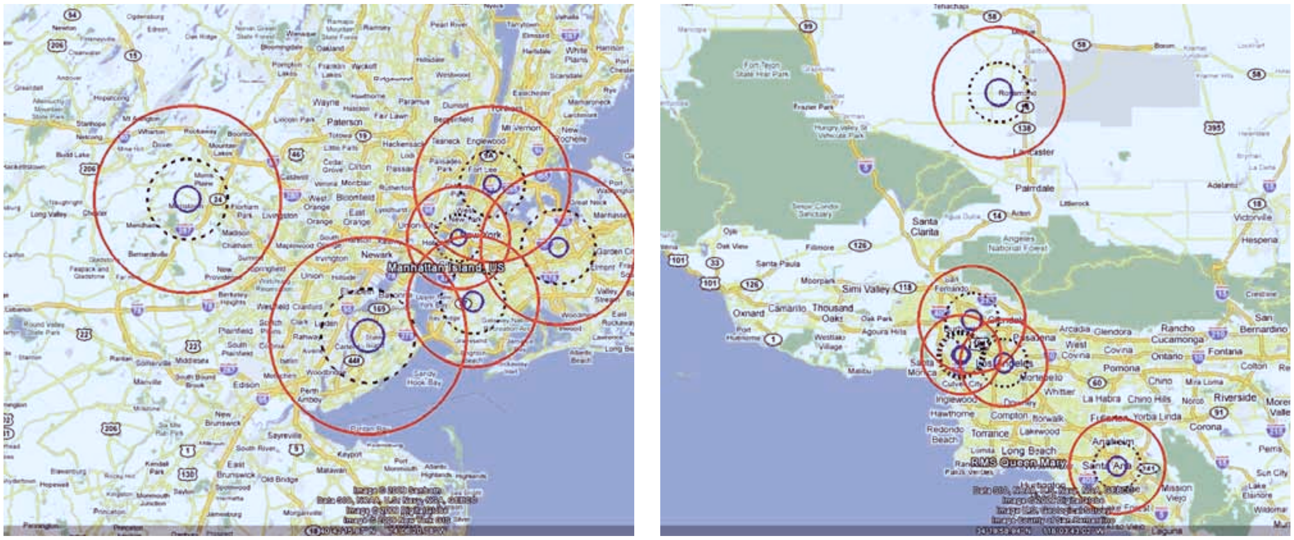
[6] US census data. Downloaded from http://www.census.gov.



(a) Manhattan

(b) Downtown Los Angeles

Figure 1: **Maps giving a visual representation of the median daily ranges of cell phone users in Manhattan and downtown Los Angeles. The radii of the inner, middle, and outer circles represent the 25^{th}, 50^{th}, and 75^{th} percentiles, respectively, of these ranges across all users in a city. Ranges for all users in a city are made to originate in a common point for clarity of display. The two maps are drawn to the same scale.**



(a) New York and New Jersey subregions

(b) Los Angeles subregions

Figure 2: **Maps giving a visual representation of the median daily ranges of cell phone users in subregions of the LA and NY metropolitan areas. The radii of the inner, dashed, and outer circles represent the 25^{th}, 50^{th}, and 75^{th} percentiles, respectively, of these ranges across all users in a subregion. Ranges for all users in a subregion are made to originate in a common point for clarity of display. The two maps have *different* scales.**

Workshop on the analysis of mobile phone networks / Tuesday May 11, 2010 (MIT, Cambridge, MA / USA)

Davy Janssens, Professor at Hasselt University
davy.janssens@uhasselt.be
Sjef Moerdijk, doctorate researcher at Hasselt University
sjef.moerdijk@rws.nl

Title: Predicting travelling activity space based on phoning activity space in social networks

For the last decade, activity-based models have set the standard for modelling travel demand. The idea behind these models is that travel demand is derived from activities that individuals and households need or wish to perform. As a result of this, travel in itself is often also derived in these models by means of simple categories of travel, such as commuting, leisure, or business, as if these activities exist separate and are self-contained.

Only recently, social networks have been added to these traditional classifications, as a new possible predictor of travel behaviour and movement. Researchers such as Urry and others (see Urry, 2006; Sheller, 2006; Larsen, 2006)) have argued that flows and meetings of people produce small worlds which require connections and meeting places; a phenomenon which is also known as the new mobilities paradigm. The paradigm states that mobility of people in our postmodern era is no longer restricted to 'corporeal mobility' in which people (and goods) have face-to-face contacts. These contacts are extended and complemented with contacts mediated by interfaces like mobile phone, e-mail, Blackberry, etc. the so called 'face-to-interface-to-face' contacts. In trying to understand this completion, one should focus on the interaction and 'interconnections' (Larsen, 2006) between these mobilities. In other words: travel and social meetings require systems of coordination using virtual and communicative travel in-between physical travel and meetings. With this understanding of social networks as a facilitator of virtual as well as physical networks and the movements within them, travel becomes a result of human networking (see for some research on how to model this Dugundji, 2008; Axhausen 2008; Arentze 2008).

ICT developments have enabled communication with distant others without physical co-presence, which makes that 'face-to-face' meetings are not a prerequisite to build or maintain social contacts any more. A large body of literature on the transportation impacts of specific ICT applications has been produced. Although recent publications from physics seem to prove the contrary, it is widely believed in transportation research that it is crucial to understand the geography of the social networks to which the travellers belong, if one wants to understand the destination choices of the travellers. Recently, Mok and Wellman (2007) have contributed to this line of research. Axhausen has shown for instance how to examine distance decay functions and market shares of face-to-face meetings, phone, e-mail and SMS (see Axhausen, 2004, 2007). Axhausen and Frei (2007) have shown geographical size and structure of global social networks, where Frei and Axhausen have shown distance decay of different new media that are used in maintaining the social network.

While these papers contribute a lot to the understanding of the spatial distributions of individuals, they do not lay emphasis on the relationship between virtual and physical travel in *social networks* on the one hand and the *corresponding spatial and temporal distributions* on the other hand.

The main research question that we therefore want to address is whether behaviour in terms of activity-space use by social network members can be explained (or predicted) by their usage of new ICT means in general and mobile phone in particular. To the best of our knowledge, this is the first study that aims to examine the relationship between a travel activity space and a phoning activity space (in short defined here as the square area that a person would have to cover if he wanted to meet distant others all over this area face-to-face, and as a fact now is having contact with by mobile phone contact (in other words: 'virtually meeting'). This the face-to-interface-to-face contact) and the travelling activity space for members that belong to the same social network (with the 'ego' as the central person, surrounded by his 'alters'). If such a relationship can be found, phoning activity geographies can be used as an approximation (or prediction) for travel geographies, leading to more and larger quantities of data (and thus potentially lower error-bounds) that can be relied upon for training and testing the derived activity-spaces in model environments, given the large inflow of spatial-temporal information which can be extracted from mobile devices. The methodology could contribute to the existing literature and research that has been carried out in the field of individual accessibility measures, and may finally be used as input for a destination choice component of activity-based travel demand models. Apart from the obvious applications in transportation models, added insight can also lead to a better understanding of the basic laws governing human mobility. Indeed, as many scientists have pointed out, the goal of social sciences is not simply to understand how people behave in large groups (as it was studied in Gonzalez et al. 2008), but to understand what motivates individuals to behave the way they do (Editorial Nature, 2008).

The main research goal that has been outlined above (degree of possible prediction of travelling activity spaces based on phoning activity spaces by social networks members), needs to be further specified in two additional specific research questions.
First of all, one obviously needs to know how the geographies of mobile phone usages by travellers that maintain social networks look like.
Secondly, one is obviously also interested in what kind of usages of mobile phone during travelling show maintenance of social networks and what usages don't show maintenance.

In the workshop we want to focus on the construction of social network and phoning geographies, relying upon existing measures of spatial distributions that are based on daily activity spaces. We want to higlight some (im)possibilities in this context. The results of a first analysis are focused on the relations between mobile phone usage and travelling. For instance how much distance respondents travel compared to their phoning distances.

We found that the following points need to be addressed further:

- The new mobilities paradigm is about corporeal, digital, virtual and even mental mobilities. All these mobilities show coordination patterns, interactions, etc. So far, researchers and scholars have focussed on corporeal travel and mobile phone interactions, partly because the data availability, partly also because of the fact that corporeal travel and mobile phone are both mobile. These spatial and temporal interactions are self-evident. But e-mail use has become mobile as well (for instance Blackberry, mobile internet via mobile phone, etc.), so the spatial and temporal interaction of corporeal travel, mobile phoning and e-mailing looks very interesting.

So we are looking forfeasible data sources to use for constructing the spatial and temporal interactions.

- Social networks are maintained by human interactions. For instance by phoning or e-mailing each other (and consequently after a while meeting each other face-to-face). We still need to find out how to present such multi-interaction manifestations, what is a useful way to present social networks and ‘their’ multi-level (or: corporeal and digital and virtual and even in a conceptual way mental!) mobilities?
- One of the main methodological issues in these humanistic studies concerns the difference between stated preferences (‘I’m use my phone five times a day’ etc.) and revealed facts (an individual using his phone six times a day). In a more general way we still need to discuss which survey methods can be validated with these revealed facts (which don’t) and how these validation should be developed.

References

Larsen, J., J. Urry, & K.W. Axhausen (2006), Mobilities, networks, geographies. Ashgate, Aldershot.

Urry, J. (2006), Sociology beyond societies: mobilities for the 21st century. London, Routledge.

Sheller, M. & J. Urry (2006), The new mobilities paradigm. Environment and Planning A 38 (2), pp. 207-226.

Frei, A. & K.W. Axhausen (2007), Size and structure of social network geographies. Arbeitsbericht Verkehrs- und Raumplanung, 439, IVT , ETH Zürich, Zürich.

Axhausen, K.W. (2004), Social networks and travel: some hypotheses . CD-ROM of the conference on progress in activity-based analysis, Maastricht.

Axhausen, K.W. & A. Frei (2007), Contacts in a shrunken world. Arbeitsbericht Verkehrs- und Raumplanung, 440, IVT , ETH Zürich, Zürich.

Mok, D. , Wellman, B. and Carrasco, J. , 2008-07-31 "Does Distance Matter in the Age of the Internet: Are Cities Losing Their Comparative Advantage?

González, M.C., Hidalgo, C.A. and Barabási, A.L. , Understanding individual human mobility patterns (2008), Center for Complex Network Research and Department of Physics and Computer Science.

Workshop on the Analysis of Mobile Phone Networks
Tuesday, May 11, 2010, MIT (Cambridge, MA)

Preliminary findings on application of mobile phone data analysis to urban studies

Fabio Manfredini, Paola Pucci, Paolo Tagliolato, Paolo Dilda*

Dipartimento di Architettura e Pianificazione (DiAP)
Politecnico di Milano
* corresponding author: *fabio.manfredini@polimi.it*

In recent years, a new approach for estimating people's movement in cities through mobile phone traffic analysis has been developed. Compared to the traditional methods of urban surveys, the use of aggregated and anonymous cellular network log files appears to be a promising tool for large-scale studies with notably smaller efforts and costs.
In particular, the use of mobile traffic phone data shows many advantages, for example:

- large data samples, proportional to the pervasiveness of mobile phone use;
- monitoring of any area, given the extent of mobile phone network coverage;
- generation of data almost in real time;
- high spatial and temporal resolution, compared to traditional and institutional data sources.

However, despite the positivist approach to the new methodology, additional evidence is needed to show how cellular network signals correlate with the actual presence of people in the city, how this source of information can be used to characterize and to map different urban situations and their occupants and how this tool could support urban planning and urban policies.
The presentation will deal with this topic by showing the results of a research carried out by a Department of Architecture and Planning of the Politecnico di Milano research team, during the 2009, for Telecom Italia, the main Italian mobile phone operator.
Purpose of this paper is to address this shortcoming by presenting the results of a survey effectuated with Telecom's cell-phone-network data in Lombardia Region (Italy) during the year 2009, as a promising approach for characterizing and mapping urban domains and their occupants, assisting traditional databases and analyses of urban dynamics.
The aim was to assess the contribution of cellular phone traffic data for understanding and analyzing urban dynamics and to propose possible uses of this kind of information in urban planning.
In order to analyze the complex temporal and spatial patterns of mobile phone activity, we were given access to data covering the whole Lombardia Region (Northern Italy), provided by Telecom in form of Erlang, a measure which describes the mobile phone activity as a function of position and time, recorded at a spatial resolution of about 250 meters, every 15 minutes in the period January -October 2009.

The first step of our research was the analysis of mobile phone activity trends for some already known urban sectors, subjects of previous research, selected on the basis of present land-use patterns of population and activities, densities and socio-economical profiles. For each of these urban situations we studied the Erlang trends during a typical week, to verify temporal and spatial patterns of cellular phone usage for similar urban sectors in terms of population and activities characteristics.

The results show that mobile phone activity patterns can provide useful information for interpreting the specific dynamics of different urban situations such as monofunctional residential zones, railway stations, urban sprawl areas, factories.

In the second step of the research, we focused on the reliability of Erlang data with respect to traditional data sources.

We therefore compared mobile phone data of the main cities of Lombardia Region and their population dynamics[1] during a typical weekday and we found a strong correlation between these variables proving the potentialities of Erlang data for describing the variability of daily changes of urban population at the municipality scale.

We also worked on mobile phone traffic data at the Milano urban region scale[2], during the 2009 International Design Week, a leading event which concentrates its activities in the Fair area and in hundreds of places within and outside the city.

We therefore performed several analysis to evaluate the potential contribution of Erlang data to describe, to represent and to manage an event, from the beginning until its conclusion. We mapped and interpreted the spatial configuration of mobile phone activity in order to highlight which parts of the city showed a significant concentration of traffic comparing the days when the event occurred with other days without events.

We defined a set of significant spatial operations between Erlang matrixes aimed at underline the territorial effects at a wider scale and at different temporal patterns of the event such as, for example, ratio between nighttime and daytime mobile phone activity, ratio between weekdays and holidays.

This type of information, if suitably interpreted, may be useful to assess the consequences of a specific event on the entire urban system in its spatial and temporal patterns and to evaluate its impacts on the whole urban system (mobility, congestion, tourism).

Urban planning traditional data sources are mainly based on static statistical surveys and are not able to catch the variability in the intensity of urban space’s use by present population.

Further analysis focused on the correlation between the intensity of telephone calls at certain times of the day with the spatial configuration of residents and workers in the Milan area. Preliminary results showed that telephone traffic data can effectively help to represent and to describe, dynamically over time, the intensity of activities and of presences at the urban scale.

Because of its spatial and temporal resolution, mobile phone data constitute an interesting and unique source of information on urban uses. Indeed, if we consider the observed and aggregated telephone traffic as the result of individual behaviors and habits, we conclude that mobile phone data can provide information, which changes over times, on urban contexts, and can bring to new interpretation of urban dynamics.

This study therefore suggests that cell-phone-network data has the potential to drastically change the way we view and understand the urban environment.

[1] The latter was obtained from a traditional mobility data-base, the 2002 Lombardia Region Origin-Destination survey, which is widely used in urban and traffic studies.

[2] The Milano Urban Region is defined as an extended region that go far beyond the traditional administrative structure. Milan is the main center of the urban region which is characterized by a stratum of dense urbanization stretched over the ancient framework while the bordering provinces have been incorporated in the strongly urbanized and enlarged urban region.

Secondly, it explores whether mobile network data can reveal the significant time-dependent variation which is missing from traditional analysis and can thus describe cities dynamically over time.

Another conclusion is that urban planning competences, together with specific knowledge on urban systems and on the distribution of activities and population within the territory are needed to correctly interpret mobile phone data and to characterize and to map urban contexts and their occupants.

In the final phase of the research activity, we presented our work to different stakeholders, belonging to private and public sectors in order to collect ideas and proposals on possible applications of this approach to different topics such as event management, civil protection, mobility monitoring, urban rhythms analysis and mapping; we have seen a great interest from most of them in terms of the potential contribution of this data to provide new decision tools for the development of action and policies in different sectors.

Main references

Ahas, R., & Mark, U. (2005). *Location based services - new challanges for planning and public administration?* Futures , 37, 547-561.

Caceres, N. Wideberg, J. P. Benitez, F. G. (2007). *Deriving origin-destination data from a mobile phone network*. Iet Intelligent Transport Systems, 1 (1), 15-26

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

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

Analyzing cell-phone mobility and social events

Francesco Calabrese*, Giusy Di Lorenzo*, Francisco Pereira*, Liang Liu*, Carlo Ratti*

Abstract

Location-based services and traffic planning may benefit from predicting leisure preferences for people living in a particular location. To make such predictions, surveys are often administered, but surveys are notoriously expensive and easily become out-of-date. Recently, researchers have proposed to automate the process of extracting preferences by analyzing user-generated content. Here we propose to analyze digital information that is *implicitly* generated by users while they carry their mobile phones. We show that, by analyzing mobility traces we are able to associate locations with travel demands for different types of social events.

1 Introduction

According to the US Federal HighwayAdministration a "special event" is an occurrence that "abnormally increases traffic demand," unlike an accident, or construction and maintenance activities, which typically restrict the roadway capacity. Planned Special Events (PSEs) include sporting events, concerts, festivals, and conventions at permanent multi-use venues. They also include less frequent public events, such as parades, fireworks displays,bicycle races,seasonal festivals,etc.

The often mentioned outburst of mobile phones during late 20th century accompanied by the more recent trend of sensors and advanced communication systems (e.g. GPS, digital cameras, Bluetooth, WiFi) allow for unforeseen amounts of data from urban areas through which to study both groups [1], individuals [2] or both [3]. On the other side, the emergence of web 2.0 techniques, especially geospatial enabled web services, such as upcoming.org, etc, enables the citizens organize and inform the audience in a real time way. The convergence of these two trends in ubiquitous computing open a new opportunity and challenge to understand the travel demand triggered by PSE.

In our work, we analyse fine grain anonymized individual mobility information for travel demand forecasting in the context of PSE [5]. Both the process followed and the data precision are thus far novel and unique to our knowledge.

2 Extracting Preferences for Social Events

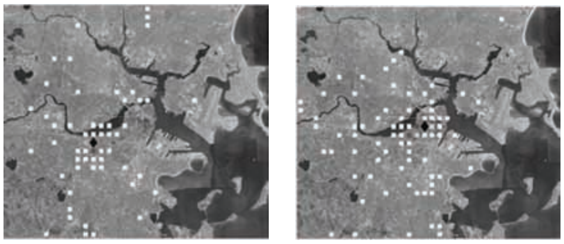
To determine the social events people living in certain areas go to, we use two datasets:

- *Location of Mobile Phones.* We process 130 millions of anonymous location estimations - latitude and longitude - from roughly 1 million mobile phones in greater Boston (an area of $15km^2$). The dataset consists of anonymous cellular phone signaling data collected by AirSage [4]. The finest-grained estimation of location we are able to achieve is represented by a cell of 500 x $500m^2$.
- *Boston Globe website.* We crawl the "Boston Globe Calendar" website to extract social events. This website is a reputable and comprehensive list of social events in Greater Boston (more than 500 daily). The events are organized in 14 categories: Arts & Crafts, Business & Tech, Community, Dance, Education/Campus, Fairs & Festivals, Food & Dining, Music, Other, Performing Arts, Shopping, Sports & Outdoors, Visual Arts, and Cinema.

By using those two datasets, we generate pairs of the type:

- `home|destination` - For each user, we determine her home location and her set of destinations, i.e., we extract trajectories and generate tuples of the form `home|destination`. Our dataset contains location estimations for mobile phones. To extract mobility traces from location estimations, for each user, we infer her stops (places where the user has stopped for more than an hour), and we then collect the trajectories that originate from those stops. A trajectory is set of subsequent physical locations visited by the user such that the time interval between two subsequent points is less than twelve hours. We pick the user's home location to be the most frequent stop at night (i.e., between 10pm and 7am).
- `destination|event` – We associate destinations with types of social events by crawling the event section of the "Boston Globe" website. We select the destinations (other than home locations) at which users stay while big social events are taking

*MIT Senseable City Lab, fcalabre@mit.edu



(a) Boston Red Sox vs. Baltimore Orioles at Fenway Park, 2009-9-9

(b) Shakespeare on the Boston Common, 2009-8-13

Figure 1: Examples of events in Boston. Figures show the locations of the events (diamond) and estimated origins distribution of people attending the events: shade from light (low) to dark (high).

place. We consider only users who stay for at least 70% of an event duration, and only events that are "unique" - no other big event is taking place within the radius of 1 km at the same time. In so doing, we are able to create tuples that associate destinations with social events. This results in 58 types of social events across 7 locations [5].

- `home|event` - We combine pairs in the previous two points (`home|destination` pairs and `destination − event` pairs) to obtain, for each home location, the list of types of social events its residents go to.

3 Methodology

Our methodology for describing events through mobility choices is based on the use of the estimated origins of people attending to the events. Figure 1 shows some examples of spatial variation of the estimated origins of people attending different events.

Sport events such as baseball games attract about double the number of people which normally live in the Fenway Park area. Moreover, those people seem to be predominantly attended by people living in the surrounding of the baseball stadium, as well as the south Boston area (Figure 1(a)).

Performing arts events such as the "Shakespeare on the Boston Common" (Figure 1(b)) which his held yearly, attract people from the whole Boston metropolitan area, and very strongly people which live in the immediate surroundings of the Boston Common (average distance lower than 500 meters). The number of people attending the event is instead about 1.5 times greaten than what it is usually found in the Boston Common.

By comparing the two images in Figures 1(a) and 1(b) it is easy to understand that most of the people attending to one type of event are most probably not attending the other type of events, showing a complementary role of sports and arts events in attracting different categories of people.

Since the origins of people attending an event are strictly related to the location and type of events, we argue that by using just this information we would be able to predict the type of event. If a relationship between origin of people and type of event is found, it would be possible to determine the abnormal and additive travel demand due to a planned event by just considering the type of that event. It would then be possible to provide a city with critical information on which to take decisions about changes in the transportation management, e.g. increasing the number of bus lines connecting certain areas of the city to the venue of the event.

In the next section we will show 8 different models that we have developed to perform the prediction of the type of event starting from the mobility data associated with it.

3.1 Prediction The task at hand is to understand the relationships between events and origins of people. Particularly, we seek for the predictive potential of events in respect to mobility phenomena. This can be seen from two perspectives: a classification task in which we want to understand how a vector of features (e.g., attendees origin distribution) predicts a classification (e.g., an event name or type); a clustering task, in which the feature vectors are distributed according to similarity among themselves.

We used the Weka open source platform [6], which contains a wide range of choices for data analysis. For classification, we use a Multilayer Perceptron, with one hidden layer and the typical heuristic of $(classes + attributes)/2$ for the number of nodes. For clustering, we apply the K-Means algorithm (with $K = \#$ event types or $K = \#$ event places). In each experiment, we used 10-fold cross-validation, in which a tenth of the dataset is left aside for testing the algorithm while using the remaining for training. This train-test process is ran 10 times (one for each tenth of the dataset).

4 Experiments

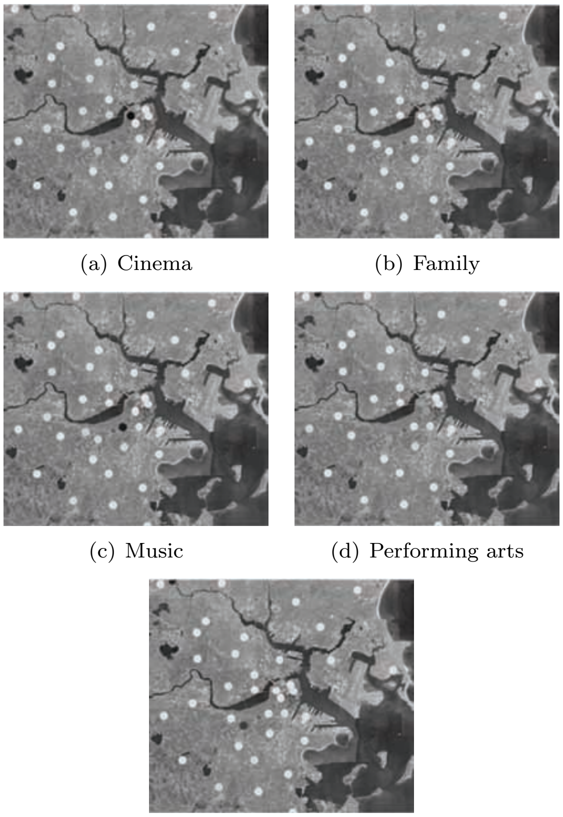
We aggregated attendees in terms of zipcode area and distance to event, discretized in 2000 bins. We did so because if we were to use a geographic coordinate of individuals, the resulting data would be sparse. Instead, by aggregating data geographically, we could find useful patterns. To avoid the strong bias towards attendees in the neighborhood of the event, we also remove those that live in the same area of the event (their home location falls in the same 500m x 500m cell of the event) because we would not be able to distinguish between event and home.

For each event, we created an *instance* that contains the corresponding attendee *origin pattern distribution*, evaluated at the level of the zipcode area (with average size of $4.5km^2$). For example, for one showing of the Shakespeare's "Comedy of Errors" at the Boston Common, we have 96 attendees (users monitored by the system, with a share of about 20% of the population) and then count the total number of people coming from each zipcode. Our goal is to test whether similar events show similar geographical patterns. More specifically, given *origin pattern distribution*, the goal is to predict the type of event.

We met this goal by testing 3 prediction models, and we measure their accuracy in terms of fraction of correctly identified event types.

Before training our algorithms, we analyzed the overall distribution of events to get the *classifier* baselines. The principle is to know the accuracy of a classifier that simply selects randomly any of the 5 event types or that always chooses the same event type, and use them as a baseline to compare for the improvement of the quality. The average value of this baseline is 23.34% (standard deviation of 4.03) for random classification. Differently, if the classifier chooses the event with highest probability (performing arts), the accuracy will be 35%. The first experiment was to use all vectors as just described, applied to a Multilayer Perceptron. The result is a surprising 89.36% of correctly classified events in the test set. From the clustering analysis, we see that mostly attendees come from the event's zipcode area, suggesting that people who live close to an event are preferentially attracted by it. To focus on effects other than close proximity, we created a new prediction model considering only people coming from zipcode different from the event's.

The result is 59.57%, which still indicates the recurrence of origin patterns for events of the same type. A clustering analysis brings the distributions that we can see in Figure 2.



(a) Cinema

(b) Family

(c) Music

(d) Performing arts

(e) Sports

Figure 2: Spatial visualization of clusters centroids. The circles correspond to the zipcode areas with value greater than zero. The shade from light (low) to dark (high) is proportional to the value.

5 Conclusions

Based on our analysis of nearly 1 million cell-phone traces we correlated social events people go to with their home locations. Our results show that there is a strong correlation in that: people who live close to an event are preferentially attracted by it; events of the same type show similar spatial distribution of origins. As a consequence, we could partly predict where people will come from for future events.

References

[1] J. Reades, F. Calabrese, A. Sevtsuk, and C. Ratti, "Cellular census: Explorations in urban data collection," *IEEE Pervasive Computing*, vol. 6, no. 3, pp. 30–38, July-September 2007.

[2] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, 2008.

[3] N. Eagle and A. Pentland, "Reality mining: sensing complex social systems," *Personal Ubiquitous Computing*, vol. 10, no. 4, pp. 255–268, 2006.

[4] Airsage, "Airsage wise technology," http://www.airsage.com/.

[5] F. Calabrese, F. Pereira, G. DiLorenzo, L. Liu, and C. Ratti, "The geography of taste: analyzing cellphone mobiloty and social events," in *International Conference on Pervasive Computing*, 2010.

[6] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques, 2nd Edition.* Morgan Kaufmann, 2005.

Paris by Night

Weekends urban activity via mobile phone data

Christophe Cariou[a], Cezary Ziemlicki[b] & Zbigniew Smoreda[b]

[a.] EVERYDATALAB, Rennes, France (christophe.cariou@everydatalab.com)
[b.] SENSe/Orange Labs, Paris, France (cezary.ziemlicki, zbigniew.smoreda@orange-ftgroup.com)

Abstract - We use mobile phone data to study the impact of two weekend festivals – Fête de la Musique and Nuit Blanche – on the city of Paris compared with a standard weekend. Using the mobile data we can study collective dynamics in the city and separate normal fluctuations from anomalous fluctuations, i.e. spatio-temporal changes caused by the two events. We also detect cultural scenes, allowing us to more precisely identify urban usages and the relative audiences for the different events in each festival.

Keywords - Mobile phone data, complex networks, urban dynamics.

Introduction

The Economist recently devoted an issue to «The Data Deluge», the vast potential of the exponential availability of massive volumes of digital data [1]. These data are increasingly used to reveal urban dynamics in ever more refined spatial and temporal terms [2]. In recent years, works were developed to reveal the pulse of the city - Milan [3], Graz [4], Rome [5], New-York [6] - and to understand individual and collective movements [7-9]. The mobile phone data were also mixed with transport data [5], business data [10], photo activity [11]... The introduction is detailed in the paper but outlined in this abstract. In particular, we present the sociological characteristics - differences, similarities and regularities - of the two events for Paris, explaining the choice of methods deployed in the paper.

From communications to paths

The mobile phone data

We have access to data provided by the Orange operator for the weekends of Fête de la Musique, an annual national music festival (June 21-22 2008) and Nuit Blanche, an art festival (October 04-05 2008), and a weekend when there was no event liable to have a major impact on the city (October 18-19 2008). For each weekend the data are anonymous and include the start and end time of the incoming and outgoing calls, the handovers (cell changes) during a call and the incoming and outgoing text messages. Finally, the data are collected by antennas in real time.

The data can be considered to be representative of the communication activity of Ile-de-France residents. Orange has a market share of around 40% and mobile penetration in Ile-de-France is around 120%. We therefore have access to the communications made entirely or partially within the city of Paris by, respectively, 1,148,552, 1,143,687 and 1,163,273 anonymous individuals during Fête de la Musique, Nuit Blanche and the standard weekend.

The mobile phone network

All the measurements were collected by Parisian antennas. The Paris mobile network is made up of 3,692 antennas on 1,365 cell sites; we only have these cell sites' geolocation. We take each cell site to cover a Voronoi region: every point in the diagram is at an equivalent distance from the cell sites [12]. The area of the city is 105km^2, while the area covered by the network is a little larger – 127km^2 – because of boundary effects. This technique breaks the city down into small areas equivalent to 300m x 300m squares.

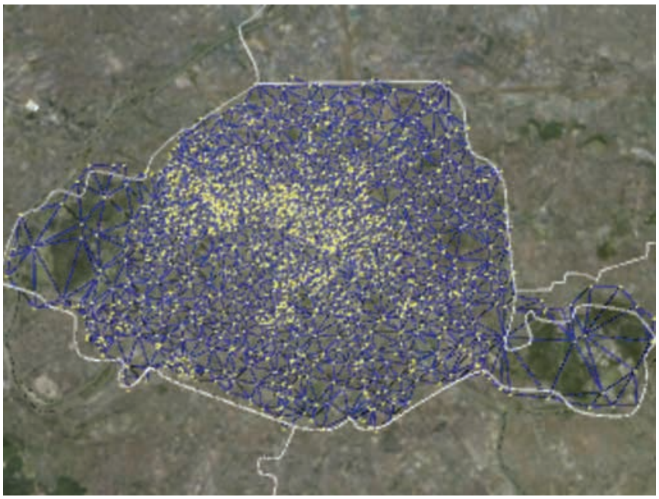


Figure 1. The mobile phone network - nodes (yellow) and edges (blue).

We constructed the Paris mobile network in the following way. We see the cell sites as nodes in the mobile network; it is important to remember that each site is made up of several antennas. The edges in the network are thus all the existing connections between the adjacent sites. The length of each edge is measured very simply by the geographic distance between the furthest sites. This gives us a spatial network of 1,365 nodes and 3,979 edges. [Figure 1]

In conclusion, we studied the city of Paris purely in the form of its mobile network.

Individual dynamics

We consider all the measurements collected to be the locations of an anonymous individual at a given moment of time, whatever the communication type. We focus on periods of fifteen minutes rather than on real time. As a consequence, we duplicate the communications which are spread over several quarter-hour periods. We then transform the communications into individual trajectories: each end of a call and start of the next call thus become an individual's trajectory, giving us a set of trajectories performed by individuals geo-located by the antenna they are using.

As we do not have the geographic location of the antenna, we make the following adjustment. The variable $P_{iik}(T)$ measures: on the one hand, the static presence of an individual using an antenna located on a cell site i, and on the other hand, the path made by an individual between antennas in a single cell site i. The variable $P_{iik}(T)$ thus measures the path made by an individual k between two antennas located in two different cell sites i and j. These two different cell sites must be adjacent. When this is not the case, we use the A* algorithm to predict trajectories [13]: the shortest path between an initial node and a final node (adjacent cell sites), taking into account the geographic distance between the two nodes. Another predictive algorithm could have been used; we chose this one for reasons of simplicity. Finally, we eliminated all the duplications of one individual during a fifteen-minute period.

In conclusion, we have, respectively, 19,355,414, 20,671,622 and 20,944,771individual trajectories for Fête de la Musique, Nuit Blanche and the standard weekend.

Collective dynamics

In this paper, we are not interested in individual dynamics but rather in collective dynamics within the city of Paris. We aggregate these trajectories in terms of nodes and edges in the mobile network:

$$P_{ij}(T) = \sum_{k} P_{ijk}(T)$$

Figure 2 shows the temporal dynamic of these trajectories for the three weekends. If the three weekends do not really differ in terms of the total number of traces, then the temporal profiles will give us relatively little information. The profiles are similar: the difference between Nuit Blanche and the standard weekend is slight, and there is nothing to suggest that it is caused by the event itself. The difference between Fête de la Musique and the standard weekend is substantially greater: the extra activity during the Saturday morning continues during the afternoon; there is a further peak between 9 and 11pm. The apparently greatest impact of Fête de la Musique is observed on Sunday; Paris' collective hangover is a repercussion of this. This considerable difference on Sunday thus explains the lower number of trajectories for Fête de la Musique. Finally, it is worth noting that the city does not switch off during the night from Saturday into Sunday, unlike the previous night.

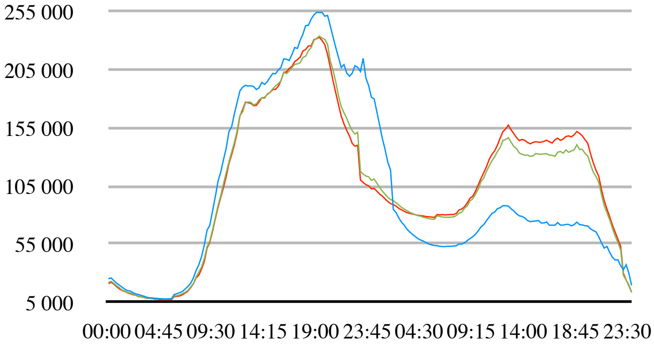


Figure 2. The total number of paths : Fête de la Musique (blue), Nuit Blanche (green) and Standard Week-End (red).

From routines to events

The impact of the two events on the city compared with a standard weekend cannot be analysed using the previous approach. We therefore think it is worthwhile to extend the analysis by distinguishing normal fluctuations from anomalous fluctuations.

Normal versus anomalous fluctuations

Mobile data have already been used to draw a distinction between fluctuations at city level [14]; we adapted it to our particular context. Indeed, we have just one standard weekend as a reference. The number of anomalous paths *AP* measures the discrepancy between urban behaviour during the events *EP* and during the course of the standard weekend *NP*:

$$AP_{ij}(T) = EP_{ij}(T) - NP_{ij}(T)$$

Anomalous behaviour will thus be defined by comparison with the standard deviation:

$$\sigma(T) = \sqrt{\frac{1}{N}.\sum_{ij} A_{ij}(T)^2}$$

We take into consideration the fact that we only have one standard weekend as a reference: the error is thus 25%. Anomalous fluctuations are defined as follows:

$$\left|AP_{ij}(T)\right| \geq 0.25.\sigma(T)$$

We also take into account the fact that the measurement taken at a cell site is much higher than between sites; we therefore perform the previous calculation separately for the two situations and then reaggregate them.

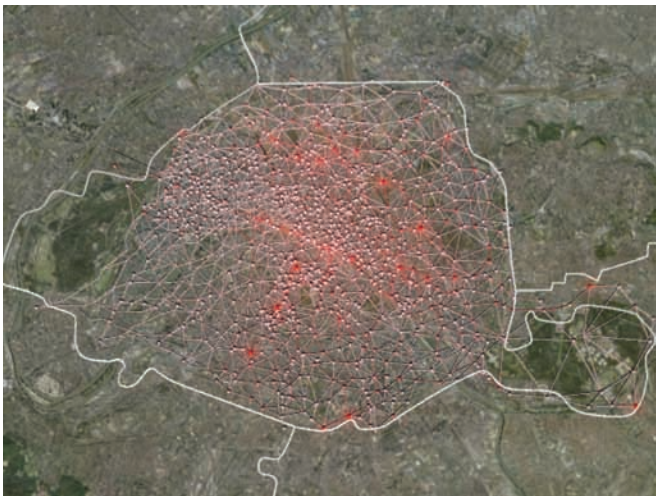


Figure 3. Standard Week-End: the normal fluctuations. Saturday, 10:45pm.

Musical and artistic dynamics

Figure 2 (red curve) and Figure 3 show the normal fluctuations seen for the city of Paris, in other words during the standard weekend. The urban dynamic is fairly conventional for a weekend: at night the city never switches off, unlike on weekdays; Sunday differs greatly from the other days of the week, particularly in the afternoon. Figure 4 shows the anomalous fluctuations caused by the two events in the city of Paris.

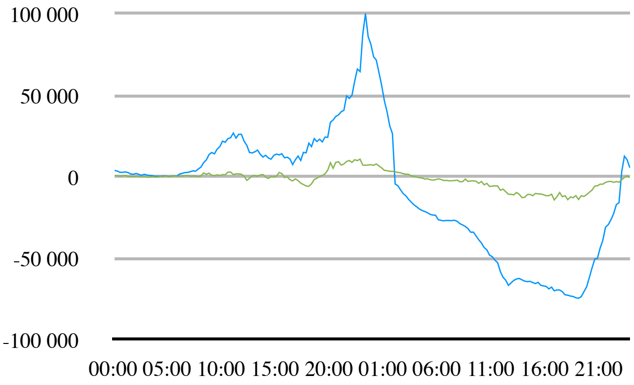


Figure 4. The net anomalous paths. Fête de la Musique (blue) and Nuit Blanche (green).

Fête de la Musique quickly and substantially upsets the normal order in the city because from the morning onwards and throughout Saturday day anomalous fluctuations account for one third of normal fluctuations. The size of the event requires organisation in terms of concert venues. The first concerts, aimed mainly at young people and families, begin in the early afternoon.

While the standard Saturday peaks between 7.15 and 7.29pm, the Saturday of Fête de la Musique peaks much later, between 11.15 and 11.29pm: there are thus as many anomalous fluctuations as normal fluctuations. During this period, Fête de la Musique doubles the number of completed trajectories. In this way the highly popular event extends the Parisian evening. This later-than-normal time marks the end of part of Fête de la Musique: the very large mass concerts for young people and families in particular.

Then, during the night and until the following morning, anomalous fluctuations still account for half of normal fluctuations. The impact is therefore still very great. It is less than previously noted, because of a high number of people leaving and a lower population density at the different venues as the night progresses. The after-effect of this event is seen on Sunday, with very strongly negative abnormal fluctuations. The city of Paris never really reawakens during Sunday, an upheaval which extends until Sunday evening.

Nuit Blanche also has a major impact on the city, but to a lesser degree: it never reaches the same magnitude as Fête de la Musique. The impact of the latter is much greater than that of the former. Figure 5 shows the correlation between the fluctuations during the two events in fifteen-minute periods. There is a very good quality purely linear relationship, with an R2 of 0.83. It is also high: the impact of Fête de la Musique is on average 6.3 times greater than that of Nuit Blanche. The two events do not radically alter the city in the same way.

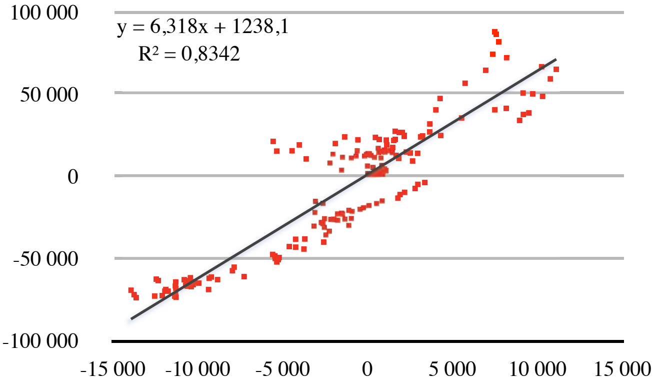


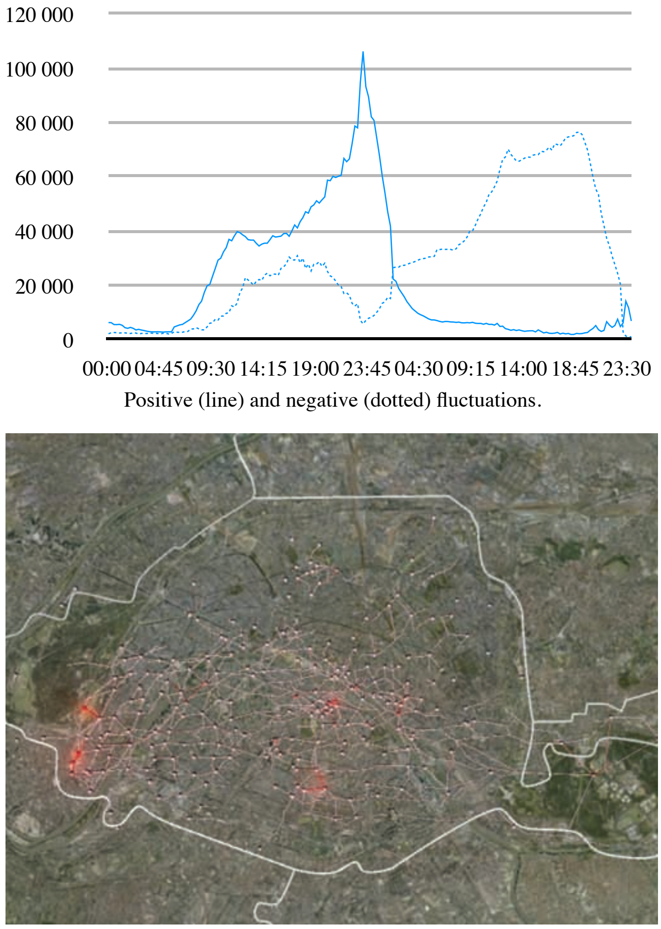
Figure 5. Correlation between fluctuations.

The conclusion which emerges, quite simply, is that the musical event is a more popular event while the artistic event is more elitist.

Urban attraction and repulsion

The above analysis does not fully reflect the impact of the two events on the normal city, because it only takes into account net fluctuations. We thus distinguish positive fluctuations from negative fluctuations. The former include the urban nodes and edges which attracted a large number of spectators during the event. Conversely, the latter show that some nodes and edges lost users by comparison with a more standard weekend. By comparing these two forms of abnormal fluctuations we identify the urban nodes and edges which attracted and repelled spectators.

Figure 7 shows these fluctuations for Fête de la Musique. The previous trends do not change, but the figure shows us that during the day, some concerts attracted a number of spectators to the detriment of other neighbourhoods. On the other hand, at the peak of Fête de la Musique, in other words during the evening and the night, the concerts attracted increasing numbers of spectators without emptying neighbourhoods. Finally, Sunday suggests that no neighbourhood benefited from Fête de la Musique: on the contrary they experienced much lower activity than normal.

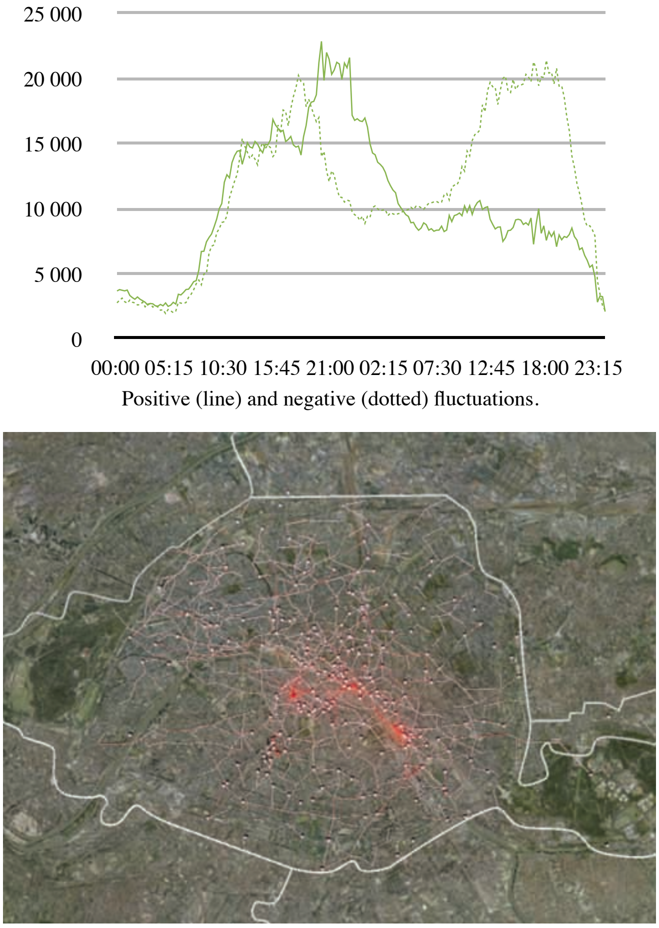


Positive (line) and negative (dotted) fluctuations.

Positive fluctuations : Saturday, 10:45pm.

Figure 7. Fête de la Musique.

Figure 8 shows these fluctuations for Nuit Blanche. The previous trends are confirmed again, but there are differences with Fête de la Musique. During Saturday daytime, null net fluctuations are caused by the attraction of some venues being offset by the repellency of others. At the peak of Nuit Blanche (Saturday evening and Sunday night), we see the same relationship as for Fête de la Musique: the arts venues attract many spectators but this is not to the detriment of other neighbourhoods. Finally, on Sunday a similar trend emerges, even if some events venues attract more individuals. However, these venues are railway stations and other stations, so it is more difficult to draw conclusions.



Positive (line) and negative (dotted) fluctuations.

Positive fluctuations : Saturday, 10:45pm.

Figure 8. Nuit Blanche.

In conclusion, the data shows that the two events substantially alter urban habits.

Detecting cultural scenes

In the section describing the specificities of the two events (not in this abstract), we suggested the similarity between a cultural scene within the mobile network and a community within a social network. We present the detection algorithm chosen for this study.

The edge betweenness

In network analysis there are various methods of detecting communities [15]: proximity within communities vs. distance between communities, where proximity and distance criteria depend on the context being studied. Two major characteristics suited to our study emerge: on the one hand, the need to focus on edges in order to take account of the mobility between the various small events in each festival; and on the other hand, the aim of starting with the city as a whole to determine scenes, rather than starting with nodes and edges and aggregating them to determine scenes. For this reason, the algorithm developed by Newman and Girvan [16-18], often used in network analysis, seemed comparatively relevant.

This algorithm is founded on the betweenness of edges rather than nodes. The betweenness indicator measures the number of paths between all the pairs of nodes in the network running along an edge in the network. With a spatial network, we can therefore take into consideration traffic flows in a transport network [19], and paths in our radio-mobile network. A high value thus represents an edge responsible for a connection between nodes, and by extension between communities. It is similar to the idea of weak ties which link two communities vs. strong ties which define the communities [20]. Granovetter's urban adaptation is used in Jane Jacobs' analyses [21].

The algorithm

The algorithm is founded on the usual measurement of betweenness in a network. Rather than taking account of all the paths running along an edge, it is based on the shortest path. The first step thus involves finding the shortest path between all the pairs of nodes in the network. In order to take account of the spatial component of the mobile network, the shortest path is determined using the A* algorithm: in this way we consider the distance between the nodes. This gives us the number of geographically shortest paths running along each edge in the network, *SP**.

The second step consists in taking account of the fact that the network is valued. The idea developed by Newman and Girvan is to convert the value of an edge into a multi-edge network: each edge with a value equal to *AP* becomes *AP* single valued edges. We adapt the analysis to take into account the fact that we also have valued nodes. We simply proportionally distribute the value of each node to its edges. This gives us the number of single paths for each edge in the network: *AP**.

By combining the two steps we can identify, at a given moment, the edge with the highest score, in other words the edge which connects the most different urban scenes. The measure is as follows:

$$\frac{SP*_{ij}(T)}{AP*_{ij}(T)}, i \neq j$$

Illustration: Nuit Blanche

We will illustrate the relevance of our approach using the example of Nuit Blanche. This event takes place in 70 well-defined venues, dispersed across the city, and linked together by the metro. Superposing the scenes and the different festival venues gives a very clear picture. The resulting partitioning is comparatively more accurate than that proposed by the official programme (four major scenes). It takes into account spectators' actual usages. It also allows us to determine relative audiences. A more detailed analysis is performed in the article.

How does music drive Paris?

The same analysis is performed for Fête de la Musique. The 557 official venues for this festival do not take account of times and musical genres. It will be expanded upon in the full article, but because of a lack of space we cannot go into it any further in this abstract.

References

[1] The Economist, February 27th - March 5th 2010.

[2] N.Nova, Les médias géolocalisés: Comprendre les nouveaux médias numériques, Paris: FYP Editions, 2009.

[3] C.Ratti, R.M.Pulselli, S.Williams, D.Frenchman, "Mobile landscapes", Environment and Planning B, p.727-748, 2006.

[4] C.Ratti, A.Sevtsuk, S.Huang, R.Pailer, Mobile landscapes: Graz in real time, Proceedings of the 3rd Symposium on LBS and TeleCartography, Vienna, 2005.

[5] F.Calabrese, C.Ratti, Real time Rome, Networks and Communication Studies 20, 2007.

[6] F.M.Rojas, C.Celdesi Valeri, K.Kloecki, C.Ratti, New York talk exchange, SA+P Press, 2009.

[7] M.C.Gonzalez, C.A.Hidalgo, A.L.Barabasi, "Undestanding individual human mobility patterns", Nature 453, 2008.

[8] C.Song, Z.Qu, N.Blumm, A.L.Barabasi, Limits of predictability in human mobility, Science, vol.327, 2010.

[9] R.Lambiotte, V.D.Blondel, C.de Kerchove, E.Huens, C.Prieur, Z.Smoreda, P.Van Dooren, "Geographical dispersal of mobile communication networks", Physica A, 2008.

[10] F.Calabrese, J.Reades, C.Ratti, " Eigenplaces:analyzing cities using space-time structure of the mobile phone network", Environment and Planning B, 2007.

[11] F.Girardin, A.Vaccari, A.Gerber, A.Biderman, C.Ratti, "Quantifying urban attractiveness from the distribution and density of digital footprints", International Journal of Spatial Data Infrastructures Research.

[12] G.Voronoi, Nouvelles applications des paramètres continues à la théorie des formes quadratiques, Journal für die Reine und Angewandte Mathematik, vol.133, p.97-178, 1907.

[13] P.E.Hart, N.J.Nilsson, B.Raphael, "A formal basis for the heuristic determination of minimum cost paths", IEEE Transactions on Systems Science and Cybernetics, vol. 4 , p. 100–107, 1968.

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

[15] S.Wasserman, K.Faust, Social networks analysis: methods and applications, New-York: Cambridge University Press, 1994.

[16] M.E.J. Newman, M.Girvan, Mixing patterns and community structure in networks, in Pastor-Satorras R., Rubi J-M., Diaz-Guilera A., Statistical Mechanics of Complex Networks, Springer, berlin, 66-87, 2003.

[17] M.E.J. Newman, M.Girvan, Finding and evaluating community structure in networks, Physical Review E69, 2004.

[18] M.E.J. Newman., Analysis of weighted networks, Physical Review E70, 2004.

[19] M.T.Gastner, Traffic flow in a spatial network model, Proceedings of the 6th International Conference on Complex Systems, 2006.

[20] M.Granovetter, "The strength of weak ties", American Journal of SOciology 78, p.1360-1380, 1973.

[21] J.Jacobs, The death and life of great American cities, New-York: Random House, 1961.

Session D

The “Friends and Family” Mobile Phone Study: Overview and Initial Report

Nadav Aharony, Cory Ip, Wei Pan, Alex (Sandy) Pentland
MIT Media Lab
Contact: {nadav, coryip, panwei, pendland}@media.mit.edu

Abstract

In this talk we will give an overview of the “Friends and Family” study – a long term mobile phone experiment, in which a graduate family community is transformed into a living-lab for investigating a broad range of issues, including individual and group identity, real world decision making, social diffusion, social health, and privacy boundaries. In the first phase of the study, starting March 2010, a 100 Android based phones are distributed to selected participants in the community, equipped with our software platform that turns them into flexible social sensors and intervention-delivery mechanisms. By the time of the conference we should be able to also present some preliminary results.

Introduction

Today’s mobile phones are becoming powerful computing and sensing platforms. We are investigating ways to help people make use of the knowledge collected by their own mobile phones, as well as aggregate data contributed by many users, to improve their lives in constructive ways. In addition, we are investigating how this data can contribute to the understanding of societal and community related issues.

In recent years our lab has developed the methodology of Reality Mining, which is defined as the collection and analysis of machine-sensed environmental data pertaining to human social behavior, and is a key component in the transformation of traditional social science into the emerging area of computational social science. To gather this information, we use both our own home-brewed sensor platforms as well as smart-phones. We have already performed two large-scale experiments using close to a hundred phones at MIT campus in recent years. One study was performed in 2005 with participants from the MIT Media Lab and the Sloan School of Management. The second study was performed at an MIT undergraduate dorm during the 2008-2009 academic year. The first study was within a population of colleagues and co-workers. The second was done with a typical undergraduate population.

Our current goal is to pick a more “realistic” living community, with a population of couples and families that also have a community life and social interactions with one another. This community is a graduate family housing community at MIT, which has more than 400 residents. We would like to equip as many of these residents as we can with Android smart-phones running our software platform for data collection and other study related applications. To begin with, we are starting with handing out approximately 100 Android OS phones to selected residents.

The broad goal for this experiment is to develop mathematical models of social behavior of individuals and groups. These include (1) models of real-world group dynamics (in contrast to group dynamics within organizations and other structured settings), and also (2) models of the way that different "things" diffuse through a social network. "Things" can refer to a broad range – it would include the spread of behaviors, attitudes, and opinions - like exercise habits, smoking habits, music tastes, or political opinions. This also includes the spread of disease like the common cold or the flu. It is likely that many of these spread through social interaction – like social influence of one person on another, or the mere physical proximity between two people. The models we develop through this study would be then used to enhance mobile applications and software programs with the ability to comprehend human social interactions, and consequently, create more relevant, immersive, and privacy aware experience for the users.

Research Methodology

In the Friends and Family study, we are using mobile phones as in-situ social sensors to map users' activity features, proximity networks, media consumption and behavior diffusion patterns. The phones are augmented with social software to periodically execute different "probes" that capture information like cell tower ids, wireless LAN (WLAN) ids, Bluetooth ids, accelerometer and compass data, call and SMS, statistics on installed phone application and media files and usage and background noise/audio features. All phone numbers as well as any open text fields are encrypted using a one-way hash function.

The study involved the following data collection components:

1. Data collection by the mobile phone.
2. Surveys to establish "ground truth".
3. Establish participant spending patterns through purchase receipts.
4. Opt-in component: Facebook application, which, analogously to the phone probes, will collect data about the participant's online social activity.

The first part of the study (approx. 30-60 days), which is currently launching, is a baseline phase, where data is collected but no feedback will be presented to the users. After this phase, we begin a series of intervention phases. A simple intervention would be exposing participants to some of the data that has been collected through their device, and observing the effect on their behavior. For example, the participant will be able to review the amount of their social activity, or information about their approximate sleep hours (which might be inferred using data from the phone's accelerometer and/or information from an alarm clock application). Not all users will necessarily be exposed to the same information or information visualization to allow comparison. Another example intervention is to allow participants to decide whether to share this information with people in their contact list. Such feedback will help us investigate the effects of being exposed to one's own data, as well as to explore privacy issues related to the sharing of this information. Participants might also be exposed to aggregate information. For example, informing them on what "people like you" are doing, or what "people you spend time with" are doing. Such feedback must be carefully aggregated and anonymized so that it is not possible to infer information about specific participants.

Research Questions

The research questions we are trying to answer include a wide range of topics. Here are some examples of different questions that we expect to be able to answer through our study:

- **In the context of individuals:** How can the sensed real-world behavioral data be used to construct a “rich identity” profile of the user, which is more detailed and dynamic than current static demographic profiles. How is media propagation related to the user's face-to-face social network? Do social connectors play a major role? How is social influence defined in this context? Are there predictive patterns in how users consume and share media or other purchase recommendations? Can mobile devices automatically infer the user’s interest clusters and social groups, as well as recommend desired privacy settings based on these patterns?

- **Groups within a community:** What can we learn about group dynamics through the data collected by the experiment? Can we infer which formal and informal social groups (ethnic, religious, shared hobbies, neighbors, parents of similar aged children, etc.) participants belong to, and which of these have more influence in different contexts? How can we use the phones and phone data to improve community operation? (An example of an idea to improve community operation would be an application showing when different common areas are occupied.)

- **Community health and wellness**: How can we use the phones and phone data to improve community health and wellness? (An example of data that can be collected to improve health and wellness would be flu propagation. An example of an idea to improve community wellness would be an app to let people know about exercise habits within the community.)

- **Questions related to privacy and data interaction**: We want to use this experiment as a platform to learn how to deal with the sensitivities of collecting and using this deeply personal data. For example we would like to explore different techniques and methodologies to protect the users’ privacy while being able to generate meaningful outputs of the system. We would also like to explore different user interfaces for privacy settings, and for visualizing this personal data to the user.

Efficient Collaborative Application Monitoring Scheme for Mobile Networks

Yaniv Altshuler
Deutsche Telekom Labs
Ben Gurion University, Israel
Email: yanival@cs.technion.ac.il

Shlomi Dolev
Computer Science Department
Ben Gurion University, Israel
Email: dolev@cs.bgu.ac.il

Yuval Elovici
Deutsche Telekom Labs
Ben Gurion University, Israel
Email: elovici@bgu.ac.il

Nadav Aharony
Media Lab
MIT, Cambridge, MA, USA
Email: nadav@media.mit.edu

I. Introduction

In this work we discuss the problem of collaborative monitoring of mobile phones applications that are suspected of being malicious. New operating systems for mobile devices allow their users to download millions of new applications created by a great number of individual programmers and companies, some of which may be malicious or flawed. The importance of defense mechanisms against an epidemic spread of malicious applications in mobile networks was recently demonstrated by Wang et. al [19]. In many cases, in order to detect that an application is malicious, monitoring its operation in a real environment for a significant period of time is required. Mobile devices have limited computation and power resources and thus can monitor only a limited number of applications that the user downloads. We propose an efficient collaborative application monitoring algorithm, harnessing the collective resources of many mobile devices. Mobile devices activating this algorithm periodically monitor mobile applications, derive conclusion concerning their maliciousness, and report their conclusions to a small number of other mobile devices. Each mobile device that receives a message (conclusion) propagates it to one additional mobile device. Each message has a predefined TTL. The algorithm's performance is analyzed and its time and messages complexity are shown to be significantly lower compared to existing state of the art information propagation algorithms. In addition, we analytically prove that the algorithm is tolerant to Byzantine attacks aimed for injecting false information into the system. The algorithm was also implemented and tested extensively in a simulated environment.

II. Scope and Paradigm

Companies that are distributing new mobile devices operating systems had created a market place that motivates individuals and other companies to introduce new applications (such as Apple's *App Store* Google's *Android Market*, Nokia's *Ovi Store* and others). The content of the marketplace is not verified by the marketplace operators and thus there is no guarantee that the marketplace does not contain malicious or severely flawed applications. Downloading a malicious application from the marketplace is not the only way that a mobile device may be infected by malicious code. This may also happen as a result of a malicious code that manages to exploit a vulnerability in the operating systems and applications or through one of the mobile phone communication channels such as Bluetooth, Wi-Fi, etc' [8], [19].

In many cases, in order to detect that an application is malicious, monitoring its operation in a real environment for a significant period of time is required. The monitored data is being processed using advanced algorithms in order to assess the maliciousness of the application [7], [10], [11].

Harnessing their collective resources, a large group of limited devices can be shown to achieve a decentralized and efficient information propagation capability. Using such a service, participating users could significantly improve their "defense utilization" — the ratio between the resources a user is required to allocate for the collaborative service, and the probability to block attack attempts.

We present a collaborative application monitoring algorithm that provides high efficiency, scalability and robustness. The algorithm is completely decentralized and no supervising authority is assumed, nor do any central of hierarchical tasks allocation or any kind of shared memory. Specifically, we show that by sending $O(\ln n)$ messages, the number of applications a device would have to monitor in order to become "vaccinated" is reduced by a factor of $O(\ln n)$. Using real-world numbers, implemented as a service executed by 1,000,000 units, assuming 10,000 new applications are released every month, we analytically show that by monitoring a single application each month and sending 4 SMS messages per day, a participating mobile device can be guaranteed to be immune for 99% of all malicious applications.

III. Related Work

Since the problem of finding the minimum energy transmission scheme for broadcasting a set of messages in a given network is known to be NP-Complete [1], flooding optimization often relies on approximation algorithms. For example, in [6], [14] messages are forwarded according to a set of predefined probabilistic rules, whereas in [13] a deterministic algorithm which approximates the connected dominating set within a two-hop neighborhood of each node is proposed.

In this work we applied a different approach — instead of a probabilistic forwarding of messages, we assign a *TTL* value for each message, using which we are able to guide the flooding process. The analysis of the system is done by

	Time	Messages
TPP using $G(n,p)$ overlay	$O\left(\frac{\zeta}{\ln n}\right)$ in most cases $O(\ln n)$	$O\left(n \ln n\right)$
Flooding	$O\left(\textit{Graph's diameter}\right)$	$O(\lvert E\rvert)$
Network Coded Flooding **[3]** using $G(n,p)$ overlay	$O\left(n^{-1} \cdot p^{-2}\right)$	$O\left(n\right)$
Neighborhood Epidemics **[4]** using $G(n,p)$ overlay	$O(n^c)$ for some constant c	$O(c \cdot n)$ for a constant c
Hierarchical Epidemics **[16]** using α-tree overlay	$O(\ln n)$	$O(\alpha \cdot n \ln n)$ branching factor α
LRTA* [9] in planar degree bounded graphs	$O(n^2)$	$O(n^2)$
SWEEP [18] in the $\mathbf{Z}^2$ grid	$O(n^{1.5})$	$O(n^{1.5})$

TABLE I

PERFORMANCE COMPARISON BETWEEN THE *TPP* ALGORITHM AND AVAILABLE STATE OF THE ART ALGORITHMS.

	Time	Messages
TPP	$O\left(\frac{\zeta}{\ln n}\right)$ in most cases $O(\ln n)$	$O\left(n \ln n\right)$
Flooding	$O\left(\ln n\right)$	$O(n^2 p)$
Network Coded Flooding	$O\left(n^{-1} \cdot p^{-2}\right)$	$O\left(n\right)$
Neighborhood Epidemics	$O(n^c)$ for some constant c	$O(c \cdot n)$ for a constant c
Hierarchical Epidemics using α-tree overlay	$O(\ln n)$	$O(\alpha \cdot n \ln n)$ branching factor α

TABLE II

PERFORMANCE COMPARISON FOR RANDOM $G(n,p)$ GRAPHS, WITH $p < O((n \ln n)^{-0.5})$.

modeling the messages as agents practicing *random walk* in a *random graph* overlay of the network.

It is well known that the basic flooding algorithm, assuming a single source of information, guarantees completion in a worse case cost of $O(n^2)$ messages and time equals to the graph's diameter, which in the case of a random graph $G(n,p)$ is approximately $O(\log n)$ [2]. Variants of flooding algorithms use various methods to improve the efficiency of the basic algorithm, such as area based methods [12] or neighborhood knowledge methods [15]. An extremely efficient flooding algorithms in terms of completion time, is the network coded flooding algorithm, discussed in [3]. In this work, a message is forwarded by any receiving vertex $\frac{k}{d(v)}$ times, while k is a parameter which depends on the network's topology. Using this method, the algorithm achieves a completion time of approximately $O(\frac{n^3}{|E|^2})$. This algorithm, however, is still outperformed by our proposed algorithm. Specifically, our algorithm performs faster in graphs with average degree of less than $O\left(\sqrt{\frac{n}{\ln n}}\right)$.

An alternative approach to be mentioned in this scope is the use of *epidemic algorithms* [17]. There exist a variety of epidemic algorithms, starting with the basic epidemic protocol [5], through *neighborhood epidemics* [4] and up to *hierarchical epidemics* [16]. In general, all the various epidemic variants has a trade-off between number of messages sent, completion time, and previous knowledge required for the protocols.

Tables I and II present a summary of the performance of the *TPP* algorithm compared to the main body of works in this domain. The results of the second table assume that the average degree of network vertices is relatively small, marking the algorithm which guarantees best performance in gray. The value of the constant ζ can be approximated as $\zeta = \Omega(\ln^2 n)$ for most real world networks.

IV. EXPERIMENTAL RESULTS

In order to examine its performance, we have implemented the algorithm and conducted extensive simulations using various scenarios. In this section we describe one example, due to space considerations. This example concerns a network of $n = 1000$ units, having access to $N = 100$ applications, one of which was malicious[1]. Each unit is assumed to download 30 random applications, monitoring 1 application every week, and allowed to send notification messages to 10 random network members. Upon completion, at least 990 network members are required to become aware of the malicious application, and that this would hold in probability of 0.999. In addition, we assumed that among the network members there are 100 adversaries, whose goal is to mislead at least 50 of the network's members to believe that some benign application is malicious.

Figure 1 shows the time (in days) and messages required in order to complete this mission, as a function of the *decision threshold* ρ[2]. We can see that whereas the adversaries succeed in probability 1 for $\rho < 3$, they fail in probability 1 for any $\rho \geq 3$. Note the extremely efficient performance of the algorithm, with completion time of ~ 260 days using only 5 messages and at most 30 monitored applications per user. The same scenario would have resulted in 100 messages per user using the conventional *flooding* algorithm, or alternatively, in 700 days and 100 monitored applications per user using a non-collaborative scheme. Figure 2 demonstrates the decrease in completion time and messages requirement as a result of decreasing the *penetration threshold*, namely — the portion of the network which we allow to be deceived by attackers. A similar example concerning the effect of changing the graph's density (namely — the number of messages sent by each unit, upon the identification of a malicious application) is given in Figure 3. Figure 4 demonstrates the evolution in the malicious application's penetration probability throughout the vaccination process.

REFERENCES

[1] M. Cagalj, J.P. Hubaux, and C. Enz, *Minimum-energy broadcast in all-wireless networks: Np-completness and distribution issues*, MOBICOM, 2002.

[2] F. Chung and L. Lu, *The diameter of sparse random graphs*, Advances in Applied Mathematics **26** (2001), 257–279.

[1] Note that the number of malicious applications does not influence the completion time of algorithm, as monitoring and notification is done in parallel. The number of message, however, grows linearly with the number of malicious applications.

[2] An application is considered "malicious" where at least ρ messages are received concerning its maliciousness from different sources. This mechanism was designed in order to prevent the injection of false information to the network

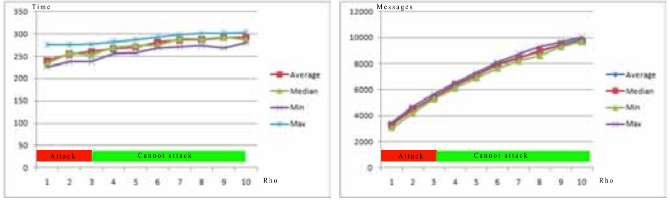


Fig. 1. An experimental result of a network of $n = 1000$ members, with $N = 100$ applications, *penetration threshold* $= 0.01$, graph density $= 0.01$ and 100 adversaries that try to mislead at least 5% of the network into believing that some benign application is malicious. Notice how changes in ρ dramatically effect the adversaries' success probability, with almost no effect on the completion time.

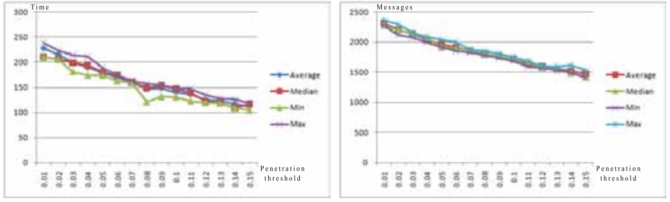


Fig. 2. The effect of decreasing the *penetration threshold* on the algorithm's completion time and number of messages ($\rho = 1$).

[3] S. Crisostomo, J. Barros, and C. Bettstetter, *Flooding the network: Multipoint relays versus network coding*, 4th IEEE Intl. Conference on Circuits and Systems for Communications (ICCSC), 2008, pp. 119–124.

[4] D. Ganesa, B. Krishnamachari, A. Woo, D. Culler, D. Estrin, and S. Wicker, *An empirical study of epidemic algorithms in large scale multihop wireless networks — technical report ucla/csd-tr 02-0013*, Technical report, UCLA Computer Science, 2002.

[5] R. Golding, D. Long, and J. Wilkes, *The refdbms distributed bibliographic database system*, In Proc. of Usenix94, 1994, pp. 47–62.

[6] Z. Haas, J. Halpern, and L. Li, *Gossip-based ad-hoc routing*, IEEE/ACM Transactions of networks **14** (2006), no. 3, 479–491.

[7] Hahnsang Kim, Joshua Smith, and Kang G. Shin, *Detecting energy-greedy anomalies and mobile malware variants*, MobiSys '08: Proceeding of the 6th international conference on Mobile systems, applications, and services (New York, NY, USA), ACM, 2008, pp. 239–252.

[8] J. Kleinberg, *The wireless epidemic*, Nature **449** (2007), 287–288.

[9] R. Korf, *Real-time heuristic search*, Artificial Intelligence **42** (1990), 189–211.

[10] R. Moskovitch, I. Gus, S. Pluderman, D. Stopel, C. Glezer, Y. Shahar, and Y. Elovici, *Detection of unknown computer worms activity based on computer behavior using data mining*, CISDA 2007. IEEE Symposium on Computational Intelligence in Security and Defense Applications, 2007, pp. 169–177.

[11] R. Moskovitch, S. Pluderman, I. Gus, D. Stopel, C. Feher, Y. Parmet, Y. Shahar, and Y. Elovici, *Host based intrusion detection using machine learning*, 2007 IEEE Intelligence and Security Informatics, 2007, pp. 107–114.

[12] S. Ni, Y. Tseng, Y. Chen, and J. Sheu, *The broadcast storm problem in a mobile ad hoc network*, In Proceedings of the ACM/IEEE International Conference on Mobile Computing and Networking (MOBICOM), 1999, pp. 151–162.

[13] L.V.A. Qayyum and A. Laouiti, *Multipoint relaying for flooding broadcast messages in mobile wireless networks*, Proceedings of HICSS, 2002.

[14] Y. Sasson, D. Cavin, and A. Schiper, *Probabilistic broadcas for flooding in wireless mobile ad-hoc networks*, Proceedings of IEEE Wireless communication and networks (WCNC), 2003.

[15] Ivan Stojmenovic, Mahtab Seddigh, and Jovisa Zunic, *Dominating sets and neighbor elimination-based broadcasting algorithms in wireless networks*, IEEE Transactions on Parallel and Distributed Systems **13** (2002), no. 1, 14–25.

[16] R. van Renesse and K. Birman, *Scalable management and data mining using astrolabe*, In Proc. of the First International Workshop on Peer-to-Peer Systems (IPTPS02), 2002.

[17] Werner Vogels, Robbert van Renesse, and Ken Birman, *The power of epidemics: robust communication for large-scale distributed systems*, SIGCOMM Comput. Commun. Rev. **33** (2003), no. 1, 131–135.

[18] I.A. Wagner, Y. Altshuler, V. Yanovski, and A.M. Bruckstein, *Cooperative cleaners: A study in ant robotics*, The International Journal of Robotics Research (IJRR) **27** (2008), no. 1, 127–151.

[19] P. Wang, M.C. Gonzalez, C.A. Hidalgo, and A.L. Barabasi, *Understanding the spreading patterns of mobile phone viruses*, Science **324** (2009), 1071–1075.

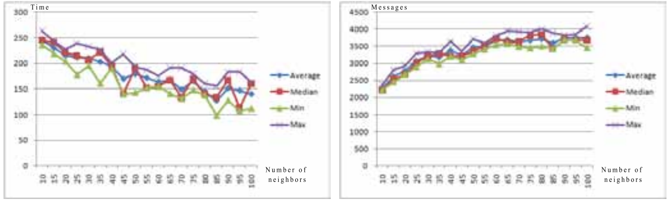


Fig. 3. The effect of decreasing the graph density on the algorithm's completion time and number of messages ($\rho = 1$).

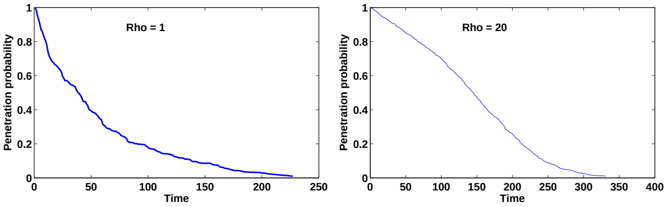


Fig. 4. The *penetration probability* of the malicious application, as a function of the time, with $\rho = 1$ (on the left) and $\rho = 20$ (on the right).

Beyond San Fancisco Cabs : Building a *-lity Mining Dataset[*]

Marc-Olivier Killijian
CNRS; LAAS
Univ. Toulouse, France
marco.killijian@laas.fr

Matthieu Roy
CNRS; LAAS
Univ. Toulouse, France
roy@laas.fr

Gilles Trédan
TU Berlin
Berlin, Germany
gilles@net.t-labs.tu-berlin.de

ABSTRACT

In this paper, we report our advances, choices and first insights in the design of a mobile phone powered data collection platform. We believe that collecting such data is vital to achieve a better understanding/modeling of several phenomenons related to human activity (e.g. mobility, social contacts, or terminal failures). However, designing such a platform raised a lot of questions that we present in this paper.

1. MOTIVATION

Context.

The ubiquity of geo-located devices (mobile/smart-phones, GPS, etc.) permitted scientists to collect datasets containing mobility information. Mining reality (and mobility) out of these datasets has drawn a lot of attention, for example for studying the spread of viruses [9, 12], for designing socially-aware network routing protocols [2, 8], and in the raising dynamical networks community. These datasets can also be used to design models of users behavior [5, 7, 17], phone usage, etc.

Background.

We are interested in building *resilient ubiquitous mobile systems*. To attain this goal, several aspects of these systems need to be investigated, from their formal foundations to their experimental evaluation. As such, we worked on distributed [16] and cooperative [10] algorithms for such systems, but also on both analytical [3] and experimental resilience evaluation [11] of mobile systems. One of the major challenge when modeling a mobile distributed system is to provide an adequate mobility and connectivity scheme, that should represent, as precisely as possible, actual interactions between *human beings*. Mobile devices are the first sensing devices that are almost coupled with their carrier, allowing to capture human activity with an unprecedented resolution. This drove us to dig in research on mobility datasets, to perform experimental and analytical evaluation based on actual mobility traces.

Indeed, as far as analytical evaluation is concerned, we needed to use parameters representing the rate at which a node encounters a cooperating peer, the rate at which it gets Internet connectivity, the rate at which it (or any peer) fails, etc. We had to choose these parameters (and their distribution law) by rule of thumb but definitely wanted to get actual and accurate data to backup these estimates. Regarding experimental evaluation, we are building an emulation platform based on small robots that carry laptops running the algorithms. At the moment we used predefined vehicular scenarios but we want the platform to be able to use proper mobility models.

Mobility and Social Models.

It is worth noticing that the whole mobile systems community struggles with classical random-* mobility models. They are too far from the application scenarios to be sound. Recent research works try to tackle this problem (e.g., in [13]). Yet, there is still a need for extending mobility models with failure models for small devices such as mobile phones. Indeed, as these devices are mass-produced and carried by users, they are more prone to failures: they move a lot, are light, and henceforth are prone to fall; they have a small battery and are prone to energy depletion; they are used for gaming or multimedia purpose, users install a bunch of applications and henceforth they are prone to software failures. For all these reasons, an ideal dataset would include information about energy consumption, operating system reboots, mobility, etc.

To the best of our knowledge, there is no dataset that satisfies this specification. Most of CRAWDAD[1] datasets were produced in a limited vicinity, either during a conference, or on a campus. Very few available datasets include precise localization data [14], most of them focused on contact traces. Unfortunately phone operators are not keen to release their data and do so to only some lucky few [7, 17]. Furthermore, there are very few data collection tools available. For example, the software used to build the Reality Mining dataset [4] is not maintained anymore [15]. This drove us to decide to build our own platform for collecting mobility, failure and energy consumption data.

2. A PLATFORM FOR COLLECTION OF MOBILITY AND SOCIAL TRACES

2.1 Operating System Platform

Recently, mobile phone platforms have drawn a lot of attention from manufacturers and users, mainly due to the

[*] Contact author: Matthieu Roy
roy@laas.fr

[1] http://crawdad.cs.dartmouth.edu/

Submitted to NetMob 2010.

opportunity to embark what has become lightweight computers in everyone s pocket. As such, current mid- to high-end phones are now equipped with large screens, GPS, high speed cellular network access, local wireless network interfaces, and various sensors (accelerometers, magnetometers, etc). As for every computing device, interaction with hardware is done through an operating system, using an API. The different phone manufacturers offer platforms that differ radically in their operating system and API choices. In this section, we describe the various options available, and explain which platform we chose and why such a choice.

Indeed, from our point of view, the rationale to choose one platform instead of another is drawn by two main reasons:

- the platform should offer a simple way to access as much important parameters on user s activity as possible,
- the platform should be desirable for users, so that users will accept to use the platform as their main phone, and run our software.

Most mid- to high-end phone on the market now provide a GPS for mobility tracing, at least one wireless network interface (bluetooth), and a large touchscreen. The difference between available platforms resides in their operating system and their development model (API and software distribution model). In our view, the best platform for performing a background collection of various parameters on a large base of users appeared to us to be the Symbian platform: information for programming is easily available, development tools are mature, the system is very stable, it can run multiple applications in parallel, and it is actively developed and maintained.

2.2 Hardware Platform

We want to provide many users with smartphones equipped with our software. This led us to find a phone (1) that can *sense* many parameters (at least GPS, WiFi, bluetooth), (2) that has a good autonomy to support the additional energy consumption necessary for our logging application, (3) that is cheap and, last but not least, (4) that is considered desirable for users.

We finally opted for the Nokia 5800 smartphone. Its retail cost is about 230€, which is relatively cheap when considering that it provides all hardware we needed, a GPS chip, a large (640∗360) touchscreen, an accelerometer, a compass, a WiFi interface and a bluetooth interface in a small form factor (about 100g). Moreover, this phone is considered a good one from the users point of view, due to its light weight, its good camera, its free navigation system and the fact that it can be used with video conference softwares such as Skype.

2.3 Design: what can be sensed

When designing our application, it appeared that the choice of the programming language would have an impact on available data that can be sensed. In Java (be it Java Micro Edition or Java Standard Edition), no platform allows a program to get access to low-level parameter such as the list of available WiFi interfaces. Thus, we had to use system-oriented API and programming language. Symbian standard interface is a C++ API that gives the programmer access to most of the capabilities of the phone. In our case, we listed the following interesting sources of information:

- GPS information. Location information is essential when building a mobility trace. This piece of information will be sampled at fixed interval. It can also generate events when passing near some predefined interest points.
- WiFi access points, and WiFi usage.
- Cellular information.
- Nearby bluetooth devices.
- Battery level. This should be sampled, and the logging software should collect power events (charging, battery warning) when they occur.
- Phone calls.
- Accelerometer. Although this parameter is not *per se* related to interactions or mobility, it may be used to detect inactivity.
- Compass.
- Proximity Sensor senses wether the surface of the phone is close to an object (a pocket, an ear)
- Light Sensor.
- Power consumption related resources: CPU usage, current intensity drawn from the battery.

2.4 Implementation considerations

In our first implementation, we limited the logging to a subset of available sources of information, as summarized in Figure 1:

Information source	**periodicity**	**fixed/variable size**
GPS	periodic	fixed
WiFi	periodic	variable
Bluetooth	periodic	variable
Battery level	periodic	fixed
Battery events	sporadic	fixed
Phone calls	sporadic	fixed
Reboots	sporadic	fixed

Figure 1: Logged sources of information

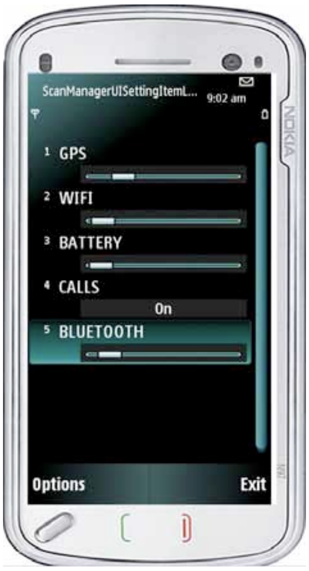
There are two reasons why we capture these sources of information only. First, all above data are needed if we want to capture user interaction with its environment and with other users, as well as failure information. Second, due to security restrictions in the Symbian OS, we would have to certify our program with Symbian Foundation if we wanted to log information such as the cell identifier the phone is connected to, or the signal strength of the 3G/HSDPA.

Logging.

Since the program runs on limited resources, much care has been taken to have the smallest possible footprint on the system. The program is divided in two: a graphical user interface, to start/stop the service and to modify logging parameters, and the logger application itself.

The graphical user interface is simple, and permits the user to stop the logging service when he/she wants, modify the frequency of scan for every data source, and turn on or off the logging for every parameter. A screen capture is shown in Figure 2.

The logger is programmed using one thread only, by using the concept of Active Objects. Active Objects is a system mechanism that provides a comprehensive way to perform



1 GPS

2 WIFI

3 BATTERY

4 CALLS
On

5 BLUETOOTH

Figure 2: GUI for managing logger frequency

multiple tasks using system calls (e.g., reading of system parameters such as GPS information) within a single thread that acts as a scheduler. Hence, the resulting program uses low system resources and optimizes battery usage.

The thread is programmed to scan all parameters shown in Figure 1, and to store all information in a text file. The chosen data format is a simple human readable line-based log, that shows for each scanned parameter the time of logging, the type of logging, and the value of the log, as shown in Figure 3.

```
Starting scan
current parameters#GPS#1#WIFI#1#BATTERY#1#CALL#1#BT#1#
2009-10-2 13:50:39#GPS#+43.59407#+1.46388#+136.49504#+25.50000
2009-10-2 13:50:39#BATTERY#71#EPoweredByBattery
2009-10-2 13:50:39#POWER#BATTERY
2009-10-2 13:50:51#BT#berimbau#00:02:26:60:08:8b
2009-10-2 13:50:57#BT#euclide#00:01:14:45:51:1d
2009-10-2 13:50:58#BT#Nokia marco#00:02:24:40:04:4d
2009-10-2 13:51:23#BT#Mattmobile#00:02:21:1f:fc:c3
2009-10-2 13:51:32#BT##00:02:25:50:00:0c
2009-10-2 13:51:39#WIFI#Network 1#60#open#Infrastructure#\
         00:23:33:78:7f:60#laas-welcome
2009-10-2 13:51:39#WIFI#Network 2#91#open#Infrastructure#\
         00:25:45:b5:75:00#laas-welcome
```

Figure 3: Example of log

Gathering.

The program that runs on the mobile stores information locally. We are currently developing a networking part, that will opportunistically send logs to a secured server in the laboratory. Although sending information to a server seems a simple task, this requires additional work, due to security risks put on users, particularly for privacy reasons. We are investigating cryptographic mechanisms that would protect users' privacy, while still permitting analysis on gathered data.

Obviously, we cannot just anonymize every data trace with a fixed random identifier, because each trace contains many personal information (localization, users interactions) that could be used to "de-anonymize" a trace [6].

3. ANALYSIS AND USE OF COLLECTED DATA

Once the privacy-preserving database is filled with user data, it can be used for different tasks: analysis of data for social networks, resilience evaluation of mobility-aware algorithms, and privacy preserving analysis of the database.

Social Links Engineering.

In the vein of recent work on social links engineering [17] [7] [5], we plan to use this database, that contains more information than operators' mobile phones logs, to derive realistic mobility models. Current mobility models are mostly random-based and thus do not take into account the fact that devices are carried by humans.

In fact, the database will not only be useful for mobility modeling but, more generally, will serve as a basis for social links analysis, and particularly to answer the following questions:

> Does there exist *social locality*?, i.e., when social links exist between a group of users, do these links imply a more frequent co-locality in the physical world ? Such a result would have a strong impact on possible extensions of current social services (MySpace, Twitter, Facebook) to geo-localized systems. Intuitively, such locality principle should exist. However, measurements would confirm or infirm this hypothesis, and would allow to develop models that efficiently capture this potential locality.
>
> Can we model expected *interaction patterns* between users, so that we can build collaborative services that are based on stable users interactions ?
>
> Are there patterns of interactions between users and the environment that are more likely to appear than others ?

Replying to these questions seems for us a prerequisite to be able to design efficient and useful services for distributed systems of mobile users.

Resilience evaluation.

Collected data will also serve as an input both for our research on analytical evaluation of resiliency of algorithms [3] and for our reduced-size experimental platform [11]. For analytical evaluation, we will compute probabilistic laws for useful evaluation parameters, such as the expected time of users interaction, the expected time between failures, etc.

For experimental evaluation, we are interested in isolating *mobility patterns* that will then be used as an input to our experimental platforms, with the final goal of performing controlled and realistic experimentation by testing different algorithms on the same mobility patterns, a task that is impossible in a real environment.

Privacy evaluation.

The last use of the collected data we foresee in the actual evaluation its privacy-preserving features. From the very start of the project, we were struggled by privacy issues. Research results [6] [1] show that simple privacy-preserving strategies do not suffice to protect users, due to the huge amount of personal identifying data traces stored.

The more traces collected, the higher the threat on users privacy. As technology evolves, eventually almost everyone will be able to dig into such information, and providing users with means to protect their data will become a priority. Performing attacks on such a database will permit us to detect wether our strategies for pseudonymation and privacy-guarantee is robust.

4. CONCLUSION

We presented the design and implementation of a mobile phone powered data collection platform. Our platform collects data that, we believe, will ease the understanding and modeling of several phenomenons related to human activity: social links, mobility, etc.

We still have to define the economical nature of data collection. Users hardly want to have their privacy exposed without a counter part. Such reward could consist of several parts: (1) an economic part, by providing a free terminal, (2) a service-oriented part, by providing services users are not used to, such as free video-conference service, and, as we hope, (3) an altruist academical part, providing users the satisfaction to participate to a research program. It is worth noticing that the chosen remuneration scheme will imply a bias on collected data that will have to be studied. . .

5. ACKNOWLEDGEMENTS

The authors would like to thank Hamdi Ayed for its implication in the development of the first version of the software.

6. REFERENCES

[1] L. Backstrom, C. Dwork, and J. Kleinberg. Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 181 190, New York, NY, USA, 2007. ACM.

[2] C. Boldrini, M. Conti, and A. Passarella. *Social-based autonomic routing in opportunistic networks*, pages 1 37. Springer, 2009.

[3] L. Courtès, O. Hamouda, M. Kaaniche, M.-O. Killijian, and D. Powell. Dependability evaluation of cooperative backup strategies for mobile devices. In *Proceedings of the IEEE International Symposium on Pacific Rim Dependable Computing*, Melbourne, VA, Australia, December 2007.

[4] N. Eagle and A. S. Pentland. CRAWDAD data set mit/reality (v. 2005-07-01). Downloaded from http://crawdad.cs.dartmouth.edu/mit/reality, July 2005.

[5] N. Eagle, A. S. Pentland, and D. Lazer. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, 106(36):15274 15278, 2009.

[6] S. Gambs, M.-O. Killijian, and M. N. del Prado. Gepeto: a geo-privacy enhancing toolkit. In *Proc. of the Int. Workshop on Advances in Mobile Computing and Applications: Security, Privacy and Trust, 24th IEEE AINA conference, Perth, Australia, April 2010.* IEEE, 2010.

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

[8] D. Hay and P. Giaccone. Optimal Routing and Scheduling for Deterministic Delay Tolerant Networks. In *WONS'09: 6th. Int. Conf. on Wireless On-demand Network Systems and Services*, pages 25 32. IEEE, 2009.

[9] M. J. Keeling and L. Danon. Mathematical modelling of infectious diseases. *British Medical Bulletin*, 92(1):33 42, DEC 2009.

[10] M.-O. Killijian, D. Powell, M. Banatre, P. Couderc, and Y. Roudier. Collaborative backup for dependable mobile applications. In *Proceedings of 2 nd International Workshop on Middleware for Pervasive and Ad-Hoc Computing (Middleware 2004)*, pages 146 149, Toronto, Ontario, Canada, October 2004. ACM Press.

[11] M.-O. Killijian and M. Roy. A platform for experimenting with mobile algorithms in a laboratory. In *ACM Conference on Principles of Distributed Computing (PODC'09)*, pages 316 317. ACM, 2009.

[12] S. Merler and M. Ajelli. The role of population heterogeneity and human mobility in the spread of pandemic influenza. *Proceedings of the Royal Society B: Biological Sciences*, 277(1681):557 565, FEB 22 2010.

[13] M. Musolesi and C. Mascolo. Mobility models for systems evaluation. In *State of the Art on Middleware for Network Eccentric and Mobile Applications (MINEMA). Springer, February 2009.*

[14] M. Piorkowski, N. Sarafijanovic-Djukic, and M. Grossglauser. CRAWDAD data set epfl/mobility (v. 2009-02-24). Downloaded from http://crawdad.cs.dartmouth.edu/epfl/mobility, Feb. 2009.

[15] M. Raento, A. Oulasvirta, R. Petit, and H. Toivonen. Contextphone: A prototyping platform for context-aware mobile applications. *IEEE Pervasive Computing*, 4(2):51 59, 2005.

[16] M. Roy, F. Bonnet, L. Querzoni, S. Bonomi, M.-O. Killijian, and D. Powell. Geo-registers: An abstraction for spatial-based distributed computing. In *Int. Conf. On Principle of Distributed Systems (OPODIS'08)*, volume 5401 of *Lecture Notes in Computer Science*, pages 534 537. Springer, 2008.

[17] C. Song, Z. Qu, N. Blumm, and A.-L. Barabasi. Limits of Predictability in Human Mobility. *Science*, 327(5968):1018 1021, 2010.

Mining Large Scale Cell Phone Data

Jean Bolot
Sprint
Burlingame, California, USA
http://jeanbolot.com/

ABSTRACT

Cell phones are ubiquitous in modern life and the call records collected by network operators are a powerful tool to study the behavior of cell phone users, and how those users use network resources, at previously impossible-to-achieve scales. In this paper we report on results from the analysis of large scale call records data, and more generally of the data generated by mobile users, at a large cellular operator. We consider in particular three kinds of data, namely social network data (who calls whom, how often, etc), location and mobility data (who is where) and spectrum data (who uses how much spectrum in which cell). We describe practical examples of insights derived from mining that data, the impact of the data on areas ranging from marketing to business models or to security, and also consider interesting research challenges ahead.

1. INTRODUCTION

The Internet has become a fundamental component of modern economies, and it provide services, starting with connectivity, that are strategic to companies, governments, families and individual users, and in general to the well functioning of modern life. A growing fraction of those services are accessed by mobile users. Indeed, the size and strategic importance of the mobile Internet, i.e. the Internet as accessed via mobile devices such as laptops or cell phones, is rapidly increasing. Recent reports indicate that the mobile Internet is ramping up in size faster than the "desktop Internet" did in the 80's and 90's; in fact, the estimated total value of the mobile data industry grew by 20% in 2009 - a year of major economic crisis when the global economy decreased by 5% - and mobile data revenues reached $284B [5]. This is now larger than the total PC Internet economy, including Internet content and advertising revenues plus all subscription fees such as monthly dial up and broadband access fees. Furthermore, the number of users of the mobile Internet (measured by the number of users accessing browser-based services on cell phones only) is estimated at between 500 million and 1 billion, almost on par with the total number of PCs connected to the Internet [5, 1]. Thus, cell phones already dominate the Internet, and their importance will continue to grow [4].

A key characteristic of cellular networks and devices is their ability to capture and analyze (at least partial) information on the behavior of mobile users. In particular, operators have routinely captured large scale location data for billing purposes, but also to improve location management or satisfy legal requirements such as E911. More recently they, as well as a number of analytics companies and academic research groups worldwide, have started analyzing a growing variety of data including social network data (who calls whom), location and mobility data (where users are when they call or use services), click-stream data (which sequence of sites users visit, or which sequence of applications and services they use), etc.

In this paper, we report on results from the analysis of such data carried out at a large cellular operator. We consider in particular three kinds of data, namely social network data, location data and spectrum usage data (who uses how much spectrum in which cell). We describe some of the insights derived from mining that data and consider some of the interesting research challenges ahead.

2. SOCIAL NETWORKS

We have analyzed a very large social network gathered from call details records, which reflects the voice and SMS interactions of more than ten million users through hundreds of millions of calls and SMS exchanges. We examined the distributions of the number of phone calls per customer; the total talk minutes per customer; and the distinct number of calling partners per customer. We found that these distributions are skewed, and that they significantly deviate from what would be expected by conventional wisdom, namely power-law and lognormal distributions.

We found instead that our observed distributions (number of calls, of distinct partners, and of total talk time) very closely fit a lesser known but more suitable distribution, namely the Double Pareto LogNormal (DPLN) distribution [6]. We found good fits over time (morning-evening, weekday-weekend) and space (US East Coast-West Coast, urban-suburban).

More importantly, we also found that our graph evolved over time in a way consistent with a generative process based on geometric Brownian motion. Furthermore, this generative process lends itself to a natural and appealing *social wealth* interpretation, and also allows for extrapolations and interpolations. We hope that our success with DPLN spurs further studies involving other datasets and their underlying generative processes. In particular, we hope that our "social wealth" interpretation and analysis will serve as an incentive for social scientists to study the large-scale evolutionary aspects of social characteristics. Indeed, we continue to collect data from our social network for longer-term analysis.

3. LOCATION AND MOBILITY

We have also analyzed call records to understand the mobility patterns of more than a million users over several thousand square miles. We made two contributions to the analysis of mobility patterns of cell phone users. First, using only coarse-grained location information, namely the location of the cell tower associated with a user at the beginning and end of each call, we examined the scaling laws of human mobility, in terms of distance and time. We found that both the distance traveled as well as the duration of calls (on periods) and pauses (off periods) are heavy tailed, in agreement with earlier results (e.g. [3]). However, we found that mobility patterns change during and in between calls, and that patterns are correlated over time, with strength of correlation dependent on activity.

Second, we developed a general technique, using tools from stochastic geometry and Bayesian statistics [7], to refine mobility models as more precise location information becomes available [11]. Thus, we can correct the distributions of distance traveled and direction as coarse location information is augmented by information such as distance to the associated cell tower, signal strength, location of neighboring cells towers, etc. To demonstrate the benefits of our technique, we first showed, using timing measurements from call records, that users are not uniformly distributed in cells. We then showed how that location information impacts the estimated distance distribution and then extended our earlier technique, illustrating the impact of increasingly more precise location information. Our approach is very general and applicable not just to cellular networks, but to other wireless networks such as wireless LANs (WiFi, ...) or adhoc networks.

4. SPECTRUM USAGE

Most existing studies of spectrum usage have been performed by actively sensing the energy levels in specific RF bands including cellular bands. Our approach has been to provide a unique, complementary analysis of cellular primary usage by analyzing a dataset collected inside a cellular network. One of the key aspects of our dataset, compared to others examined in related spectrum analysis, is its scale - it consists of data collected over three weeks at hundreds of base stations. We dissected this data along different dimensions to characterize and model primary usage as well as understand its temporal and spatial variations. Our analysis revealed several results that are relevant if Dynamic Spectrum Access (DSA) approaches are to be deployed for cellular frequency bands. For example, we found that call durations show significant deviations from the often-used exponential distribution. Though this can complicate the modeling of primary usage, we found that a random walk process, which does not use call durations, can be used for modeling the aggregate cell capacity. Another novel result we found is that spatial spectrum usage is highly non-uniform, especially during periods of high load, with clusters of sectors whose intra-cluster usage patterns are correlated.

We also considered the more fundamental problem of whether or not spectrum sensing is actually a viable approach to estimate when and how much secondary users can take advantage of available capacity. Indeed, sensing mechanisms that estimate the occupancy of wireless spectrum play a cricital role in enabling non-interfering secondary usage. The problem of designing such mechanisms is, therefore, crucial to the success of approaches based on Dynamic Spectrum Access. We developed key insights into this problem by empirically investigating the design of sensing mechanisms applied to check the availability of excess capacity in CDMA voice networks. We focussed on power-based sensing mechanisms since they are arguably the easiest and the most cost-effective.

We made three main contributions [9]. First, we found that accurate single sensor spectrum sensing is essentially unachievable, i.e. power at a single sensor is too noisy to help us accurately estimate unused capacity. However, we also found that there are well-defined signatures of call arrival and termination events. Using these signatures, we showed that we can derive lower bound estimates of unused capacity that are both useful (non-zero) and conservative (never exceed the true value). Finally, we used a combination of measurement data and analysis to deduce that multiple sensors are likely to be quite effective in eliminating the inaccuracies of single-sensor estimates.

5. FUTURE RESEARCH: BUSINESS MODELS

The capture and availability of large scale cell phone data has enabled, and will continue to enable, a wide range of new services. For example, in the case of location and mobility data, the capture and availability of such data has enabled the development of many location-based or location-aware services, and indeed an rapidly increasing number of such services is now available, ranging from navigation to location-aware advertising, friend finder, etc, and many more are announced or launched on a daily basis. However, this location data, since it enables new services and new economic activities, is seen as economically valuable. This raises the question then of how valuable it is, and how to quantify that value. This is precisely the goal of our recent research.

Using insights from cell phone data, we have developed an analytic framework, namely models and the techniques to solve them, to help quantify the economics of location information [2]. Our aim has been to derive models which can be used as decision making tools for entities interested in or involved in the location data economics chain, such as mobile operators or providers of location aware services (mobile advertising, etc). We considered in particular the fundamental problem of quantifying the value of different granularities of location information, for example how much more valuable is it to know the GPS location of a mobile user compared to only knowing the access point, or the cell tower, that the user is associated with. We have used our approach to derive insights into what is arguably the quintessential location-based service, namely proximity-based advertising.

To our knowledge, our work is the first one to present and analyze economic models which can help understand the economic value generated by mobile users with location based services, for different granularities of location information in wireless networks. We believe that the work provides an important first step towards a general analysis of not just the data itself, but also of the business models enabled by large scale cell phone data.

6. REFERENCES

[1] T. Ahonen, *Mobile as the 7th Mass Media*, London, UK: futuretext, 2008.

[2] F. Baccelli, J. Bolot, “Modeling the economic value of location and preference data of mobile users”, submitted for publication, Nov. 2009.

[3] M. Gonzalez, C. Hidalgo, and A. Barabasi, “Understanding individual human mobility patterns”, *Nature*, vol. 453, pp. 479–482, 2008.

[4] S. Keshav, “Why cell phones will dominate the future Internet, *Computer Communications Review*, vol. 35, no. 2, April 2005.

[5] M. Meeker, *The Mobile Internet Report*, Chichester: Morgan Stanley, Dec. 2009.

[6] M. Seshadri, S. Machiraju, A. Sridharan, J. Bolot, J. Leskovec, C. Faloutsos, “Mobile call graphs: Beyond power-law and lognormal distributions”, *Proc. ACM KDD Conference on Knowledge Discovery and Data Mining*, Las Vegas, Aug 2008.

[7] D. Stoyan, W. Kendall, and J. Mecke, *Stochastic Geometry and its Applications*, Wiley, 1995.

[8] D. Willkomm, S. Machiraju, J. Bolot, A. Wolisz, “Primary users in cellular networks: A large-scale measurement study”, *Proc. IEEE Symp. Dynamic Spectrum Access Networks (Dyspan)*, Chicago, IL, Oct. 2008.

[9] D. Willkomm, S. Machiraju, A. Wolisz, “The problem of spectrum sensing in cellular networks”, submitted for putblication, Mar 2010.

[10] H. Zang, J. Bolot, “Mining call and mobility data to improve paging efficiency in cellular networks”, *Proc. ACM Mobicom ’07*, Montreal, Canada, Sept. 2007.

[11] H. Zang, F. Baccelli, J. Bolot, “Bayesian inference for localization in cellular networks”, *Proc. IEEE Infocom 2010*, San Diego, CA, Apr. 2010.