Mobility and Sociocultural Events in Mobile Phone Data Records

Nicolas B. Ponieman [a] Carlos Sarraute [a,*]
Martin Minnoni [a] Matias Travizano [a]
Pablo Rodriguez Zivic [b] Alejo Salles [c]
[a] *Grandata Labs*
Buenos Aires, Argentina
E-mail: {nico, charles, martin, mat}@grandata.com
[b] *Computer Science Dept., Universidad de Buenos Aires*
E-mail: prodriguez@dc.uba.ar
[c] *Physics Dept. & Instituto de Cálculo, Universidad de Buenos Aires and CONICET*
E-mail: alejo@df.uba.ar

The massive amounts of geolocation data collected from mobile phone records have sparked an ongoing effort to understand and predict the mobility patterns of human beings. In this work, we study the extent to which social phenomena are reflected in mobile phone data, providing some proof-of-principle examples. We illustrate in various ways how these events are reflected in the data, and show how information about the events can be used to improve predictability in a simple model for a mobile phone user's location. We further propose a method for the automatic detection of such events and discuss their relation to the social fabric as derived from mobile phone communications.

Keywords: human mobility, human predictability, social phenomena

1. Introduction

Mobile phone operators have access to an unprecedented volume of information about users' real-world behavior. The records of calls and messages exchanged between their users provides a deep insight into the interactions and activities of millions of individuals. The social graph induced by mobile communications has provided a rich field to apply social network analysis to real-world problems. For instance, we can highlight the use of community detection techniques (see [2,7,13,21]); and the more recent advances in detecting the evolution of communities in dynamic networks (taking into account the evolution of the social graph over time) in [1,17].

A key aspect of the data collected by mobile phone operators that has attracted considerable attention in recent years is the information about how people are moving in the real world. In fact, mobile phone records can be considered as the most detailed information on human mobility across a large part of the population [19]. The study of the dynamics of human mobility using the collected geolocations of users, and applying it to predict future users' locations, has been an active field of research [6,12]. In particular, this information can be used to validate human mobility models (as the authors of [14] did with the information from a location-based social networking site); and to study the interplay between individual mobility and social networks [20].

The study of human mobility can be applied to domains as diverse as city planning and traffic engineering (e.g. to optimize the public transportation system and the roads network); public health (e.g. to allow health officials to track and predict the spread of contagious diseases); or to guide humanitarian relief after a large-scale disaster (see [12] wherein the authors study population movements after the Haiti 2010 earthquake). Using real-world data to understand human mobility is critical to such applications. On the business side of applications, mobile carriers are seeking for new revenue streams based on the anonymized and aggregated analysis of their subscribers' mobility data [11].

In this work, we study the extent to which social phenomena are reflected in mobile phone data, focusing in particular in the cases of urban commute and major sports events. The rest of the paper is organized as follows. Section 2 describes the

*Corresponding author: *charles@grandata.com*.

AI Communications
ISSN 0921-7126, IOS Press.

real-world data source that we used for our experiments. In Section 3, we present a simple model to predict the location of a mobile phone user, that we used as baseline of predictability. In Section 4, we illustrate how urban commute can be observed in the data, and compute basic metrics. The mobility pattern associated with a sports event is exposed in Section 5, where we also discuss how exogenous information about social events can be used to improve the predictability of the simple model. In Section 6, we show that the information from social events is actually already contained in the data, and illustrate this by implementing a mechanism for automatically detecting events from the call data records alone. We further include the information from the topology of the network to argue for a 'herd behavior' of people when attending sociocultural events, which constitutes the basis for an effective tagging of users, for which we give an outline. Finally, Section 7 concludes the paper, and discusses ideas for future work.

2. Mobile Data Source

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

For our purposes, each record is represented as a tuple $\langle x, y, t, d, l \rangle$, where user x is the caller, user y is the callee, t is the date and time of the call, d is the direction of the call (incoming or outgoing, with respect to the mobile operator client), and l is the location of the tower that routed the communication. The temporal granularity used in this study is the hour, justified by the findings in [18,19].

From the operator's data, it is possible to have direct information on mobile phone usage patterns, as can be seen in Figure 1, which shows the volume of communications according to the day of the week and the hour. The expected contrast between weekend and workweek is evident. More interesting information is given by the communications peaks during the morning (around 11 a.m.) and the afternoon (around 18 p.m.) from Monday through Friday, which depend on the working habits in Argentina. The pattern for Monday appears in Figure 1 as a mixture between weekdays and weekends, which is explained by the fact that most public holidays in the period studied were on this day of the week.

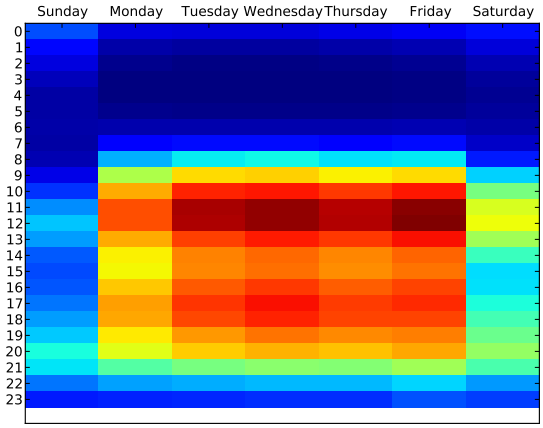


Fig. 1. Call distribution according to the day of the week and the hour, averaged over a period of five months.

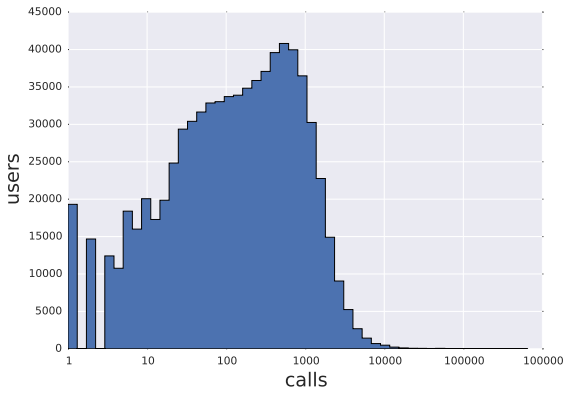


Fig. 2. Distribution of users according to their total number of calls.

Further insight into the dataset can be gained from Figs. 2 and 3, where we show the distribution of users according to the total number of calls, and the distribution of call durations. These constitute a sanity check for our data.

3. Mobility Model

To predict a user's position, we use a simple model based on previous most frequent locations,

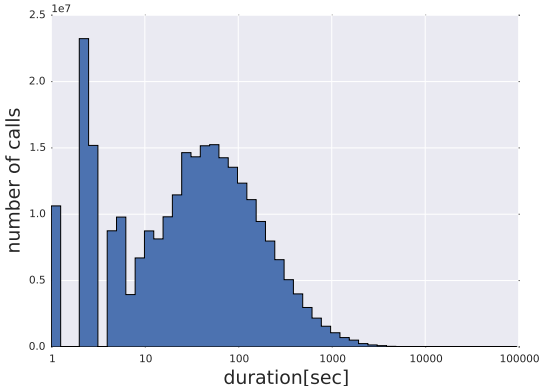


Fig. 3. Distribution of users according to the duration of calls.

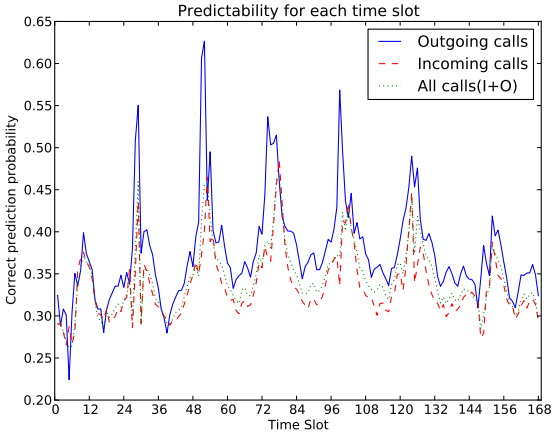


Fig. 4. Users' location predictability by time slot. Blue: Outgoing calls. Red: Incoming calls. Green: All calls. We considered the week starting at Sunday, so the first time slot corresponds to Sunday from midnight to 1 a.m., whereas time slot 168 corresponds to Saturday from 11 p.m to midnight.

and compute its correct prediction probability (i.e. accuracy). In order to find these locations, we split the week in time slots, one for each hour (as in Figure 1), totalizing $7 * 24 = 168$ slots per week. We then predict a user's location as the most visited in the prior equivalent time slots of the week. Since humans tend to have very predictable mobility patterns [8,10,19], this simple model turns out to give a good predictability baseline, achieving an average of around 35% correct predictions for a validation period of 2 weeks (and training with 15 weeks of data), including peaks of over 50% predictability. This model was used as a baseline in [3], with which our results agree. In Figure 4 we show the average predictability for all time slots (considering the week from Sunday to Saturday).

Although the kind of periodic behaviour observed in the figure is widely explained in the literature, it is important to make a few remarks about the results obtained:

- It is clear that predictability is at least 25% higher during weekdays (Monday - Friday) than during the weekend.
- During the night, people have a peak of predictability, corresponding to the time they typically spend at home.
- Predictability is slightly higher when computed from outgoing calls than when computed from incoming calls.

One way to improve predictability would simply consist in restricting the analysis to frequent callers, but our aim here is predicting positions for as many users as possible. Another possible way to improve the accuracy of this simple model would be by clustering antennas, since we are considering each antenna as a different location. Although this seems to be a reasonable choice, real life situations do not adjust perfectly to this schema. While a user is at her house, she might be using more than one antenna, and we are considering that she is in two different places. On the other hand, a user might use the same antenna while she is in different locations, like her workplace and the nearby restaurant where she eats lunch. Some of these problems are pathological, and can not be tackled due to the poor space resolution given by antennas (as contrasted, for example, with GPS information). However, several problems can be solved by clustering antennas, where those clusters would represent real locations for users.

4. Urban Commute

The phenomenon of commuting is prevalent in large metropolitan areas (often provoking upsetting traffic jams and incidents), and naturally appears in mobile phone data. For instance, in [9] the authors study commute distances in Los Angeles and New York areas. Mobile data allows for the quantification of this phenomenon by providing indicators whose direct measurement is unfeasible. Figure 5 shows the call distribution for each antenna in the area of interest, averaged over a whole month. We include a series of call patterns illustrating the Buenos Aires commute in Figure 6.

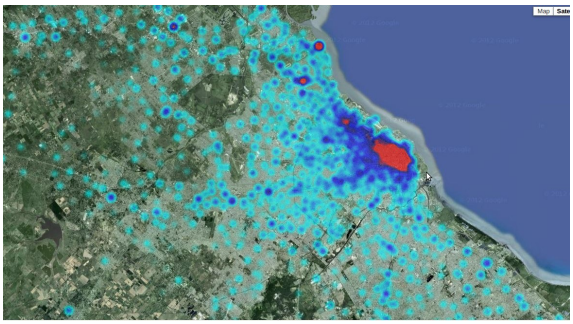


Fig. 5. Antenna call distribution in Buenos Aires city and its surroundings (the Greater Buenos Aires). Note that the color scale is different than the one used in Figure 6.

Red color corresponds to a higher number of calls, whereas blue corresponds to an intermediate number of calls and light blue to a smaller one.

From the data, we can estimate the radius of the commute (RoC - the average distance travelled by commuters). To do so, we take into consideration the two most frequently used antennas as the important places for each user (home and work, see [5]). We proceed first to define a night time, where users are usually at their houses (9 p.m. - 5 a.m. during weekdays) and a day time where users are usually at their workplace (12 p.m. - 4 p.m. during weekdays). We then count how many times each user makes or receives calls in each of those two time spans.

To simplify the procedure, we take a square area roughly corresponding to the country capital (the autonomous city of Buenos Aires), which is separated by a political boundary from the rest of the large metropolitan area (the Greater Buenos Aires). Around 3 million people live in Buenos Aires city, whereas around 13 million people live in the Greater Buenos Aires, which is among the top 20 largest agglomerations of the world by population.

A large part of the individuals working in Buenos Aires city live in the Greater Buenos Aires, and commute every day. Surveys and estimations state that more than 3 million people commute to Buenos Aires city every day. Thereby, we define a user as a *commuter* if she makes most of her night time calls from outside the city and most of her day calls from inside the city. To perform the experiment, we chose a threshold $\tau = 80\%$, meaning that at least τ of a user's night time calls must be made from outside the city and at least τ of day time calls must be made from inside the city in order to be considered a commuter.

After defining commuters, their home and work locations have to be found in order to compute the radius of commute. We define their home to be the antenna with the highest number of calls from outside the city during night time, and analogously define their work to be the antenna with highest number of calls inside the city during day time. We consider the distance between those two main antennas as the RoC for each user.

Having made the preceding definitions and assumptions, we compute an average RoC of 7.8 km (as a comparison, the diameter of the city is about 14 km, and the diameter of the considered metropolitan area is 90 km).

We also computed a random RoC assuming users' locations were randomly distributed in the region of interest and, once more, defining as commuters users that live outside the city but work inside. The result was a randomized average RoC of 32.9 km. This confirms an intuitive idea: people's living and working places tend to be closer than what a random distribution would predict.

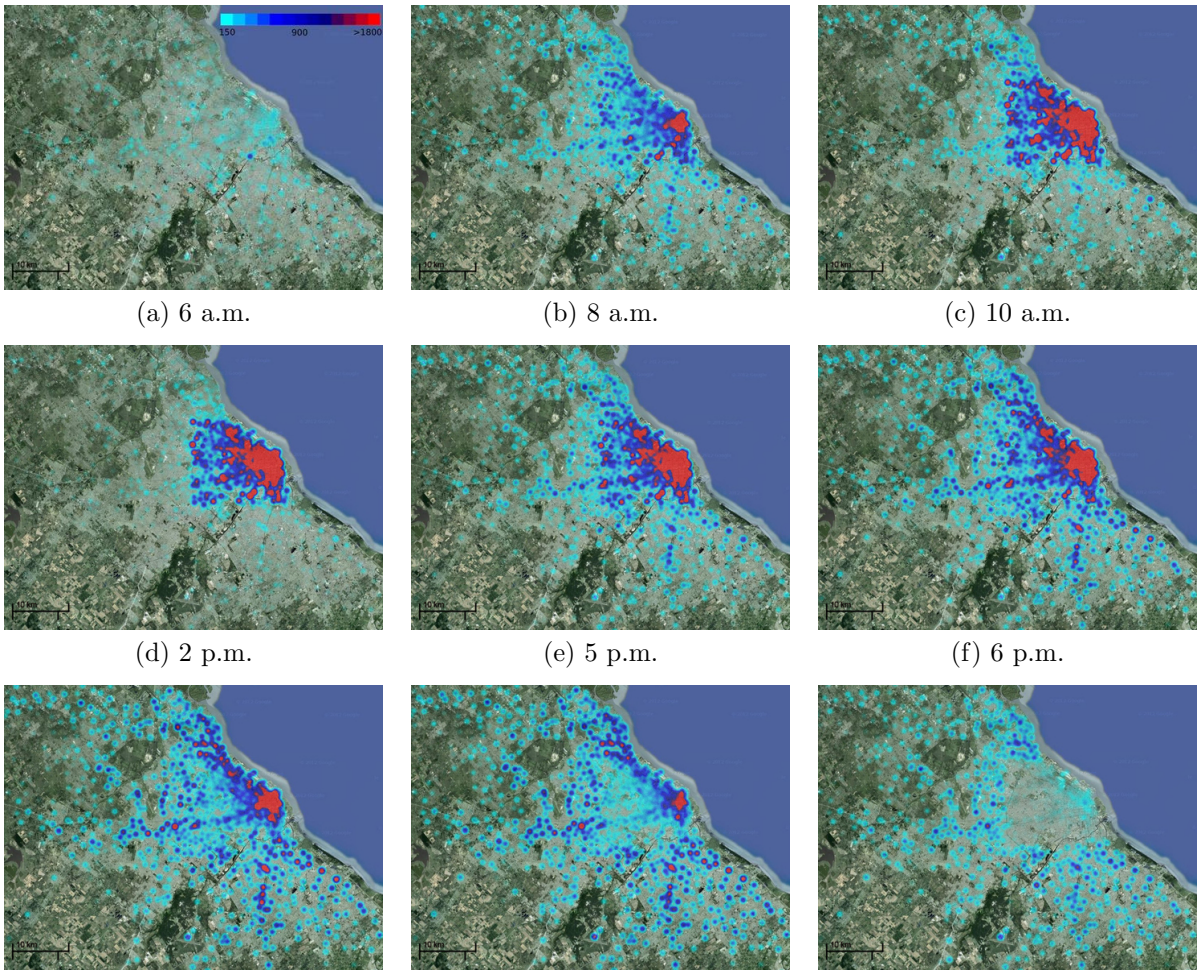
5. Sports Events

As in the urban commute case, we study human mobility in sports events as seen through mobile phone data. In Figure 7, we show how assistants to a Boca Juniors soccer match converge to the stadium in the hours prior to the game, and disperse afterwards. Average attendance to Boca Juniors home matches is 42000 people.

Note that postselecting the users attending the event necessarily produces the effect of having no calls outside the chosen area during the match. However, the convergence pattern observed is markedly different from the one seen for the same time slot of the week on a day with no match, as shown in Figure 8.

5.1. Improving Predictability with External Data

So far, our results illustrate how social events appear in the analysis of mobile phone data. This can be in turn used to improve the mobility model. Social relations among individuals have been used to improve predictability in mobility models before, as in [3], where social links learned from the



(a) 6 a.m. (b) 8 a.m. (c) 10 a.m.

(d) 2 p.m. (e) 5 p.m. (f) 6 p.m.

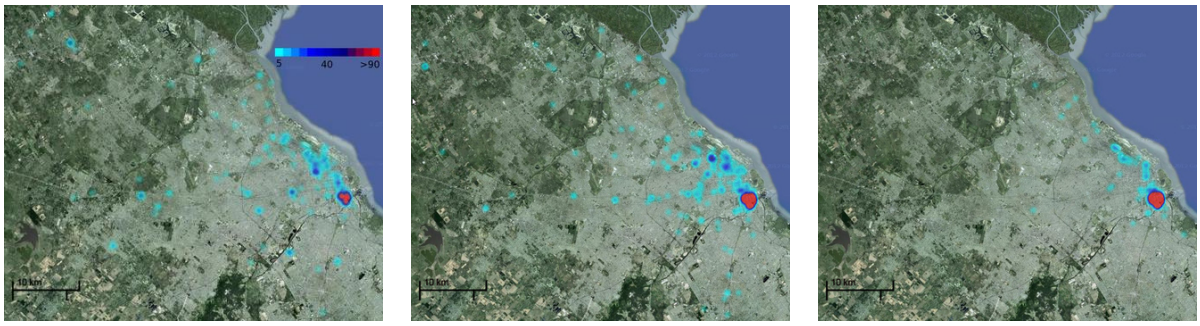
(g) 7 p.m. (h) 8 p.m. (i) 10 p.m.

Fig. 6. Commute to Buenos Aires city from the surrounding areas on a weekday, for different hours. The color scale can be seen in image (a) where numbers represent estimated number of people in a circle from the corresponding color. Image (g) – corresponding to 7 p.m. – clearly shows the major roads and highways connecting the city center to the North, West and South suburbs.

mobile phone records are used to this end. In this section, we show how an external data source can be used to improve the model.

We illustrate this effect using as proof of concept the case study of soccer matches. By taking the soccer fixture, we tag users as "Boca Juniors fans" if they make calls using antennas around the stadium and during the time slots of Boca matches for three selected consecutive matches (which include both home and away matches), which can be considered as part of our training set for the new approach. Using this tagging, we can dramatically improve predictability for this group of Boca fans, even predicting locations that had never been visited by a user before, 1000 km away from her usual location.

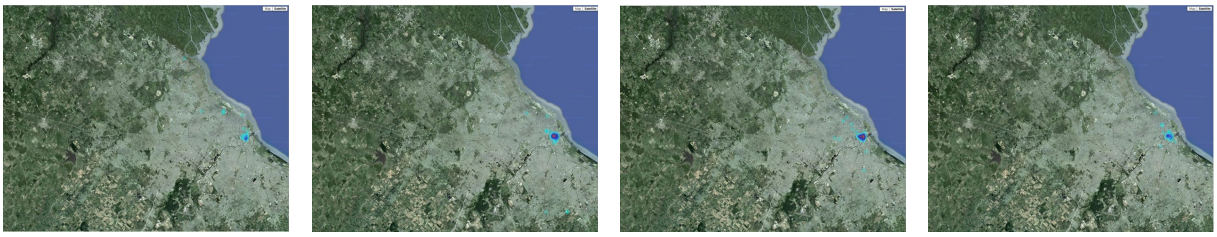
In the basic model, we predicted a user's location in a particular time slot to be her most frequent location in that particular time slot in the training set, whereas in this enriched model, we predict the stadium location (as a cluster of the antennas surrounding it) in case the user is a Boca Juniors fan and we are making predictions on a day where Boca plays a match on that stadium. To evaluate the basic model, we use 15 weeks of data for training purposes, as described in section 3. For the enriched model, we use the same training data, adding the previously mentioned social



(a) 5 hours before (b) 2 hours before (c) 1 hour before

(d) 1 hour after (e) 2 hours after (f) 3 hours after

Fig. 7. Convergence to Boca Juniors stadium on hours prior to a soccer match, and dispersal after its end. The color scale can be seen in image (a) where numbers represent estimated number of people in a circle from the corresponding color.



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

Fig. 8. Similar to figure 7 on a day with no match. The time references are relative to the time of the match in fig. 7 and are provided for ease of comparison.

information (i.e. tagging users) on three consecutive Boca matches in that period. The evaluation is made on the same testing data set in both cases, consisting on the three days where Boca plays the next matches.

The predictability of the model for these tagged users considering the fixture data rises for the days where there is a Boca Juniors match to 38% – which doubles the 19% accuracy achieved by our previous model for the same set. Moreover, the initial model is only able to make predictions in 63% of events in the given set (as a consequence of a lack of information from the training set data), whereas the socially enriched model tries to predict 100% of the events during match days, which make the previous results even more significant.

In order to understand these results, we illustrate with a few examples where the enriched model outperforms the simple model. On one hand, the simple model would never predict a user's location on a different city or in an unvisited location, whereas the enriched model would do so if the user is a Boca fan, and Boca has an away match in that city. On the other hand, if Boca usu-

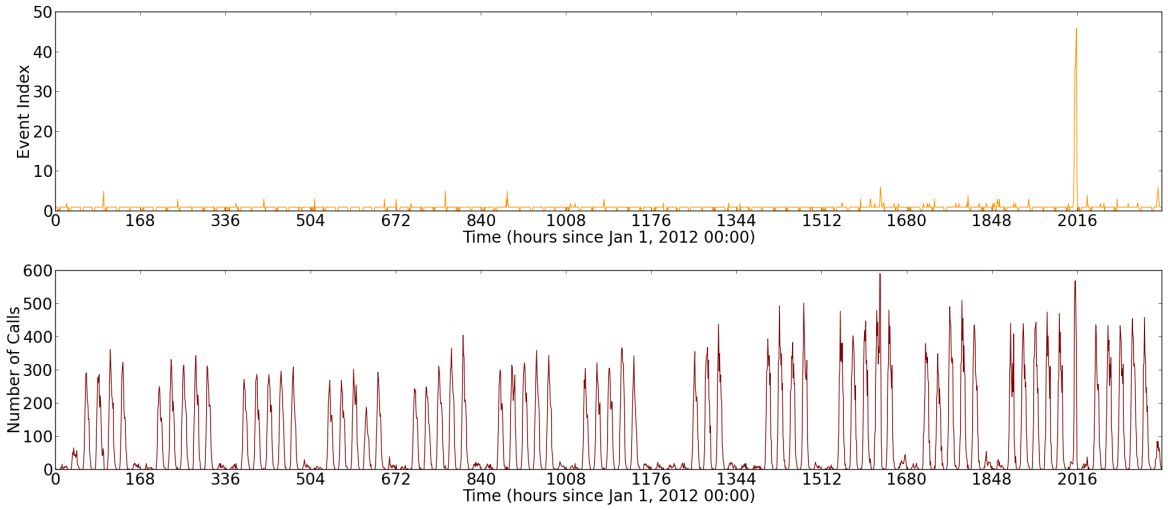


Fig. 9. Event index E and volume of calls registered in a cell tower near Plaza de Mayo for each time slot (hour) during a period of 3 months. A massive event was detected on March 24th, 2012 (a demonstration at Plaza de Mayo).

ally plays matches on Sundays at 7 p.m. a Boca fan could have a stadium antenna as its most frequent antenna for that particular time slot. Consequently, the simple model would predict her to be in that location for any other Sunday. However, the enriched model wouldn't do so for away matches, and can even take into account that season is over, and therefore predict the location of the following most frequent antenna. We note that although we have illustrated the improvement in predictability only in cases of massive events, there is no requirement for this in principle. As long as one has a satisfactory tagging of users, which could be achieved automatically by collating a big number of small events (like users who systematically go to local movie releases), or even manually, the improvement in location predictability would be warranted.

6. Automatic Event Detection

In the previous sections, we showed how social events show up in mobile phone data, and how information about these events can be used to improve the predictability in human mobility models. This information was exogenous in the sense that we had to consult sources other than the cell phone data in order to find out the location of the Boca stadium, or the time of the matches, for instance. In this section, we show how we can use the actual phone data in an endogenous manner in order to detect past events automatically (as [4,15] did with social media such as Twitter). The central idea here is that the events we are interested in involve a large amount of people, and hence change the usage patterns of antennas nearby, as evidenced in their traffic time series.

We consider the number of calls per hour registered in each antenna, according to the week w_i ($0 \leq i < n$, where n is the number of weeks in the studied dataset), the day of the week d_j ($0 \leq j \leq 6$) and the hour of the day h_k ($0 \leq k \leq 23$). For each time slot (w_i, d_j, h_k), we denote the number of calls as $C_{(i,j,k)}$ and we compute an *event index* $E_{(i,j,k)}$ given by the ratio

$$E_{(i,j,k)} = \frac{C_{(i,j,k)}}{\frac{1}{n-1}\sum_{0\leq \ell<n,\, \ell\neq i} C_{(\ell,j,k)}}. \qquad (1)$$

To understand the rationale of the event index, we should notice that the traffic of calls in two different antennas usually differ. Moreover, even for the same antenna, the traffic of calls can vary widely among the different days of the weeks, or hours of the day. Thus, a fair way of determining whether the traffic received by an antenna in a particular timeslot is extraordinary or not, is to compare with the traffic for the same antenna, in the same day and hour, for different weeks.

Figure 9 shows (a) the event index E and (b) the volume of calls registered in a cell tower near Plaza de Mayo (in Buenos Aires downtown) for each time

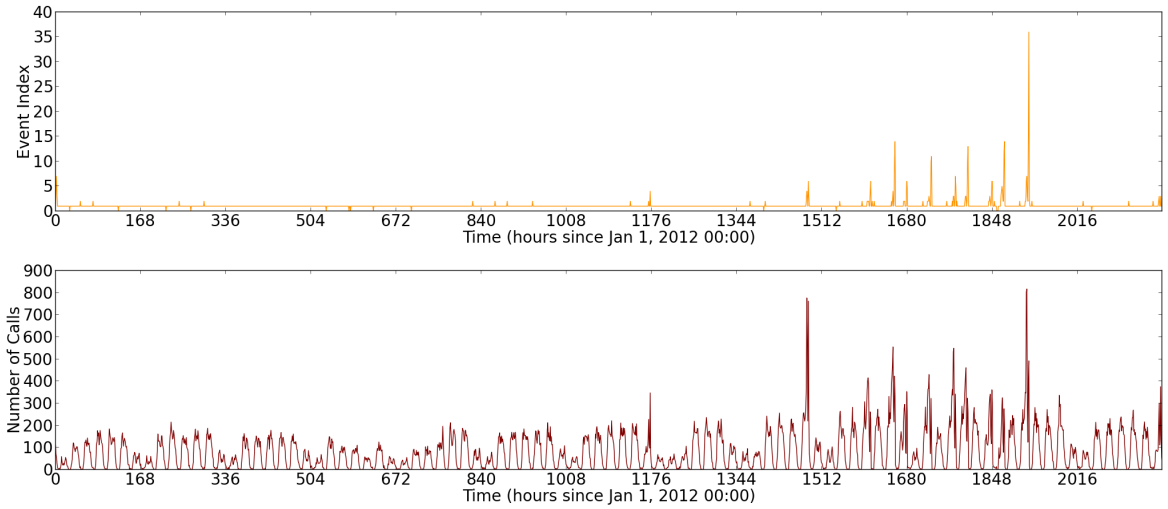


Fig. 10. Event index E and volume of calls registered in a cell tower near River Plate Stadium for each time slot (hour) during a period of 3 months. Several massive events were detected during March 2012 (the Roger Waters concerts).

slot (hour) during a period of 3 months (January 1 to March 31, 2012). The peak in E corresponds to a demonstration in remembrance of the beginning of the military regime that ruled the country for six years from 1976.

Figure 10 shows (a) E and (b) the volume of calls registered in a cell tower near the River Plate stadium, during the same period of 3 months. We can identify the series of Roger Waters concerts that took place during March 2012.

As was just illustrated, we can automatically detect events through peaks in the event index E as defined in equation (1), to which we can afterwards assign a meaning manually as in the two examples provided. At this point, we can proceed in the way we did with the Boca match, by tagging users according to their interests and improving their location predictability. The advantage in using an automatic procedure for event detection lies in that events most singular in terms of the E peak are those which involve larger usage fluctuations, and thus may allow for the tagging of larger amounts of users.

The autodetected events can also be used for a different purpose without going through the manual labeling stage. Given that users tend to assist events of the same kind (as was exemplified in the Boca case), we can tag users with similar patterns of participation in events as having common interests, and then when exogenous information about a user becomes available, we can propagate it to user with similar interests. We will illustrate the plausibility of this procedure by showing how events concentrate clusters of users in the network.

In Fig. 11 we make use of the network topology to plot the conditional probability of an individual being in an event as a function of the number of contacts of the individual present in the event. For each number of contacts $n_c \in [1, 17]$ we define this probability as the number of users present in the event that have n_c contacts also present in the event, divided by the total number of users having n_c contacts present in the event. We clearly see how the probability increases with the number of contacts, a clear manifestation of the 'herd behav-

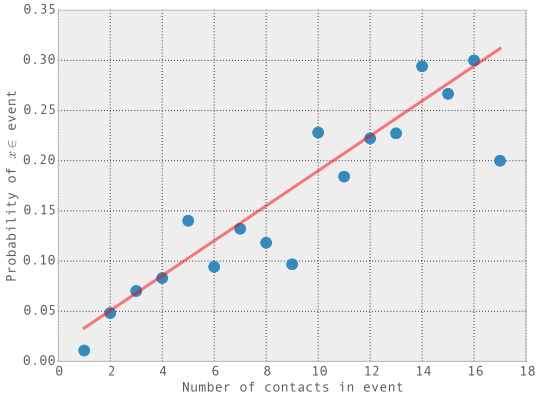


Fig. 11. Conditional probability of a user being in an event given that n_c of his/her contact were present in the event.

ior' in people's attendance to sociocultural events, which enables the propagation of tags as described above. Further exploration of this phenomenon is performed in [16].

7. Conclusions and Future Work

We illustrated how social phenomena can be studied through the lens of mobile phone data, which can be used to quantify different aspects of these phenomena with great practicality. Furthermore, we showed how including external information about these phenomena can improve the predictability of human mobility models.

Although we showed this in a specific case as a proof of concept experiment, we note that this procedure can be extended to other settings, not restricted to sports but including cultural events, vacation patterns and so on (see [12] for a specially relevant application). The tagging obtained is useful on its own and is of great value for mobile phone operators. The big challenge in this line of work is to manage to include external data sources in a systematic way.

In this respect, we showed how automatic event detection can be performed from mobile phone data and then manually tagged, and drafted a procedure for the automatic tagging of users' interests. By combining these techniques, we can build an expanding network of common interests which nourishes both from massive events (tagged or otherwise) and individual user tags arising from their phone usage pattern. Our challenge now is to fully develop the automatic user tagging procedure delineated in section 6 and set up the actual population-scale interest network.

References

[1] Thomas Aynaud and Jean-Loup Guillaume. Static community detection algorithms for evolving networks. In *WiOpt'10*, pages 513–519. IEEE, 2010.

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

[3] Eunjoon Cho, Seth A Myers, and Jure Leskovec. Friendship and mobility: user movement in location-based social networks. In *ACM SIGKDD*, pages 1082–1090. ACM, 2011.

[4] Freddy Chong Tat Chua and Sitaram Asur. Automatic summarization of events from social media. 2013.

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

[6] Manlio De Domenico, Antonio Lima, and Mirco Musolesi. Interdependence and predictability of human mobility and social interactions. *Nokia Mobile Data Challenge Workshop*, 2012.

[7] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3):75–174, 2010.

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

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

[10] Shan Jiang, Joseph Ferreira, and Marta C González. Clustering daily patterns of human activities in the city. *Data Mining and Knowledge Discovery*, pages 1–33, 2012.

[11] Jessica Leber. How wireless carriers are monetizing your movements. *MIT Technology Review*, 2013.

[12] Xin Lu, Linus Bengtsson, and Petter Holme. Predictability of population displacement after the 2010 haiti earthquake. *Proceedings of the National Academy of Sciences*, 109(29):11576–11581, 2012.

[13] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

[14] Tommy Nguyen and Boleslaw K Szymanski. Using location-based social networks to validate human mobility and relationships models. In *Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on*, pages 1215–1221. IEEE, 2012.

[15] Alan Ritter, Oren Etzioni, Sam Clark, et al. Open domain event extraction from twitter. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1104–1112. ACM, 2012.

[16] Carlos Sarraute, Jorge Brea, Javier Burroni, Klaus Wehmut, Artur Ziviani, and Ignacio Alvarez-Hamelin. Social events in a time-varying mobile phone graph. In *Fourth International Conference on the Analysis of Mobile Phone Datasets (NetMob)*, 2015.

[17] Carlos Sarraute and Gervasio Calderon. Evolution of communities with focus on stability. In *Third International Conference on the Analysis of Mobile Phone Datasets (NetMob)*, 2013.

[18] Chaoming Song, Tal Koren, Pu Wang, and Albert-László Barabási. Modelling the scaling properties of human mobility. *Nature Physics*, 6(10):818–823, 2010.

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

[20] Dashun Wang, Dino Pedreschi, Chaoming Song, Fosca Giannotti, and Albert-Laszlo Barabasi. Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1100–1108. ACM, 2011.

[21] Qinna Wang and Eric Fleury. Community detection with fuzzy community structure. In *Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on*, pages 575–580. IEEE, 2011.