Social Event Detection in Massive Mobile Phone Data Using Probabilistic Location Inference

Vincent Traag, Arnaud Browet, Francesco Calabrese, Frédéric Morlot

Social Event Detection in Massive Mobile Phone Data Using Probabilistic Location Inference

V.A. Traag
ICTEAM
Université catholique de Louvain
Louvain-la-Neuve, Belgium
vincent.traag@uclouvain.be

A. Browet
ICTEAM
Université catholique de Louvain
Louvain-la-Neuve, Belgium
arnaud.browet@uclouvain.be

F. Calabrese
IBM Research
Dublin, Ireland
fcalabre@ie.ibm.com

F. Morlot
Orange Labs
Issy-les-Moulineaux, France
frederic.morlot@orange-ftgroup.com

***Abstract*—The unprecedented amount of data from mobile phones creates new possibilities to analyze various aspects of human behavior. Over the last few years, much effort has been devoted to studying the mobility patterns of humans. In this paper we will focus on unusually large gatherings of people, i.e. unusual social events. We introduce the methodology of detecting such social events in massive mobile phone data, based on a Bayesian location inference framework. More specifically, we also develop a framework for deciding who is attending an event. We demonstrate the method on a few examples. Finally, we discuss some possible future approaches for event detection, and some possible analyses of the detected social events.**

I. Introduction

Over the last decade many new data sources have arisen that can be used in the social sciences, ranging from online social networks, such as Facebook or Twitter, to huge mobile phone data, promising a completely new approach in the social sciences [1], [2]. This unprecedented amount of data on social behavior can be, and has been used to study the behavior of human beings. Data from mobile phones have been used to analyze many dynamics [3] such as mobility behavior of people [4]–[6], uncovering highly regular work-office patterns [6], [7], communities in mobile phone networks [8], [9], the geography of calling behavior [10], showing a (gravitational like) effect of distance on the probability of a link [11], and the so-called strength of weak ties [12].

In this paper we will focus on detecting social events in a massive mobile phone data set. One can think of events such as rock concerts and sports finals, but also of events such as protests or emergencies. The idea here is to focus on the non-routine behavior of people, unlike earlier approaches [13], [14]. For example, in an office district, many people will be in a single place around the same time, but this does not constitute any social event.

Earlier analysis of events showed common geographical profiles for certain types of events, and suggested a proximity effect [13]. Such events will presumably have a significant impact on urban transit, so are important for urban planning [15], [16]. Another approach focused on anomalies to detect emergency events [14]. Moreover, people seem to behave differently at social events [17]. This highlights the importance of detecting social events, in order to analyze them.

II. Data

Because the methodology we will develop here is partly motivated by the type of data available, we will first briefly introduce them. The original data set we propose to analyze consists of all the calls of a large mobile phone company in a European country. For each call, we have an identifier (properly anonymized) for the person making the call (caller) and for the person receiving the call (callee). For both caller and callee we also have available the cell tower identifier at the time the call started. Coupled with the location of all the antennas of the company, we can infer some position of the users. We included both text messages and actual calls in our analysis. The relevant data cover 14 months for about 5.75 million users and around 900 million calls and text messages.

We perform a selection of the users based on their calling behavior. In order to be able to correctly identify users' locations, we need sufficiently regular connections to the network, which can be expressed in terms of the time between two calls [18]. We impose that 80% of the time, a user will be involved in a new call less than one day after his last one. Based on this selection, we keep around 55% of the users, while keeping around 87% of the total number of calls and text messages.

III. Location Inference

To be able to extract accurate and meaningful information from this raw data, we use a simplified probabilistic framework, based on the work of [19]. The most important reason for taking a probabilistic modeling approach is the somewhat erratic antenna jumps. It frequently happens that a user switches from neighbouring antennas while making several calls, although it is unrealistic the user is actually moving (because he would move too fast to be realistic). Furthermore, since we expect usage to be more intense than usual when events take place, multiple antennas will probably serve customers at the event location due to load balancing. Finally, our method can be seen as a smoothened Voronoi tessellation, thereby dealing automatically with these type of phenomena.

We denote by x the position of the user and by X_i the position of antenna i. We will denote the probability to be

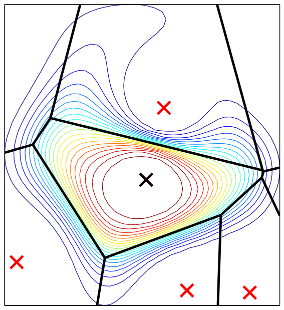


Fig. 1. Probability density $\psi_i(x)$ (represented by the level curves) for a particular antenna i (the central black 'X'), showing neighboring antennas (the red 'X's) and the local Voronoi tessellation (in dark lines). The density can be seen as a smoothened Voronoi tessellation, where there is also some (small) probability to be connected to antenna i when the user is in another Voronoi cell.

connected to antenna i given position x with

$$\begin{aligned}\phi_i(x) &= \Pr(a = i|x) \\ &= \int_0^\infty e^{-r} \prod_{j \neq i} \left(1 - \exp\left(-r \frac{\|x - X_j\|^\beta}{\|x - X_i\|^\beta}\right)\right) dr, \quad (1)\end{aligned}$$

where β is a parameter representing how quickly the signal decays [19]. As stated, this can be seen as a smooth approximation of the Voronoi tessellation, in which a user will always connect to the closest antenna. To see this, assume that the distance $d_i = \|x - X_i\| < d_j = \|x - X_j\|$ is smaller for i than for j. Then the fraction $d_i/d_j < 1$, and the probability $\phi_j(x) \to 0$ for $\beta \to \infty$. Hence, the probability to connect to any antenna j, while there is another antenna i closer, goes to 0 for $\beta \to \infty$, congruent with the Voronoi tessellation. An example of this probability density is shown in Fig. 1. Since we are interested in the probability $\psi_i(x)$ to be present at x upon connecting to antenna i, using Bayes' rule (without prior information), we obtain

$$\psi_i(x) = \frac{\phi_i(x)}{\int_\mathcal{D} \phi_i(x) dx}. \quad (2)$$

IV. Event Detection

In this section we will introduce the methodology to detect social events based on the calling patterns of users. We define social events as exceptionally large gatherings of people who are ordinarily not present at a specific location. There are a few key ingredients in this definition that need to be made clear: (1) presence at a location; (2) ordinary presence at a location; and (3) exceptionally large gatherings. We will define these concepts more clearly and formally in the following subsections.

We will first explain how we define the probability to be present at a certain location. Then we will define the ordinary probability of a user to be present at a certain location, and use both these probabilities to define a measure of attending a (possible) event. Finally, we will specify how we decide whether there is an event taking place or not at a certain location at a certain time.

A. Presence Probability

We will be looking for an event in region $\mathcal{A}$, at some starting time t_s and ending time t_e during a specific week w. Let us denote by $\mathcal{X}_\mathcal{A}$ those antennas who cover the region $\mathcal{A}$. Furthermore, let τ be the time window $[t_s, t_e]$ of the potential event and τ_v be the same time window during week $v = 1, \dots, W$. We can then select all calls that took place within the time window τ_w at antennas in $\mathcal{X}_\mathcal{A}$. Furthermore, a user u has made calls at antennas $i_1, \dots, i_{C_u} \in \mathcal{X}_\mathcal{A}$ at time $t_1, \dots, t_{C_u}$ in τ_w. We are interested in the probability $\Pr_p$ a user u was present at $\mathcal{A}$ during time τ_w.

For the exact time t_c, for a specific call c, the probability a user was in $\mathcal{A}$ is clear from the previous section, and we denote it by $\Psi_{\mathcal{A},c}(t_c) = \int_\mathcal{A} \psi_{i_c}(x)dx$. We now have to infer somehow the probability to be present at $\mathcal{A}$ at some time $t \neq t_c$. Keeping it simple, we assume a person leaves a particular location at a constant rate γ for $t > t_c$, without any probability of returning. Similarly, we assume a person to arrive at a constant rate γ for $t < t_c$. This constant rate assumption then yields

$$\Psi_{\mathcal{A},c}(t) = e^{-\gamma|t - t_c|} \int_\mathcal{A} \psi_{i_c}(x)dx. \quad (3)$$

We have chosen γ such that there remains only 1% of the original probability 15 minutes after and before the call c. Taking all calls into account, and normalizing by the theoretical maximum, it follows that

$$\Pr_p(u, \mathcal{A}, \tau_w) = \frac{1}{t_e - t_s} \frac{\int_{t_s}^{t_e} \max_c \Psi_{\mathcal{A},c}(t)dt}{\max_{i \in \mathcal{X}_\mathcal{A}} \psi_i(\mathcal{A})}. \quad (4)$$

B. Ordinary Probability and Probability of Attending

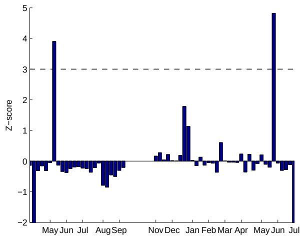
Based on the same idea we used to compute the presence probability $\Pr_p(u, \mathcal{A}, \tau_w)$, we derive the average probability a user u was in $\mathcal{A}$ within a certain time window τ *for all weeks different from* w. This probability will be called ordinary probability $\Pr_o$ to reflect the fact that it concerns ordinary behavior, i.e. regular mobility pattern independent of the event that may occur at week w. This can be defined as

$$\Pr_o(u, \mathcal{A}, \tau_w) = \frac{1}{W - 1} \sum_{\substack{v=1 \\ v \neq w}}^{W} \Pr_p(u, \mathcal{A}, \tau_v), \quad (5)$$

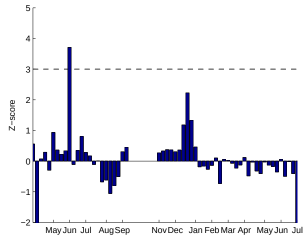
This probability captures how regularly this particular user was in the area of interest $\mathcal{A}$ during the time window specified by τ, at other weeks than w. We then define the probability the user was attending the event as

$$\Pr_a(u, \mathcal{A}, \tau_w) = \Pr_p(u, \mathcal{A}, \tau_w) \left(1 - \Pr_o(u, \mathcal{A}, \mathcal{T})\right), \quad (6)$$

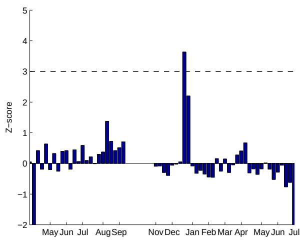
where higher values indicate a higher degree of certainty the user was attending an event on week w compared to its ordinary behavior during all other weeks.



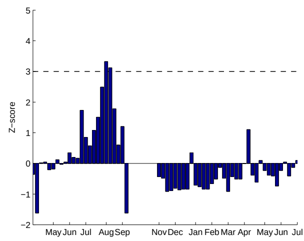
(a) Stadium



(b) Park



(c) Rural area



(d) Tourist destination

Fig. 2. Z-scores of number of attendees for four different areas of interest: (a) the national football stadium; (b) a large city park; (c) a rural area; and (d) a tourist destination. For the national football stadium, the two finals clearly stand out, and for the park the music festival stands out. The rural area shows only one event, namely Christmas evening, while the tourist destination shows two events during the summer.

C. Event Detection

In order to classify wether a user is attending an event or not, we select a cut-off value Pr_a^* such that only 1% of the users have such a high attendance probability on average over all weeks. We then say a user has attended a social event on week w whenever $\mathrm{Pr}_a(u, \mathcal{A}, \tau_w) > \mathrm{Pr}_a^*$.

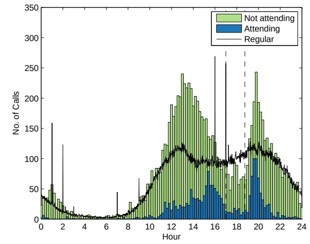
Let us denote by n_w the number of users that according to this decision rule have attended for a week w. Then, let μ denote the average number of attendees, and by σ the standard deviation of the number of attendees. We then state that an event has taken place, whenever

$$\frac{n_w - \mu}{\sigma} = z > 3, \tag{7}$$

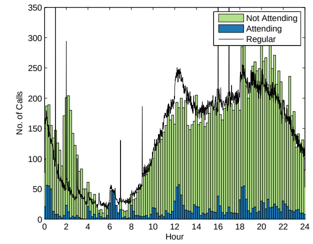
which is known as a z-score. Since n_w seems to be normally distributed, when one removes the outliers (which will most likely be our events), the above condition simply states that the probability to see so many possible attendees given the normal distribution of n_w is only about 1% with $z > 3$, hence they really represent unusually large gatherings of people. We will use this threshold of $z > 3$ in the remainder of the paper.

V. Results

We will now demonstrate the method on a number of different examples: (1) the national football stadium; (2) a city park; (3) a rural area; (4) a touristic area. For the first location we know what matches were being played, in particular the finals of the national football cups. We know that (at least) one big music festival took place in the large city park. We included a remote rural area, for which we expected to find



(a) Stadium



(b) Park

Fig. 3. Figure (a) shows the number of calls made by those who have attended the first final according to our method and those who haven't. At the time of the match (between 17-19h), the drop is most visible for those who attend, as expected. Figure (b) shows the number of calls for those attending and those not attending the music festival. The festival lasted several days, with music continuing late into the night (or early in the morning).

no events. The popular tourist destination is included to see how the method is affected by a strong seasonal trend.

The results of the z-scores for the football stadium are displayed in Fig. 2(a). The two peaks represent exactly the two finals played, and our method then seems to detect these two finals correctly. Both finals were played at a Sunday between 17–19h, and we used the data between 15–21h to detect the event. A more detailed analysis of the first final shows that the number of calls drop during the hours 17–19h, as shown in Fig. 3(a), but more stronger for those who are attending the event according to our method. The results for the second final are similar.

We can also compare the number of calls to the average on this weekday, which is represented by the thin dark line. The number of calls of the non-attendees is larger than usual just before and just after the match. So, this suggests that some people could be incorrectly classified as non-attendees.

The z-scores for the large city park are displayed in Fig. 2(b). We observe only one clear peak, which corresponds to the date of the music festival. We again consider the difference between the calls of those who attended (according to our method) and those who do not, as displayed in Fig. 3(b). The most striking feature is that the calls of attendees does not really seem to be increasing or decreasing throughout the day. More specifically, those who attend the music festival seem to be mainly the ones who call during the night. Given the nature of the festival, probably many young people continue to party into the night.

For the rural area we unexpectedly found one event. Upon closer examination, this specific day surprisingly turned out to be Christmas evening. Probably many family members gather, who would normally be elsewhere in the country. Indeed for the other locations this week also shows a somewhat higher z-score (although not very high). The tourist destination shows signs of two events during the summer, and it is clear what is the high season and what the low season. For the tourist destination it is quite normal there are relatively many people during the summer who are not there often, namely to spend their holidays.

These four examples suggest our method is capable of detecting events, although one should take care in interpreting the results. Looking into the calling dynamics for specific days of the event suggests that our classification of attending and not attending may work well, although it remains difficult to assess the performance exactly.

VI. Conclusion & Discussion

This method gives a probabilistic framework to detect events, and determine which users participated in the event. Based on a simple Bayesian location inference framework, we have suggested how we can indicate which users are likely to have attended the event, and when and where any events happened. We have demonstrated this method on a few examples, using limited data, namely only positions of the antennas. Still, it remains difficult to validate the method without additional information.

However, considering a simple Voronoi method, not using such a probabilistic framework, already seems to provide some indication whether there is an event or not. However, it can easily misinterpret which people are actually attending. Therefore, we might consider the following improvement. We first detect social events using a simple Voronoi method, but use the more refined method suggested here to decide which people actually participated in the event. So using the Voronoi approach we obtain a coarse-grained view of which events happened, while our method gives a more fine-grained view of who is attending, and could provide a more accurate estimate of the exact location. This would speed up the algorithm, making it more feasible to detect events across the whole country with reasonable accuracy.

Once we have detected a number of events, we can analyze them more closely. Since we are able to (re)trace the steps of attendees, we can analyze their mobility behavior and their calling behavior. People attending any events are showing, by definition, mobility behavior that is different from their routine. So the non-routine mobility behavior of people seems to be correlated with each other. Furthermore, we can analyze whether this correlation is different for people that call each other. It can be expected for example, that most people will not go to social events on their own, but rather meet with friends. It might also be possible to investigate any possible word-of-mouth effect. Finally, the methodology might be useful in situations of crowd management or emergency detection. One of the more interesting directions of research will be that although people are behaving differently from ordinary, they tend to do so together.

Acknowledgment

The authors acknowledge support from a grant “Actions de recherche concertées — Large Graphs and Networks” of the “Communauté Française de Belgique” and from the Belgian Network DYSCO (Dynamical Systems, Control, and Optimization), funded by the Interuniversity Attraction Poles Programme, initiated by the Belgian State, Science Policy Office. The authors would like to thank Paul van Dooren, Vincent Blondel and Gautier Krings for some helpful comments and discussion.

References

[1] D. J. Watts, “A twenty-first century science,” *Nature*, vol. 445, p. 489, Jan 2007.

[2] D. Lazer, A. S. Pentland, L. Adamic, S. Aral, A. L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, and M. V. Alstyne, “Life in the network: the coming age of computational social science,” *Science*, vol. 323, p. 721, Feb 2009.

[3] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, M. A. D. Menezes, K. Kaski, A.-L. Barabási, and J. Kertész, “Analysis of a large-scale weighted network of one-to-one human communication,” *New J Phys*, vol. 9, p. 179, Jun 2007.

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

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

[6] C. Song, Z. Qu, N. Blumm, and A.-L. Barabasi, “Limits of predictability in human mobility,” *Science*, Jan 2010.

[7] N. Eagle and A. Pentland, “Eigenbehaviors: identifying structure in routine,” *Behav Ecol Sociobiol*, vol. 63, pp. 1057–1066, May 2009.

[8] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, “Fast unfolding of communities in large networks,” *J Stat Mech-Theory E*, vol. 2008, pp. P10008+, Oct 2008.

[9] G. Palla, A.-L. Barabasi, and T. Vicsek, “Quantifying social group evolution,” *Nature*, vol. 446, pp. 664–667, Apr 2007.

[10] V. D. Blondel, G. Krings, and I. Thomas, “Regions and borders of mobile telephone in Belgium and in the Brussels metropolitan zone,” *Brussels Studies*, vol. 42, 2010.

[11] R. Lambiotte, V. D. Blondel, C. de Kerchove, E. Huens, C. Prieur, Z. Smoreda, and P. V. Dooren, “Geographical dispersal of mobile communication networks,” *Physica A*, vol. 387, no. 21, pp. 5317–5325, 2008.

[12] J. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A. Barabási, “Structure and tie strengths in mobile communication networks,” *P Natl Acad Sci USA*, vol. 104, no. 18, p. 7332, 2007.

[13] F. Calabrese, F. Pereira, G. D. Lorenzo, and L. Liu, “The geography of taste: Analyzing cell-phone mobility and social events,” *Pervasive Computing*, Jan 2010.

[14] J. Candia, M. González, P. Wang, T. Schoenharl, G. Madey, and A.-L. Barabasi, “Uncovering individual and collective human dynamics from mobile phone records,” *J Phys A-Math Gen*, vol. 41, p. 224015, Jun 2008.

[15] C. Ratti, S. Williams, D. Frenchman, and R. Pulselli, “Mobile landscapes: using location data from cell phones for urban analysis,” *Environ Plann B*, vol. 33, no. 5, p. 727, 2006.

[16] J. Reades, F. Calabrese, and C. Ratti, “Eigenplaces: analysing cities using the space- time structure of the mobile phone network,” *Environ Plann B*, vol. 36, no. 5, pp. 824–836, 2009.

[17] J. P. Bagrow, D. Wang, and A.-L. Barabsi, “Collective response of human populations to large-scale emergencies,” *PLoS ONE*, vol. 6, p. e17680, 03 2011.

[18] A.-L. Barabási, “The origin of bursts and heavy tails in human dynamics.,” *Nature*, vol. 435, pp. 207–11, May 2005.

[19] H. Zang, F. Baccelli, and J. Bolot, “Bayesian inference for localization in cellular networks,” in *Proc 29th Conf Info Comm*, pp. 1963–1971, IEEE Press, 2010.