Detecting and understanding big events in big cities

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

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

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

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

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

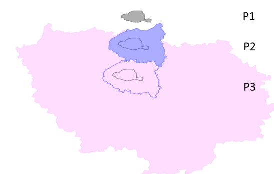

Figure 1: Administrative partitioning of Paris area: P1, P2, and P3.

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

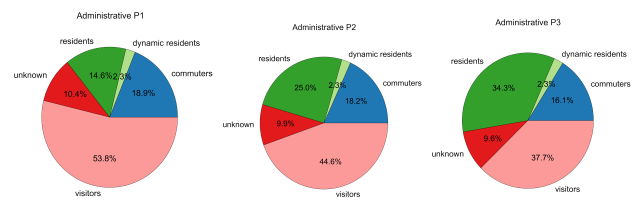

Figure 2: Variation of the composition of the population in the three Administrative areas of Paris.

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

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

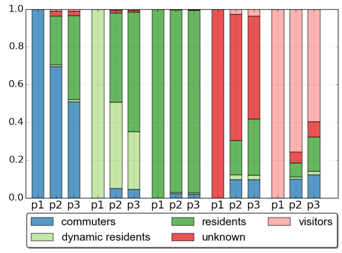

Figure 3: Multi-Classification of each category of users over the 3 areas.

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

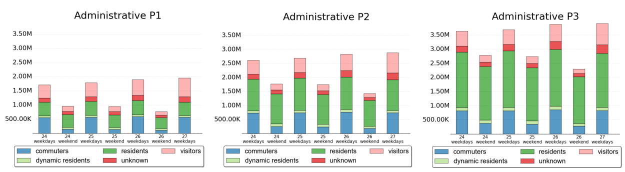

Figure 4: Weekly distribution of the presences in the three Administrative areas of Paris.

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

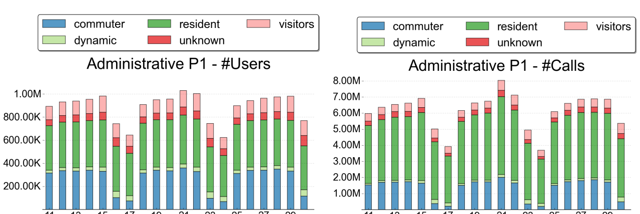

Figure 5: Daily distribution of presences and calls in P1 in the month of June.

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

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

Conclusions

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

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

References

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

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

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

habit models that underlie the sociometer, 2014.

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

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

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