This is a pre-copy-editing, author-produced PDF of an article accepted for publication in '11th International Conference on Computers in Urban Planning and Urban Management' following peer review.

TOWARDS ESTIMATING THE PRESENCE OF VISITORS FROM THE AGGREGATE MOBILE PHONE NETWORK ACTIVITY THEY GENERATE

Fabien GIRARDIN
SENSEable City Laboratory
Massachusetts Institute of Technology
77 Massachusetts Avenue, Cambridge,
MA 02139, U.S.A
Tel: 1-617-2537926
Fax: 1-617-2588081
E-mail: fabieng@mit.edu

Andrea VACCARI
SENSEable City Laboratory
Massachusetts Institute of Technology
77 Massachusetts Avenue, Cambridge,
MA 02139, U.S.A
Tel: 1-617-2537926
Fax: 1-617-2588081
E-mail: avaccari@mit.edu

Alexandre GERBER
AT&T Labs-Research
180 Park Avenue
Bldg 103, Room B133
Florham Park, NJ 07932 , USA
Tel: 1-973-3607086
Fax: 1-9733608871
E-mail: gerber@research.att.com

Assaf BIDERMAN
SENSEable City Laboratory
Massachusetts Institute of Technology
77 Massachusetts Avenue, Cambridge,
MA 02139, U.S.A
Tel: 1-617-2537926
Fax: 1-617-2588081
E-mail: abider@mit.edu

Carlo RATTI
SENSEable City Laboratory
Massachusetts Institute of Technology
77 Massachusetts Avenue, Cambridge,
MA 02139, U.S.A
Tel: 1-617-2537926
Fax: 1-617-2588081
E-mail: ratti@mit.edu

Abstract: This paper illustrates how aggregated mobile phone network activity logs provide anonymous information that reveals valuable insight into the presence of tourists visiting a city. Technologies supporting pervasive services like cellular networks have the potential to generate vast quantities of detailed subscriber data, raising important privacy concerns. But they also provide to researchers and city planners an unprecedented opportunity to understand the presence and movement of physical communities. We demonstrate how aggregated communication records offer the opportunity to note where people are transmitting and receiving information. The emergence of these spatial imprints obtained through novel technological means has a significant potential to support urban studies and particularly the optimization of tourism strategies, plans and marketing tools. In this study, we examine the use of locally and non-locally registered mobile phones in the vicinity of the "Waterfalls" public exhibit in New York City in 2008. We study aggregated statistics (i.e. number of calls) related to the network sectors covering the exhibit and its proximity. With the future contribution of traditional survey techniques, such as field counts, to calibrate these mobile phone network measurements, we aim to develop techniques to estimate the aggregate movements and location of visitors through time and space, while assuring their privacy.

Keywords: urban dynamics, location-based services, crowd estimation, urban attractiveness

This is a pre-copy-editing, author-produced PDF of an article accepted for publication in '11th International Conference on Computers in Urban Planning and Urban Management' following peer review.

1. INTRODUCTION

Traditional methods, such as manual counts or personal surveys to identify the presence of visitors and tourists in a city are often expensive and result in limited empirical data. Today, thanks to the emergence of ubiquitous digital technologies, new data sources are available. Information produced by the interaction with wireless and online services has helped to create and define new methods of observing, recording, and analyzing activity logs, and therefore human dynamics, in the context of the city (O'Neill et al., 2006). These devices create, in essence, an opportunity to examine geographically-aware imprints from a digital sensor web (Goodchild, 2007), that may reveal collective social behaviors with unprecedented detail (Eagle and Pentland, 2006). These data present an opportunity in tourism statistics to build more efficient ways of collecting aggregate information of visitors' activities. Indeed, tourists have many ways of leaving electronic imprints: prior to their visits they generate server log entries when they consult digital maps (Fisher, 2007) or travel web sites (Wöber, 2007); during their visit they leave imprints on wireless networks (Ahas et al., 2007) whenever they use their mobile phones or the motorway when they use their credit cards to pay the tolls (Houée and Barbier, 2008); and after their visit they may publish online reviews and photos (Girardin et al., 2008). Additional data-mining operations can complement the statistics on collective accommodation, customized services for citizens and visitors, allow accurate timing of service provision that can be based on demand, and more synchronous management of service infrastructure. But the social advantages of these applications are in conflict with important privacy concerns. Researchers and developers in this area must take conscientious, principled, and evident care to protect the subscriber's, resident's and visitor's privacy. In this paper, we consider highly aggregated, non-personally identifiable digital imprint records generated by mobile phones using the AT&T wireless network to make or receive calls.

Previous research has shown that the wide diffusion of mobile phones and the widespread coverage of mobile phone wireless networks in urban areas have made these technologies efficient tools to study both groups (Ratti et al., 2006) and individuals (González et al., 2008). For example, the analysis of mobile data for vehicle traffic analysis (see Yim (2003) for a review) has led to the awareness of traffic conditions in real-time. Some other efforts correlated with limited success cellular network signals with the actual presence of vehicles and pedestrians in the city (Sevtsuk and Ratti, 2007). In a case study of tourism dynamics in Estonia, Ahas et al. (2007) proved that the sampling and analysis of passive mobile positioning data is a promising source for tourism research and management. They showed that this type of aggregated data is highly correlated with accommodation statistics in urban touristic areas. Overall, the advantage of mobile phone network activity logs over traditional tourism statistics is its superior spatial and temporal precision and breadth.

While that work highlighted the tourism dynamics of an entire country (Estonia) with general-purpose sources (for instance to obtain more detailed geographical information such as urban versus rural tourism), our study takes place at the scale of a city neighborhood. In this paper, we report on an ongoing study that aims at estimating the presence of tourists. In a case study, we use aggregated and anonymized records of mobile phone network activity to develop a technique to describe and quantify the attendance of the Waterfalls exhibition in New York City in 2008. In the remainder of the paper, we present the types of data collected, the methodology to analyze them, and early results to assess the promises of our approach. Finally we conclude with a discussion on the limitations and future works.

This is a pre-copy-editing, author-produced PDF of an article accepted for publication in '11th International Conference on Computers in Urban Planning and Urban Management' following peer review.

2. CONTEXT

The New York City Waterfalls was a public art project of four man-made waterfalls rising from New York Harbor between June 26 and October 13, 2008. This large investment required the assessment of the economical impact of the event. While traditional methods (e.g. manual people count and surveys) provide a quantification of visitors at a specific area on a limited time, we hypothesized that digital imprints can help estimate the presence of people. Therefore, we decided to investigate the relationship between mobile phone network activity and empirical manual counts performed on site at the different vantage areas of the exhibit. A first step was to characterize over space and time the data generated by cell phone network activities, and to compare those at the official vantage areas for the various waterfalls with activity at the main tourist attractions in the vicinity, such as the World Trade Center site, Wall Street, City Hall and the Brooklyn Bridge (Figure 1).

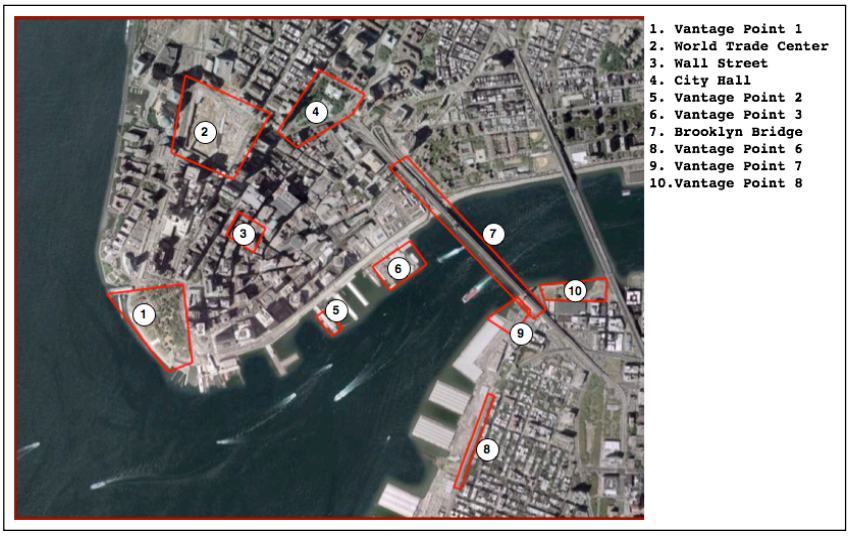

Figure 1. Defined in red, the main vantage areas and attractions in proximity to the New York Waterfalls exhibit

3. DATA

Our collaboration with the mobile operator AT&T granted us access to anonymized hourly aggregated records of network activity generated by mobile phone users making and receiving phone calls through the AT&T cellular network in lower Manhattan and the Brooklyn waterfront areas from August 2007 to August 2008. To ensure the complete privacy of AT&T's mobile users, these data were provided in accordance with AT&T's privacy policies and our approach of collecting and analyzing aggregate network usage data fully complies with the 2002 Directive of the European Parliament and Council on privacy (Poullet, 2006). The use of aggregated statistics does not present traces of the individual, like their identity or trajectory: indeed, the study only estimates the number of mobile phone digital imprints in a given area of the city at a given time, thus avoiding privacy issues raised by other

methodologies (González et al., 2008). The section of the network under study is made of Base Transceiver Stations (BTS) covering the lower part of Manhattan and the west part of Brooklyn. A BTS represents the elementary unit of the infrastructure that is used to wirelessly connect users' devices to the mobile phone network and is dedicated to provide connectivity to a specific geographic region, called a sector. Each observation is constituted by the number of calls and text messages originated and terminated within each sector, Table 1 provides a detailed description of the meaning of each data type.

Table 1. Description of the datasets of mobile phone network activity

Data type	Description
Aggregated calls	Number and duration of calls originated and terminated in the area
Aggregated Call Detail Records (CDR)	A CDR is the record produced by a telephone exchange. These aggregated records provide information about each call or text, such as its BTS, hour, home location and direction. The home location of the mobile phone user is determined based on the area code of the handset's mobile telephone number for U.S.-registered phones and the country code associated with the mobile telephone number for foreign-registered phones.[1]

It should be noted that the data can be biased and contain spatial noise. In particular along the river, some network activities on one waterfront are handled by the infrastructure on the other side of the river thus affecting the ability to capture fine-grained information of the location of the mobile phone users[2]. Moreover, several factors directly affect the capacity of estimating the network activity at specific areas such as the areas of interests we defined. For instance, the dimension of a wireless sector can greatly vary, depending on the built environment it has to cover, and sectors can partially overlap and serve similar areas.

4. METHODOLOGY

The methodology that we propose starts with a preprocessing of the data to overcome the inherent problems of resolution and reliability of the mobile phone network statistics. It then implements different approaches to the analysis of the data that allow the extraction of different temporal and spatial patterns of the network

[1] Note that because many U.S. mobile phone customers may have handset telephone numbers that do not reflect the state where they currently reside, or foreign visitors may acquire domestic U.S. SIM cards and mobile telephone numbers for the duration of their stay in the U.S., there will be some inaccuracy in our inferences about the actual home locations of mobile phone users.

[2] Indeed, because this study identifies user locations only to the BTS that serves the area, it is intrinsically less precise than a study that may make use of GPS information available from certain mobile handsets.

activity at each of the areas of interest we aim at estimating. We believe that this case study may eventually lead to inferences about the actual presence of tourists based on network activity.

4.1 Radio map

A wireless coverage is divided in sectors. A base transceiver station (BTS) controls each of these overlapping geographical area. Their coverage depends on the location of the BTS that propagates the signal of the network to the mobile phones. A standard solution calculates a Voronoi diagram to define a unique section covered by the best serving BTS. However this approach does not respect the rather small areas represented by the areas of interest. Therefore, we use a propagation model characterized by the location, height, azimuth, Effective isotropic radiated power (EIRP), type of antenna, and frequency served of the BTS and some conditions of the physical environment (e.g. presence of a river; the presence of skyscrapers is not taken into account) to generate a radio map of the study area formed by around 3000 non-overlapping partitions. For each element of the partition, we computed its statistics by summing the weighted contributions of all the sectors overlapping the specific element of the partition. As a result, each partition reports on an estimate of the network activity generated in the area that it covers.

4.2 Network activity at specific areas

The generated partition statistics of the radio map allow us to estimate network activity within the boundaries of the defined tourist attractions. These areas of interest can be as large as the Financial District in Lower Manhattan or as small as Pier 1, a pier renovated as vantage point for the Waterfalls exhibit. To do so, we calculated the relative weights of each sector contributing to each element of the partition, and the relative weights of each partition contributing to each area of interest. Then we recomputed the data into partition-based statistics and then area-centric statistics (Figure 2).

$$Calls_{POI_i} = \sum \left(\frac{Area_{POI_i}}{Area_{PART_k}} \cdot Calls_{PART_k} \right)$$

for each partition element k overlapping POI i.

$$Calls_{PART_i} = \sum \left(\frac{Area_{PART_i}}{Area_{SECTOR_k}} \cdot Calls_{SECTOR_k} \right)$$

for each sector k overlapping partition element i.

Figure 2. Equations summarize the computation performed to estimate the number of calls per area of interest.

Following this procedure, we were able to improve the spatial resolution of the data at the potential cost of a reduced reliability of the statistics caused by too simplistic assumptions such as assuming that statistics apply uniformly over the sectors. Nevertheless, it allows us to have an understanding of visitor density in specific parts of the exhibit and its vicinity. The remainder of the paper discusses the density in space (e.g. spatial distribution of visitors) and time (e.g. patterns and seasonality) and similarities of vantage points and areas of interests in the vicinity of the Waterfalls exhibit.

This is a pre-copy-editing, author-produced PDF of an article accepted for publication in '11th International Conference on Computers in Urban Planning and Urban Management' following peer review.

4.3 Spatial distribution of locals and visitors

For the week of August 10 to 17, 2008 AT&T provided hourly aggregates per area of interest indicating the number of phone calls and text messages per mobile phone registration location category (foreign country or US state). From this data, we were able to count how much activity involved mobile phones registered in New York; how much activity involved mobile phones registered in the United States, but outside of New York; and how much activity involved mobile phones registered outside the United States. While it may not be always true, it is reasonable to assume that locals generate the majority of calls from mobile phone registered in New York, and that visitors generate the majority of calls that involve mobile phones registered outside New York or outside of the US. With this assumption, we were able to map the overall presence of locals and visitors on an average weekday and weekend (Figure 3). It becomes clear that visitors enjoy the lower Manhattan waterfront, particularly on the weekend, and reveal very little presence in Brooklyn. On the other hand locals seem to enjoy the Brooklyn waterfront on the weekend.

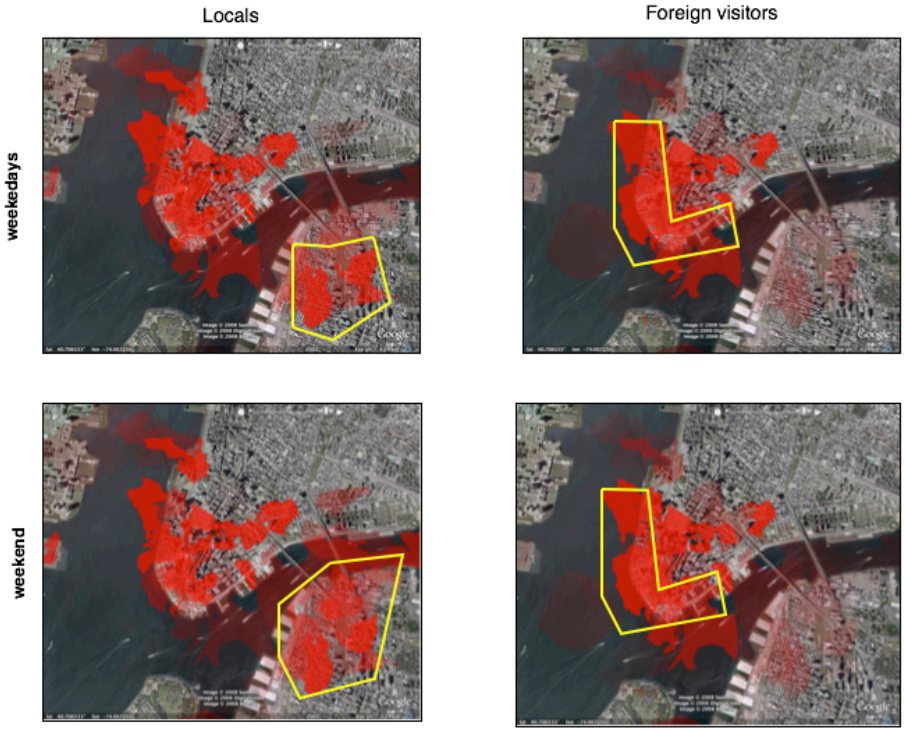

Figure 3. Spatial distribution of locals (New Yorkers) and foreign visitors in the neighborhood of the Waterfall. New Yorkers generate network traffic activity in the financial district and in Brooklyn, a neighborhood not attracting many foreigners compared to the waterfront of lower Manhattan.

This is a pre-copy-editing, author-produced PDF of an article accepted for publication in '11th International Conference on Computers in Urban Planning and Urban Management' following peer review.

Analysis of the ratio of locals' calls versus visitors' calls at each area of interest can provide an estimate of their relative density. Figure 4 shows for each area of interest the relative call quantities from the three user groups (locals, US visitors and foreign visitors) during the workdays and the weekends. The first set of histograms show that weekends are characterized by a general reduction of call activity from nearly all user groups. The bottom row of bar charts shows the percentage variation in the fraction of total activity contributed by each user group, and generally shows that the fraction of activity attributable to visitors relative to locals rises on the weekends.

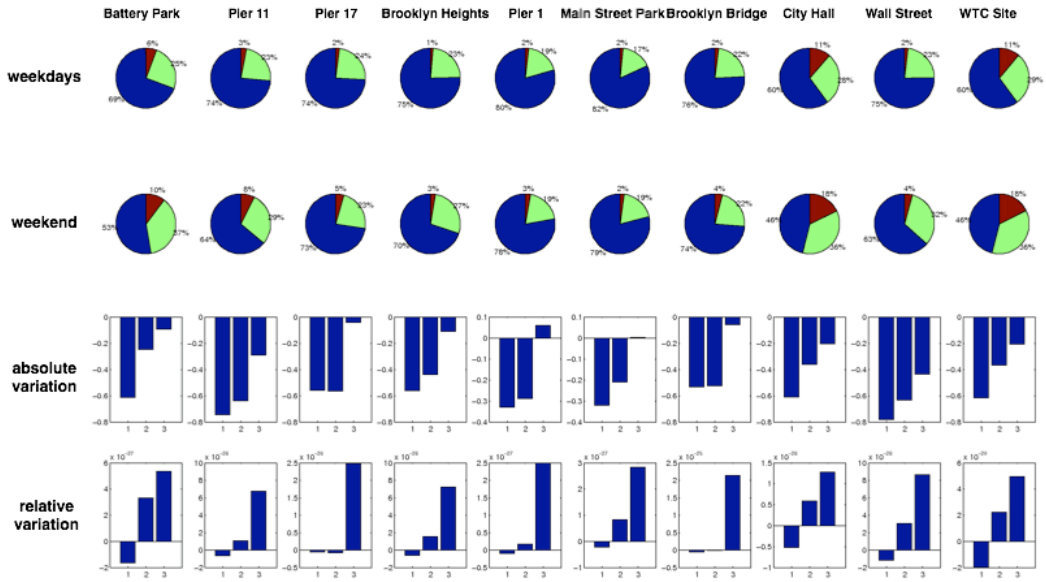

Figure 4. CRS and percentage variations of locals' (blue,1), US visitors' (green,2) and foreign visitors' calls (red,3) at the vantage points and main areas of interests between week days and the weekend.

These statistics show an average growth of foreign visitors of 88% across all the vantage points between weekdays and the weekend, while the growth over the other areas of interests is 77%. However, the average relative number of US visitors' calls over the vantage points shows a growth of only 15%, versus a growth of the other areas of interest of 31%. This means that foreign visitors tend to generate more activity at the vantage points rather than the other areas of interest. However, US visitors show a reverse dynamic.

4.4 Comparing the network activity of the areas of interest

While the one-week snapshot of aggregate CDR data makes it possible to study the spatial presence of locals and visitors, it doesn't provide insights on their temporal patterns and seasonality. To understand these major trends, we plotted stacked bar charts that display the daily patterns of total activity for each area of interest between August 2007 and August 2008. For each data set, we generated two versions of the charts: one which represents the absolute values of total activity, useful to understand both the behavior of each area of interest and the overall trend, and one which represents the relative total activity of each area of interest with respect to the others, useful to compare variations in behavior between areas of interest.

This is a pre-copy-editing, author-produced PDF of an article accepted for publication in '11th International Conference on Computers in Urban Planning and Urban Management' following peer review.

Daily patterns

Figure 5 shows the daily density of phone call activity, i.e. the average number of phone calls originated or terminated in each area of interest in one day, per unit of surface. A first observation about the data is that there is a year-long positive trend in the overall phone calls. Most of the growth corresponds to an increased activity in Vantage Point 6 and Vantage Point 8 that almost doubled their activity. Moreover, it is possible to note a strong weekly seasonality, which causes the phone call activity over the weekends to drop (with slight variations among the different areas of interest) to about half of the activity of the workdays. Beside the normal spikes that correspond to national festivities (i.e. Thanksgiving, Christmas, New Year's Eve, Easter, and the 4th of July), the noise in the data does not allow us to detect changes in the phone call activity of the vantage points during and after the opening of the waterfalls. This suggests that a finer radio map should be generated to extract the signals of a very specific area over a short period of time.

However, the absolute density reveals a clear weekly seasonality; in particular, while some areas of interest tend to have less activity over the weekends, others tend to increase their activity, a divergence that can help to understand the kind of traffic generated in the areas of interest (e.g. work-related versus leisure-related).

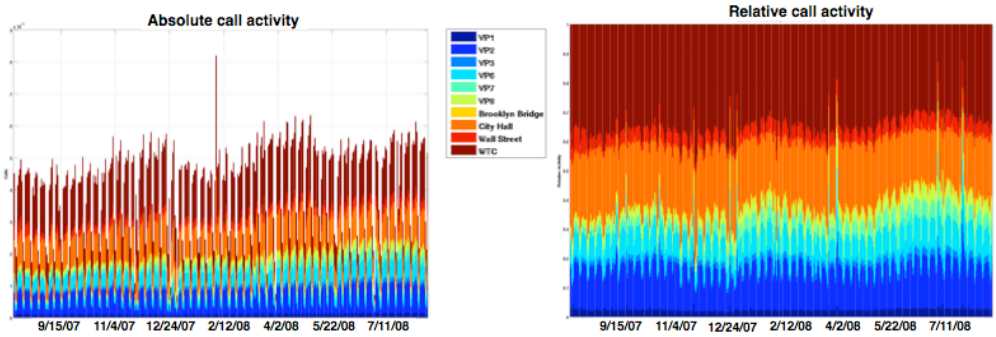

Figure 5. Daily absolute (left) and relative (right) density of phone calls per area of interest. Note: y values have been multiplied by a constant factor.

Seasonal patterns

The data set also indicates that there are differences in the seasonal patterns of the areas of interest. Figure 6 shows the average phone call activity throughout the day (per slots of three hours each, from midnight to 3 am, from 3 am to 6 am, and so forth): all the areas of interest present the same behavior at night and during the mid-afternoon, while they have different behaviors during the evening. The average phone call activity per day throughout the week shows a similar behavior during the workdays among the areas of interest, while they vary during the weekends when some of them drop to an activity that is about 30% of the workdays, and others maintain an activity of about 60%. Finally the average activity per month throughout the year presents an overall higher variability of behaviors, but they still tend to have clearly different activities from August to November.

This is a pre-copy-editing, author-produced PDF of an article accepted for publication in '11th International Conference on Computers in Urban Planning and Urban Management' following peer review.

Co

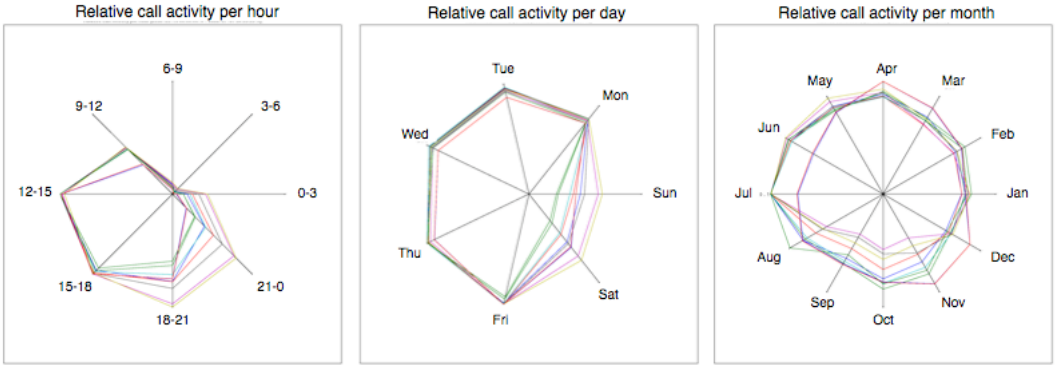

Figure 6. Seasonal patterns of the phone activity between areas of interest: per slots of three hours throughout the day on the left, per day throughout the week in the center, per month throughout the year on the right (period August 2007 to August 2008).

The seasonal patterns show clearly that there are periods of the day (in particular late evening) when the areas of interest exhibit different behaviors. It is possible to gain more insight into these differences by plotting a matrix of the scatter plots of the density of activity for each pair of areas of interest, where each point represents the daily call activity of a Saturday or a Sunday (Figure 7). Analyzing the scatter plots we can identify areas of interest that present similar trends during the day, which are locations that have either strong or weak activity concurrently. Identifying the pairs that have a correlation coefficient higher than 0.8 we highlight that the vantage points that cluster together, and don't cluster with the other locations.

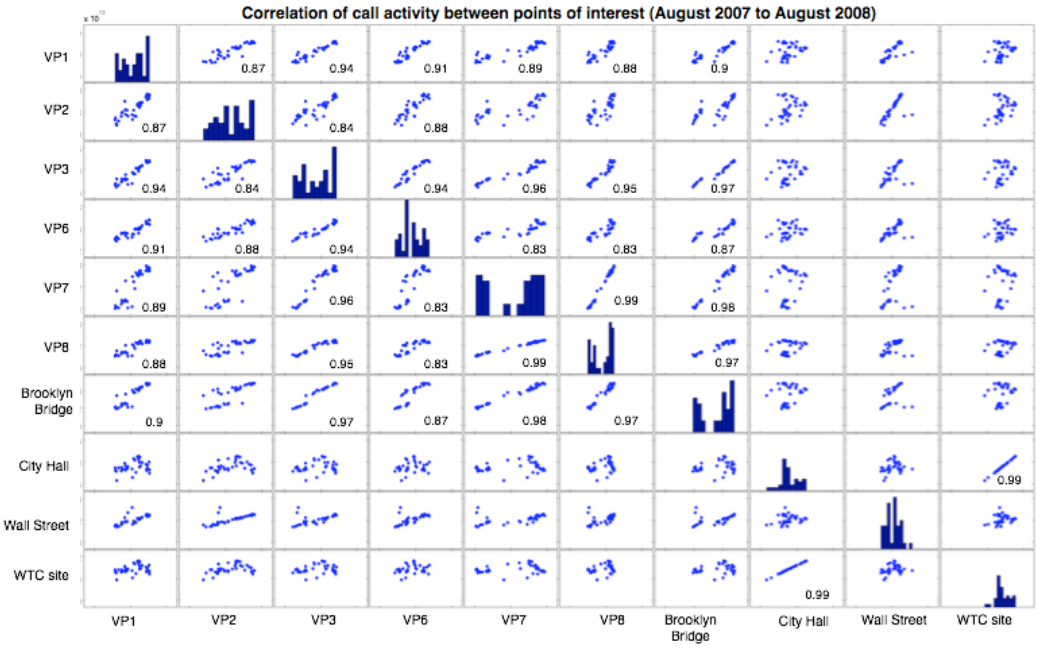

Figure 7. Correlation of the daily call activity between the areas of interest. Correlation coefficient higher than 0.8 are displayed.

5. CONCLUSION AND FUTURE WORKS

The worldwide adoption of mobile phones provides an unprecedented opportunity to study the aggregate patterns of urban dwellers. Complementing traditional surveys (e.g. gate counts, accommodation surveys) and other digital footprints (i.e. credit cards transactions, subway turnstiles) that register usage patterns in fixed locations, mobile phones travel along with people as they move from place to place, thus offering the potential to describe more richly how places are used, and yet avoiding the deployment of ad hoc and costly measurement infrastructures and resources. These new types of digital imprints, combined with traditional methods, provide a comprehensive insight into the tourism dynamics in a city or neighborhood.

The generation of a fine grain radio map of the neighborhood of the New York City Waterfalls public art exhibit enables the analysis of the aggregate cell phone network activity at vantage points and areas of interests. CDR data reveals the spatial distribution of non-personally identifiable local users and visitors in those places over a given time. When collected over time, such as in this case study, temporal and seasonal patterns can be extracted from these aggregated records of initiated and received mobile phone call and text messages. The correlation of these network activities between the different areas of interest suggests which areas manifest similar work patterns or touristic interests. The ability to understand the evolution of the distribution of visitors in specific places and the temporal aspects of the network activity at those places offers the possibility of more precise estimation of the popular success of different city events, exhibitions and areas of interest. Our future work will aim at comparing the spatiotemporal data derived from mobile phone network activity logs with ground truth data from extensive head counts and surveys performed in the field. Based on the characteristics of the data set, we expect the future challenges to involve having a continuous and consistent ground truth data and identifying areas of interest that match the dimension of areas covered by the radio map partition. More specifically, partitions should not be too small because their estimated network activity would be poisoned by the predominant activity of bigger neighboring sites, and neither too big in order to avoid degenerating into areas that are served by too many stations. Together with dimensions, other physical characteristics affect the quality and reliability of our analysis: the distribution and quality of the network stations, and the presence of obstacles such high-rise buildings, are also important factor to keep into consideration. In consequence, our technique to estimate the radio maps of the area need to go beyond the assumption that statistics apply uniformly over the sector. One solution could come from the measurements of network signals on site to refine the signal degradation of each BTS and the boundaries of their partitions.

REFERENCES

O'Neill, E., Kostakos, V., Kindberg, T., gen. Schieck, A. F., Penn, A., Fraser, D. S., and Jones, T. (2006). Instrumenting the city: Developing methods for observing and understanding the digital cityscape. **In Ubicomp**, pages 315–332.

Goodchild, M. F. (2007). Citizens as voluntary sensors: Spatial data infrastructure in the world of web 2.0. **International Journal of Spatial Data Infrastructures Research**, 2:24–32.

Eagle, N. and Pentland, A. S. (2006). Reality mining: sensing complex social

systems. **Personal Ubiquitous Computing**, 10(4):255–268.

Fisher, D. (2007). Hotmap: Looking at geographic attention. **IEEE Trans. Vis. Comput. Graph.**, 13(6):1184–1191

Wöber, K. (2007). Similarities in information search of city break travelers — a web usage mining exercise. **In Information and Communication Technologies in Tourism 2007**, pages 77–86.

Ahas, R., Aasa, A., Silm, S., and Tiru, M. (2007). Mobile positioning data in tourism studies and monitoring: Case study in tartu, estonia. In Marianna Sigala, L. M. and Murphy, J., editors, **Information and Communication Technologies in Tourism 2007**, Ljubljana, Slovenia. Springer Vienna.

Houée, M. and Barbier, C. (2008). Estimating foreign visitors flows from motorways toll management system. **In 9th International Forum on Tourism Statistics**.

Girardin, F., Calabrese, F., Fiore, F. D., Ratti, C., and Blat, J. (2008). Digital footprinting: Uncovering tourists with user-generated content. **IEEE Pervasive Computing**, 7(4):36–43..

Ratti, C., Pulselli, R. M., Williams, S., and Frenchman, D. (2006). Mobile landscapes: Using location data from cell-phones for urban analysis. **Environment and Planning B: Planning and Design**, 33(5):727 – 748.

González, M. C., Hidalgo, C. A., and Barabási, A.-L. (2008). Understanding individual human mobility patterns. **Nature**, (453):779–782.

Yim, Y. (2003). The state of cellular probes. Technical report, California Partners for Advanced Transit and Highways (PATH)

Sevtsuk, A. and Ratti, C. (2007). Mobile surveys. **In Urbanism of Track 2007**. Delft University Press.

Poullet, Y. (2006). Eu data protection policy. the directive 95/46/ec: Ten years after. **Computer Law & Security Report**, 22(3):206–217.

This is a pre-copy-editing, author-produced PDF of an article accepted for publication in '11th International Conference on Computers in Urban Planning and Urban Management' following peer review.

First author biography

Fabien Girardin is a Ph.D. candidate in Computer Science and Digital Communication in the Interactive Technologies Group (Department of Information and Communication Technologies) at the Universitat Pompeu Fabra in Barcelona, Spain. He is also affiliated with the Senseable City Lab (Department of Urban Studies and Planning) at the Massachusetts Institute of Technology (MIT) in Boston, USA. His current investigation explores the co-evolution of people with ubiquitous technologies in urban environments. This research is financially supported by Barcelona Media.