Urban Sensing Using Mobile Phones Network Data: A Survey of Research

FRANCESCO CALABRESE
IBM Research Ireland, Dublin, Ireland
LAURA FERRARI
Dipartimento di Scienze e Metodi dell'Ingegneria, University of Modena and Reggio Emilia, Italy
VINCENT D. BLONDEL
Universit catholique de Louvain, Louvain-la-Neuve, Belgium

The recent development of networks is producing an unprecedented wealth of information. There is an increasing interest in analyzing such data both from telecoms and other stakeholders points of view. In this survey, we outline some examples of data that can be collected from telecommunication networks as well as their strengths and weaknesses. We introduce techniques for dealing with anonymity, limitations in granularity, and pre-processing of such data to infer patterns related to human activities in the city. Each of these techniques will be described in terms of assumptions and limitations with state of the art examples that use real telecommunication dataset. Finally, we provide an overview of the challenges currently being faced in this field.

Categories and Subject Descriptors: ... [**...**]: ...

General Terms: Algorithms, Experimentation, Measurement

Additional Key Words and Phrases: ...

1. INTRODUCTION

Over the past decade the development of networks has produced an unprecedented wealth of information reflecting various aspect of urban life. These digital traces are valuable sources of data in capturing the pulse of the city in an astonishing degree of temporal and spatial detail, and could be used to make urban systems more efficient.

Telecom operators gather massive amount of data about how their users interact or occupy the city's infrastructure. In fact the International Telecommunication

Authors' address:
Francesco Calabrese, IBM Research Ireland, Damastown Industrial Park, Mulhuddart, Dublin 15, Ireland;
Laura Ferrari, University of Modena and Reggio Emilia, Dipartimento di Scienze e Metodi dell'Ingegneria, via Amendola 2, 42122 Reggio Emilia, Italy
Vincent D. Blondel, Avenue Georges Lematre 4, bote L4.05.01, B-1348 Louvain-la-Neuve, Belgium;

Union (ITU, 2011) estimates that at the end of 2011 there were 6 billion mobile subscriptions, with a global penetration of 87%, and 79% in the developing world. Every mobile phone leaves digital traces while interacting with the his infrastructure. Each phone can be seen as a mobile sensor that allows to define the geographic position of the subscriber holder almost in real time. Telecom operators are aware of the potential of such data and they have recently started to experiment with new business models in which they would generate revenues not only from their final customers (mobile phone users) but also from upstream customers such as traffic analysis, social networking, and advertising companies. As a result, they are sharing aggregate mobile data with various research communities (Review, 2010). Recently, massive datasets about cellphone users have been exploited in a variety of urban-related applications, including understanding mobility patterns (González et al., 2008; Isaacman et al., 2010), the use of urban spaces (Reades et al., 2007), travel demand during special events (Calabrese et al., 2010), social network structure (Onnela et al., 2007) and geographical dispersal of mobile communications (Lambiotte et al., 2008).

In this survey, we describe the types of data that can be collected from telecommunication networks and consider their strengths and weaknesses in terms of accuracy, level of details and applications. In particular, Section 2 shows what telecoms data can tell about urban dynamics. Section 3 outlines a brief overview of the mechanisms at the basis of mobile phone data generation and introduces what can be done with each type of data. Using running examples based on telecommunication datasets, Section 4 both presents some filtering and processing techniques necessary to deal with this data. Finally, Section 5 provides an overview of the challenges currently being faced in this field and Section 6 concludes.

2. MOBILE PHONE NETWORK DATA FOR URBAN ANALYSIS

It is well known that 50% of the globe's population lives in urban areas, that cover only the 0.4% of the Earth's surface (Fund, 2007). 70% are projected to do so by 2050. From one side, such urbanization opens great opportunities for improving people lifestyles, from the other side there is the need to prevent a potential economic, health and environmental disaster (Manyika et al., 2011). Pervasive technologies datasets are a way to understand how people use the city's infrastructure from the point of view of mobility (transportation mode), consumption (energy, water, waste) and environmental impact (noise, pollution). In fact, this kind of information offers new insights about the city (see for example the *Villevivante* project [1]), which are of great interest both from an economic and political perspective. In particular, urban planning can benefit from the analysis of personal location data. Decisions that can be improved by analyzing such data include the mitigation of traffic congestion and planning for high-density development. Urban transit and development planners will increasingly have access to a large amount of information about peak and off-peak traffic hotspots, volumes and patterns of transit use with which they can potentially cut congestion and the emission of pollutants. By drilling down into this wealth of data, urban planners will be more informed when they make decisions on anything from the placing and sequencing of traffic lights

[1] *http://villevivante.ch*

to the likely need for parking spaces. Singapore's public transportation [2] is already using ten-year demand forecasts partly based on personal location data to plan transit needs. Thus, understanding the urban dynamics allows both to improve services and create feedback loops with citizens to reduce energy consumption and environmental impact. Figure 1 shows how pervasive technologies datasets fit in this scenario. The human behavior of people in a city reflects how citizens use the built environment, the natural environment and the services offered by a city. Pervasive technologies are able to capture human behaviors and produce related datasets that contain very useful information for planning and management.

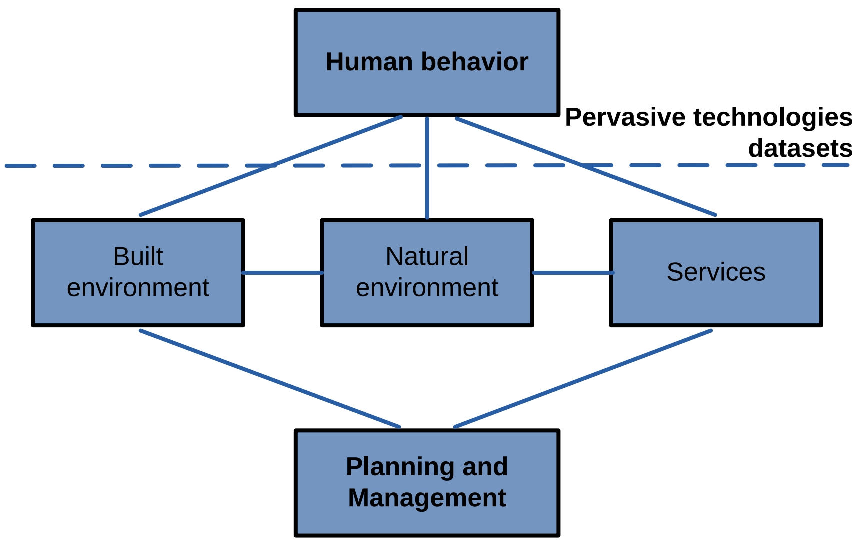


Fig. 1. Schema reflecting the role of pervasive technologies data sets in an urban scenario.

An important aspect not yet covered in this section is privacy concerns. Every country has its own regulations that telecommunication operators have to comply with. The main worry arising from the use of mobile phone network data is the fact that phone users' movements are continuously monitored, particularly in cases where such personal location data are made available to applications whose beneficiaries are third parties. As an example, the European Directive 2002/58/EC [3] regulates the treatment of personal data and protection of intimacy in the electronic communications sector. Article 14 of this Directive includes a description of location data, stating that: *"Location data may refer [...] to the identification of the cell in the network in which the mobile terminal is located at a given moment*

[2] *www.onemotoring.com.sg/publish/onemotoring/en/on_the_roads/traffic_management.html*

[3] **Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications)**, *http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32002L0058:en:HTML*

or to the time at which the localization information has been registered." Article 9 of this Directive also supplies regulations covering location data, as follows: *"In the event that location data can be processed [...] such data may only be processed if they are made anonymous, or with the prior consent of the users or clients, to the extent and for the time necessary to provide a value-added service."* Thus, in order to be compliant with regulations, all the data used for the research in this field (see the list of references) has been released by telecom operators so that it is impossible to associate the location data with actual cell-phone users.

In the field of urban analysis, mobile phone network data has been used in several topics:

(1) **Estimating population distribution.** With this regard, the use of mobile phone network data is twofold: *(i)* estimate where people live and *(ii)* estimate how population density changes over time, i.e. identify regions densely populated during particular days of the week and hours of the day. In particular, from one side the focus is on identifying locations meaningful to users. Authors n (Ahas et al., 2010; Isaacman et al., 2011) introduce a model for determining the geographical location of home and work places, while the paper in (Nurmi and Bhattacharya, 2008) describes and evaluates a non-parametric Bayesian approach for identifying places from sparse GPS traces (given the generic approach of the methodology, it can be easily applied to mobile phone network data). From the other side, the focus is on analyzing how the density of people changes over time. For example, in (Sohn et al., 2006; Sevtsuk and Ratti, 2010; de Jonge et al., 2012) authors explore how coarse-grained GSM data from mobile phones can be used to recognize high-level properties of user mobility and daily step count. The work in (Krisp, 2010) shows how calculating and visualizing mobile phone density assist fire and rescue services. Moreover, in (Soto et al., 2011) the information derived from the aggregated use of cell phone records is used to identify the socioeconomic levels of a population.

(2) **Estimating types of activities in different parts in the city.** During the week, the call activity of a residential region, a commercial or a business is different. It may be possible to derive a classification from the call activity profile of a region, thus allowing to classify regions as "residential", "commercial" or "business". For example, the work in (Girardin et al., 2009) provides a case study where aggregate and anonymous cell phone network activity data and georeferenced photos from Flickr allow to track the evolution of the attractiveness of different areas of interest in New York. Other works try to focus on the specific land use of a city. For example, in (Soto and Frias-Martinez, 2011) authors use Voronoi tessellation to automatically identify land uses from call detail record databases. The work is focused on the following types: industrial parks and office areas, commercial and business areas, nightlife areas, leisure and transport hubs, residential areas. In (Cáceres et al., 2012) authors analyze when and where people use greenspaces and how such behavior differes from urban areas. In (Reades et al., 2007) authors monitor the dynamics of Rome and obtain clusters of geographical areas measuring cell phone towers activity.

(3) **Estimating commuting patterns and mobility.** Using the cell phone ids, timestamp and location data of an event (call, sms, internet usage) it is possible

to estimate commuters mobility in predefined regions. For example, the work in (González et al., 2008) shows how the widespread coverage of mobile phone wireless networks in urban areas makes possible to track both groups and individuals. Authors in (Calabrese et al., 2011) use the real-time data collected from mobile phones to monitor the vehicular traffic status and the movements of pedestrians in Rome, Italy. Using an algorithm to analyze opportunistically collected mobile phone location data, the authors of (Calabrese et al., 2011) estimate weekday and weekend travel patterns of a large metropolitan area with high accuracy.

(4) **Analyzing social events and social networks.** The increase availability of mobile phone usage data sets in recent years has led to a number of studies also related to social events and the geography of social networks. In particular, (Ferrari et al., 2012; Traag et al., 2011) try to detect the underway of a social events from passive mobile phone networks data while the work in (Calabrese et al., 2010) analyse the movements of people during special events. Moreover, the geography of social networks has been exploited from a statistical perspective (Lambiotte et al., 2008), to derive a geography of mobile communications based on the relative frequency of communications as well as their average duration (Blondel et al., 2010), to study social radius of influence at both communication and mobility scale (Calabrese et al., 2011,?).

Moreover, mobile phone network data has been used not only in research works but also in running products based on both aggregated and raw data. A first group of applications deal with the issue of using mobile phone network data to derive urban traffic. Traditional companies (such as Inrix, *www.inrix.com* and Delcan, *http://delcantechnologies.com*) use traffic collection methods based on locating GPS-enabled vehicles and mobile devices. The use of mobile phone network data in order to leverage traffic information, enables to handle more data nodes (given the huge number of mobile phones subscribers), and therefore higher resolution than traditional traffic collection methods that are based on a relative small group of GPS-enabled vehicles. Thus, an increasing number of mobile phones operators are making partnerships with external companies [4] that can provide real-time services using traffic information. For example, Cellint (*www.cellint.com*) provides a worldwide service using mobile signaling data to locate the cars on the road. Such data is then analyzed to provide immediate incident detection (such as road sensors), as well as travel time and local speed over short segments (e.g. 200 meters in urban areas and 500 meters in other areas) for all the roads within a covered area. Intellimec is a similar company (*www.intellimec.com*) that provides real-time traffic and incidents information in the Canada area. Another company that leverage mobile phone network data to provide traffic information is Airsage (*www.airsage.com*), who aggregates signaling data from cellular networks to provide real-time speed and travel times for major roads. The company currently provides real-time location and traffic data in almost every city in the USA. Airsage also tries to provide unprecedented insight into the behavior of consumers at specific locations and at

[4] see for example the partnership between Vodafone and TomTom, *http://enterprise.vodafone.com/discover_global_enterprise/case_studies/tomtom.jsp*

different times during the day. This data can be used to understand locations, behaviors, and movements that are vital information for advertisers, corporations, commercial carriers, departments of transportation, and urban planners and, more in general, for any group that needs real-time geo-targeted information to plan, build and grow. Other applications focus on using mobile phone network data to provide services based on a more "social" aspect. For example Sense Networks [5] is commercializing Macrosense, a machine-learning technology model that aggregates historical and real-time mobile phone location data to, for instance, identify the best street corners from which to hail a taxi. Sense Networks' first application for consumers was CitySense, a tool designed to answer the question "Where is everyone going right now?". CitySense shows the overall activity level of the city, hotspots and places with unexpectedly high activity, all in real time. The tool uses also Yelp and Google to show what venues are operating at those locations. CabSense, another Sense Network application realised in early 2010, offers users an aggregated map generated by analyzing tens of millions of data points that rank street corners by the number of taxicabs picking up passengers every hour or every day of the week. From a complementary perspective, other research works try to combine mobile phone network data and social networks. For example, Mr.Typ (Mobile and Real-Time Yellow Pages [6]) and Social Telescope [7] offer a prototype of a platform for the flexible use of several forms of mobility data (GSM, mobile social networks data and wifi localizations). These websites provide a search engine for places where the rank is based on the average number of people that visit the area. Such a ranking takes advantage of telecoms data and of other data from different social networks (in the first prototype) or only on social networks (in the second prototype). For a more detailed description of Social Telescope see (Shankar et al., 2012). In this context, mobile phone network data has the following potentials: *(i)* offer the possibility to study micro and macro behaviors; and *(ii)* truly reflects human behavior given the fact that data is becoming more and more available thanks to the increasing adoption of mobile technologies.

The big issue shared by all these works is to compare the obtained results with other data sets in order to validate them.

To this regard, comparative data sets are useful to:

(1) Validate findings extracted from analysis of the mobile phone network data;

(2) Define scaling factors to extend results to the overall population;

(3) Augment information about urban space, which is useful to extract higher level patterns.

Table 2 outlines the main comparative data sets useful to validate the results obtained from mobile phone network data and highlights their pros and cons.

[5] *www.sensenetworks.com*
[6] *www.mrtyp.it*
[7] *www.socialtelescope.com*

Type	Pros	Cons
Census and Surveys	Very refined spatial resolution	Often outdated
Land use	Different categories	Different spatial units
Points of interest	Very refined categories	Different sources of data may provide different categories for the same points of interest

In particular:

Census and Surveys Data. Census and surveys data provide dataset related to very different areas: demography, health, education, government and security, communication and transport, etc. (see for example the 2010 US Census [8]). Such data set can be used to: *(i)* validate home and working areas; *(ii)* validate city patterns such as hotspots, commuting, traffic flows, etc.; *(iii)* validate land use. The main advantage of this kind of data is the very refined spatial resolution which is often the census block. The main disadvantages are that they are updated usually only every 5/10 years. Moreover, only some questions are asked thus providing only a partial view of human behavior.

Land Use. Global land use data sets (e.g., *http://data.giss.nasa.gov/landuse/*) offer access to a number of datasets that characterize an area based on its planned use. Different categories have been defined such as country codes, population density, cultivation intensity, etc. The main disadvantages are the possibly different spatial units in which they are aggregated.

Points of Interests. Points of interests are a list of businesses and important places to visit in a city. Usually every point of interest is characterized by a category and a location. There are many possible different sources: Yellow Pages, Yelp, Google Places etc. which might provide different information. As an example, the “A60”, a famous rooftop bar in Manhattan can be categorized as “Bar” by one source and as “Nightlife” by another source. In most comparisons, categories are aggregated in super-categories (e.g., bar and restaurants are aggregated in the super-category “Food”).

There are some challenges and limitations in comparing different datasets. The main one is that different collection periods and different spatial units introduce difficulties in comparing datasets. For example, census data is aggregated at block, track or country level while mobile phone network data is aggregated at cell tower level.

3. MOBILE PHONE NETWORK DATA GENERATION

When a mobile phone is switched on, always notifies its position in terms of the actual cell where it is currently located. The notification of the mobile phone position can be triggered by *events* (call, sms, or internet usage) or by updates of the *network* (for a more detailed description of the technologies and standards used to derive the position of mobile phones see (Wang et al., 2008)).

[8] *http://2010.census.gov*

Event-Driven Mobile Phone Network Data Today, there are two primary sources of these data: communication and internet usage. Most telephone networks generate Call Detail Record (CDR) that are data record produced by a telephone exchange documenting the details of a phone call or sms passed through the device. A CDR is composed of data fields that describe the telecommunication transaction such as the user id of the subscriber originating the transaction, the user id receiving the transaction, the transaction duration (for calls), the transaction type (voice or sms), etc. Each telecommunication operator decides which information is emitted and how it is formatted. As an example, there could be the timestamp of the end of the call instead of the duration. Figure 2(a) shows an example of a CDR log: in this case the telecommunication operator decides to emit the user id (as a hash string), the International Mobile Subscriber Identity (IMSI, an identification code used to individually represent cell phones on the GSM and UMTS networks), the id of the cell where the user is connected and the timestamp of the log.

user hash	IMSI	cell id	timestamp
6fb175825f09bf	22201	662188114	1330944127195
6cd347681a76fd	22201	662188114	1340718433219
6fb175825f09bf	22201	564389331	1330944127195

(a)

cell id	lat	lon
662188114	44.658885	10.925102
564389331	44.701606	10.628872

(b)

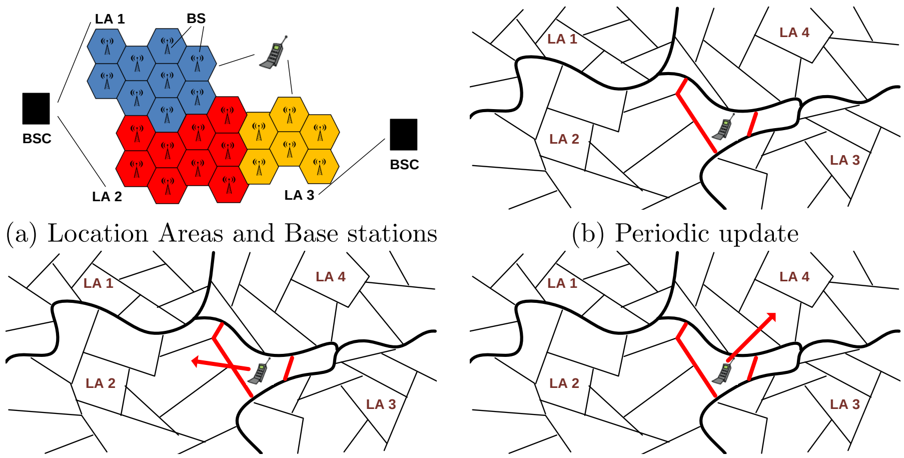
Fig. 2. **(a)** Example of a CRD log: anonymized user id, International Mobile Subscriber Identity (IMSI), cell id and timestamp in millisecond from epoch; (b) Cell location information.

The second source of data is internet usage. In telecommunications, an IP Detail Record (IPDR) provides information about Internet Protocol (IP)-based service usage and other activities. The content of the IPDR is determined by the service provider, the Network/Service Element vendor, or any other community of users with authority for specifying the particulars of IP-based services in a given context. Examples of IPDR data fields are: user id, type of the website, time of event, number of bytes transmitted, etc. It is important to note that the margin of error in this case varies widely according to whether the device to which the IP address is attached is mobile, and to the density and topology of the underlying IP network.

Both communication and internet usage can be associated to the cell phone towers used during the interaction.

Network-Driven Mobile Phone Network Data A cellular network is a radio network of individual cells, known as base stations. Each base station covers a small geographical area which is part of a uniquely identified location area. By integrating the coverage of each of these base stations, a cellular network provides a radio coverage over a much wider area. A group of base stations is named a

Location Area (LA), or a routing area. A LA is a set of base stations that are grouped together to optimise signalling (see Figure 3(a)).



(a) Location Areas and Base stations (b) Periodic update

(c) Handover (d) Mobility Location Update

Fig. 3. (a) Location area and base stations; (b) Periodic update; (c) Handover; (d) Mobility Location Update.

Typically, tens or even hundreds of base stations share a single Base Station Controller (BSC). The BSC handles allocation of radio channels, receives measurements from the mobile phones, controls handovers from base station to base station.

In such a context, three different types of location update can happen:

(1) **Periodic Update**, which is generated on a periodic base and provides information on which cell tower the phone is connected to (see Figure 3 (b)).
(2) **Handover**, which is generated when a phone involved in a call moves between two cell areas (see Figure 3 (c)).
(3) **Mobility location update**, which is generated when the phone moves between two Location Areas (see Figure 3 (d)).

Another important aspect is how the user's location can be detected. Location information can be extracted as part of the interaction data between the mobile phone and the telecommunication infrastructure. In most cases it is represented by the cell tower position or the cell sector to which the mobile phone is connected. Figure 2(a) shows an example of a CDR location information, represented by the *cell id* field. Figure 2(b) maps each *cell id* in the corresponding latitude and longitude coordinates.

In particular, triangulated location can be estimated having access to data collected at lower levels in the network. The format of such data is given by standard documentation provided by networks operators (as an example, see the 3gppp standard documentation, (3gp, 2012)). The principal techniques are the following:

(1) **Timing Advance (TA)**, which is a value that corresponds to the length of time a signal takes to reach the cell tower from a mobile phone. Since the

user hash	longitude	latitude	uncertainty	timestamp
4ba232e4d96f47dc94f7441e87c164fb	16	81	56	1246759931
4ba232e4d96f47dc94f7441e87c164fb	06	09	252	1246759922
4ba232e4d96f47dc94f7441e87c164fb	99	95	208	1246760034

Table I. Example of cell tower location information obtained using propagation models: compared with Figure 2, such table shows an additional information represented by the uncertainty field.

users are at various distances from the cell tower and radio waves travel at the finite speed of light, the precise arrival time can be used by the cell tower to determine the distance to the mobile phone (see Figure 4(a)).

(2) **Received Signal Strength (RSS)**,which is a measurement of the power present in the signal received by cell towers from one another. Because the power levels at the start of the signal transmission are well known and the power drop in signal in open spaces is well defined, RSS can be used to estimate the distance between a mobile phone and the surrounding cell towers (see Figure 4(b)).

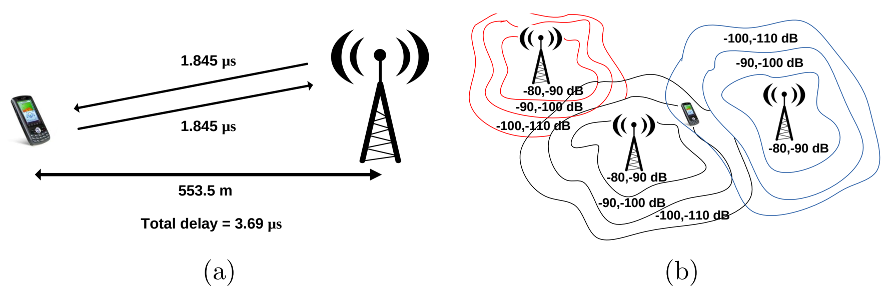


Fig. 4. Estimating the mobile phone location information: (a) Time Advance and (b) Received Signal Strength techniques.

It is important to note that with these methodologies the accuracy of the mobile phone position is around 500m in urban areas. An accuracy of 150m in urban areas can be obtained using propagation models and irradiation diagrams; such techniques estimate the mobile phone position by finding the point that minimizes the mean square error between measured and estimated mean power received by all base stations. Table I shows an example of the cell tower location information obtained using propagation models and irradiation diagrams; the main difference is represented by the uncertainty field that gives an estimation of the accuracy of the mobile phone position.

The kind of data explained in this section has been used to mine how people move and behave in cities both from a spatial and temporal point of view (see Section 2). In the next section we will show, from a general perspective, advantages, disadvantages and potential applications of these kind of data.

3.1 Data Aggregation

Service providers in each country have different rules and restrictions as to what kind of data can be exchanged through their network. Individual data is rarely available in real time even for service providers, but is usually available the day after if additional hardware is not installed on mobile phones. Moreover, the use of individual data can lead to privacy concerns (as explained in Section 2). The same data can be aggregated at different spatial and temporal scales. For example, mobile phone network data can be aggregated at cell tower level by considering: the number of calls, Erlang (total communication time), the number of sms, the number of handovers, the number of location updates, etc.

Aggregated data can be more easily accessible in real time or with low delay. Moreover, regarding the volume, aggregated data can be easily manageable, while individual data might be difficult to manage. A possible solution to this regard would be to analyze only a subset of users but this would rise the problem of selecting a good and representative sample.

Table summarizes advantages, disadvantages as long as applications for each type of mobile network data.

	Advantages	**Disadvantages**	**Applications**
		Aggregated Data	
Aggregated cell tower statistics	Easy to manage, possibly in real time	No information on users' mobility	Land use estimation, population density estimation
Aggregated CDR with cell tower location information	Easy to manage	No individual interaction information	Connection between places, Regional partitioning
		Individual Data	
Individual CDR	Individual communication patterns	Large dataset, mostly not real time	Social network analysis
Individual CDR with cell tower location information	Individual communication and mobility patterns	Large dataset, mostly not real time	Mobility analysis between large areas
Individual Event-driven triangulated location	Individual mobility patterns, possibly in real time	Large dataset, possibly need for special hardware to access data	Origin destination, transportation mode
Individual Network-driven data	Individual mobility patterns, possibly in real time	Large dataset, possibly need for special hardware to access data	Useful for mobility analysis between large areas

Table II. Advantages, disadvantages and possible applications for each type of mobile phone network data.

In particular:

—**Aggregated data at cell tower level.** Such data can contain or not contain the information regarding the cell tower location information. In particular, *(i) aggregated cell tower statistics* that do not contain location information are easy to manage and can be available in real-time. Such data can be used both for the land use and the population density estimation; *(ii) aggregated CDR with cell tower location information* have the disadvantage of not presenting the information regarding the interaction between individuals.

—**Individual data.** Such data is provided for each individual and can contain or not contain cell tower location information. The main disadvantage shared by all this type of data is that they are represented by large datasets sometimes not in real-time. In addition, there is sometimes the need of additional hardware to access the data. In particular, *(i) individual CDR* provides the information regarding individual communication patterns that can be used for social network analysis; *(ii) individual CDR with cell tower location information* provides information regarding individual and mobility patterns. Such data can be used to analyze the mobility between large areas; *(iii) individual event-driven triangulated location* provides information regarding individual mobility patterns possibly in real-time, but in order to access the data, there is possibly the need for special hardware to access the data. It can be used for origin-destination and transportation node analysis. *(iv) individual network-driven data* presents the same features of the above type of data and can be used for mobility analysis between large areas.

4. TECHNIQUES FOR MOBILE PHONE NETWORK DATA ANALYSIS

In this section we will show several techniques for mobile phone network data analysis that have been used in research works (some of them are briefly introduced in (de Jonge et al., 2012)). Each technique will be described in terms of assumptions and limitations with a running example using real mobile phone network datasets. First, we will describe some filtering techniques necessary to reduce rawness in the data. Then, we will describe a list of features that can be extracted from mobile phone network data as long as the necessary processing techniques.

4.1 Filtering Techniques

In order to mine mobile phone network data to derive human mobility patterns in cities, several techniques are needed to reduce both the spatial uncertainty and the noisiness of the raw data. The main issues to this regard are *(i)* assigning the user to a specific location and *(ii)* identifying when the user stops in a location or is simply passing through it.

—**Assigning the user to a specific location.** State of the art works in the area suggest two main solutions:

(1) *Assign the user to the centroid of the cell area.* As shown in Section 3, each CDR produced by a mobile phone is associated to a cell whose location is known by the mobile phone operator. In (González et al., 2008) authors first divide the area under investigation with a Voronoi tessellation technique based on the cell tower locations, then they assign the user position to the centroid of the corresponding Voronoi cell. A different approach is shown in (Girardin et al., 2009), where the user location is assigned to the best serving cell. The computation is made on simulated coverage and takes into account both the cell sector and propagation models.

(2) *Assign the user a probability to be in a given location.* This second solution introduces uncertainty in assigning a user to a location. For example the work in (Traag et al., 2011) uses a propagation model to assign a user a probability of being connected to a particular cell tower. The main advantage is that this solution takes into consideration the fact that multiple towers might be covering the same location.

—**Stop detection.** Another important issue is determining which places are important to the user, i.e., in which places the user stops for a reasonable time period. Given the rawness of mobile phone network data, the same event can e registered as consecutive events associated to different close by locations. The solutions proposed so far to improve accuracy in the raw mobile phone network data can be divided in two groups:

(1) **Solutions that leverage on consecutive location data,** where consecutive measurements which are close enough can be collapsed in a unique single measurement. within a given spatial and temporal window are averaged all together. For example, in (Calabrese et al., 2010) the authors fixed both a spatial S_{th} and a temporal T_{th} threshold in order to detect stops, i.e., two consecutive stops $stop_i$ and $stop_j$ can be collapsed in the same stop if $distance(d_{stop_i}, d_{stop_i}) < S_{th}$ and $(t_{stop_i} - t_{stop_j}) > T_{th}$.

(2) **Solutions that leverage on historical location data,** where historical location data is used to help understanding which places are important for the user. For example, the work in (Isaacman et al., 2011) uses clustering techniques (in particular the Hartigan's algorithm) on a dataset spanned over 78 days with the aim of identifying which places are important to the users such as home and work location.

4.2 Processing Techniques

In this section we will describe the kind of analysis that can be done on mobile phone network data and the corresponding necessary processing techniques as shown in the state of the art works. In particular, we have divided the analysis on the kind of analysis that can be done on individual data and on aggregated data.

INDIVIDUAL DATA

Using individual mobile phone data, several features have been analyzed. In particular:

(1) **Home and work location estimation.** Using CDR with location information, some works (Calabrese et al., 2011; Isaacman et al., 2011) have been focused on estimate the home and work location of the users. In order to increase the precision in estimating the location, a dataset consisting of several days of mobile phone network data for each user has been used. Necessary information in the raw data are: *(i)* the number of times a cell tower was contacted by the user; *(ii)* the length (in terms of time) of stay in a location. In particular, home location has been determined as the most frequented place during evenings, while work location as the most frequented place during weekday mornings/afternoons and excluding the home location and places with a high number of evening events. Data has been validated using US census population estimates at census track level.

(2) **Daily mobility estimation.** In (González et al., 2008), the authors tried to infer daily trips using the distance between any two different visited locations. In (Isaacman et al., 2011), the daily range of mobility related to where people live has been analyzed. Moreover, origin and destination of trips can be mapped (see for example (Calabrese et al., 2011)) thus allowing to count the number

of trips for any time of the day and to analyze the attractiveness of an area (measured as the number of different places people come from). In (Couronne et al., 2011) users has been clustered on the basis of how often they move using spatio-temporal analysis.

(3) **Analyzing how social events impact mobility in the city.** Using both CDR with location information and individual event-driven triangulated location data, some works (Calabrese et al., 2010; Traag et al., 2011) tried to model and predict non-routine origin-destination flows (e.g., mobility flows generated by the attendance to an event) in the city. The aim of these works is twofold: *(i)* improve event planning and management (e.g., predict the effect of an event on the urban transportation, adapt public transit -schedules and routes- to accommodate additional demand, etc.); *(ii)* improving location based services, for example recommending social events (see (Quercia et al., 2010)). In this last work authors build a recommender system that analyzes users' mobile phone network data with the aim of suggesting events based on the users' whereabouts patterns. In particular, the authors analyzed also the "cold start problem", i.e., the kind of events that can be suggested to a user that has no location history.

(4) **Integrating social and mobility information.** As shown in Section 2, mobile phone network data has been mined also to integrate calling and location pattern in order to help inferring face-to-face meetings. In (Calabrese et al., 2011) authors discovered that people calling while connected to the same cell tower (co-location) are a good proxy for face-to-face meetings. In particular, they discovered that people tend to interact much more just before and after this event, and the number of inferred face-to-face meetings decreases with the users' home distance. From the call interactions the authors are able to predict when and where people will be meeting.

AGGREGATED DATA

As shown in Section 3 compared to individual data, aggregated data is much more easy to manage and can be possibly available in real time. In the following we will show the techniques that have been applied to mobile phone network data in the state of the art works.

(1) **Land use inference.** Starting from aggregated cell tower statics, it is possible to understand activities in the city from telecommunication usage patterns. This can augment existing built environment data collection and analysis methods (census, business registrations, etc.) at low cost and with very low latencies. Categories of activities can be considered. In particular, classical time series analysis can be performed (for example, the Principal Component Analysis technique has been used in (Reades et al., 2007) or the Dynamic Time Warping technique in (Yuan and Raubal, 2012)) and clustering of time series can classify places based on usage (like the Fuzzy C-Means technique proposed in (Soto and Frias-Martinez, 2011)).

(2) **Space partitioning.** Using CDR with location information, it is possible to partition the space based on the level of human interactions. In particular, partitioning at different scale has been analyzed:

—*Regional Partitioning.* Mobile phone users location at call time can be used to infer origin and destination of the calls, thus allowing to model the effect of geography on human mobility and interactions. Using network analysis, in (Lambiotte et al., 2008) authors find that human interactions decrease as distance increases following a gravity-like behavior. Exception emerges are mainly due to: geographical features (e.g., rivers, see for example (Ratti et al., 2010)), administrative borders and cultural differences.

—*City scale partitioning.* At the city scale, interaction events can be aggregated to create a network of places where nodes are locations (e.g., cell towers) and edges between nodes exists if interactions happens between people connected to the two cell towers. The weighted graph can be partitioned in communities using standard network analysis tools (modularity optimization). Researches can detect: *(i)* which areas in the city are most connected; *(ii)* where interaction borders exist (see (Blondel et al., 2010)); *(iii)* how borders change over time (see (Walsh and Pozdnoukhov, 2011)).

—*Country scale partitioning.* In this case, CDR with location information have been aggregated at a country scale. Users' home country has been assigned with the most frequent country where calls are made. Then, CDR have been aggregated on the basis of users' home country. Some interesting resulting data shown in (Calabrese et al., 2011) are the interactions between countries, in particular: state boundaries emerge in most of the cases, metropolitan areas (e.g., NYC, LA) define new regions, some area merge as level of the interaction is higher than expected. Starting from raw data, authors had to take some actions: *(i)* normalization in order to deal with operator share not being equal for every area and *(ii)* filtering of countries with a too low number of customers or share (to preserve representativeness of the sample).

5. OPEN CHALLENGES

In this paper we have shown how mobile phone network data can be used to gain insights on urban patterns. In dealing with this type of data, some challenges still remain open:

(1) **Limitations of event-driven data** In order to analyze certain types of urban patterns, it is important to have very frequent location data. As explained in Section 2, event-driven data are generated only when the user takes some action, i.e., sends an SMS, makes a call, etc. Thus, the location of the user might not be updated very frequently. Some approaches proposed so far to solve this problem are:

—*Sampling only highly active users.* This solution might be effective since high communication (e.g., calling someone or sending an SMS) has been found to be correlated to high mobility (Couronne et al., 2011). The main problem to this regard is how to choose users that represents a good sample of citizens' behavior.

—*Sampling smartphone internet usage data.* Given the high penetration of smartphones (Manyika et al., 2011), another option is to use the internet usage to derive location data. The main pros is that such kind of data generally presents the lower inter-event time (Calabrese et al., 2010), but smartphone users' behavior can not represent a general sample of citizen's behavior today.

—*Network-driven data.* Given the low frequency of users' localization updates, a better type of data could be network-driven data. In particular, periodic sampling is independent on events but is not too good for short term mobility. Another alternative could be mobility-based sampling that is good for analyzing mobility between large areas such as Location Areas.

(2) **Limitations in spatial accuracy.** It might be important to have very precise location data for certain types of applications, such as to determine the accurate location, the route undertaken by the user or the transportation modes. As shown in Section 2, mobile phone network data does not provide accurate localization. Some solutions proposed so far are:

—*Look at history for recurring locations.* This can help in smoothing irregularities in the location data, allowing to assign a wrong (because of the low accuracy in the localization) position to the nearest recurring location.

—*Look at handover during calls.* Handoff patterns are relatively stable across different routes, speeds, directions, phone models, and weather conditions (Becker et al., 2011), thus allowing to derive the trajectories of mobile devices using also CDR data with a low frequency of localization update.

(3) **Managing uncertainties.** Looking at the previous open challenges, it is clear that the uncertainties in the user's status in time and space can be relatively large. This is due to both the low frequency of user's localization update and the spatial resolution of mobile phone network data. Thus, it is important to provide reliable and uncertain-aware results. One proposed solution in estimating uncertainties in users' position. For example, in (Couronne et al., 2011) the authors try to estimate the bias of user behavior in mobile phone data taking into account the imprecision of data, with a trigonometric approach to describe both mobility values and uncertainty: theta and norm.

(4) **Finding comparative datasets.** Traditional city data (e.g., census and surveys) are collected using different methods, sampling time and collection years. This makes difficult to compare results obtained analyzing mobile phone network data with these traditional datasets. Proposed alternatives are:

—*Self-reported data.* Self-reported data can make an additional value to traditional data since they might be accurate, not outdated and with a correct sampling time to make comparisons. An example of self-reported data is the one that can be obtained from Flickr (`www.flickr.com`), that is used for example in (Girardin et al., 2008) to mine tourists patterns in Rome.

—*Social networking data.* Similar to the previous one, social networking data provides specific information regarding the places visited by the users. There are a plethora of location-based social networks such as Foursquare (`https://foursquare.com`), Twitter (`https://twitter.com`), Facebook Places (`www.facebook.com/about/location`), etc. that provides public access to their own data. The works in (Bawa-Cavia, 2011; Ferrari et al., 2011) provide examples of using such data for urban analysis.

(5) **Dealing with privacy and anonymity.** Using individual mobile phone network data, even if anonymized, it is possible to detect important information

from users (e.g., home and work location). Authors in (Zang and Bolot, 2011) show that in the cases where more than one top location can be identified, anonymity is not preserved. Two proposed solutions (Krumm, 2009) so far are:

—*location obfuscation*, which consists in non reversible ways to slightly alter the location such that it does not reflect the real location of the user, but still contains enough information to provide a satisfactory service. See (Wightman et al., 2011) for more information regarding the evaluation of several location obfuscation techniques;

—*k-anonymity for trajectories*, which ensures that each individual trajectory can only be released if there are at least $k1$ distinct individuals whose associated trajectories are indistinguishable from the former (see (Gedik and Liu, 2008) for more detailed information).

(6) **Mobility/communication interplay.** The interplay between telecommunications and physical location is still a challenge. In some cases it has been suggested that telecommunications may be a substitute for physical interaction (Albertson, 1977). In other cases conflicting hypotheses have been made, including those of a complementary (Mok et al., 2010), neutral (Choo et al., 2010) or reinforcing (Sasakia and Nishiib, 2010) effect. Regarding mobile phone network data, the work in (Calabrese et al., 2011) investigates the relationship between people's calls and their physical location. In (Wang et al., 2011) the authors mine the similarities between people's movements (as collected by the mobile phone network) and social networks.

(7) **Real Time data acquisition and processing.** Many urban sensing applications (e.g., traffic monitoring, event management, etc.) are useful if results are presented in real time or near-real time. The problem is that usually mobile phone network data is first acquired and then pushed to databases, thus it is not usually available in real time (see Section 3). Since the quantity of mobile phone network data produced everyday is massive, there is the need of ad-hoc algorithms and platforms to process such data in real time. Proposed solutions are streaming platforms able to deal with different types of data in real time (see for example (Gasparini et al., 2011; Kaiser and Pozdnoukhov, 2011)).

6. CONCLUSIONS

This article discusses the current state of the art and open challenges in the emerging field of mobile phone network data for urban sensing. The primary obstacle to this new field is not a lack of infrastructure: millions of people already carry phones and telecoms operators already have the needed infrastructure. Rather, the technical barriers are related to performing privacy-sensitive reasoning with noisy and sparse data. Research is still particularly needed in: *(i)* inferring behavioral patterns; *(ii)* building analytics and systems to process massive datasets and automatically extract patterns; *(iii)* building control systems able to make use of inferred patterns to optimize city services. Mobile phone network data will ultimately provide both micro- and macroscopic views of cities and help understand citizens' behaviors and patterns.

REFERENCES

2011. International Telecommunication Union. http://www.itu.int/ITU-D/ict/.

2012. The Mobile Broadband Standard, checked May. http://www.3gpp.org.

Ahas, R., Silm, S., Jrv, O., Saluveer, E., and Tiru, M. 2010. Using mobile positioning data to model locations meaningful to users of mobile phones. *Journal of Urban Technology 17,* 1, 3–27.

Albertson, L. A. 1977. Telecommunications as a travel substitute: Some psychological, organizational, and social aspects. *Journal of Communication 27,* 2, 32–43.

Bawa-Cavia, A. 2011. Sensing the urban: Using location-based social network data in urban analysis. In *Proceedings of the First Workshop on Pervasive Urban Applications*. Purba '11. San Francisco, CA, USA.

Becker, R. A., Caceres, R., Hanson, K., Loh, J. M., Urbanek, S., Varshavsky, A., and Volinsky, C. 2011. Route classification using cellular handoff patterns. In *Proceedings of the 13th international conference on Ubiquitous computing*. UbiComp '11. ACM, New York, NY, USA, 123–132.

Blondel, V., Krings, G., and Thomas, I. 2010. Regions and borders of mobile telephony in belgium and around brussels. *Brussel Studies* 42.

Cáceres, Small, C., Rowland, J., and Urbanek, S. 2012. Exploring the use of urban greenspace through cellular network activity. In *Proceedings of the Second Workshop on Pervasive Urban Applications*. Purba '12. Newcastle, UK.

Calabrese, F., Colonna, M., Lovisolo, P., Parata, D., and Ratti, C. 2011. Real-time urban monitoring using cell phones: A case study in Rome. *IEEE Transactions on Intelligence Transportation Systems 12*, 141–151.

Calabrese, F., Dahlem, D., Gerber, A., Paul, D., Chen, X., Rowland, J., Rath, C., and Ratti, C. 2011. The connected states of america: Quantifying social radii of influence. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*. 223 –230.

Calabrese, F., Di Lorenzo, G., Liu, L., and Ratti, C. 2011. Estimating origin-destination flows using mobile phone location data. *Pervasive Computing, IEEE 10,* 4 (april), 36 –44.

Calabrese, F., Pereira, F., Lorenzo, G. D., Liu, L., and Ratti, C. 2010. The geography of taste: Analyzing cell-phone mobility and social events. In *Pervasive Computing*, P. Floren, A. Krger, and M. Spasojevic, Eds. Lecture Notes in Computer Science, vol. 6030. Springer Berlin / Heidelberg, 22–37.

Calabrese, F., Smoreda, Z., Blondel, V. D., and Ratti, C. 2011. Interplay between telecommunications and face-to-face interactions: A study using mobile phone data. *PLoS ONE 6,* 7 (07), e20814.

Choo, S., Lee, T., and Mokhtarian, P. L. 2010. Do transportation and communications tend to be substitutes, complements, or neither?: U.s. consumer expenditures perspective, 1984-2002. *Transportation Research Record: Journal of the Transportation Research Board 2010/2007*, 121–132.

Couronne, T., Olteanu, A.-M., and Smoreda, Z. 2011. Urban mobility: Velocity and uncertainty in mobile phone data. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*. 1425 –1430.

Couronne, T., Smoreda, Z., and Olteanu, A.-M. 2011. Chatty mobiles: Individual mobility and communication patterns. *NetMob 2011*.

de Jonge, E., van Pelt, M., and Roos, M. 2012. Time patterns, geospatial clustering and mobility statistics based on mobile phone network data. *Statistics Netherlands, Division Methodology and Quality Sector Methodology Heerlen*.

Ferrari, L., Mamei, M., and Colonna, M. 2012. People get together on special events: Discovering happenings in the city via cell network analysis. In *PerCom Workshops*. 223–228.

Ferrari, L., Rosi, A., Mamei, M., and Zambonelli, F. 2011. Extracting urban patterns from location-based social networks. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*. LBSN '11. ACM, New York, NY, USA, 9–16.

Fund, U. N. P. 2007. 2007 State of World Population . `http://www.unfpa.org/swp/swpmain.htm`.

Gasparini, L., Bouillet, E., Calabrese, F., Verscheure, O., O'Brien, B., and O'Donnell, M. 2011. System and analytics for continuously assessing transport systems from sparse and noisy observations: Case study in dublin. In *Intelligent Transportation Systems (ITSC), 2011 14th International IEEE Conference on*. 1827 –1832.

Gedik, B. and Liu, L. 2008. Protecting location privacy with personalized k-anonymity: Architecture and algorithms. *IEEE Transactions on Mobile Computing 7,* 1 (Jan.), 1–18.

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

Girardin, F., Vaccari, A., Gerber, A., and Ratti, C. 2009. Quantifying urban attractiveness from the distribution and density of digital footprints. *Journal of Spatial Data Infrastructure Research* 4, 175–200.

González, M., Hidalgo, C., and Barabási, A.-L. 2008. Understanding individual human mobility patterns. *Nature 453,* 7196 (June), 779–782.

Isaacman, S., Becker, R., Cáceres, R., Kobourov, S., Martonosi, M., Rowland, J., and Varshavsky, A. 2011. Identifying important places in people's lives from cellular network data. In *Proceedings of the 9th international conference on Pervasive computing*. Pervasive'11. Springer-Verlag, Berlin, Heidelberg, 133–151.

Isaacman, S., Becker, R., Cáceres, R., Kobourov, S., Rowland, J., and Varshavsky, A. 2010. A tale of two cities. In *Proceedings of the Eleventh Workshop on Mobile Computing Systems & Applications*. HotMobile '10. ACM, New York, NY, USA, 19–24.

Kaiser, C. and Pozdnoukhov, A. 2011. Modelling city population dynamics from cell phone usage data streams. *NetMob 2011*.

Krisp, J. M. 2010. Planning fire and rescue services by visualizing mobile phone density. *Journal of Urban Technology 17,* 1, 61–69.

Krumm, J. 2009. A survey of computational location privacy. *Personal Ubiquitous Comput. 13,* 6 (Aug.), 391–399.

Lambiotte, R., Blondel, V., Dekerchove, C., Huens, E., Prieur, C., Smoreda, Z., and Vandooren, P. 2008. Geographical dispersal of mobile communication networks. *Physica A: Statistical Mechanics and its Applications 387,* 21 (Sept.), 5317–5325.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., and Byers, A. H. 2011. Big data: The next frontier for innovation, competition and productivity. In *McKinsey Global Institute*.

MOK, D., WELLMAN, B., AND CARRASCO, J. 2010. Does distance still matter in the age of the internet? *Urban Studies 47,* 13, 2747–2783.

NURMI, P. AND BHATTACHARYA, S. 2008. Identifying meaningful places: The non-parametric way. In *Pervasive Computing*, J. Indulska, D. Patterson, T. Rodden, and M. Ott, Eds. Lecture Notes in Computer Science, vol. 5013. Springer Berlin / Heidelberg, 111–127.

ONNELA, J.-P., SARAMÄKI, J., HYVONEN, J., SZABÓ, G., LAZER, D., KASKI, K., KERTÉSZ, J., AND BARABÁSI, A.-L. 2007. Structure and tie strengths in mobile communication networks. *Proc. Natl. Acad. Sci. (USA*, 7332.

QUERCIA, D., LATHIA, N., CALABRESE, F., DI LORENZO, G., AND CROWCROFT, J. 2010. Recommending social events from mobile phone location data. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*. 971 –976.

RATTI, C., SOBOLEVSKY, S., CALABRESE, F., ANDRIS, C., READES, J., MARTINO, M., CLAXTON, R., AND STROGATZ, S. H. 2010. Redrawing the map of great britain from a network of human interactions. *PLoS ONE 5,* 12 (12), e14248.

READES, J., CALABRESE, F., SEVTSUK, A., AND RATTI, C. 2007. Cellular census: Explorations in urban data collection. *IEEE Pervasive Computing 6*, 30–38.

REVIEW, T. 2010. Mobile Data: A Gold Mine for Telcos. `http://www.technologyreview.com/news/419101/mobile-data-a-gold-mine-for-telcos/`.

SASAKIA, K. AND NISHIIB, K. 2010. Measurement of intention to travel: Considering the effect of telecommunications on trips. *Transportation Research Part C: Emerging Technologies 18*, 36–44.

SEVTSUK, A. AND RATTI, C. 2010. Does urban mobility have a daily routine? learning from the aggregate data of mobile networks. *Journal of Urban Technology 17,* 1, 41–60.

SHANKAR, P., HUANG, Y., CASTRO, P., NATH, B., AND IFTODE, L. 2012. Crowds replace experts: Building better location-based services using mobile social network interactions. In *PerCom*. 20–29.

SOHN, T., VARSHAVSKY, A., LAMARCA, A., CHEN, M., CHOUDHURY, T., SMITH, I., CONSOLVO, S., HIGHTOWER, J., GRISWOLD, W., AND DE LARA, E. 2006. Mobility detection using everyday gsm traces. In *UbiComp 2006: Ubiquitous Computing*, P. Dourish and A. Friday, Eds. Lecture Notes in Computer Science, vol. 4206. Springer Berlin / Heidelberg, 212–224.

SOTO, V. AND FRIAS-MARTINEZ, E. 2011. Robust land use characterization of urban landscapes using cell phone data. In *Proceedings of the First Workshop on Pervasive Urban Applications*. Purba '11. San Francisco, CA, USA.

SOTO, V., FRIAS-MARTINEZ, V., VIRSEDA, J., AND FRIAS-MARTINEZ, E. 2011. Prediction of socioeconomic levels using cell phone records. In *Proceedings of the 19th international conference on User modeling, adaption, and personalization*. UMAP'11. Springer-Verlag, Berlin, Heidelberg, 377–388.

TRAAG, V., BROWET, A., CALABRESE, F., AND MORLOT, F. 2011. Social event detection in massive mobile phone data using probabilistic location inference. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*. 625 –628.

WALSH, F. AND POZDNOUKHOV, A. 2011. Spatial structure and dynamics of urban communities. In *Proceedings of the First Workshop on Pervasive Urban Applications*. Purba '11. San Francisco, CA, USA.

WANG, D., PEDRESCHI, D., SONG, C., GIANNOTTI, F., AND BARABASI, A.-L. 2011. Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM*

SIGKDD international conference on Knowledge discovery and data mining. KDD '11. ACM, New York, NY, USA, 1100–1108.

Wang, S., Min, J., and Yi, B. 2008. Location based services for mobiles: Technologies and standards. In *IEEE ICC*. Beijing.

Wightman, P., Coronell, W., Jabba, D., Jimeno, M., and Labrador, M. 2011. Evaluation of location obfuscation techniques for privacy in location based information systems. In *Communications (LATINCOM), 2011 IEEE Latin-American Conference on*. 1 –6.

Yuan, Y. and Raubal, M. 2012. Extracting dynamic urban mobility patterns from mobile phone data. In *Proceedings of the 7th international conference on Geographic Information Science*. GIScience'12.

Zang, H. and Bolot, J. 2011. Anonymization of location data does not work: a large-scale measurement study. In *Proceedings of the 17th annual international conference on Mobile computing and networking*. MobiCom '11. ACM, New York, NY, USA, 145–156.