Urban Sensing Using Mobile Phone 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 telecommunication networks is producing an unprecedented wealth of information and, as a consequence, an increasing interest in analyzing such data both from telecoms and other stakeholders' points of view. In particular, mobile phone datasets offer access to insights into urban dynamics and human activities at an unprecedented scale and level of details, representing a huge opportunity for research and real world applications. This paper surveys the new ideas and techniques related to the use of telecommunication data for urban sensing. We outline the data that can be collected from telecommunication networks as well as their strengths and weaknesses with a particular focus on urban sensing. We survey existing filtering and processing techniques to extract insights from this data, and summarise them to provide recommendations on which datasets and techniques to use for specific urban sensing applications. Finally, we discuss a number of challenges and open research areas currently being faced in this field. We strongly believe the material and recommendations presented here to become increasingly important as mobile phone network datasets are becoming more accessible to the research community.

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

General Terms: Algorithms, Experimentation, Measurement

Additional Key Words and Phrases: ...

1. INTRODUCTION

Over the past decade the development of digital networks has produced an unprecedented wealth of information reflecting various aspect of urban life. These

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;

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.

The International Telecommunication Union 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 (ITU, 2011). Every mobile phone leaves digital traces while interacting with its infrastructure. Thus, each phone can be seen as a mobile sensor that allows to detect 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 have started sharing aggregated mobile data with various research communities (Technology Review, 2010). Thanks to that, 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). More recently, research challenges have also been proposed by Orange[1], Telefonica[2] and Telecom Italia[3], where operators have released telecommunication data to the wide research community, and are now accessible and studied by hundreds research laboratories around the world. Clearly, using mobile phone data for urban sensing could have great impact in developing countries, where specific sensors (such as traffic sensors) are rarely put in place. A recent Primer from the UN Global Pulse organisation summarises the latest research examples addressing developing countries[4].

While several research works have been done on using different types of mobile phone network data for specific purposes, each work has been done on a specific flavour of the data (different accuracy, granularity, aggregation level), and so it is difficult to understand whether a particular technique could indeed be applied to a different dataset, and which results that would provide. At the same time, if a researcher or practitioner is interested in building a specific urban sensing application, it is difficult for him/her to figure out which particular mobile phone network data set would be the most suitable, and which techniques should be applied to the data to achieve the specific goal. This paper surveys the new ideas and techniques related to the use of telecommunication data for urban sensing, with the specific goal to help researchers and practitioners to navigate the variety of mobile phone network datasets and associated processing techniques, that have been presented in the literature to build urban sensing applications. More specifically, Section 2 shows what telecoms data can tell about urban dynamics. Section 3 outlines the mechanisms at the basis of mobile phone data generation. Section 4 surveys the

[1] `http://www.d4d.orange.com`
[2] `http://dynamicinsights.telefonica.com/674/the-details`
[3] `http://www.telecomitalia.com/tit/en/bigdatachallenge.html`
[4] `http://www.unglobalpulse.org/Mobile_Phone_Network_Data-for-Dev`

filtering and processing techniques proposed so far to extract insights from this data, and summarises them to provide recommendations on which datasets and techniques to use for specific applications. 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), and 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 (e.g. transportation mode), consumption (e.g. energy, water, waste) and environmental impact (e.g. noise, pollution). In fact, this kind of information offers new insights about the city (see for example the *Villevivante* project[5]), 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 to the likely need for parking spaces. As an example, Singapore's public transportation is already using ten-year demand forecasts partly based on personal location data to plan transit needs[6], and is continuing investing in this direction through the Future Urban Mobility initiative[7]. 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.

One important pioneering work in the field of community dynamics sensing using cellphone data has been conducted within the *"Reality Mining"* project[8]. Reality mining deals with the collection and analysis of machine-sensed environmental data pertaining to human social behavior, with the goal of identifying predictable patterns of behavior. Mobile phones (and similarly innocuous devices) are used for data collection, opening social network analysis to new methods of empirical stochastic modelling (More and Lingam, 2013). The Reality Mining project collected data by asking volunteers to carry cellphones programmed to measure and store sensor

[5] http://villevivante.ch

[6] http://www.onemotoring.com.sg/publish/onemotoring/en/on_the_roads/traffic_management.html

[7] http://smart.mit.edu/research/future-urban-mobility/future-urban-mobility.html

[8] http://realitycommons.media.mit.edu/

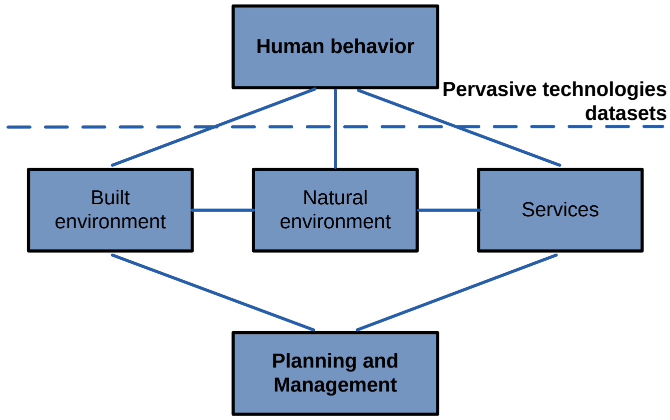


Fig. 1. Schema reflecting the role of pervasive technologies data sets in an urban scenario.

data. In this survey instead we focus on data opportunistically collected by the telecom operators by product of their own operations, without the requirements of people to carry specific devices or agree to install or enable specific features on their phones. Of course, privacy is a real issue in using this kind of technology. In fact 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 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 regulates the treatment of personal data and protection of intimacy in the electronic communications sector[9]. 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 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 was 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 research efforts:

(1) **Estimating population distribution.** With this regard, the use of mobile phone network data is twofold: *(i)* estimate where people live and *(ii)* estimate

[9] **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`

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 that are meaningful to users. Authors in (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 can 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. In (Reades et al., 2007) authors monitor the dynamics of Rome and obtain clusters of geographical areas measuring cell phone towers activity. Other works try to focus on the specific land use of a city. For example, in (Soto and Frias-Martinez, 2011) the authors use time series analysis to automatically identify land uses from aggregated 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.

(3) **Estimating mobility patterns.** Using the cell phone id, timestamp and location data of an event (call, sms, internet usage) it is possible to estimate commuters mobility in predefined regions. Several groups of researchers did an extensive work in this field. To name a few, the *"Barabasi Lab"* [10] has an open project on "Individual Mobility Patterns". 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. In (Song et al., 2010) authors investigated to what degree is human behavior predictable with results indicating that the development of accurate predictive models is a scientifically grounded possibility, with potential impact on our well-being and public health. Moreover, they analyzed several aspects of mobility patterns ranging from human trajectories (Song et al., 2010) to migration (Simini et al., 2012) and road usage patterns (Wang et al., 2012).

[10] `http://www.barabasilab.com`

Several other important works in this area have been conducted by the *"MIT Senseable city lab"* [11]. Their works aim to investigate and anticipate how digital technologies are changing the way people live and their implications at the urban scale. In particular, in their first works authors used the real-time data collected from mobile phones to monitor the vehicular traffic status and the movements of pedestrians in Rome, Italy (Calabrese et al., 2011). Finally, (Becker et al., 2013) characterised human mobility in several US cities to derive insight into a variety of important social issues such as evaluating the effect of human travel on the environment.

(4) **Analyzing local events.** The increasing availability of mobile phone usage data sets in recent years has led to a number of studies also related to local events and the interplay with mobility. In particular, several works tried to infer the human patterns of mobility during emergencies and special events (Bagrow et al., 2011; Calabrese et al., 2010; Ferrari et al., 2012; Lu et al., 2012; Traag et al., 2011).

(5) **Analyzing the geography of social networks.** The impact of geography on social interactions 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) and to study social radius of influence at both communication and mobility scale (Calabrese et al., 2011).

Mobile phone network data has been used not only in research works but also in running products based on both aggregated and individual data. A first group of applications deals with the issue of using mobile phone network data to derive urban traffic. Traditional companies (such as Inrix[12] and Delcan[13]) 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 telecom operators are making partnerships with external companies that can provide real-time services using traffic information, see for instance the partnership between Vodafone and TomTom[14]. For example, Cellint[15] 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[16] that provides real-time traffic and incidents information in Canada. Another company that leverage mobile phone network

[11] `http://senseable.mit.edu`
[12] `http://www.inrix.com`
[13] `http://delcantechnologies.com`
[14] `http://enterprise.vodafone.com/insight_news/case-study/tomtom.jsp`
[15] `http://www.cellint.com`
[16] `http://www.intellimec.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

Table I. Pros and cons of the main comparative data sets.

data to provide traffic information is Airsage[17], which 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 insight into the behavior of consumers at specific locations and at different times during the day. A similar approach has been taken by Telefonica, with the Smart Steps product[18], which uses anonymised and aggregated mobile network data to provide insights representative of the total population in each area and time.

Other applications focus on using mobile phone network data to provide services based on a more "social" aspect. For example Sense Networks[19] 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 showed 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.

The examples above show how 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 validate the extracted insights. 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 I outlines the main comparative data sets commonly used to validate the results obtained from mobile phone network data and highlights their pros and cons.

[17] http://www.airsage.com
[18] http://dynamicinsights.telefonica.com/488/smart-steps
[19] http://www.sensenetworks.com

In particular:

Census and Surveys. Census and surveys provide dataset related to very different areas: demography, health, education, government and security, communication and transport, etc. (see for example the 2010 US Census[20]). Such data set can be used to: *(i)* validate home and working areas; *(ii)* validate city patterns such as hotspots, commuting, traffic flows; *(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 offer access to a number of datasets that characterize an area based on its planned use (e.g., the NASA Global Land Use Datasets[21]). 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.

Finally, another limitation in the use of mobile phone data to estimate urban dynamics, is due to the potential biases in differential ownership of phones among different demographic groups. A recent study however (Wesolowski et al., 2013) has showed that for the purpose of estimating human mobility, mobile phone data from a large telecom operator in Kenya seemed robust to biases in phone ownership across different geographical and socioeconomic groups. While this study does not automatically generalises to any mobile phone network dataset, it shows that for large enough samples, the biases have a low impact on the extracted mobility patterns.

In the next section we will discuss how telecommunication networks generate the mobile phone datasets, and their features.

3. MOBILE PHONE NETWORK DATA GENERATION

When a mobile phone is switched on, it regularly 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

[20] `http://2010.census.gov`
[21] `http://data.giss.nasa.gov/landuse/`

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)).

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 records 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. Table II shows an example of a CDR log, as well as a mapping between cell ids and locations.

originating_id	originating_cell_id	terminating_id	terminating_cell_id	timestamp	duration
24393943	10121	17007171	10121	24031517	29
24393943	5621	17007171	2721	25141136	38
24393943	17221	17007171	2521	25534630	188
24393943	31041	17007171	5111	32440483	111
24393943	10121	17007171	9411	33152308	145
24393943	6321	17007171	20921	33431903	132
24393943	7041	17007171	10021	33435718	17
24393943	7021	17007171	14321	34160370	53

(a)

cell id	lat	lon
10121	44.658885	10.925102
17221	44.701606	10.628872

(b)

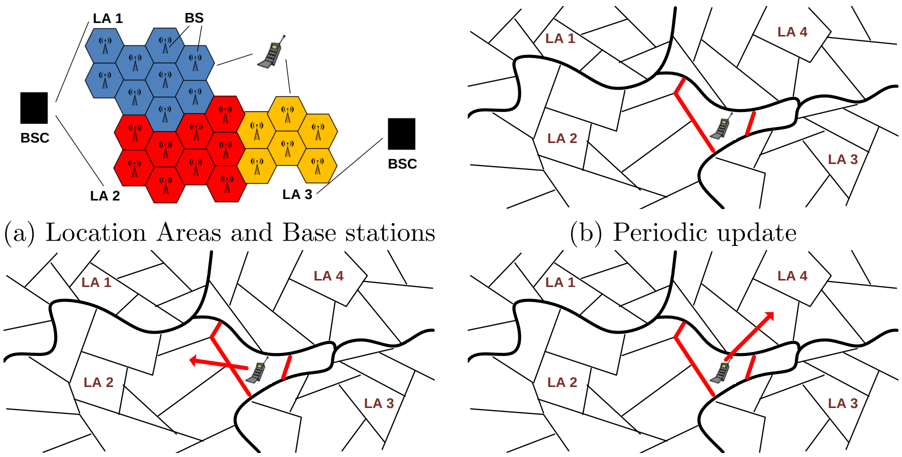
Table II. **(a)** Example of a CDR log: anonymized originating and terminating user id, originating and terminating cell id, timestamp and call duration (b) Cell location information.

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 2(a)).



(a) Location Areas and Base stations

(b) Periodic update

(c) Handover

(d) Mobility Location Update

Fig. 2. (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, 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 2 (b)).
(2) **Handover**, which is generated when a phone involved in a call moves between two cell areas (see Figure 2 (c)).
(3) **Mobility location update**, which is generated when the phone moves between two Location Areas (see Figure 2 (d)).

Location updates also happen when the phone changes type of connectivity it uses to access the telecommunication infrastructure (e.g., from 2G to 3G). The frequency of these updates strongly depend on how the operator has deployed the different connectivity technologies.

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. Table II(a) shows an example of a CDR location information, represented by the *cell id* field. Table II(b) maps each *cell id* to 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

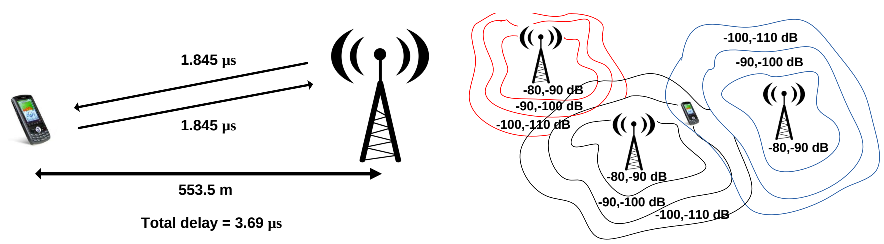
user hash	longitude	latitude	uncertainty	timestamp
4ba232e4d96f47dc94f7441e87c164fb	16	81	56	1246759931
4ba232e4d96f47dc94f7441e87c164fb	06	09	252	1246759922
4ba232e4d96f47dc94f7441e87c164fb	99	95	208	1246760034

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

documentation provided by networks operators (see the 3gpp standard documentation[22]). 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 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 3(a)).

(2) **Received Signal Strength (RSS)**, which is a measurement of the power present in the signal received by cell towers surrounding the phone. 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 3(b)).



(a) (b)

Fig. 3. 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 III 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 (for instance in meters) of the mobile phone position.

[22] `http://www.3gpp.org/`

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. 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, see Freeman (2005)), 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 data 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.

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)). First, we will describe some filtering techniques necessary to reduce noise in the data. Then, we will describe a list of features that can be extracted from mobile phone network data as well as the necessary processing techniques.

4.1 Filtering Techniques

In order to mine mobile phone network data to derive human 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 operator provided the best serving cell map, which associates to each location in a grid, what is the cell tower that best covers that location. 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 at a specific location, given the fact he/she is 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 be 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 consecutive location data,** where consecutive measurements which are close enough can be collapsed in a unique single measurement. 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 $(t_{stop_i} - t_{stop_j}) > T_{th}$. A similar approach was also used in (Jiang et al., 2013).

(2) **Solutions that leverage 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 spanning 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 summarise different techniques proposed in the literature to process mobile phone network data and extract insights into urban dynamics. These techniques have been categorised based on the aggregation level provided by the datasets.

4.2.1 *Individual data processing.* Mobile phone data at the individual level has been used in several applications:

(1) **Home and work location estimation.** Using CDR with location information, some works (Calabrese et al., 2011; Isaacman et al., 2011) have focused on estimating the home and work location of the users. The technique used to this goal involves selecting, for each user, a dataset consisting of several days of mobile phone network data. 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. Home location is then determined as the most frequented place during evenings (where an evening is characterised by a time internal to be specified), 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. Please note that when applying the technique to different countries, the time intervals to be used to identify evening and morning periods might have to be adjusted based on the working habits of the country, as discussed for instance in (Berlingerio et al., 2013).

(2) **Mobility estimation and applications.** By connecting the sequence of visited locations for each user and use that as an estimation of mobility, several researchers have proposed applications for mobility study. In (González et al., 2008), the authors proposed a technique to infer daily trips using the distance between any two different visited locations. Distance between the most two distant visited locations has been used in (Isaacman et al., 2011) as a measure of daily range of mobility. By grouping users' mobility by origin and destination of trips, Origin-Destination matrices can be inferred, and used to analyze the attractiveness of an area (measured as the number of different places people come from), see for example Calabrese et al. (2011). In (Couronne et al., 2011) users has been clustered on the basis of how often they move using spatio-temporal analysis. In Schneider et al. (2013), the authors associated daily mobility networks extracted from the sequence of trips in a day, called *motifs*, with trip chains extracted from travel diary surveys, and tried to associate a trip purpose by examining semantic-enriched land users surrounding destinations of individual's motifs (Jiang et al., 2013).
In (Berlingerio et al., 2013) authors have further exploited frequent travel patterns found in the mobile phone data in order to come up with recommendations for improving public transportation systems, by recommending the introduction of new routes in areas which experience high travel demand, which is however not matched by the current transit network.
Finally, using mobility patterns extracted from CDRs, researchers have defined a model that describes how diseases could spread across the country (Lima et al., 2013). This has led to testing the effect of information campaigns in containing the disease spread.

(3) **Integrating social and mobility information.** A first group of works tries to understand the interplay of mobility patterns and social ties (Cho et al., 2011; Crandall et al., 2010; Pan et al., 2013; Wang et al., 2011) 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; Wu et al., 2008) 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.

4.2.2 *Aggredated data processing.* 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 statistics, 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. Several categories of activities can be considered. Classical time series analysis

is initially 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.** Mobile phone users location at call time can be used to infer local of callers and callers, thus allowing to model the effect of geography on human interactions. Using network analysis, in (Lambiotte et al., 2008) authors found that human interactions decrease as distance increases following a gravity-like behaviour. Exceptions emerge and are mainly due to: geographical features (e.g., rivers, see for example (Ratti et al., 2010)), administrative borders and cultural differences. Using aggregated CDR with location information, it is possible to measure the level of human interactions between places. This has led to several works focused on how to best cluster areas based on these interactions. 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 happen between people connected to the two cell towers. The weighted graph can be partitioned in communities using standard network analysis techniques (e.g., modularity optimization). Through that, researches can detect which areas in the city are most connected, and where interaction borders exist, see (Blondel et al., 2010). An important aspect to take into account while performing this study is the mobile phone penetration and share of the operator in each area under analysis. Indeed, if such share is not uniformly distributed over the entire area under analysis, the resulting interactions network might be distorted. This was one of the problems addressed in (Calabrese et al., 2011) when dealing with regional partitioning at the level of the entire USA. Starting from CDR data with location information aggregated at the county level, the 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 counties with a too low number of customers or share (to preserve representativeness of the sample). More recently, new methods have also been proposed to estimate the significance of the association between geographical divisions of the population originating in ethics, language, religious or political differences (Bucicovschi et al., 2013). A study has been data on the Ivory Coast, to take into account the 60 local majority languages spoken.

(3) **Event detection.** Looking at time series of call tower-to-tower communications, researchers have proposed a visual analytics tool to characterise events (van den Elzen et al., 2013). This tool identifies clusters of cell towers having similar call behaviour to detect events. The characterisation can be further refined by introducing individual data to identify whether mobile phone users are unusually found in the specific location where the mobile traffic anomaly was detected (Traag et al., 2011). A Threshold to detect these outliers has to be imposed and tuned based on partial ground truth on historical events.

Based on what discussed above, Table IV summarises our recommendations on which datasets and processing techniques should be used to develop specific urban sensing applications.

Application	Preferred dataset	Processing techniques	Observations
Estimating population density	Individual CDR with cell tower location information	Home location determination	Test different temporal thresholds for home location determination
Estimating types of activities in different parts of the city (e.g. land use)	Aggregated cell tower statistics	Time series clustering (e.g. Fuzzy K-means)	Can be improved with the help of external data, e.g. POIs
Estimating mobility patterns (Origin/Destination matrices)	Individual CDR with cell tower location information	Home and work location determination, mobility estimation	Evaluate feasibility to map match on transport network
Estimating mobility patterns (Traffic monitoring)	Individual Event-driven triangulated location	Mobility estimation, Mode inference	Evaluate availability of data in real time.
Detecting events	Individual or Aggregated CDR with cell tower location information	Mobility estimation, Event identification	Test detection thresholds on partial ground truth
Analysing the geography of social networks (Regional partitioning)	Aggregated CDR with cell tower location information	Modularity partitioning	Test different definitions of weights on edges
Analysing the geography of social networks (communication-mobility interplay)	Individual CDR with cell tower location information	Mobility estimation, Social network analysis	Use reciprocate calls to identify social ties. Use location at call time to identify co-location

Table IV. Urban sensing applications and associated datasets and processing techniques

5. OPEN CHALLENGES

In this paper we have shown how mobile phone network data can be used to gain insights into urban dynamics. 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 3, 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 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 does not always represent a general sample of citizen's behavior.

—*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 to use 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 3, 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 the nearest recurring location to a noisy position (because of the low accuracy in the localization), see (Isaacman et al., 2011).

—*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.

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

—*Self-reported data.* Self-reported data can provide additional value compared to traditional data since they might be more spatially accurate, not outdated and with a frequent sampling time to make comparisons. An example of self-reported data is the one that can be obtained from Flickr[23] 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[24] that provides public access to their own data and have been recently used to support urban analysis, see for instance (Noulas et al., 2013).

(5) **Dealing with privacy and anonymity.** The sharing of mobility data raises serious privacy concerns. Mobility data can reveal the mobility behavior of the

[23] `http://www.flickr.com`
[24] `http://foursquare.com`

people: where they are going, where they live, where they work, their religion preferences, etc. All this information refers to the private personal sphere of a person and so may potentially reveal many facets of his/her private life. As a consequence, this kind of data has to be considered personal information to be protected against undesirable and unlawful disclosure. A recent study by (de Montjoye et al., 2013) showed that knowing four spatio-temporal points are enough to uniquely identify 95% of the individuals. Thus sophisticated techniques should be designed to protect the privacy of individuals. Many privacy enhancing technologies for mobility data have been proposed by the scientific community, see (Giannotti and Pedreschi, 2008) for a review on privacy in mobility data. In particular, 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 $k - 1$ distinct individuals whose associated trajectories are indistinguishable from the former (see (Gedik and Liu, 2008) for more detailed information).

Very recently, (Mir et al., 2013) have also proposed a method, validated against billions of location samples from a real telecommunication network, to generate synthetic Call Detail Records to capture the mobility patterns of real metropolitan populations while preserving privacy.

This is just the tip of the iceberg. The concerns that people have over the collection of this data will naturally extend to any analytic capabilities applied to the data, even the ones which try to preserve users' privacy. Users of data mining should start thinking about how their use of this technology will be impacted by legal issues related to privacy. A critical evaluation of data mining and privacy was released in a report saying that data mining *"may be the most fundamental challenge that privacy advocates will face in the next decade.."*, (Cavoukian, 1998). The report looks at data mining and privacy in the context of the international "fair information practice" principles.

These collisions between data mining and privacy are just the beginning. Over the next few years we should expect to see an increased level of scrutiny of data mining in terms of its impact on privacy. The sheer amount of data that is collected about individuals, coupled with powerful new technologies such as data mining, will generate a great deal of concern by consumers. Unless this concern is effectively addressed, we expect to see legal challenges to the use of data mining technologies.

(6) **Mobility/communication interplay.** Studying 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 effect (Sasakia and Nishiib, 2010). 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. Still a lot of work has to be done in this area to fully characterise the real interplay.

(7) **Real Time data acquisition and processing.** Many urban sensing applications (e.g., traffic monitoring, event management) 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 for ad-hoc algorithms and platforms to process such data in real time. Some proposed solutions are based streaming platforms able to deal with different types of data in real time, see for example (Kaiser and Pozdnoukhov, 2013).

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. Telecom operators are nowadays generating terabytes of records of potential use for urban sensing. 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. Privacy is also a very sensitive issue, that had to be addressed. 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/`.

Ahas, R., Silm, S., Järv, 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.

Bagrow, J. P., Wang, D., and Barabási, A.-L. 2011. Collective response of human populations to large-scale emergencies. *PLoS ONE 6,* 3.

Becker, R., Cáceres, R., Hanson, K., Isaacman, S., Loh, J. M., Martonosi, M., Rowland, J., Urbanek, S., Varshavsky, A., and Volinsky, C. 2013. Human mobility characterization from cellular network data. *Commun. ACM 56,* 1 (Jan.), 74–82.

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.

Berlingerio, M., Calabrese, F., Lorenzo, G., Nair, R., Pinelli, F., and Sbodio, M. 2013. Allaboard: A system for exploring urban mobility and optimizing public transport using cellphone data. In *Machine Learning and Knowledge Discovery in*

Databases, H. Blockeel, K. Kersting, S. Nijssen, and F. Železný, Eds. Lecture Notes in Computer Science, vol. 8190. Springer Berlin Heidelberg, 663–666.

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

Bucicovschi, O., Douglass, R. W., Meyer, D. A., Ram, M., Rideout, D., and Song, D. 2013. Analyzing social divisions using cell phone data. In *Proceedings of the 3rd International Conference on the Analysis of Mobile Phone Datasets (NetMob'13)*. Boston, Massachusetts, USA.

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., Xiaoji, C., Rowland, J., Rath, C., and Ratti, C. 2011. The connected states of america: Quantifying social radii of influence. In *Privacy, security, risk and trust*. SocialCom '11.

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. Floréen, A. Krüger, 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.

Cavoukian, A., Ed. 1998. *Data Mining: Staking a Claim on Your Privacy.* Ontario.

Cho, E., Myers, S. A., and Leskovec, J. 2011. Friendship and mobility: User movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '11. ACM, New York, NY, USA, 1082–1090.

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 *SocialCom/PASSAT'11*. 1425–1430.

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

Crandall, D. J., Backstrom, L., Cosley, D., Suri, S., Huttenlocher, D., and Kleinberg, J. 2010. Inferring social ties from geographic coincidences. *Proceedings of the National Academy of Sciences 107,* 52 (Dec.), 22436–22441.

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*.

de Montjoye, Y.-A., Hidalgo, C. A., Verleysen, M., and Blondel, V. D. 2013. Unique in the Crowd: The privacy bounds of human mobility. *Scientific Reports 3*.

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.

FREEMAN, R. L. 2005. *Fundamentals of Telecommunications.* John Wiley.

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

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.

GIANNOTTI, F. AND PEDRESCHI, D., Eds. 2008. *Mobility, Data Mining and Privacy - Geographic Knowledge Discovery.* Springer.

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.

JIANG, S., FIORE, G. A., YANG, Y., FERREIRA, JR., J., FRAZZOLI, E., AND GONZÁLEZ, M. C. 2013. A review of urban computing for mobile phone traces: Current methods, challenges and opportunities. In *Proceedings of the 2Nd ACM SIGKDD International Workshop on Urban Computing.* UrbComp '13. ACM, New York, NY, USA, 2:1–2:9.

KAISER, C. AND POZDNOUKHOV, A. 2013. Enabling real-time city sensing with kernel stream oracles and mapreduce. *Pervasive and Mobile Computing 7,* 5, 708–721.

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., DE KERCHOVE, C., HUENS, E., PRIEUR, C., SMOREDA, Z., AND DOROREN, P. V. 2008. Geographical dispersal of mobile communication networks. *Physica A: Statistical Mechanics and its Applications* 387, 5317 – 5325.

LIMA, A., DOMENICO, M. D., PEJOVIC, V., AND MUSOLESI, M. 2013. Exploiting Cellular Data for Disease Containment and Information Campaigns Strategies in Country-wide Epidemics. In *Proceedings of the 3rd International Conference on the Analysis of Mobile Phone Datasets (NetMob'13).* Boston, Massachusetts, USA.

LU, X., BENGTSSON, L., AND HOLME, P. 2012. Predictability of population displacement after the 2010 haiti earthquake. *Proceedings of the National Academy of Sciences*.

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.*

MIR, D. J., ISAACMAN, S., CACERES, R., MARTONOSI, M., AND WRIGHT, R. N. 2013. Dp-where: Differentially private modeling of human mobility. In *Big Data, 2013 IEEE International Conference on*. 580–588.

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

MORE, J. AND LINGAM, C. 2013. Current trends in reality mining. *IRJES 2,* 2 (2), 35–39.

NOULAS, A., MASCOLO, C., AND FRÍAS-MARTÍNEZ, E. 2013. Exploiting foursquare and cellular data to infer user activity in urban environments. In *MDM*. 167–176.

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.

PAN, W., GHOSHAL, G., KRUMME, C., CEBRIAN, M., AND PENTLAND, A. 2013. Urban characteristics attributable to density-driven tie formation. *Nature Communications 4,* 1961.

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,* 3, 30–38.

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.

SCHNEIDER, C. M., BELIK, V., COURONNE, T., SMOREDA, Z., AND GONZALEZ, M. C. 2013. Unravelling daily human mobility motifs. *J. R. Soc. Interface 10,* 84.

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.

SIMINI, F., GONZÁLEZ, M. C., MARITAN, A., AND BARABÁSI, A.-L. 2012. A universal model for mobility and migration patterns. *Nature 484,* 7392, 96–100.

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.

SONG, C., KOREN, T., WANG, P., AND BARABÁSI, A.-L. 2010. Modelling the scaling properties of human mobility. *Nature Physics 6,* 10, 818–823.

SONG, C., QU, Z., BLUMM, N., AND BARABÁSI, A.-L. 2010. Limits of predictability human mobility. *Science 327,* 2, 1018–1021.

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.

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

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.

van den Elzen, S., Blaas, J., Holten, D., Buenen, J.-K., van Wijk, J. J., Spousta, R., Miao, A., Sala, S., and Chan, S. 2013. Exploration and Analysis of Massive Mobile Phone Data: A Layered Visual Analytics approach. In *Proceedings of the 3rd International Conference on the Analysis of Mobile Phone Datasets (NetMob'13)*. Boston, Massachusetts, USA.

Wang, D., Pedreschi, D., Song, C., Giannotti, F., and Barabási, 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, P., Hunter, T., Bayen, A. M., Schechtner, K., and Gonzalez, M. C. 2012. Understanding road usage patterns in urban areas. *Sci. Rep 2*.

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

Wesolowski, A., Eagle, N., Noor, A. M., Snow, R. W., and Buckee, C. O. 2013. The impact of biases in mobile phone ownership on estimates of human mobility. *J. R. Soc. Interface 10,* 81.

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.

Wu, L., Waber, B. N., Aral, S., Brynjolfsson, E., and Pentland, A. 2008. Mining face-to-face interaction networks using sociometric badges: Predicting productivity in an it configuration task. In *ICIS*. Association for Information Systems, 127.

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.