Users in the Urban Sensing Process: Challenges and Research Opportunities

Thiago H. Silva[†◊*], Clayson S. F. de S. Celes[◊], João B. Borges Neto[◊], Vinícius F. S. Mota[◊], Felipe D. da Cunha[◊], Ana P. G. Ferreira[◊], Anna I. J. T. Ribeiro[◊], Pedro O. S. Vaz de Melo[◊], Jussara M. Almeida[◊], Antonio A. F. Loureiro[◊]

[◊] *Universidade Federal de Minas Gerais*
Department of Computer Science
Belo Horizonte, MG, Brazil

[†] *Universidade Tecnológica Federal do Paraná*
Department of Informatics
Curitiba, PR, Brazil

Abstract

The popularization of portable devices such as smartphones and the worldwide adoption of social media services make it increasingly possible to be connected and share data anywhere, anytime. Data from this process represent a new source of sensing, which is called participatory sensor network (PSN). In this scenario, people participate as social sensors voluntarily providing data that capture their experiences of daily life. This large amount of social data can provide new valuable forms to obtain information that is currently not available within the same global reach and be used to improve decision-making processes of different entities (e.g., people, groups, services, and applications). The objective of this chapter is to discuss participatory sensor networks, presenting an overview of the area, challenges, and opportunities. We aim to show that PSNs (e.g., Instagram, Foursquare, and Waze) can act as valuable sources of large scale sensing, providing access to important characteristics of city dynamics and urban social behavior, more

[*] Corresponding author. Email: thiagohs@dcc.ufmg.br. Address: Av. Antônio Carlos, 6627. Universidade Federal de Minas Gerais, ICEX. Pampulha. CEP 31270-010. Belo Horizonte, MG, Brasil. Tel.: +55 31 34095863. Fax.: +55 31 34095858. Thiago is now with Universidade Tecnológica Federal do Paraná, where he finished this work.

quickly and comprehensively. This chapter start studying the properties of PSN. Next, it discusses how to work with PSN, showing its applicability in the development of more sophisticated applications. In addition, it discusses several research challenges and opportunities in this area.

Keywords: Urban computing, location-based social networks, participatory sensor networks, city dynamics, urban social behavior, social media, challenges, opportunities

Contents

1. Introduction

The study of urban data provided by users in Participatory Sensor Networks (PSNs) is a recent research area. PSNs allow large scale observation of people's actions in (almost) real time during long periods of the time. With that, PSNs have the potential to become a fundamental tool to better understand urban human interaction in the future. Data from PSNs can increase our knowledge over different aspects of our life in urban scenarios, which can be useful in the development of more sophisticated applications in several segments, such as, in the urban computing area [82].

Furthermore, PSNs have the potential to complement traditional Wireless Sensor Networks (WSNs) [2] in several aspects. While WSNs are designed to sense limited size areas, such as forests and volcanoes, PSNs can reach areas of varying size and scale, such as large cities, countries, or even the planet. Additionally, a WSN is more subject to failure, since its operation depends on proper coordination of actions of its sensor nodes that have severe power, processing, and memory constraints. On the other hand, PSNs are formed by autonomous and independent entities, i.e., humans with their mobile devices. This makes the sensing task highly resilient to individual failures.

The objective of this chapter is to discuss the concept of participatory sensor networks, presenting an overview of the area, research trends and the main challenges. We aim to show that the PSNs (e.g., Instagram[1], Foursquare[2], and Waze[3]) can act as valuable sources for large scale sensing, providing access to important features of city dynamics and urban social behavior quickly and comprehensively. First, we analyze the properties of PSN data studied on various systems. Next, we discuss how to work with PSN data, showing its applicability in the development of more sophisticated applications. Furthermore, we discuss several challenges and research opportunities related to participatory sensor networks.

The remainder of this chapter is organized as follows. Section 2 discusses the emerging concept of participatory sensor networks. Section 3 presents the properties of PSN. Section 4 discusses how to work with PSN data, including how to obtain them. Section 5 presents challenges and opportunities about current research topics related to PSNs. Finally, Section 6 presents

[1] http://www.instagram.com.

[2] http://www.foursquare.com.

[3] http://www.waze.com.

the conclusions.

2. Participatory Sensor Networks

There are several ways to get urban data, among them we can mention the emerging participatory sensor networks (PSNs) [12, 82]. Section 2.1 presents the definition of PSN; Section 2.2 discusses the functioning of PSN, while Section 2.3 illustrates examples of PSNs.

2.1. What is Participatory Sensor Network?

Participatory sensor network rely on the idea of participatory sensing [12], and can be defined as a system that supports a distributed process of gathering data about personal daily experiences and various aspects of the city. Such a process requires the active participation of people using portable devices to voluntarily share contextual information and/or make their sensed data available, i.e., the users manually determine how, when, what, and where to share the sensed data. Thus, through PSNs we can monitor different conditions of cities, as well as the collective behavior of people connected to the Internet in (almost) real time [82].

PSNs have become popular thanks to the increasing use of portable devices, such as smartphones and tablets, as well as the global adoption of social media services. Therefore, a central element of a participatory sensor network is a user with a portable computing device. In this scenario, people participate as social sensors, voluntarily providing data on a particular aspect of a place that implicitly capture their experiences of daily life. These data can be obtained with the aid of sensing apparatus, for example, sensors embedded in smartphones (e.g., GPS, accelerometer, microphone, and so on) or by human sensors (e.g., vision). In the latter case, data are subjective observations produced by the users [82].

PSNs provide unprecedented opportunities to access sensing data on a global scale. This large amount of data ease the gathering of information that is not promptly available with the same global reach, and can be used to improve the processes of decision making of different entities (e.g., individuals, groups, services, and applications).

It is worth mentioning that several terms defined recently, for example, *Humans as Data Sources* and *Ubiquitous Crowdsourcing* reflect basically the idea of participatory sensor networks [91, 67, 34]. It is also important to mention that the term opportunist sensing [54], which is a type of sensing that users also uses portable computing devices in the sensing process, can lead to confusion with the term participatory sensing. Participatory sensing differs from opportunistic sensing mainly by the user participation, where in the latter case the data collection stage is automated without user participation [55, 54]. Opportunistic sensing supports the sensing process of an application without requiring user efforts, determining automatically when the devices should be used to meet specific demands of the applications. Thus, applications can take advantage of the sensing capabilities of all devices of users of the system without the need of human intervention in this process [55].

2.2. The Functioning of PSN

Similarly to traditional wireless sensor networks, data sensed in a PSN is sent to the server, or "sink node", where the data can be accessed (using, for example, APIs, such as the API of Instagram[4]). But unlike WSNs, PSNs have the following characteristics: (a) sensor nodes are autonomous mobile entities, i.e. a person with a mobile device; (b) the cost of the network is distributed among the sensors, providing a global scale; (c) the sensing depends on the willingness of people to participate in this process; and (d) sensor nodes do not have severe limitations of energy.

PSNs have the potential to complement WSNs in several aspects. Traditional wireless sensor networks are designed to sense areas of limited size, such as forests and volcanoes. In contrast, PSNs can reach areas of different sizes and scale, such as large cities, countries or even the planet [82]. Additionally, a WSN is subject to failure, since its operation depends on proper coordination of actions of its sensor nodes that have severe power, processing, and memory constraints. Since PSNs are formed by autonomous and independent entities, human beings, the sensing task becomes more robust to individual failures. Obviously, PSN also bring several new challenges, for example, its success is directly linked to the popularity of smartphones and social media services.

Figure 1 illustrates the idea of PSN consisting of users with their mobile devices sending sensed data about their locations for systems in the Internet. The figure shows sharing activities (represented by dots in the cloud) of four users in three different moments in time, labeled as "Time 1", "Time 2", and

[4] http://instagram.com/developer.

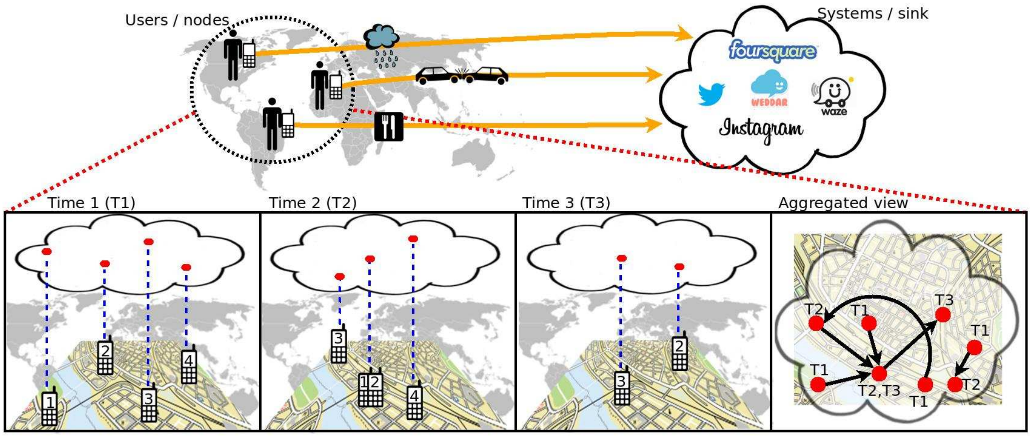


Figure 1: Ilustration of participatory sensor network (image from [82]).

"Time 3". Note that a user does not participate necessarily in the system at all times. After a certain time, this data can be analyzed in different ways. For example, the bottom rightmost part of the figure shows, by an aggregate view, a directed graph in which nodes/vertices represent the locations where the data has been shared, with edges connecting locations that were shared by the same user. Using this graph we can extract, for instance, mobility patterns of users that can be used to perform load management more efficiently in urban infrastructure of mobile networks. In fact, knowledge discovery in PSNs goes together with the use of graph/networks theory [29, 71, 70].

2.3. Examples of PSNs

Location-based social networks, which are a special kind of social media that combine online social networks[5] features and the possibility of share data with spatial and temporal information[6] can be considered the most popular examples of PSNs. It is possible to find several examples of such systems already deployed on the Internet, such as Waze, which serves to report traffic conditions in real time; Foursquare to share where the user is visiting; and Instagram, to send real time images to the system. In particular, Instagram can be seen as one of the most popular PSN, with 200 million

[5] Virtual platform that built and reflects social relations of real life among people.

[6] Data type that allows, for example, building location-based services.

users [44]. When considering this network, the sensed data is a picture of a specific place. We can extract information of such data in various ways. One possibility is to visualize in real time how is the situation in a certain area of the city.

Note that all described systems above are composed of an online social network. However, there are several examples of PSNs that do not contain online social networks. For example, Weddar[7] to report weather conditions, NoiseTube[8] to share noise level in a given region of the city, or Colab[9] for sharing various problems of cities.

Some other types of social media, such as Twitter[10], which allows its users to share personal updates in texts up to 140 characters, known as “tweets”, may also be examples of PSNs. Twitter is considered an example of PSN because the content shared on it may also enable the monitoring of various aspects of cities, as well as the collective behavior of people in near real time. For example, people could use their portable devices to share tweets containing real time information about demonstrations or accidents in the city. Beyond these examples, we can also mention GarbageWatch [15] to monitor garbage aspects of a city. This example is particularly interesting because it illustrates that the use of the Web is not mandatory in a PSN. Sensed data can be sent to a specific application running on the Internet but outside the Web.

3. Properties of PSN

Many questions arise from the concept of participatory sensor networks (PSNs). Among them, one key question is: what are the properties of PSN? Answering this question helps us to understand, for instance, what are the limitations of PSNs and what type of applications we can use data from PSNs.

As data provided by PSNs can be complex, a key step in any investigation is to characterize the data collected in order to understand their challenges and usefulness. Thus, in this section we study the properties of three participatory sensor networks for location sharing, namely, Foursquare, Gowalla

[7] http://www.weddar.com.

[8] http://noisetube.net.

[9] http://www.colab.re.

[10] http://www.twitter.com.

Table 1: Description of used datasets.

Location sharing services		
System	*# check-ins*	Interval
Foursquare1	≈5 milion	April 2012 (1 week)
Foursquare2	≈12 milion	Feb2010-Jan2011
Foursquare3	≈4 milion	May 2013 (2 weeks)
Gowalla	≈6 milion	Feb2009-Oct2010
Brightkite	≈4 milion	Apr2008-Oct2010
Photo sharing services		
System	*# of Photos*	Interval
Instagram1	≈2 milion	Jun2012-Jul2012
Instagram2	≈2 milion	May 2013 (2 weeks)
Traffic alert services		
System	*# of alerts*	Interval
Waze	+212 thousands	Dec2012-Jun2013

and Brightkite[11]. In addition, we also study a PSN for photo sharing, particularly Instagram, as well as a PSN for traffic alert sharing (Waze).

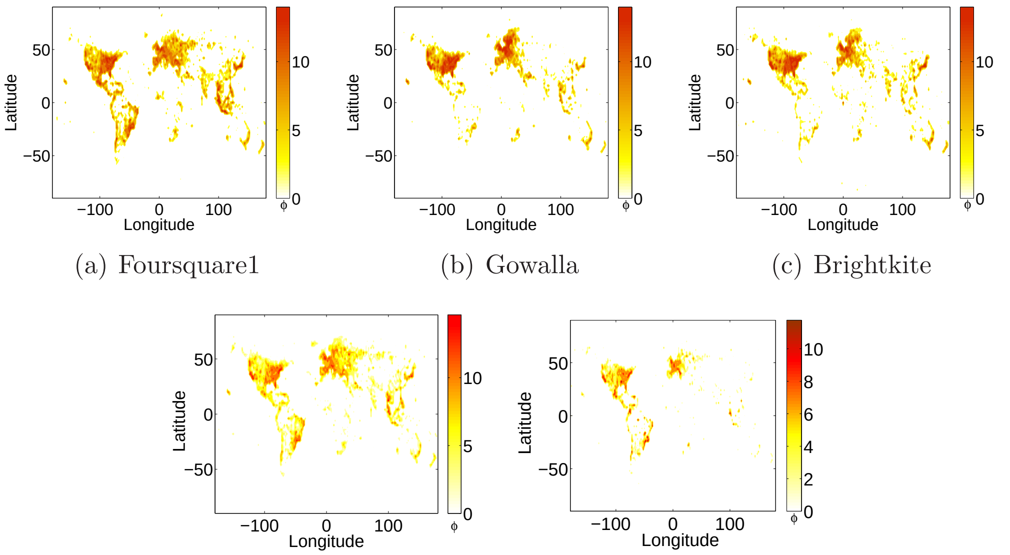
The rest of this section is organized as follows. Section 3.1 describes the datasets of the PSNs used in this chapter. Next, Section 3.2 analyzes the coverage of these PSNs in different spatial granularity. Section 3.3 discusses the frequency that nodes share data on individual regions of our dataset. Section 3.4 discusses the seasonality in the sensing process. Finally, Section 3.5 studies the behavior of the nodes of the PSNs.

3.1. Data Description

Table 1 displays all datasets considered in the analysis performed in this section. The data were collected through Twitter because in addition to plain text users can also share other types of contents, for instance, photos, check-ins, or traffic alerts, from an integration with Instagram, Foursquare, or Waze. In this case, Instagram photos, Foursquare check-ins, or Waze alerts announced on Twitter become available publicly, which by default does not happen when the data is published solely in the analyzed systems. As we can see in Table 1, the data reflect different periods. Furthermore, the datasets include a significant amount of data: Over 30 million records considering all sources.

Each sensed data (photo, check-in, or alert) consists of GPS coordinates (latitude and longitude), the data sharing time, and the id of the user who

[11] Gowalla and Brightkite are not in operation currently.



(a) Foursquare1 (b) Gowalla (c) Brightkite

(d) Instagram1 (e) Waze

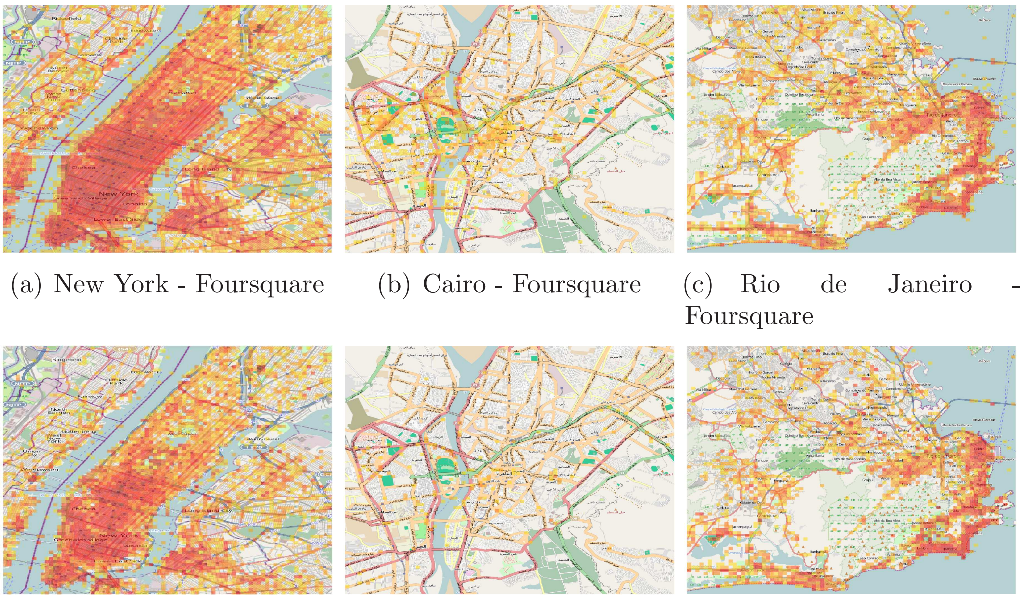
Figure 2: Coverage of PSNs. Number of data n per pixel indicated by the value of ϕ shown in the figure, where $n = 2^{\phi} - 1$ (images from [84, 87]).

shared the data. Foursquare1 dataset has extra information about the type of place: category (e.g., food) and a local unique identifier. More information about these specific datasets and how they were obtained can be found in [17, 83, 85, 86, 87]. Section 4.1, however, discuss how to obtain data from PSNs.

3.2. Network Coverage

In this section, we study the coverage of PSN at different spatial granularity, starting from the all globe, then cities and, finally, specific areas of a city. Figure 2 shows the global coverage in different PSNs: Foursquare (Foursquare1 dataset, Figure 2a); Gowalla (Figure 2b); Brightkite (Figure 2c); Instagram (Instagram1 dataset, Figure 2d); and Waze (Figure 2e). Data from these figures represent a heatmap of user participation: darker colors represents a larger number of shared data in a given area. As we can see, the coverage is very comprehensive and has a planetary scale.

Now we evaluate the participation of users in several large cities located in different regions, but we show the results only for some of them: New York,



(a) New York - Foursquare

(b) Cairo - Foursquare

(c) Rio de Janeiro - Foursquare

(d) New York - Instagram

(e) Cairo - Instagram

(f) Rio de Janeiro - Instagram

Figure 3: Spatial coverage of Foursquare and Instagram in 3 populous cities around the world (images from [84, 85]).

Rio de Janeiro and Cairo. Figure 3 shows a heatmap of sensing activity for each of these cities. Again, darker colors represent larger numbers of data in a given area. We observe a high coverage for some cities, as shown in Figure 3a and 3d (New York). However, as we can see in Figure 3b and 3e, the sensing in Cairo, city that also has a high number of inhabitants, is significantly lower. Such difference in coverage can be explained by several factors. Besides economic aspects, cultural differences can have a significant impact on the adoption and use of these considered systems in Cairo [6].

In addition, we can observe that the coverage in some cities, as in Rio de Janeiro (Figures 3c and 3f), is far more heterogeneous when compared with New York coverage. This is probably because of particular geographical aspects, i.e., large green areas and large portions of water. Rio de Janeiro has the largest urban forest in the world, located in the middle of the city, and many hills of difficult access for humans. These geographical aspects limit the sensing coverage. In addition, the points of public interest, such as tourist spots and shopping centers, are unevenly distributed around the city.

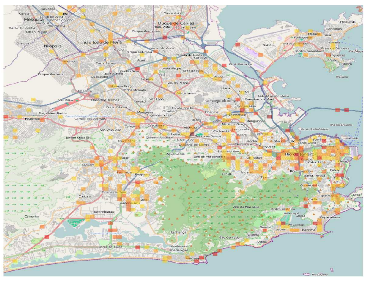


Figure 4: Spatial coverage of PSN for traffic alerts sharing in Rio de Janeiro (image from [87]).

There are large residential areas with few points of this type, while other areas have high concentrations of these points.

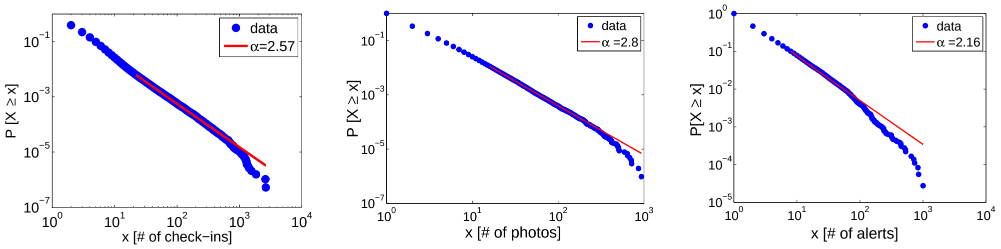
The spatial coverage of data of PSN for traffic alerts is not as comprehensive as PSNs for location and photo sharing. This can be seen in Figure 4, which shows the number of alerts in different regions of Rio de Janeiro by a heatmap. One factor that may help to explain this result is the population of users of the dataset of traffic alerts, which is smaller than the others studied. Another factor is that users may have fewer opportunities to share traffic alerts compared to opportunities to share photos or check-ins.

As the activity of participation can be quite heterogeneous within a city, we analyze the coverage of PSN in specific areas of a city. To have an id of a specific area of the city for datasets of Instagram and Waze, we propose to divide the area of the cities into smaller rectangular spaces, as in a grid[12]. We call each rectangular area of a *specific area* within a city. We consider that a specific area has the following definition: $1{\cdot}10^{-4\circ}$ (latitude) $\times$ $1{\cdot}10^{-4\circ}$ (longitude). This represents an area of approximately 8×11 meters in New York and 10×11 meters in Rio de Janeiro. For other cities, the areas can also vary slightly, but not enough to significantly affect the analysis.

Figure 5 shows the Complementary Cumulative Distribution Function (CCDF) of the number of shared data (check-ins, photos, or alerts) by specific area of all locations in our datasets. First, note that, in both cases, a power law[13] describes well this distribution. This implies that in most of the specific

[12] Note that in selected areas borders are not considered.

[13] Mathematically, a quantity x follows a power law if it can be obtained from a prob-



(a) Foursquare (b) Instagram (c) Waze

Figure 5: Distribution of number of data in specific areas in log-log scale (images from [84, 84, 87]).

areas there are few shared data, while there are a few areas with hundreds of shared data. These results are consistent with the results presented in [73], work that studied the participation of users in location sharing systems. In the analyzed systems, it is natural that some areas have more activity than others. For example, in tourist areas the number of photos shared tends to be higher than in a supermarket, although a supermarket is usually a popular site. If a particular application requires a more comprehensive coverage, it is necessary to encourage users to participate in places they normally would not. Micro-payments or scoring systems are examples of alternatives that could work in this case. We discuss these opportunities in Section 5.3.

We show that a PSN may have a global scale coverage. However, this coverage can be quite uneven, where large areas are practically uncovered. With that in mind, Figure 6 shows the percentage of different locations where users shared data in a given time interval in Instagram and Foursquare[14], which have 598,397 and 725,419 unique places, respectively. The maximum percentage of distinct places that have data shared on it per hour is less than 3% for all systems. This indicates that the instant coverage of these PSNs is very limited when we consider all locations that could be sensed on the planet (considering all the locations already sensed at least once). In other words, the probability of a random specific area be sensed at a random time is very low.

ability distribution $p(x) \propto x^{-\alpha}$, where α is a constant parameter known as exponent or scale parameter, and it is a value typically between $2 < \alpha < 3$ [19].

[14] We consider the datasets Instagram2 and Foursquare3 because they represent the

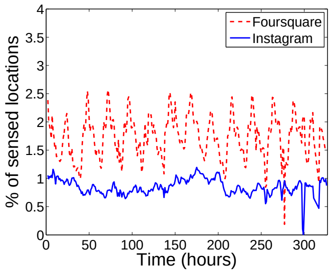


Figure 6: Percentage of specific areas sensed over time (image from [82]).

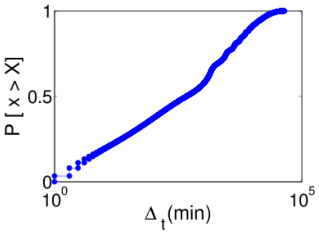
3.3. Sensing Interval

PSNs are very scalable because their nodes are autonomous, that is, users are responsible for their own operation and functioning. As the cost of infrastructure is distributed among the participants, this huge scalability and coverage is achieved more easily. The success of this type of network is to have sustainable and high quality participation. In other words, the sensing is efficient since users are kept motivated to often share their resources and sensed data.

This motivates the study of frequency that users perform data sharing in PSNs. In [85, 87, 84] the authors show that there are times when a lot of data are shared in interval of few minutes and times when there is no sharing for hours. This may indicate that the majority of data sharing occurs at specific intervals, probably related to the routine of people. For example, photo sharing in restaurants tends to happen more in the lunch and dinner hours. Applications based on this type of sensing should consider that user involvement can vary significantly over time.

Figure 7 shows the Cumulative Distribution Function (CDF) of the interval between photos shared by the same user on a popular specific area. We can see that a significant portion of users perform consecutive photos sharing in a short time interval. For example, about 20% of all observed photo sharing occurs within 10 minutes. This suggests that users tend to share more than one photo in the same area. Noulas et al. [73] also noted that a significant number of check-ins on Foursquare are performed within a short time. For example, more than 10% of check-ins occur within 10 minutes.

same time interval.



(a) Instagram

Figure 7: Cumulative distribution of the time interval between photos shared in a popular specific area (image from [85]).

3.4. Routines and Data Sharing

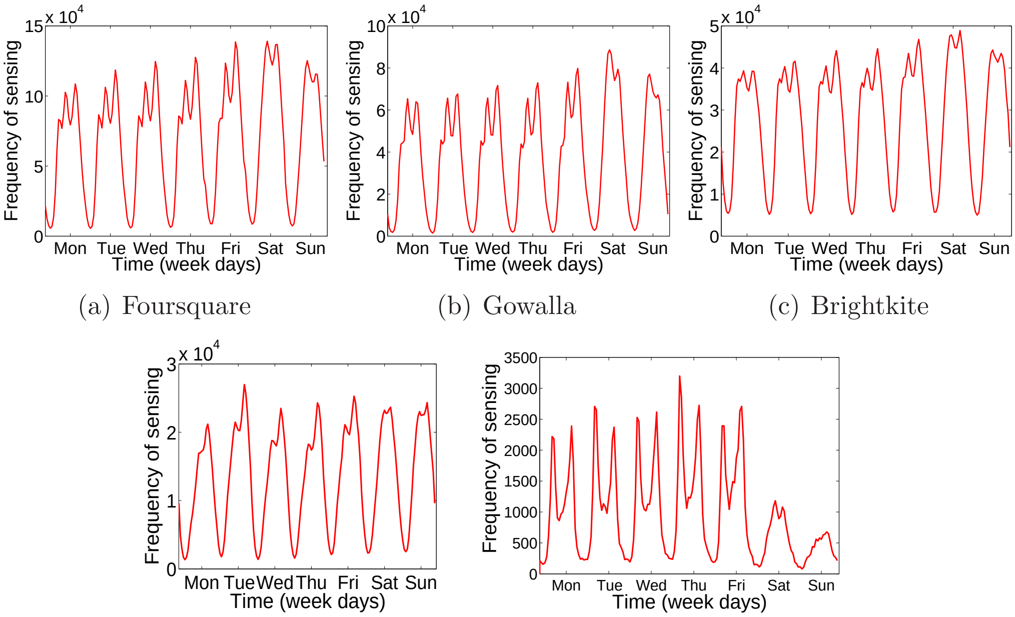
We analyze now how the routine of humans affect data sharing. Figure 8 shows the weekly data sharing pattern in all types of PSNs analyzed[15]. As expected, data shared in PSNs have a diurnal pattern, which implies that during the night the sensing activity is quite low.

Considering weekdays, we can see a slight increase in activity throughout the week, with few exceptions when there is a peak of activity. The study of Cheng et al. [17], who analyzed systems for location sharing, also observed this same behavior without any day as an exception.

We can also note that some activity peaks vary throughout the day according to the purpose of the PSN. As we can see in Figure 8, in PSN for location sharing (Figures 8a–c) there are three peaks evident around the breakfast, lunch, and dinner time. This was also noted by Cheng et al. [17]. In PSN for photo sharing (Figure 8d) there are only two obvious peaks occurring around lunch and dinner time. And in the case of PSN for traffic alerts sharing (Figure 8e) there are also two obvious peaks, one around 7:00am and 8:00am, and another around 6:00pm, coinciding with typical times of highest traffic intensity.

Figure 9 shows the temporal sharing pattern for Instagram and Foursquare considering all datasets. This figure shows the average number of data shared per hour during weekdays (Monday to Friday) and during weekend (Saturday and Sunday). Analyzing different patterns of behavior for weekdays and

[15] The sharing time was normalized according to the location where the data was shared, making use of geographic information of the location.



(a) Foursquare

(b) Gowalla

(c) Brightkite

(d) Instagram

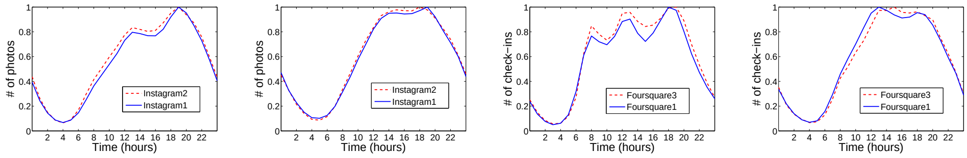
(e) Waze

Figure 8: Data sharing pattern during weekdays (images from [84, 85, 87]).

weekend we can see that the pattern is significantly different. Note that the peaks observed on weekdays are not evident on weekends. The lack of well defined routine on weekends is one of the possible explanations for that. Moreover, differences between the results for weekdays and weekends are related to the type of the analyzed system. For example, as on weekends many people do not need to drive, it is natural to expect a lower volume of data in Waze.

Surprisingly, we see that each sharing pattern is very similar, despite the huge gap between the samples (approximately one year). This happens for weekdays and weekends, suggesting that user behavior in both systems tends to remain consistent over time. This is an interesting and important result because it indicates that we can use different datasets of similar purposes.

We now show how the routines impact on the sharing behavior during the week. For this analysis, we consider the datasets of Instagram and Foursquare



(a) Instagram – weekday (b) Instagram – weekend (c) Foursquare – weekday (d) Foursquare – weekend

Figure 9: Temporal sharing pattern on Instagram and Foursquare (images from [86]).

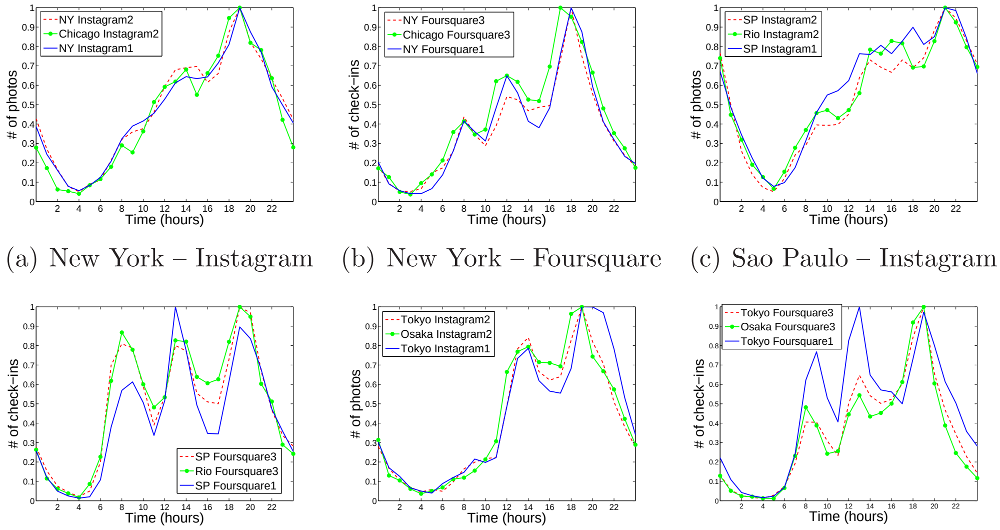
for New York, Sao Paulo, and Tokyo. The results are shown in Figure 10[16]. In all figures we display data from datasets of the same period (Instagram2 and Foursquare3) for two cities in the same country, and data from a dataset referring to a previous period (Instagram1 and Foursquare1) to one of these cities, as a comparison reference.

First, note the distinction between the curves of each city in the same system (e.g., in Instagram Figures 10a, 10c, and 10e) and in different systems (e.g., Figures 10a and 10b for New York). Then note that the sharing pattern for each city in the same country is quite similar, which may be a consequence of the cultural patterns of inhabitants of those countries. That is, in some way, a signature of cultural aspects, illustrating once again the potential of this type of data for the study of city dynamics and urban social behavior.

3.5. Node Behavior

In this section we analyze the performance of PSN nodes (i.e., users) regarding data sharing. Figure 11 shows the distribution of the number of data (photos and alerts) shared by each user in our database. As we can see, the distribution has a heavy tail, meaning that the participation of users can be very uneven. For example, about 40% of users contributed with only a picture during the period considered, while that 17 % and 0.1 % of users contributed with more than 10 and 100 pictures, respectively. It is natural that this variability occurs for several reasons. For instance, some users may give more importance to privacy questions than others. A heavy tail is also observed in the distribution of the number of check-ins, as shown by Noulas

[16] Each curve is normalized by the maximum number of shared content in a specific region representing the city.

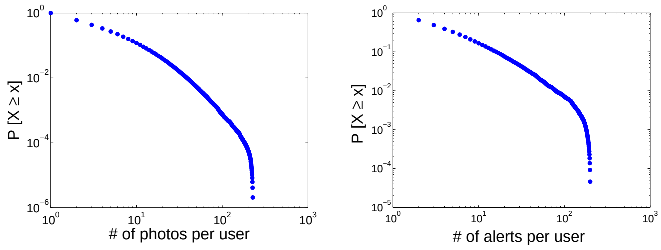


(a) New York – Instagram (b) New York – Foursquare (c) Sao Paulo – Instagram

(d) Sao Paulo – Foursquare (e) Tokyo – Instagram (f) Tokyo – Foursquare

Figure 10: Temporal sharing pattern for Instagram and Foursquare to New York, Sao Paulo, and Tokyo during weekdays (images from [86]).

et al [73]. About 20% of users performed only one check-in, 40% above 10, while about 10 % performed more than 100 check-ins.



(a) Instagram (b) Waze

Figure 11: Distribution of the number of data shared by users (images from [85, 87]).

3.6. Discussion

In this section we studied the properties of PSNs for location sharing, photo sharing, and traffic alert sharing. These PSNs have several properties in common: (i) global scale; (ii) highly unequal frequency of data sharing, both spatially and temporally, which is highly correlated with the typical routine of people; (iii) user participation in the number of shared data and where such data are shared can vary significantly; and (iv) temporal sharing pattern for the same type of system do not vary considerably over time.

The properties identified here show the potential of PSNs to conduct several studies on city dynamics and the urban social behavior, as discussed in the next section. Moreover, the understanding of the user behavior is the first step to model it. With models that explain user behavior we might predict actions and develop better systems for load capacity planning.

It is important to point out some possible limitations of our datasets. First, they reflect the behavior of a fraction of the citizens of the city. Some of our datasets are based on data shared by users of Foursquare, Instagram and Waze on Twitter. Therefore, data is skewed to the citizens that use these systems. Second, our datasets are based on a limited sample of data. This means we have just a sample of activities. External factors, such as bad weather may have affected the total number of data we collect for some places, especially in outdoor locations. Therefore, before drawing conclusions with PSN data, it is highly recommended comparing the results with data obtained in a traditional way (offline), as done, for example, in [89].

4. Working with PSN Data

In this section we discuss how to work with PSN data. In Section 4.1 we present how to obtain data from PSN. Next, we discuss some approaches to extract and generate contextual information from data of participatory sensor networks. These studies are grouped into two classes: Understanding city dynamics (Section 4.2); and Social, Economic, and Cultural patterns (Section 4.3).

4.1. Data Collection

In this section we introduce three main ways to collect data in PSNs: APIs, Web crawler, and applications.

4.1.1. APIs

The web is full of sources of data, among them PSN, representing a huge opportunity to researchers in several areas to collect large-scale data and extract knowledge from them.

Some PSNs provide APIs that could be used to collect data. Through this process is possible to obtain data from PSNs that can be used in other applications or in specific analysis. Several popular PSNs, such as Twitter, Instagram, and Foursquare, have APIs to access data shared by users. However, it is common to have different rules for their use.

Basically, there are two main ways of working with APIs: (1) Based on streaming; (2) Based on requests. The API based on streaming allows the collection of data in (almost) real time that they are published in a PSN. Twitter Streaming API, for instance, allows collecting in almost real time public tweets. On the other hand, an API based on requests make data available upon request, which typically includes specifics demands, such as all the last 10 tweets from a user. Both methods can suffer limitations about the amount of data that can be provided. For example, Flickr API allows 5,000 requests per hour, and the Twitter API might make available approximately 1% from all the total public tweets. This might prevent some kind of analysis that needs a large number of samples, for instance, in one hour period.

In fact, the use of APIs is a popular way to obtain data. Data collected from APIs, such as Twitter API, was used in many ways, ranging from the measurement of users influence in an network [16] to predict earthquakes [79].

An example of the use of Twitter Streaming API, written in (pseudo) Python and using TwitterAPI library[17], is showed in Algorithm 1. This algorithm accesses tweets searching by keyword "foursquare". As we can see, in few lines of code is possible to collect data from Twitter. Figure 12 illustrates this result with two tweets: *tweet1* and *tweet2*.

Some PSNs offer APIs, but with restrict access. This is the case for Foursquare, where few data is possible to be collected without user agreement. Most data available through this API are related to places, such as: tips, location, and pictures.

These limitations encourage the collection of data using alternative ways. For instance, in [89] the authors collected data about Foursquare check-ins through public messages shared at Twitter. This is possible because Foursquare allows users to share check-ins in Twitter. This procedure is shown in Figure 12. This picture shows a tweet that came from Foursquare

[17] https://github.com/geduldig/TwitterAPI.

Algorithm 1 Example of Twitter data collection

```
from TwitterAPI import TwitterAPI        ▷ Library that ease the interaction with the Twitter API

twitter_api = TwitterAPI(consumer_key = ‘XX’,
consumer_secret = ‘XX’,
access_token_key = ‘XX’,
access_token_secret = ‘XX’)        ▷ A registration in the API website provides the credentials needed

filters = {‘track’: [‘4sq’]}        ▷ Searching tweets with the keyword “foursquare”
stream = twitter_api.request(‘statuses/filter’, filters)

for item in stream.get_iterator() do
    print item[‘text’]        ▷ Display the tweet text
end for
```

and has an URL that represents a web page with more information about the check-in announced. In the example, the page represents a check-in performed at a cafe. To obtain more data about the check-in in this page it is used another data collection technique called Web Crawler, introduced in the next section.

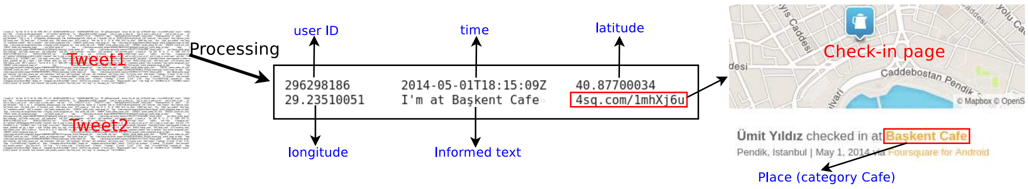


Figure 12: Steps for Foursquare data collection through tweets.

4.1.2. Web Crawler

Not all data sources available on the Internet provides direct access to their data through APIs. For this reason, it is necessary to use other strategies to obtain data. One of them is called Web Crawler, which are programs that analyzes Web pages searching for relevant data [3]. A Web crawler access some predefined Web pages and retrieve data from them.

Data collection through Web crawler depends on the data source structure that we desire to obtain data, and the approach chosen. The data source structure contains the data that we want in the Web page. For instance, the content of some HTML tags. With this, the construction of a Web crawler demands typically text mining to the extraction of the desired data. However, other non conventional ways of data extraction is possible as well.

For example, in [94] the authors built a Web crawler to collect information about traffic by taking screenshots of maps, such as Bing Maps[18], containing this information. More details about this procedure are provided in [94].

4.1.3. Applications

Another way to collect data is creating applications in existing platforms. Some popular websites, such as Facebook[19] and Instagram allow the creation of applications inside their platforms. In this way, developers can offer services using data that are shared in those apps.

Facebook, for instance, does not allow the collection of data about their users directly by APIs or Web Crawlers. However, it is possible to create applications in the Facebook platform for this purpose. When a Facebook user install an application and authorizes it to manipulate his/her data, the application can obtain diverse information, such as the shared content with his/her friends. Next we illustrate some initiatives in this direction.

In [69] the authors used this approach for data collection. They created an application for Facebook specifically to collect data that allows the study of behavior of people using this type of application. Another example was the application used in [102]. The authors created an application for Facebook that obtain the last likes[20] given by the user to draw a personality profile.

It is also possible to create applications that do not depend on platforms. This is the case of the PSN NoiseTube [65]. The authors created an application that enable users to report noises levels in the city. These data allow the identification, for instance, of areas in the city with level of noise above of the limits of the law. Another example is Colab, cited before. Besides that, there is a platform called *ohmage*[21] that ease the construction of applications to obtain data of participatory sensing.

With data from PSNs, that could be obtained using one of the approaches mentioned, we can extract knowledge using different strategies, as is discussed in the following sections.

[18] http://www.bing.com/maps.

[19] http://www.facebook.com.

[20] A *like* is a user interaction with Facebook in which he demonstrates that enjoyed a shared item.

[21] http://ohmage.org.

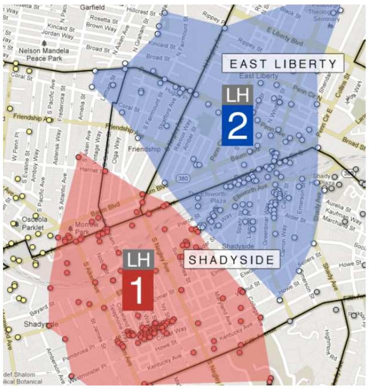


Figure 13: “Livehoods” found in New York (image from [22]).

4.2. Understanding City Dynamics

Information obtained from PSNs have the power to change our perceived physical limits, as well as help to better understand the city dynamics. This section focuses on the presentation of studies in this direction.

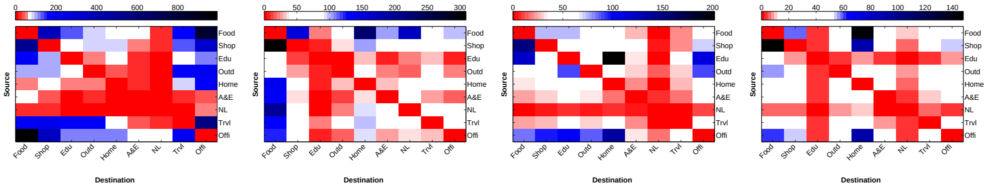
Cranshaw et al. [22] proposed a model to identify different areas of a city that reflect current patterns of collective activities, introducing new limits for neighborhoods. The idea is to expose the dynamic nature of local urban areas, considering the spatial proximity (derived from geographical coordinates) and social proximity (derived from the distribution of check-ins) locations.

For this, the authors used data from Foursquare and developed a model that groups similar places considering social and spatial characteristics. Each cluster represents different geographic boundaries of neighborhoods. The grouping method is a variation of spectral clustering proposed in [72].

Figure 13 shows two clusters (or “livehoods”, name used by the authors), found in New York, represented by the numbers 1 and 2. In this figure black lines indicate the official city limits. Note that the limits of the clusters are quite different from the original limits. To try to validate these results the authors used results of interviews with residents of the city. According to the collected answers these and other clusters were expected.

In [90] the authors proposed a technique called City Image, which provides a visual summary of the city dynamics based on the movements of people. This technique explores urban transition graphs to map the movements of users between city locations. An urban transition graph is a directed weighted graph $G(V, E)$, where a node $v_i \in V$ is the category of a specific location (for example, $food$) and a directed edge $(i, j) \in E$ marks a transition between two categories. That is, there is an edge from node v_i to the node v_j if at least one user shared data at a given place categorized by v_j after sharing data at a given place categorized by v_i. The weight $w(i, j)$ of an edge is the total number of transitions that occurred from v_i to v_j. Only consecutive data shared by the same user within 24 hours starting at 5:00 are considered in calculating a transition.

City Image is a promising technique that allows a better understanding of city dynamics, helping the visualization of common routines of its citizens. Each cell in the City Image represents how favorable is a transition from a certain category in a certain place (vertical axis) to another category (horizontal axis). Red colors represent rejection, blue colors represent favorability, and white color is indifference. We exemplify the City Image technique for two cities [22]: Sao Paulo (Figures 14a and 14b); and Kuwait (Figures 14c e 14d). In both cases, we consider weekday during daytime, which is the typical period of routines, and weekend during the night, which is a representative period of leisure activities (out of routine).



(a) SP (Daytime - weekday)

(b) SP (Night - weekend)

(c) KU (Daytime - weekday)

(d) KU (Night - weekend)

Figure 14: Images produced with the City Image technique to Sao Paulo (SP) and Kuwait (KU) at different times. Abbreviations of category of places (names used by Foursquare): Arts & Entertainment (A&E); College & Education (Edu); Great Outdoors (Outd); Nightlife Spot (NL); Shop & Service (Shop); and Travel Spot (Trvl) (images from [82]).

First, note that transitions to $office$ (workplaces) are more likely to occur on weekdays and during the day for both cities, as expected. However, note that the images of the city of Sao Paulo and Kuwait also have significant differences that reflect cultural differences between the two cities. Note, for

[22] Using data from the dataset Foursquare1.

example, the image representing transitions on weekend during the night (Figure 14d) shows the lack of favorable transitions to $nightlife$ category in Kuwait. This is not the case for Sao Paulo (Figure 14b), where the $food \rightarrow nightlife$ transition is highly favorable to happen. This suggests that in Sao Paulo people like to go to places related to food ($food$) before going to nightclubs ($nightlife$). In Kuwait, instead, people are probably more favorable to perform the transitions $shop \rightarrow food$ and $food \rightarrow home$ in the evenings of the weekend.

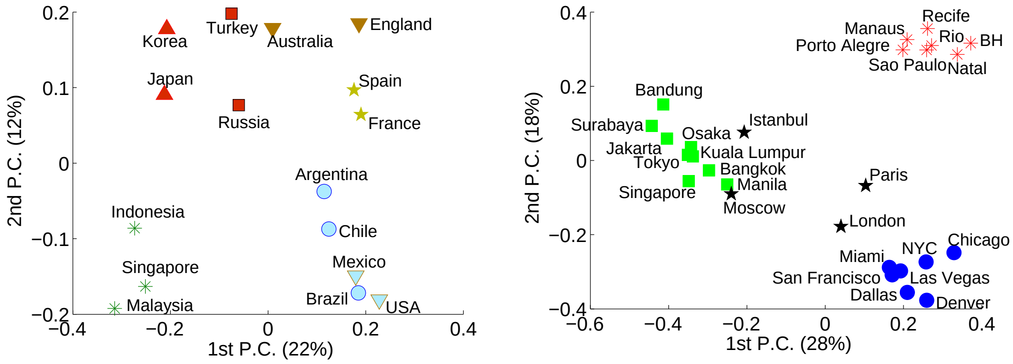
Techniques to provide easy to interpret visualizations of routines of inhabitants of a city, such as those mentioned here, are valuable tools to help urban planners to better understand the city dynamics and, therefore, make more effective decisions.

4.3. Social, Economic, and Cultural Patterns

PSN data can also be used to study social, economic, and cultural patterns of inhabitants of cities. In order to better understand social patterns from data of PSN, Quercia et al. [75] studied how virtual communities, observed in the analyzed systems, resemble real-life communities. The authors tested whether sociological theories established in social networks of real life are valid in these virtual communities. They found, for example, that social brokers on Twitter are opinion leaders who venture "tweeting" on different topics. They also found that most users have geographically local networks, and the influential ones express not only positive emotions, but also negative.

To carry out this work, the authors applied network metrics that the literature has found to be related to social relations, such as reciprocity and network constraint [75]. The reciprocity r is the proportion of edges in a network that are bidirectional $L^{<->}$ relative to the total number of edges L: $r = \frac{L^{<->}}{L}$. Considering a social network focused on a specific node ("ego") and vertices and edges to whom the ego is directly connected, low reciprocity values could indicate, for example, a social network of a celebrity. Network constraint measures the opportunities to become influential (brokerage opportunities). A high network constraint value means fewer opportunities. The authors used Burt formulation [13] in that case.

In addition, by studying the social behavior of specific areas, one of the first questions that arise is: how different a culture is from other? We know that eating and drinking habits can describe strong cultural differences. Based on this, in [89] the authors propose a new methodology for identifying cultural boundaries and similarities between societies, considering eating



(a) Countries

(b) Cities

Figure 15: Groups found using the methodology for culture separation. Each symbol reflects a group (images from [89]).

and drinking habits. For this, they used check-ins from Foursquare to represent the user's preferences regarding to what he/she eats and drinks locally, for example, in a particular city.

This analysis surprisingly says a lot about the differences and similarities between cultures. For this, the authors study the correlation between check-ins data in different types of restaurants for various cities around the world. They observed that cities of the same country, where the inhabitants often have similar culture, have the strongest correlations with respect to restaurant preferences. In addition to preferences for food and drink categories, it is also possible to see differences in the times when people go to restaurants and share data. These analyzes allowed the proposition of a methodology for identifying similar cultures, which can be applied in regions of varying sizes, such as countries, cities, or even neighborhoods [89]. This methodology uses a partitioning-based clustering algorithm ($k-means$ [39]), and the principal component analysis technique [47]. The results for countries and cities are illustrated in Figures 15a and 15b, showing how similar cultures are well separated. These figures use the first and second Principal Component (P.C.) to show the results. However, to obtain the results we considered all components.

The investigation of the cultural differences between different cities and countries is valuable in many areas and can assist various applications. For example, as culture is an important aspect for economic reasons, the identification of similarities between places that are geographically separated may be required for companies with business in a country that want to assess the compatibility of preferences between different markets.

Related to the economics of the cities, in [49] the authors studied the problem of optimal allocation of retail stores in the city. They used Foursquare data to understand how the popularity of three retail chain stores in New York is defined in terms of number of check-ins.

The authors evaluated a diverse set of features, modeling spatial and semantic information about the places and patterns of user movement in the area around the analyzed site. They observed that the presence of places that attract many users naturally, such as a train station or airport, as well as retail stores of the same type, defining a local commercial competition area, are the strongest indicators of popularity.

4.4. Final Considerations

PSNs provide updated information on places, as well as opinions and preferences of its members. Moreover, they have the potential of access the above data in (almost) real time, reaching a large number of regions of the globe. This section discussed several studies that serve as examples of how to work with PSN data. The information obtained by these studies can be useful for the development of more intelligent services and applications related to the study of city dynamics and urban social behavior.

For example, understanding the pattern of behavior in certain places in the city, as well as the identification of behaviors outside the expected pattern, can be useful for load capacity planning of an urban mobile network. Studies that aim to provide solutions to ease mobile data offloading can have great benefits by using this information as a tool to reduce surprises at current demands, as well as new demands that may arise as the city is in constant changes. Other research opportunities (and challenges) are discussed in the next section.

5. Challenges and Opportunities

This section presents current research topics related to participatory sensor networks. For each of them we also be discuss the challenges associated and opportunities for research.

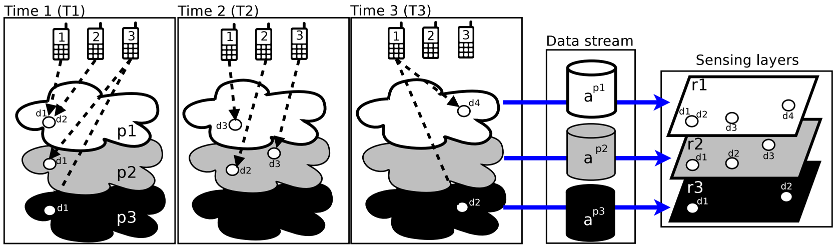


Figure 16: Data sharing illustration in three PSNs over time, resulting in sensing layers (image from [88]).

5.1. Sensing Layers

5.1.1. Preliminaries

A sensing layer consists of data describing specific aspects of a geographic location. The concept of sensing layer is quite broad: it represents data, with its attributes, from a particular data source, for example, a particular PSN. Each PSN provides access to data related to a certain aspect of a predefined geographic region (for instance, traffic conditions, pictures of places, etc.), and thus each single PSN can be represented as a sensing layer [88].

In addition to PSNs, other examples of data sources are: data available on the web not generated by users, for example, weather conditions provided by the company The Weather Channel[23] or data from traditional wireless sensor networks. We discuss here the concept of sensing layers to PSNs. However, all concepts discussed can be used for other data sources associated with predefined geographical regions, with necessary adaptations.

Figure 16 illustrates the concept of sensing layers. This figure shows data shared in three different PSNs (p1, p2 and p3) by four different users in different time instants. As discussed in Section 2, these data should be collected (for example, using an API) and processed, step that includes analysis and data standardization. Each plane in the figure represents a sensing layer of a specific region, for example, Manhattan in New York, with data from three different sources. Thus, the illustrated sensing layers are: check-ins (r1), from, for instance, Foursquare; traffic alerts (r2) from, for example, Waze; and picture of places (r3), from Instagram, for example.

In one layer each data has the following attributes: instant t when the

[23] http://www.weather.com.

	Attributes (m)		
Timestamp (t)	Area (a)	User (u)	Specialty data (s)
T1	a_1	1	"Times square"
T1	a_1	2	"Times square"
T2	a_2	1	"Fifth Av."
T3	a_4	1	"Statue of Liberty"

(a) Foursquare PSN

	Attributes (m)		
Timestamp (t)	Area (a)	User (u)	Specialty data (s)
T1	a_1	3	"Traffic Jam"
T2	a_2	2	"Accident"
T2	a_3	3	"Police control"

(b) Waze PSN

	Attributes (m)		
Timestamp (t)	Area (a)	User (u)	Specialty data (s)
T1	a_1	3	"photo data"
T3	a_4	1	"photo data"

(c) Instagram PSN

Table 2: Data stream describing users activity in three different PSNs: Foursquare, Waze, and Instagram [88].

data was shared; location a where the data was shared; specialty s of the layer (e.g., a picture or a alert about traffic); and the id u of user who shared the data.

5.1.2. Framework for the Integration of Multiple Layers

In this section we present the general idea of a framework to work with multiple sensing layers defined in [88]. Each user u can share unlimited data in any PSN p. Each j-th data d_j shared in a PSN p_k has the form $d_j^{p_k} =< t, m >$, where t refers to the moment when the user u shared data in p_k and m is a tuple containing the attributes of this data, i.e., $m = (a, u, s)$, as described above.

Data shared in a PSN can be seen as a data stream B. The authors defined that a data stream B^{p_k} consists of all data shared by users in a PSN p_k in a given time interval. Thus, B^{p_k} is used to represent a sensing layer r_{p_k}. Table 2 shows the data of the sensing layers that have been shared in the three PSNs considered in Figure 16.

To work with layers we need to represent them in a *work plan*, which contains one or more layers. This work plan is a combination of data from the layers that we want to work with. Making this combination of data depends on the layer functionality, what it captures. Various structures can be used for this task, in [88] the authors used a data dictionary, chosen for its simplicity which facilitates the understanding of the concepts.

5.1.3. Challenges and Opportunities

There are several challenges to handle data from multiple layers simultaneously, some of the main ones are described below.

1. **Data Combination:** In order to combine data we have to ensure that they are consistent across all layers. This is a mandatory condition for the correct extraction of information. For example, to combine data shared by the same user on different layers can be a problem in PSNs, because the same user may participate in different layers with different IDs. Let's assume we want to combine data from a single user that contributed in the check-ins layer and in the picture of places layer. Since the data of these layers are from independent systems, users have different IDs. One way to try to get around this problem is to check other systems in order to map the user ID of a layer in another. We know, for example, that users of Foursquare and Instagram tend to be also Twitter users. Thus, the combination process could use the identification used on Twitter. Without data management techniques that allow developers to combine data from multiple heterogeneous sensing layers, it becomes very hard to meet important requirements of urban computing applications, such as quickly respond to user queries about real time traffic conditions.
2. **Validity of the data:** Different layers can refer to data valid for different time intervals. This is natural because some data sources provide data in (almost) real time, while others do not. For example, an alert shared in Waze may refers to a traffic situation that can not exist five minutes later. However, census data are generally valid for a long time, months or years, until the next census be published. We have to be aware of all these issues when designing new applications.
3. **Modeling:** There are also opportunities regarding the modeling of sensing layers because in the same layer the entities can have different relationships between them. To illustrate this opportunity, consider the check-ins layer. As illustrated above, this layer may be used to represent urban mobility considering the relationship between places and people, being useful for understanding, for example, the frequency of transition

between different places. Another possibility is to modify the problem modeling, for example, to study the preferences of individuals. In this case, the entity to be analyzed becomes the user. Note that data from the same layer can be modeled in different ways to answer different questions. The framework presented briefly here (discussed in more details in [88]), provides basic support for this issue. However, there are several opportunities for extending that framework to offer more sophisticated services.

5.2. Temporal Dynamics of PSNs

5.2.1. Preliminaries

The study of PSN data emerge as a powerful resource for understanding city dynamics [82]. Most of the studies found in the literature represent data shared in PSNs as static structures, disregarding the temporal dynamics. Despite being an acceptable strategy this procedure may result, in many cases, in loss of important information in some scenarios.

To better illustrate this problem, consider the graph presented in Figure 17. This figure represents a static graph resulting of aggregated data from a certain day, where each vertex represents a Point of Interest (PoI) and the weighted edges represent the number of times people moved from one PoI to another (in any order). From this, note that the top 3 most popular transitions are $A-C$, $D-E$, and $A-B$. However, this observed information might present differences when a temporal perspective is considered.

When we partition the same dataset in three different intervals, as shown by Figure 18, we can see that the graph topology, as well as the weights of the edges change considerably throughout the time. Note first that in the second time interval we observe a disconnect graph, i.e., transitions $B-D$, $C-D$ and $C-F$ do not occur. This information could not be obtained using the static graph (Figure 17). Furthermore, note also that the weights of the edges change over time in the dynamic model. Observe in the first time interval that the top 3 most popular transitions are $D-E$, $D-F$, and $E-F$, while in the third time interval the top 3 most transitions are $A-B$, $A-C$, and $B-C$, information significantly different from the one obtained with a static model. This type of analysis can be useful to extract a more precise human behavior in the cities [41]. In this regard, the following are

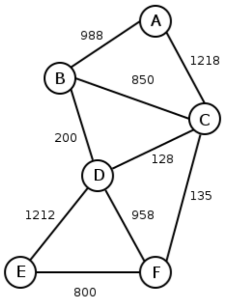


Figure 17: Example of static graph representation.

some efforts that try to exploit the temporal dimension in data analysis from PSNs.

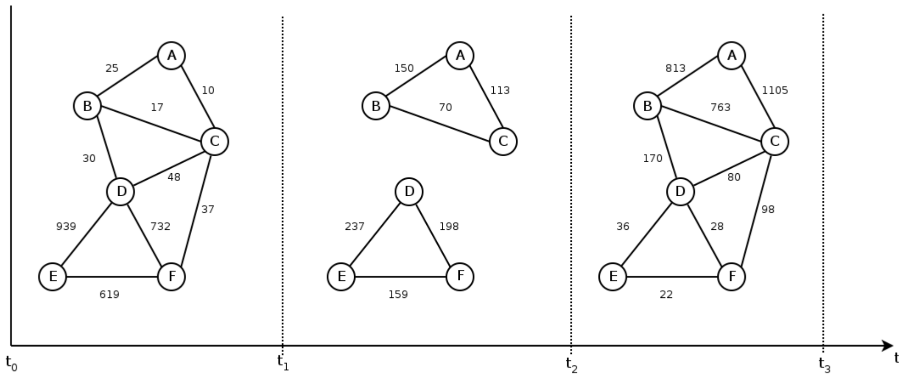


Figure 18: Example of graphs in different periods of time.

Bannur and Alonso [5] have analyzed data from Facebook check-ins to understand the temporal user participation in various categories of places (e.g., restaurants, cinemas, and get-away). The authors have defined a metric, called polarity, which represents the relationship between the number of check-ins of a category in a given region and season, and the total number of check-ins in the same region during all year. Figure 19 shows the change in polarity of the category get-away among USA states throughout the four seasons. The polarity is represented by a heat map. The intensity ranges from low (light color) to high polarity (dark color). As we can see, during winter and spring, states with high temperatures have a much higher polarity compared to those with low temperature. On the other hand, during the summer, states with low temperature such as Alaska and Montana appear as states with high polarity. This type of analysis is interesting to explain certain human behaviors based on seasonal phenomena. For example, in the fall, Nebraska has a high polarity. Nebraska is subject to tornadoes and thunderstorms during the summer and spring, whereas in winter suffer from ice storms, thereby influencing human behavior in the category of places get-away.

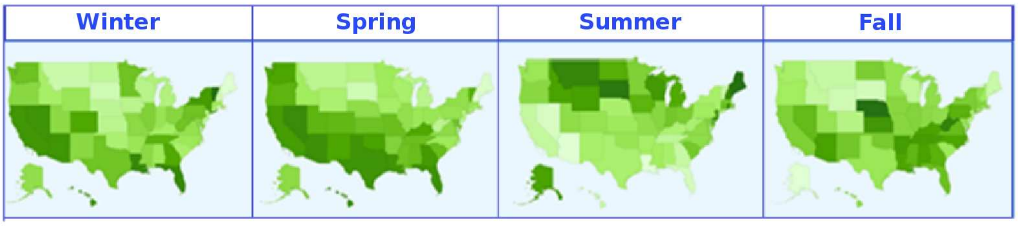


Figure 19: Visualization of check-ins for get-away category in every USA state over the four seasons (image from [5]).

Zhang et al [105] have analyzed urban activities from Foursquare data considering temporal dynamics. For this, they studied activities of groups of users with similar characteristics, considering the categories of places visited by them. Considering first the whole dataset without any separation of periods of the day, i.e., aggregated data, activities on the category of places Food are prevailing. However, analyzing the same dataset partitioned in different periods of the day (morning, afternoon, and evening), there is a greater distinction between the prevailing activities. For instance, activities on Food places in the afternoon are not as popular as in the morning and evening. In the aggregated view they do not notice this difference. This approach is interesting to show that certain activities may be performed significantly only at a certain time of day, but when they are analyzed disregarding the temporal aspect, important insights about users' behavior may be missed.

The City Image technique, presented in Section 4.2, is another initiative that exploit the temporal dimension. This dimension is applied for partitioning the data on weekdays/weekend during different periods of the day, from that the authors performed analysis on the partitioned data. With the help of this technique it is possible to see that there is significant variation of popular types of activities during different periods of the day. Moreover, the results of when applying this technique without considering the time dimension is quite different from those when considering it [83].

5.2.2. Challenges and Opportunities

The related work described above provide evidences about the advantages of using temporal information of data obtained from PSNs. However, if on one hand the investigation of temporal dynamics of PSNs is an opportunity to obtain information closer to the real network behavior, on the other hand, new challenges arise when we consider temporal dimension to the study, as described below:

1. **Temporal information:** An initial issue is how to represent and store temporal information. Since data can be from many sources, we face problems related to inconsistency, redundancy, and granularity to extract relevant temporal information. In addition, there are open questions about the validity of the information obtained. For instance, how long this information will be useful? or how often should the information be updated?;
2. **Time windows:** Studies that analyze the temporal aspect often partition the data in time intervals, e.g., morning, afternoon, and evening, called time windows. However, the proper definition of the time window size is a problem, it is necessary to define a window size that captures relevant dynamics. In this case, there are many opportunities for new approaches that consider time windows with flexible and dynamic sizes;
3. **Dynamic participation:** Since the structure of a PSN is composed of autonomous nodes (people), it is sensitive to the participation of these nodes over time. This brings a range of challenges related to the evolution of user participation in these networks, some examples are: identification of periodic behavior, detection of outliers, and activity tracking. In this direction there are several opportunities for the development of new techniques/approaches;
4. **Modeling:** Typically, data from PSNs are represented as a set of entities, for example, users or points of interest, and their relationships, e.g., transitions or communication. As the contribution of these data can vary greatly over time a model based on static graphs may not be enough to capture this dynamism. For example, data from Foursquare have spatiotemporal information, such as positioning of users and the

moments of interaction. Therefore, a challenge is to model spatiotemporal dynamics in order to better understand, for example, user participation. In this direction, temporal graphs [52] appears as a promising alternative that can be used to understand spatiotemporal dynamics. In a temporal graph, relations between entities can be modeled as edges that can be created and destroyed over time. This is useful, for instance, to understand temporal aspects of interactions between users with certain places in the city.

5. **Data visualization:** The use of visualization techniques that helps the understanding of network behavior is fundamental to assist in decision making. Thus, visualizations that explore temporal dynamics of PSNs are of paramount importance. For example, a proper visualization of the transitions of users in the city over time is useful to planners and other professionals who need to make decisions related to urban planning.

5.3. Incentive Mechanism for PSN

5.3.1. Preliminaries

Selfish, altruistic, and cooperative behavior of human beings were extensively studied in philosophy, psychology, economics, and, recently, in the context of computer science [68]. Selfishness can be defined as the act of benefiting oneself instead of another. On the other hand, altruism favors others instead of oneself [61]. Incentive mechanisms aim to engage users to cooperate with others. Cooperation occurs when an individual devotes an effort that implies a cost in some collective activity expecting some benefit. Unlike altruism, in cooperation the individual expects some benefit greater than costs [8].

Cooperation is a key point for PSN since it relies on users' willingness to collect, process, and transmit the sensed data [57]. The cooperation among PSN participants reflects directly on the quality and quantity of the sensed data, and hence in improving services offered by PSNs.

However, as PSNs consume resources of users' devices, they may be reluctant to contribute to the network. There are several reasons that can make a user benefit, but not collaborate with PSNs, such as to save battery power, data transmission costs, or even privacy issues [57].

Thus, incentive mechanisms aim to increase the engagement for users cooperate with the PSN. In recent years, academy and industry have been proposed dozens of incentive mechanisms [35]. The motivation for cooperation can be extrinsic, in which participants receive a direct reward for participating, or intrinsic, in which participants must be satisfied psychologically [50].

As extrinsic mechanisms reward participants through payments, real or virtual, we will refer to these mechanisms as *Reward-based*. Intrinsic mechanisms are based on transforming the sensing task in a more enjoyable and challenging task for the user by adding common game elements, such as contest among users, badges, and trophies. Therefore, we will address these mechanisms as *Gamification-based*[24].

5.3.2. User Cooperation

Cooperation in the context of PSNs depends on the relationship between the cost and benefit to participate on them. Fitzek et al. [32] claim that cooperation will occur whenever a participant of the network has the feeling that the benefit is higher than the cost of collaborating. This benefit could be many things, for example, the quality of the information that a PSN can offer.

Figure 20 illustrates important aspects about user cooperation in a PSN. A user request information of a PSN using a mobile network (e.g., 3G), while obtaining its location using a GPS network and collects new information using sensors from a portable device. Next, the user transmits the sensed data to the PSN. To accomplish this task, we can list as costs: power consumption; data transmission over the mobile network; and the effort to perform the sensing task. Meanwhile, the user obtains as benefits: updated real-time information; and the feeling of helping other participants in the network (for altruistic users). Note that, a user with resource constraints or a selfish user could obtain information from the PSN without collaborating with new data to it.

There are also situations in which the benefit for cooperative behavior is unclear. For instance, PSNs that aim to gather information about pollution or the health of the individual [12]. In these examples, participants may not have access to real-time information and the beneficiaries of the information gathered would be public authorities and health centers, respectively. In these situations, incentive mechanisms act as a "driving force" to encourage user cooperation.

[24] The use of game features as incentive mechanisms to perform tasks is known in the literature as *gamification*.

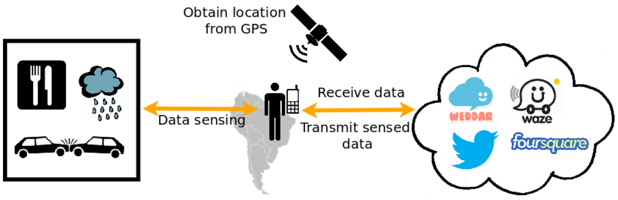


Figure 20: Important aspects about user cooperation in an PSN.

5.3.3. Reward-Based Incentive Mechanisms

Reward-based incentive mechanisms rely on the assumption that participants will not contribute or will contribute only for a short period to the PSN if the benefits are less than their expectations [57].

In some mechanisms, users negotiate with the PSN platforms how much they will receive for the sensed data before sending them, in others the platform decides how much is going to be paid for data sent by users. In any case, these mechanisms aim to improve the quality of data and minimize sensing costs.

Yang et al. [101] proposed two incentive mechanisms: *MSensing Platform-Centric* and *MSensing Auction*. In the *MSensing Platform-Centric*, PSNs has a limited budget to spend with sensing tasks. The PSN announces the reward for a certain task and each participant receives a reward proportional to the time dedicated to the task. One problem with this model is that if the number of active participants increases the reward for each participant decreases.

In the *MSensing Auction*, the PSN platform announces a set of tasks and each user chooses a subset. For each task chosen, users must submit a tuple, *task-bid*, to the platform, where the “bid” is the value of reward he/she wants to receive to perform the task. After receiving the offerings from users, the PSN platform selects a set of users as the winners of the auction, and these users will perform the tasks. A problem encountered in this type of mechanism is the explosion of the incentive costs [57]. This problem can derail the mechanism due the high cost that might be expended by the platform. In addition, if the winner is always the user that offer the lowest price, this user may be discouraged to continue sensing data for the PSN, due to the low values received. Incentive mechanisms based on game theory attempt to overcome the issues above by aiming to achieve the equilibrium of the system, i.e., maximize the gain for the user, while minimizing the costs for the platform [100, 57, 101].

Reddy et al. performed small scale experiments to evaluate the effect of payments for cooperation in participatory sensing [76]. The authors concluded that incentives work better when micro payments are combined with other factors such as user altruism and competition. In addition, they showed that a fair payment for all participants kept them motivated for longer time than low payments.

Indeed, payments can be counterproductive in some cases, as shown by Kamenica in [48]. After reviewing several studies in psychology and economics on the effect of payment as incentive mechanism, the author concluded that, in many cases, paying a high or too low value proved to be counterproductive to induce participants into collaborative behavior.

5.3.4. Gamification-Based Incentive Mechanism

Gamification can be defined as the use of elements (and design) of games in non-related game contexts [26]. Examples of such elements are: score tables; trophies or medals to reward users who perform a given task; challenges; avatars; difficulty levels; and social networks to check how is the performance of "friends" in the task development, with that, who performed more tasks, for instance, have higher ranking.

Unlike traditional games, gamification uses game elements with the purpose to perform tasks non related to a game [96]. These tasks may be, for instance, to improve a skill, encourage fitness, or, in the context of PSNs, engage users to contribute sensed data for a longer period.

As an example of participatory sensor network that uses gamification-based mechanism, we can mention Waze. In order to keep information up to date, Waze requires active participation of users, i.e., participants must manually report situations observed, such as car accidents. In Waze, game elements are represented by the use of avatars and a score system. In this case, more cooperative users achieve special avatar or badges. As a result, Waze help to raise the total number of shared data, and the quality of traffic information for all users. The reasons that lead a user to cooperate with Waze can range from simple altruistic motivation to social rewards (score and badges) given by gamification.

5.3.5. Challenges and Opportunities

An incentive mechanism is efficient if it recruits more participants to a PSN and keeps them active in the network. In order to encourage users to become active participants a PSN also faces social and psychological challenges. In this section, we present some of the main challenges for the proposition of incentive mechanisms for PSNs:

1. **Costs of incentive mechanisms based on monetary reward**: For the success of monetary incentive mechanisms, it must be considered the costs for PSN platforms, as well as the earnings for participants. PSN platforms can limit these costs by defining a maximum value to be paid to the active network participants. However, finding and deciding a value that minimizes the cost to the platform and, at the same time, motivates the users requires further investigation [35].
2. **Combination of different strategies**: The majority of the proposals to encourage cooperation in PSNs focuses on only one strategy. However, as observed by Reddy et al., combine more than one incentive mechanism simultaneously may achieve better results [76].
3. **Proposal validation**: Authors commonly validate their proposals of incentive mechanisms using a theoretical approach or small controlled experiments. However, these strategies may not be able to predict with high accuracy the participation of users over time on the platform. Although there are already successfully PSNs on the market using gamification as incentive mechanism, there is no guarantee that would work on other PSNs. Investigate which elements work (or not work) for certain types of PSNs requires also further investigation.

5.4. Quality of Data from PSN

5.4.1. Preliminaries

Quality of data is a widely studied topic by the scientific community, and the topics range from the definitions of metrics to qualitatively assess a given data to solutions that ensure the generation and recovery of data with quality.

According to dictionaries, the term *quality*, by itself, can be used to express both a characteristic possessed by someone or something, or a degree of excellence in a given subject. Adapting this to our context, we need to think of quality as a property that a given data, originated from a PSN, has or has not, and also as a way to evaluate the degree of confidence we can have on this data.

This means that we need to represent quality in a more technical way. From a computer perspective, quality refers to the correct comply of some requirements for a system. Therefore, to evaluate if these requirements are met, some metrics that summarize the main characteristics of such system have to be defined.

Generally, data collected from the PSNs, after processed, are used to extract contextual information, which are essential to the context aware systems [27]. Thus, one way to evaluate quality of data from a PSN is by the expected quality of contextual information they provide, which can be defined by the concept of Quality of Context (QoC) [11].

Buchholz e Schiffers [11] define QoC as any information that describes the quality of the inferred context, which is consistent with our needs. The authors also argue about the differences between QoC and Quality of Service (QoS) and Quality of Device (QoD) concepts. QoS refers to any information that describes how well a service operates, and QoD is related to any information about the technical properties and capabilities of a given device. Thus, the authors propose the following QoC metrics to measure quality: *precision*, *probability of correctness*, *trust-worthiness*, *resolution*, and *up-to-dateness*.

Contextualizing these metrics to a PSN scenario, *precision* of the data is related to how well it reflects the current state of a specific phenomenon or locality. *Probability of correctness* denotes the probability that a given data is correct, this metric can be seen as a statistic that reflects an *a priori* knowledge of the data or the user that generated it. *Trust-worthiness* is similar to the probability of correctness, but it is used to classify the quality of the user that generated the data. *Resolution* denotes the granularity of the information, that, as discussed in Section 3, may represent the details of coverage of a particular region. Finally, *up-to-dateness* describes the age of the data, being essential to assess its validity when there are real time requirements.

In the same direction, Li et al. [62] extended this QoC definition to evaluate the quality of the data from pervasive environments. By investigating the challenges of providing data with quality in such environments, they proposed three additional metrics to assess the source of these data: *currency*, *availability*, and *validity*. *Currency* is related to the previously discussed up-to-dateness, it represents the temporal utility of the data, from the moment it is created until it becomes worthless. *Availability* measures the capability of an entity to provide data when the information about a region is needed. In the context of PSNs, this can be expressed as the expectation that the data is generated by a user. *Validity* is defined as a set of rules that can be used to validate the generated data, according to a previous knowledge about the type of the data and the behavioral pattern of the users.

With the computational representation of quality, we can describe some challenges and opportunities when dealing with the quality of data from PSNs.

5.4.2. Challenges

Following the aforementioned discussion, we can summarize the main expected requirements for data generated by PSNs in two aspects: (i) data reliability; and (ii) users credibility. Data reliability means the confidence we can assign on data we have. The credibility of the users generating these data is another important aspect of quality of data. Thus, some of the main challenges that may affect quality of data in PSNs, impacting on the metrics previously presented, are:

1. **Sample representativeness**: This challenge is related to how representative a sample is about a specific phenomenon, based on the amount of collected data. Due to its high relevance, this is a widely discussed aspect in several studies that deal with data sampling. From the PSN point of view, as discussed in Section 3, the collected data may represent a portion of the population of a city and the extracted information is based on this sample. Depending on the sampling, it is possible that the inferred information do not represent correctly the analyzed phenomenon. Then, as previously mentioned, before inferring conclusions by the data sampled from PSNs, it is necessary to compare them with other sources, for instance with data collected in offline mode;
2. **Sensing Errors**: Another challenge that might affect the precision of PSN data is the occurrence of possible sensing errors in the user's portable device. For instance, a Global Positioning System (GPS) might be poorly calibrated, generating data whose accuracy is beyond the acceptable range for this type of data. Although some errors may seem totally tolerable, depending on the application it is possible that it demands a high precision for its correct functioning;
3. **Subjectivity of interpretation**: This challenge concerns different meanings that may exist about the data, one for the user that generated it and another for whom will use it. For example, it is possible to find data that was misclassified and shared in a PSN. Foursquare, for instance, allows the definition of a category to a new added place, even if this definition is not the most appropriated, and the system must correct it subsequently. Another example is the case of Weddar, a system

that allows its users to share their interpretations of current weather conditions. While a user may interpret a shared temperature as the temperature inside his house, another user may interpret it as the temperature in a park in the same region. In this case, the interpretation of these two users may be quite different;

4. **Absence of structure**: Data shared in PSNs, in some cases, are composed by free text, without a semantic structure or encoding. This freedom given to users allows them to post whatever they want, even wrong information, and in different formats. For instance, a user could describe a traffic accident in a foreign language or by using slang through a microblogging like Twitter. Thus, the processing of these data is complex and prone to errors, since there is the possibility, for example, of data duplicity, i.e., the same data being identified as distinct, or distinct data interpreted as the same due to differences in filled fields;
5. **Pollution of data**: Pollution of data is related to the possibility of data to be incorrect due to the users malicious behavior [20, 67]. We can find this malicious behavior in several social activities, and the same can also occur in PSNs. As an example, users in PSN for traffic alerts sharing like Waze, can generate false alerts of traffic jams or accidents in order to discourage other users to use certain streets of his/her route. Malicious behavior may result in false positives on the detection of social patterns or events.

5.4.3. Opportunities

An important research topic that is affected by the quality of data in PSNs is the one related to techniques of processing and knowledge extraction from these data. One possible approach to handle this problem is to model the data as a time series and extract knowledge by signal processing techniques [56]. However, in some cases, data from PSNs may not follow a constant pattern in order to ease this processing. As previously mentioned, data from PSNs are subject to problems of subjectivity of interpretation and absence of structure, and may result in errors during the learning of the pattern and other properties of a certain phenomenon.

An interesting approach to solve these conflicts of interpretation is given by [36]. The authors improved some data classification algorithms by using

Mechanical Turk (Mturk)[25]. In this service users are financially rewarded for completed tasks, e.g., solving doubts raised by classification algorithms. The combination of computational processing and human intelligence offer important research opportunities to the participatory sensing scenarios.

Another opportunity is to evaluate the reliability of a given user in a PSN, since data generated by reliable users will be probably more reliable as well. One possible direction in this sense is related to the identification of behavioral patterns of users in the PSN. As shown in Figure 8, when a sufficient amount of data is aggregated, it is clearly possible to identify these patterns in the shared data for different week days. Assuming this previous knowledge as a reference of the expected pattern by users in a given PSN, one possibility would be compare the behavior of a suspected user with this reference pattern. For instance, users with a sharing pattern quite different from the majority may represent an unreliable user (e.g., a malicious robot).

That approach can be characterized as a kind of collaborative filtering technique [1]. This is a strategy used by recommending systems when there is no previous knowledge about the user in which it should recommend an item. For instance, by using the preferences of other similar users, assuming that their preferences are also similar.

Other alternatives to assess the quality of data in PSNs are based on the analysis of the reputation of the users generating it, aiming to increase the credibility with their good behavior. Huang et al. [43] propose the computation of a user reputation score, which is related to the trustworthiness of his contributed data. They compute this score for each user based on an outlier detection algorithm, that uses a consensus-based technique to identify the users that deviates most from the consensus data of all users.

Mashhadi and Capra [67] propose an extension to the previous work of [43] to estimate the quality of users contributions considering their credibility. They consider the contribution of points of interest by users and define regularity functions with respect to the mobility pattern of these users, and a reputation function considering their reliability based on previous contributions with the system. The feasibility of this proposal is based on studies demonstrating that urban users exhibit a high level of regularity in their daily activities. This regularity, represented as the frequency of repetitions of locations, is the pattern that helps in the identification of the credibility

[25] http://www.mturk.com.

of a user.

Several strategies discussed here point towards solutions to the two mentioned requirements of PSNs, i.e., reliability of data and credibility of the users. However, an important point emphasized by [33] is that such aspects are less related to the precision of the data itself and more about which information, or perspective, their users have about those data. In other words, there is still a lot of subjectivity about the notion of the quality of data shared in PSNs. Thus, a strategy that focuses on dealing with the quality of this data must consider the needs of each application and try to attend their requirements specifically.

5.5. PSNs and Vehicular Networks

5.5.1. Preliminaries

Vehicular Networks (VANETs) offer a range of opportunities for urban monitoring and data sharing on various aspects of the traffic. Vehicular networks do not have common constraints of wireless sensor networks, such as energy, bandwidth, and memory constraints, which allows a more accurate sensing and a larger amount of data collected. Furthermore, vehicles can contain sensors that are not commonly available in portable devices used in PSNs.

Another important aspect of VANETs is the coverage. Vehicles move through the whole city using streets and avenues. Because of this spread mobility, vehicular networks can capture several city details. All these features make VANETs an important data source that can complement data from PSNs, in order to better understand the urban phenomena.

Vehicular applications can be used in numerous scenarios. For instance, in VANETs there are diverse events to be monitored, such as potholes, traffic jam, car accidents, and presence of animals on the road. Thus, in this section, we present studies that focus on three main issues: monitoring general traffic events; the use of data of VANETs to study routines of people; and the study of traffic jams. We also discuss various challenges associated with these issues.

5.5.2. Monitoring Events

In VANETs vehicles can cooperate among themselves to collect data, which enables the identification of events, and propagate them to interested parties. Thus, these data can directly influence vehicle route, making drivers, in many cases, redefine their trajectories. Cunha et al. [23] presented a service for event monitoring and data dissemination, which considers vehicles mobility patterns. Thus, when a vehicle detect an event in the region where it is located, it propagates this information to other vehicles, warning of dangers ahead. In addition, this broadcast takes into account the interactions between vehicles, selecting those that guarantee greater coverage in the data dissemination.

Another possible solution used to sense events with vehicular networks is presented by Lee et al. [60], solution known as MobEyes. The goal of this solution is to use of vehicles equipped with sensors to collect data about roads and other vehicles nearby them. However, due to the amount of data generated, we can associate some data filters, and, thus, only the most relevant data will be stored and forwarded to the sink. In this scenario, algorithms that control collection and data delivery to the sink should be aware of peculiarities and restrictions of VANETs.

With different goals, Lee et al. [59] presented FleaNet, a platform for submitting queries in vehicular networks. Vehicles receive and submit queries about various traffic issues. For example, a mobile user detects an accident and shares a picture to the next vehicle. Differently, a market or a store can disseminate alerts offering promotions to nearby users in vehicles. In addition, a user can submit queries in the network, looking for neaby type places or attractions.

5.5.3. Routines and Behaviors

Considering the mobility of vehicles and their daily routes, it is possible to extract several cultural features of the users routine, such as their interests and the most popular places in the city. Based on this, the study proposed by Cunha et al. [24] presents an analysis of GPS traces describing mobility of vehicles in the city. From the traces, it is possible to identify similar behavioral patterns on the network and better understand routines performed in the cities. The greater the number of records about vehicles, the better will also be the quality of the characterized data. However, obtain these data is not trivial because users must allow the monitoring of their vehicles.

Similarly, Fiore et al. [31] present an analysis of vehicle mobility in order to characterize the traffic in a city through the understanding of flows and places visited. Based on the analysis of a trace of mobility, the authors demonstrate how the use of a real information about the mobility of vehicles[26] can help in the evaluation of the performance of protocols for VANETs. From the analysis, they showed that it is possible to improve the performance of protocols and better understand the traffic distribution in the city.

5.5.4. Traffic Management

The literature presents several models that deal with traffic jam. Some of them only make traffic jam detection (e.g., [46, 97]), others make traffic jam prediction (e.g., [51, 53]), and others that make both taks (e.g., [42, 66]). They are different mainly due to the following aspects: (i) time horizon to predict future jams; (ii) techniques used in the model; (iii) data sources that have been used.

Regarding to the time horizon, we have the following two categories: (i) models for *Short-term traffic flow forecasting (STFF)*, which predict the traffic behavior for the next 5 min untill 1 hour [92]; (ii) jam prediction for the next 1 hour (at least) are named *Long-term traffic flow forecasting (LTFF)*. Models that predict traffic jam for the next 15 min or 30 min ahead are much more interesting and useful than others, since this is a reasonable time interval that can be used to make a decision. Despite having several STFF and LTFF prediction models, the usage of PSN can improve the models' accuracy depending on which social variables have been used. Since PSN data are associated with habits and routines of users, the challenge is how to obtain and how to use such data in real time, mainly for STFF.

The most used techniques for traffic jam prediction are: Seasonal AutoRegressive Integrated Moving Average (SARIMA), multi-variate AutoRegressive Integrated Moving Average (ARIMA), Bayesian networks, *fuzzy* clustering, identification of traffic patterns, genetic algorithms, neural networks, Support Vector Machines (SVM), historical average, non-parametric regression, Kalman filter, and ant colony.

Such approaches differ in terms of the data source used to detect or to predict future jams, such as: GPS traces, tracking smartphones movements, online maps, data from sensors on roads, weather, seasons, traffic incidents, and social sensing. Sensed data on roads are the most used information by these models, followed by information regarding GPS traces, weather, and seasons. For instance, the Clearflow project from Microsoft Research,

[26] The authors use the Cologne trace: http://kolntrace.project.citi-lab.fr.

described in [42], uses practically all the mentioned sources. In the product offered by Intellione company[27], they use only data about smartphones movements in mobile networks to detect traffic jam (no prediction is done). In order to make traffic prediction, Song et al. [103] use the combination of several simple predictors through a genetic algorithm, achieving a forecast with a higher accuracy.

5.5.5. Challenges and Opportunities

Vehicular networks and PSNs have several possibilities of integration, which brings several challenges and opportunities that we describe next.

1. **Event tracking:** There are several initiatives that use PSN data to detect events [80, 58, 7], and the area of event tracking in vehicular networks may benefit from some of these initiatives. Furthermore, there are events hard to be identified in a vehicular network that could be reported in PSNs, as we discussed in Section 5.1. Note also that we could suggest routes in order to avoid events in the city, or even to promote the encounter of the most visited spots.
2. **Data Availability:** As we discussed above, particularly in Section 4, PSNs data can be very useful for the study of habits and routines of city inhabitants. This is an important information to vehicular networks, as mentioned in Section 5.5.3. However, users in vehicular networks may not provide the information of the visited places, a problem that can also occurs in PSNs. We can minimize this problem by stimulating the contribution of users. Another way to minimize this problem is to use data available from PSNs and vehicle networks together, which serves as a way to complement the information of movement of users.
   Regarding to traffic jam problem, generally, the more data sources used in a model the better the performance because more information will be used to improve its inferences. The problem is that not all data sources are correlated and relevant to the prediction of congestion. Thus, the inclusion of a new data source requires a characterization with respect to traffic performance.
   Moreover, as mentioned in Section 3, the data input can be very unequal in different areas of a city. If we do not have enough data in all regions, which regions will take benefit from this information?

[27] http://www.intellione.com.

3. **Detection/Prediction of traffic jam:** Generally, PSN data are underexploited in traffic jam detection/forecasting models. Some of the closest studies in this direction are: [87, 94]. Tostes et al. have analyzed traffic conditions using two types of PSN data, from Foursquare and Instagram. As we mentioned in Section 3, PSN data provide valuable information to better understand city dynamics. For instance, a geolocated message, whether on Foursquare, Instagram or, Twitter, can be used to better understand traffic conditions. In fact, Tostes et al. [94] observed that check-ins (from Foursquare) or photos (from Instagram) are well correlated with intense traffic conditions and can be used to design more efficient traffic prediction models. Besides that, imagine that a user share data at home and then commute to work and, for some reason, he/she shares another piece of geolocated data. Regardless if it was the same social network or not, there is an intrinsic information in the time interval between these data that may be related with the traffic behavior. If the traffic is more congested, this interval between the shared data may be higher than the travel time without traffic jam, which is easily calculated by the distance and maximum speed on roads. When analyzing this aspect for many users, the results could be powerful hints about the traffic. The authors have also raised up several questions on this direction, such as: (i) how to collect data from online maps in real time?; (ii) is it possible to use PSN data as a predictor characteristic for intense traffic jam?

5.6. Other Challenges and Opportunities Related to PSNs

5.6.1. Data Sampling

It is important to point out some challenges regarding to data sampling. PSN data are biased towards citizens who use them. For example, the dataset discussed on Section 3 are based on data shared by users of Foursquare, Instagram, and Waze on Twitter. Therefore, biased towards the citizens who use those systems, who are likely to be under 50 year-old, and especially those between 18-29 year-old, owners of smartphones, and urban dwellers [9, 28]. Consequently, urban areas with older and poorer populations tend to have fewer data.

There are some initiatives in the literature regarding methods and techniques to identify and recruit suitable candidates in support for data collection, most of them focusing on the selection of participants to minimize a certain cost. For instance, Reddy et al. [77] developed a framework to help to identify well-suited participants for data collections based on geographic and temporal availability as well as participation habits. In a similar direction, Hachem et al. [38] use a mobility model to predict users' future locations. Based on the predicted results they aim to select a minimal number of mobile users, expecting to cover a certain percentage of the target area. However, despite those efforts, there are still open challenges. For example, there is a lack of mechanisms that consider users with specific characteristics, such as a certain age, gender, or race. This type of selection is important, for instance, to the study of urban social behavior.

Another challenge related to data sampling in PSNs is that users might not share data at all of their destinations, for instance in love hotels and strip clubs. Thus, datasets from PSNs might offer a partial view of citizens habits. Besides that, external factors, such as bad weather conditions, might affect the total number of data to be collected for some places, especially outdoor locations. This means that we might only have access to a sample of data that could be shared under regular conditions. New mechanisms to deal with this type of situation have to be developed.

5.6.2. Large Volume of Data

Another important issue is to deal with a large volume of data that PSNs can offer, imposing challenges for storage, processing, and indexing in real time using tools of traditional database management and data processing applications. Fortunately, research on the challenges imposed by this huge amount of data (also known as big data) is very active, and recently, in conjunction with cloud computing solutions, advanced considerably [45, 78].

PSNs may offer big data that grows quickly. For this reason, storage platforms have to be distributed, scalable, secure, consistent, and fault-tolerance [40]. Recently, some services were proposed to store and manage large amounts of data covering some of these requirements. For instance, Amazon Simple Storage and Service (Amazon S3)[28] and Microsoft Azure Storage[29] provide solutions to store and retrieve large amount of data, where files can be replicated across multiple geographical sites to improve redundancy and availability. These services rely on available technologies, such as Google File System (GFS) and Hadoop Distributed File System (HDFS), to

[28] http://aws.amazon.com/s3.

[29] https://azure.microsoft.com/en-us/services/storage.

store a large volume of data across multiple machines.

Besides that, data from PSNs may have different formats (i.e., structured, semi-structured, and unstructured). Consider an application for transit monitoring, like Waze. In this type of PSN, users can share observations about accidents or potholes. Since users use an application designed for a specific purpose, the sensed data is structured. Instead, if a user uses a microblog (e.g, Twitter), the sensed data would be unstructured (e.g., message sent by user X: “traffic now is very slow near the main entrance of campus”). With that, data modeling using traditional relational model may be hard. This motivates the adoption of new alternatives, such as NoSQL databases, which allow storing and retrieving large volume of distributed data [14]. NoSQL databases are non-relational, highly distributable, and schema-free, making them increasingly used in big data and real time web applications, such as PSNs [4].

Another issue when working with PSN is the processing of a large volume of data in real time. For this task, one important aspect is how to distribute computation. MapReduce model is the first major contribution on data-processing for a parallel, distributed algorithm on a cluster [25]. Currently, this model in combination with HDFS form the Hadoop core. Hadoop[30] is a project that allows the distributed processing of large datasets across clusters of computers. Alternatively, Apache Spark[31] is a fast and general engine for large-scale data processing, and it is appropriated to applications that reuse a working dataset across multiple parallel operations, such as iterative machine learning algorithms and interactive data analysis tools [104]. With that, new algorithmic paradigms for processing, based, for example, on the mentioned parallel platforms, should be designed and specific data mining techniques should be created accordingly to manipulate, for instance, large urban transition graphs (as those mentioned in Section 4.2), with millions or billions of nodes/edges [37].

5.6.3. Privacy

Working with data from PSNs may impose threats to users’ privacy. For instance, these data could be used to infer users’ personal behavior and preferences, such as common visited locations, lifestyle, and health condition,

[30] https://hadoop.apache.org.

[31] http://spark.apache.org.

thus, not assuring freedom from the intrusion of others in their private life or affairs [64]. With that, an important challenge is guarantee user privacy while working on potentially sensitive data from PSNs.

Data privacy has been discussed in several studies, ranging from methods that allow participants control their privacy preferences to anonymization techniques for data privacy-preserving [10, 30, 63, 64, 74, 81, 93, 95, 99]. Particularly, these anonymization techniques aim to protect privacy by anonymizing data fields such that sensitive information cannot be pinpointed to an individual record [21]. Anonymization can be achieved through several ways: creating alias that avoid user identification; aggregating data from several users, thus making hard to identify an individual user; hiding sensitive locations; and injecting randomness into the data to create data perturbation [18]. A challenge related to this last approach has to design special data mining methods to derive knowledge from anonymized data [98]. Furthermore, the development of anonymization mechanisms have to consider the tradeoff between anonymity and data fidelity.

As mentioned previously, a dataset from a PSN might be used to create an urban transition graph representing users' trajectories in a given period of time. Therefore, another example of challenge is to prevent leakage of private information of individuals, while mining and releasing frequent patterns of these graphs. There are some initiatives to deal with this sort of problem. For example, Shen and Yu [81] propose an algorithm for privacy-preserving mining of frequent graph patterns.

Another important aspect to protect data privacy in PSNs is to consider security issues in the data transmission. Sensitive data have to be encrypted before they are shared by users, preventing a malicious user to obtain this sensitive data. Although data encryption helps to protect data privacy, it also obsoletes the traditional data utilization service based on plain text keyword search. A number of studies were proposed for privacy preserving database encryption while enabling some traditional functions, for instance, query using SQL [63, 99, 30, 95]. However, despite these efforts, this challenge still open and further research is to be conducted to make the proposed approaches more practically feasible. As an example, Li et al. [63] intend to study how to provide some practical data publishing methods suitable for their proposed framework to deal with privacy-preserving data queries. Besides, another opportunity is to improve the performance of the proposed approach for certain situations.

6. Conclusion

In this chapter, we show that PSNs provide unprecedented opportunities to access sensing data on a global scale. In this sense, we present a detailed view of properties of these data, as well as their usefulness in developing smarter services to meet people's needs in several areas. In addition, we discuss some of the key challenges related to PSNs, ranging from incentive mechanisms for users of PSNs, to the use of PSN data for the development of more sophisticated applications. We also highlighted several opportunities related to the use of PSN data, for example, when considering the temporal dynamics of the data.

References

[1] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *Knowledge and Data Engineering, IEEE Transactions on*, 17(6): 734–749, June 2005. ISSN 1041-4347. doi: 10.1109/TKDE.2005.99.

[2] I.F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci. Wireless sensor networks: a survey. *Computer Networks*, 38(4):393 – 422, 2002. ISSN 1389-1286.

[3] S. Anbukodi and K.M. Manickam. Reducing web crawler overhead using mobile crawler. In *Proc. of ICETECT'11*, pages 926–932, Nagercoil, India, March 2011.

[4] Paul Andlinger. *RDBMS dominate the database market, but NoSQL systems are catching up*. DB-Engines, Nov 2013. `http://db-engines.com/en/blog_post/23`.

[5] Sushma Bannur and Omar Alonso. Analyzing temporal characteristics of check-in data. In *Proc. of WWW Companion '14*, pages 827–832, Seoul, Korea, 2014. ISBN 978-1-4503-2745-9. doi: 10.1145/2567948.2579041. URL `http://dx.doi.org/10.1145/2567948.2579041`.

[6] F. Barth. *Ethnic groups and boundaries: the social organization of culture difference*. Scandinavian university books. Little, Brown, 1969.

[7] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

[8] Samuel Bowles and Herbert Gintis. *Origins of human cooperation. In: Genetic and cultural evolution of cooperation.* MIT Press Cambridge, MA, 2003.

[9] Joanna Brenner and Aaron Smith. 72% of online adults are social networking site users. `http://goo.gl/HTgNy3`, August 2013.

[10] A.J. Bernheim Brush, John Krumm, and James Scott. Exploring end user preferences for location obfuscation, location-based services, and the value of location. In *Proc. of Ubicomp '10*, pages 95–104, Copenhagen, Denmark, 2010. ACM. ISBN 978-1-60558-843-8. doi: 10.1145/1864349.1864381. URL `http://doi.acm.org/10.1145/1864349.1864381`.

[11] Thomas Buchholz and Michael Schiffers. Quality of context: What it is and why we need it. In *Proc. of OVUA'03*, Geneve, Switzerland, 2003.

[12] J. Burke, D. Estrin, M. Hansen, A. Parker, N. Ramanathan, S. Reddy, and M. B. Srivastava. Participatory sensing. In *Proc. of Workshop on World-Sensor-Web (WSW'06)*, pages 117–134, Boulder, USA, 2006.

[13] Ronald S Burt. *Structural Holes: The Social Structure of Competition.* Harvard University Press, 1992.

[14] Rick Cattell. Scalable sql and nosql data stores. *SIGMOD Rec.*, 39 (4):12–27, May 2011. ISSN 0163-5808. doi: 10.1145/1978915.1978919. URL `http://doi.acm.org/10.1145/1978915.1978919`.

[15] CENS/UCLA. *Participatory Sensing / Urban Sensing Projects.* http://research.cens.ucla.edu/.

[16] M. Cha, H. Haddadi, F. Benevenuto, and K.P. Gummadi. Measuring user influence in twitter: The million follower fallacy. In *Proc. of ICWSM'10*, Washington, USA, 2010.

[17] Zhiyuan Cheng, James Caverlee, Kyumin Lee, and Daniel Z. Sui. Exploring Millions of Footprints in Location Sharing Services. In *Proc. of ICWSM'11*, Barcelona, Spain, 2011.

[18] Delphine Christin, Andreas Reinhardt, Salil S Kanhere, and Matthias Hollick. A survey on privacy in mobile participatory sensing applications. *Journal of Systems and Software*, 84(11):1928–1946, 2011. doi: 10.1016/j.jss.2011.06.073.

[19] Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Rev.*, 51(4):661–703, November 2009. ISSN 0036-1445. doi: 10.1137/070710111. URL `http://dx.doi.org/10.1137/070710111`.

[20] Alberto Coen-Porisini and Sabrina Sicari. Improving data quality using a cross layer protocol in wireless sensor networks. *Comput. Netw.*, 56(17):3655–3665, November 2012. ISSN 1389-1286. doi: 10.1016/j.comnet.2012.08.001. URL `http://dx.doi.org/10.1016/j.comnet.2012.08.001`.

[21] G. Cormode and D. Srivastava. Anonymized data: Generation, models, usage. In *Data Engineering (ICDE), 2010 IEEE 26th International Conference on*, pages 1211–1212, March 2010. doi: 10.1109/ICDE.2010.5447721.

[22] Justin Cranshaw, Raz Schwartz, Jason I. Hong, and Norman Sadeh. The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City. In *Proc. of ICWSM'12*, Dublin, Ireland, 2012.

[23] Felipe D. da Cunha, Guilherme Maia, Aline Carneiro Viana, Raquel A. F. Mini, Leandro A. Villas, and Antonio Alfredo Ferreira Loureiro. Socially inspired data dissemination for vehicular ad hoc networks. In *Proc. of MSWiM'14*, pages 81–85, Montreal, Canada, 2014.

[24] Felipe Domingos da Cunha, Aline Viana, Thiago Assis de Oliveira Rodrigues, Raquel Mini, and Antonio Alfredo Ferreira Loureiro. Extracao de propriedades sociais em redes veiculares. In *Proc. of SBRC 2014 - WP2P+*, Florianopolis, Brasil, may 2014.

[25] Jeffrey Dean and Sanjay Ghemawat. Mapreduce: Simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, January 2008. ISSN 0001-0782. doi: 10.1145/1327452.1327492. URL `http://doi.acm.org/10.1145/1327452.1327492`.

[26] Sebastian Deterding, Dan Dixon, Rilla Khaled, and Lennart Nacke. From game design elements to gamefulness: defining gamification. In *International Academic MindTrek Conference: Envisioning Future Media Environments*, pages 9–15. ACM, 2011.

[27] A. K. Dey and G. D. Abowd. Towards a Better Understanding of Context and Context-Awareness. In *Proc. of CHI 2000 Workshops*, The Hague, The Netherlands, 2000.

[28] Maeve Duggan and Aaron Smith. Social media update 2013, Jan 2014. http://goo.gl/JhuiOG.

[29] David Easley and Jon Kleinberg. *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge University Press, 2010.

[30] Sergei Evdokimov and Oliver Gunther. Encryption techniques for secure database outsourcing. In Joachim Biskup and Javier Lopez, editors, *Computer Security – ESORICS 2007*, volume 4734 of *Lecture Notes in Computer Science*, pages 327–342. Springer Berlin Heidelberg, 2007. ISBN 978-3-540-74834-2. doi: 10.1007/978-3-540-74835-9_22. URL http://dx.doi.org/10.1007/978-3-540-74835-9_22.

[31] Marco Fiore, Jose M. Barcelo-Ordinas, Oscar Trullols-Cruces, and Sandesh Uppoor. Generation and analysis of a large-scale urban vehicular mobility dataset. *IEEE Transactions on Mobile Computing*, 13 (5):1–1, 2014. ISSN 1536-1233. doi: http://doi.ieeecomputersociety.org/10.1109/TMC.2013.27.

[32] Frank HP Fitzek, Janus Heide, Morten Videbæk Pedersen, and Marcos Katz. Implementation of network coding for social mobile clouds [applications corner]. *Signal Processing Magazine, IEEE*, 30(1):159–164, 2013.

[33] Andrew J. Flanagin and Miriam J. Metzger. The credibility of volunteered geographic information. *GeoJournal*, 72(3-4):137–148, August 2008. ISSN 0343-2521. doi: 10.1007/s10708-008-9188-y. URL http://link.springer.com/10.1007/s10708-008-9188-y.

[34] R.K. Ganti, Fan Ye, and Hui Lei. Mobile crowdsensing: current state and future challenges. *Communications Magazine, IEEE*, 49(11):32

–39, november 2011. ISSN 0163-6804. doi: 10.1109/MCOM.2011.6069707.

[35] H. Gao, C. Liu, W. Wang, J. Zhao, Z. Song, X. Su, J. Crowcroft, and K. Leung. A survey of incentive mechanisms for participatory sensing. *Communications Surveys Tutorials, IEEE*, PP(99):1–1, 2015. ISSN 1553-877X. doi: 10.1109/COMST.2014.2387836.

[36] Mihai Georgescu, Dang Duc Pham, Claudiu S. Firan, Ujwal Gadiraju, and Wolfgang Nejdl. When in doubt ask the crowd: Employing crowdsourcing for active learning. In *Proc. of WIMS'14*, pages 12:1–12:12, Thessaloniki, Greece, 2014. ACM. ISBN 978-1-4503-2538-7. doi: 10.1145/2611040.2611047. URL `http://doi.acm.org/10.1145/2611040.2611047`.

[37] F Giannotti, D Pedreschi, A Pentland, P Lukowicz, D Kossmann, J Crowley, and D Helbing. A planetary nervous system for social mining and collective awareness. *The Eur. Phy. Jour. Special Topics*, 214 (1):49–75, 2012.

[38] Sara Hachem, Animesh Pathak, and Valérie Issarny. Probabilistic Registration for Large-Scale Mobile Participatory Sensing. In *PerCom 2013 - IEEE International Conference on Pervasive Computing*, Californie, United States, March 2013. Elsevier. URL `https://hal.inria.fr/hal-00769087`.

[39] John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Applied statistics*, pages 100–108, 1979.

[40] Ibrahim Abaker Targio Hashem, Ibrar Yaqoob, Nor Badrul Anuar, Salimah Mokhtar, Abdullah Gani, and Samee Ullah Khan. The rise of big data on cloud computing: Review and open research issues. *Information Systems*, 47:98 – 115, 2015. ISSN 0306-4379. doi: http://dx.doi.org/10.1016/j.is.2014.07.006. URL `http://www.sciencedirect.com/science/article/pii/S0306437914001288`.

[41] Petter Holme and Jari Saramki. Temporal networks. *Physics Reports*, 519(3):97 – 125, 2012. ISSN 0370-1573. doi: http://dx.doi.org/10.1016/j.physrep.2012.03.001. URL `http://www.sciencedirect.com/science/article/pii/S0370157312000841`. Temporal Networks.

[42] Eric Horvitz. Predictive analytics for traffic, 2015. http://research.microsoft.com/en-us/projects/clearflow/.

[43] Kuan Lun Huang, Salil S. Kanhere, and Wen Hu. Are You Contributing Trustworthy Data? The Case for a Reputation System in Participatory Sensing. In *Proceedings of the 13th ACM international conference on Modeling, analysis, and simulation of wireless and mobile systems - MSWIM '10*, pages 14—-22, New York, New York, USA, 2010. ACM Press. ISBN 9781450302746. doi: 10.1145/1868521.1868526. URL `http://portal.acm.org/citation.cfm?doid=1868521.1868526`.

[44] Instagram. Instagram today: 200 million strong. http://blog.instagram.com/post/80721172292/200m, April 2014.

[45] H. V. Jagadish, Johannes Gehrke, Alexandros Labrinidis, Yannis Papakonstantinou, Jignesh M. Patel, Raghu Ramakrishnan, and Cyrus Shahabi. Big data and its technical challenges. *Commun. ACM*, 57 (7):86–94, July 2014. ISSN 0001-0782. doi: 10.1145/2611567. URL `http://doi.acm.org/10.1145/2611567`.

[46] P. Jain and M. Sethi. Fuzzy based real time traffic signal controller to optimize congestion delays. In *Proc. of ACCT'12*, pages 204–207, Jan 2012. doi: 10.1109/ACCT.2012.55.

[47] I. T. Jolliffe. *Principal Component Analysis*. Springer, second edition, 2002.

[48] Emir Kamenica. Behavioral economics and psychology of incentives. *Annu. Rev. Econ.*, 4(1):427–452, 2012.

[49] Dmytro Karamshuk, Anastasios Noulas, Salvatore Scellato, Vincenzo Nicosia, and Cecilia Mascolo. Geo-spotting: Mining online location-based services for optimal retail store placement. In *Proc. of KDD '13*, pages 793–801, Chicago, Illinois, USA, 2013. ACM. ISBN 978-1-4503-2174-7. doi: 10.1145/2487575.2487616. URL `http://doi.acm.org/10.1145/2487575.2487616`.

[50] Nicolas Kaufmann, Thimo Schulze, and Daniel Veit. More than fun and money. worker motivation in crowdsourcing–a study on mechanical turk. In *Proc. of Amer. Conf. on Inf. Sys.*, 2011.

[51] Qing-Jie Kong, Yanyan Xu, Shu Lin, Ding Wen, Fenghua Zhu, and Yuncai Liu. Utn-model-based traffic flow prediction for parallel-transportation management systems. *Intelligent Transportation Systems, IEEE Transactions on*, 14(3):1541–1547, Sept 2013. ISSN 1524-9050. doi: 10.1109/TITS.2013.2252463.

[52] Vassilis Kostakos. Temporal graphs. *Physica A: Statistical Mechanics and its Applications*, 388(6):1007–1023, 2009.

[53] Satoshi Kurihara. Traffic-congestion forecasting algorithm based on pheromone communication model. *Ant Colony Optimization - Techniques and Applications*, 2013.

[54] N.D. Lane, E. Miluzzo, Hong Lu, D. Peebles, T. Choudhury, and A.T. Campbell. A survey of mobile phone sensing. *Comm. Mag., IEEE*, 48 (9):140 –150, sept. 2010. ISSN 0163-6804. doi: 10.1109/MCOM.2010.5560598.

[55] Nicholas D. Lane, Shane B. Eisenman, Mirco Musolesi, Emiliano Miluzzo, and Andrew T. Campbell. Urban sensing systems: Opportunistic or participatory? In *Proc. of HotMobile '08*, pages 11–16, Napa Valley, California, 2008. ACM. ISBN 978-1-60558-118-7.

[56] B. P. Lathi and Roger. Green. *Essentials of digital signal processing.* Cambridge University Press, Cambridge, UK, 2014. ISBN 978-1107059320.

[57] Juong-Sik Lee and Baik Hoh. Dynamic pricing incentive for participatory sensing. *Pervasive and Mobile Computing*, 6(6):693–708, 2010.

[58] Ryong Lee and Kazutoshi Sumiya. Measuring geographical regularities of crowd behaviors for twitter-based geo-social event detection. In *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks*, pages 1–10, San Jose, USA, 2010. ACM.

[59] Uichin Lee, Joon-Sang Park, E. Amir, and M. Gerla. Fleanet: A virtual market place on vehicular networks. In *Proc. of Mobiquitous'06 - Workshops*, pages 1–8, San Jose, USA, 2006. doi: 10.1109/MOBIQ.2006.340433.

[60] Uichin Lee, Biao Zhou, M. Gerla, E. Magistretti, P. Bellavista, and A. Corradi. Mobeyes: Smart mobs for urban monitoring with a vehicular sensor network. *Wireless Commun.*, 13(5):52–57, 2006. doi: 10.1109/WC-M.2006.250358.

[61] D K Levine. Modeling altruism and spitefulness in experiments. *Review of economic dynamics*, 1(3):593–622, 1998.

[62] Fei Li, S. Nastic, and S. Dustdar. Data quality observation in pervasive environments. In *Proc. of IEEE CSE'12*, pages 602–609, Nicosia, Cyprus, Dec 2012. doi: 10.1109/ICCSE.2012.88.

[63] Jin Li, Zheli Liu, Xiaofeng Chen, Fatos Xhafa, Xiao Tan, and Duncan S. Wong. L-encdb: A lightweight framework for privacy-preserving data queries in cloud computing. *Knowledge-Based Systems*, 79:18 – 26, 2015. ISSN 0950-7051. doi: http://dx.doi.org/10.1016/j.knosys.2014.04.010. URL `http://www.sciencedirect.com/science/article/pii/S0950705114001324`.

[64] Qinghua Li and Guohong Cao. Privacy-preserving participatory sensing. *Communications Magazine, IEEE*, 53(8):68–74, 2015. ISSN 0163-6804. doi: 10.1109/MCOM.2015.7180510.

[65] Nicolas Maisonneuve, Matthias Stevens, Maria E Niessen, and Luc Steels. Noisetube: Measuring and mapping noise pollution with mobile phones. In *Information Technologies in Environmental Engineering*, pages 215–228. Springer, 2009.

[66] G. Marfia and M. Roccetti. Vehicular congestion detection and short-term forecasting: A new model with results. *Vehicular Technology, IEEE Tran. on*, 60(7):2936–2948, Sept 2011. ISSN 0018-9545. doi: 10.1109/TVT.2011.2158866.

[67] Afra J. Mashhadi and Licia Capra. Quality Control for Real-time Ubiquitous Crowdsourcing. In *Proc. of UbiCrowd'11*, pages 5–8, Beijing, China, 2011.

[68] J Miao, O Hasan, S Ben Mokhtar, L Brunie, and K Yim. An investigation on the unwillingness of nodes to participate in mobile delay tolerant network routing. *International Journal of Information Management*, 33(2):252–262, 2013. ISSN 02684012. doi:

10.1016/j.ijinfomgt.2012.11.001. URL `http://linkinghub.elsevier.com/retrieve/pii/S0268401212001338`.

[69] Atif Nazir, Saqib Raza, and Chen-Nee Chuah. Unveiling facebook: A measurement study of social network based applications. In *Proc. of IMC '08*, pages 43–56, Vouliagmeni, Greece, 2008.

[70] M. E. J. Newman. The Structure and Function of Complex Networks. *SIAM Review*, 45(2):167–256, 2003.

[71] Mark Newman. *Networks: an introduction*. Oxford University Press, Inc., 2010.

[72] Andrew Y Ng, Michael I Jordan, Yair Weiss, et al. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 2:849–856, 2002.

[73] Anastasios Noulas, Salvatore Scellato, Cecilia Mascolo, and Massimiliano Pontil. An Empirical Study of Geographic User Activity Patterns in Foursquare. In *Proc. of ICWSM'11*, Barcelona, Spain, 2011.

[74] T. Pontes, G. Magno, M. Vasconcelos, A. Gupta, J. Almeida, P. Kumaraguru, and V. Almeida. Beware of what you share: Inferring home location in social networks. In *Proc. of ICDMW*, pages 571–578, Brussels, Belgium, 2012.

[75] Daniele Quercia, Licia Capra, and Jon Crowcroft. The social world of twitter: Topics, geography, and emotions. In *Proc. of ICWSM'12*, Dublin, Ireland, 2012.

[76] Sasank Reddy, Deborah Estrin, Mark Hansen, and Mani Srivastava. Examining micro-payments for participatory sensing data collections. In *Proc. of Ubicomp '10*, pages 33–36, Copenhagen, Denmark, 2010. ACM. ISBN 978-1-60558-843-8. doi: 10.1145/1864349.1864355. URL `http://doi.acm.org/10.1145/1864349.1864355`.

[77] Sasank Reddy, Deborah Estrin, and Mani Srivastava. Recruitment framework for participatory sensing data collections. In *Proceedings of the 8th International Conference on Pervasive Computing*, Pervasive'10,

pages 138–155, Berlin, Heidelberg, 2010. Springer-Verlag. ISBN 3-642-12653-7, 978-3-642-12653-6. doi: 10.1007/978-3-642-12654-3_9. URL http://dx.doi.org/10.1007/978-3-642-12654-3_9.

[78] Daniel A. Reed and Jack Dongarra. Exascale computing and big data. *Communications of the ACM*, 58(7):56–68, June 2015. ISSN 0001-0782. doi: 10.1145/2699414. URL http://doi.acm.org/10.1145/2699414.

[79] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proc. of WWW'10*, pages 851–860, Raleigh, USA, 2010.

[80] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proc. of WWW'10*, pages 851–860, Raleigh, USA, 2010. IW3C2.

[81] Entong Shen and Ting Yu. Mining frequent graph patterns with differential privacy. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pages 545–553, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2174-7. doi: 10.1145/2487575.2487601. URL http://doi.acm.org/10.1145/2487575.2487601.

[82] T.H. Silva, P.O.S. Vaz De Melo, J.M. Almeida, and A.A.F. Loureiro. Large-scale study of city dynamics and urban social behavior using participatory sensing. *Wireless Communications, IEEE*, 21(1):42–51, Feb 2014.

[83] Thiago H. Silva, Pedro O. S. Vaz de Melo, Jussara M. Almeida, and Antonio A. F. Loureiro. Visualizing the invisible image of cities. In *Proc. IEEE CPScom'12*, pages 382–389, Besancon, France, 2012.

[84] Thiago H. Silva, Pedro O. S. Vaz de Melo, Jussara M. Almeida, and Antonio A. F. Loureiro. Challenges and opportunities on the large scale study of city dynamics using participatory sensing. In *Proc. of IEEE ISCC'13*, pages 528–534, Split, Croatia, July 2013.

[85] Thiago H. Silva, Pedro O. S. Vaz de Melo, Jussara M. Almeida, Juliana Salles, and Antonio A. F. Loureiro. A picture of Instagram is worth more than a thousand words: Workload characterization and

application. In *Proc. of DCOSS'13*, pages 123–132, Cambridge, USA, 2013.

[86] Thiago H. Silva, Pedro O. S. Vaz de Melo, Jussara M. Almeida, Juliana Salles, and Antonio A. F. Loureiro. A comparison of foursquare and instagram to the study of city dynamics and urban social behavior. In *Proc. of UrbComp'13*, pages 1–8, Chicago, USA, 2013.

[87] Thiago H. Silva, Pedro O. S. Vaz de Melo, Aline Viana, Jussara M. Almeida, Juliana Salles, and Antonio A. F. Loureiro. Traffic Condition is more than Colored Lines on a Map: Characterization of Waze Alerts. In *Proc. of SocInfo'13*, pages 309–318, Kyoto, Japan, Nov 2013.

[88] Thiago H. Silva, Pedro Vaz de Melo, Jussara Almeida, Aline Viana, Juliana Salles, and Antonio Loureiro. Participatory Sensor Networks as Sensing Layers. In *Proc. of SocialCom'14*, Sydney, Australia, 2014.

[89] Thiago H. Silva, Pedro O. S. Vaz de Melo, Jussara M. Almeida, Mirco Musolesi, and Antonio A. F. Loureiro. You are What you Eat (and Drink): Identifying Cultural Boundaries by Analyzing Food & Drink Habits in Foursquare. In *Proc. of ICWSM'14*, Ann Arbor, USA, 2014.

[90] Thiago H. Silva, Pedro O. S. Vaz de Melo, Jussara M. Almeida, Juliana Salles, and Antonio A. F. Loureiro. Revealing the city that we cannot see. *ACM Trans. Internet Technol.*, 14(4):26:1–26:23, December 2014. ISSN 1533-5399. doi: 10.1145/2677208.

[91] Mani Srivastava, Tarek Abdelzaher, and Boleslaw Szymanski. Human-centric sensing. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 370(1958):176–197, January 2012. ISSN 1471-2962. doi: 10.1098/rsta.2011.0244. URL `http://dx.doi.org/10.1098/rsta.2011.0244`.

[92] Shiliang Sun, Guoqiang Yu, and Changshui Zhang. Short-term traffic flow forecasting using sampling markov chain method with incomplete data. In *Proc. of Intelligent Vehicles Symposium*, pages 437–441, June 2004. doi: 10.1109/IVS.2004.1336423.

[93] Eran Toch, Justin Cranshaw, Paul Hankes Drielsma, Janice Y. Tsai, Patrick Gage Kelley, James Springfield, Lorrie Cranor, Jason Hong,

and Norman Sadeh. Empirical models of privacy in location sharing. In *Proc. of Ubicomp'10*, pages 129–138, Copenhagen, Denmark, 2010. ACM. ISBN 978-1-60558-843-8. doi: 10.1145/1864349.1864364. URL `http://doi.acm.org/10.1145/1864349.1864364`.

[94] A. I. J. Tostes, T. H. Silva, F. Duarte-FIgueiredo, and A. A. F. Loureiro. Studying traffic conditions by analyzing foursquare and instagram data. In *Proc. of ACM PE-WASUN'14*, Montreal, Canada, 2014.

[95] Cong Wang, Qian Wang, and Kui Ren. Towards secure and effective utilization over encrypted cloud data. In *Distributed Computing Systems Workshops (ICDCSW), 2011 31st International Conference on*, pages 282–286, June 2011. doi: 10.1109/ICDCSW.2011.16.

[96] Kevin Werbach and Dan Hunter. *For the win: How game thinking can revolutionize your business.* Wharton Digital Press, 2012.

[97] N. Wisitpongphan, W. Jitsakul, and D. Jieamumporn. Travel time prediction using multi-layer feed forward artificial neural network. In *Proc. of CICSyN'12*, pages 326–330, Phuket, Thailand, July 2012. doi: 10.1109/CICSyN.2012.67.

[98] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, and Wei Ding. Data mining with big data. *Knowledge and Data Engineering, IEEE Transactions on*, 26(1):97–107, 2014.

[99] ZongDa Wu, GuanDong Xu, Zong Yu, Xun Yi, EnHong Chen, and YanChun Zhang. Executing {SQL} queries over encrypted character strings in the database-as-service model. *Knowledge-Based Systems*, 35:332 – 348, 2012. ISSN 0950-7051. doi: http://dx.doi.org/10.1016/j.knosys.2012.05.009. URL `http://www.sciencedirect.com/science/article/pii/S0950705112001530`.

[100] Xiaojuan Xie, Haining Chen, and Hongyi Wu. Bargain-based Stimulation Mechanism for Selfish Mobile Nodes in Participatory Sensing Network. *Proc. of IEEE SECON'09*, pages 1–9, 2009. doi: 10.1109/SAHCN.2009.5168911.

[101] D. Yang, G. Xue, X. Fang, and J. Tang. Crowdsourcing to smartphones: incentive mechanism design for mobile phone sensing. In *Proc.*

of Mobicom'12, pages 173–184, Istanbul, Turkey, 2012. ISBN 978-1-4503-1159-5. doi: 10.1145/2348543.2348567. URL http://doi.acm.org/10.1145/2348543.2348567.

[102] Wu Youyou, Michal Kosinski, and David Stillwell. Computer-based personality judgments are more accurate than those made by humans. *Nat. Acad. of Sci.*, 112(4):1036–1040, 2015.

[103] Song Yuecong, Hu Wei, and Bi Guotang. Combined prediction research of city traffic flow based on genetic algorithm. In *Proc. of ICEMI'07*, pages 3–862–3–865, Aug 2007. doi: 10.1109/ICEMI.2007.4351054.

[104] Matei Zaharia, Mosharaf Chowdhury, Michael J. Franklin, Scott Shenker, and Ion Stoica. Spark: Cluster computing with working sets. In *Proceedings of the 2Nd USENIX Conference on Hot Topics in Cloud Computing*, HotCloud'10, pages 10–10, Berkeley, CA, USA, 2010. USENIX Association. URL http://dl.acm.org/citation.cfm?id=1863103.1863113.

[105] Ke Zhang, Qiuye Jin, Konstantinos Pelechrinis, and Theodoros Lappas. On the importance of temporal dynamics in modeling urban activity. In *Proc. of UrbComp'13*, pages 7:1–7:8, Chicago, Illinois, 2013. ISBN 978-1-4503-2331-4. doi: 10.1145/2505821.2505825. URL http://doi.acm.org/10.1145/2505821.2505825.

List of Terms and Acronyms

Glossary

A

altruism It is the act of favoring others instead of oneself. 36

Availability Measures the capability of an entity to provide data when the information about a region is needed. 42

C

City Image Technique that provides a visual summary of the city dynamics based on the movements of people. 25

Cooperation Occurs when an individual devotes an effort that implies a cost in some collective activity expecting some benefit. 37

Currency Represents the temporal utility of the data, from the moment it is created until it becomes worthless. 42

G

gamification The use of game features as incentive mechanisms to perform tasks. 37

P

precision Defines how well certain data reflects the current state of a specific phenomenon or locality. 42

Probability of correctness Denotes the probability that a given data is correct. 42

PSN Participatory sensor network rely on the idea of participatory sensing, and can be defined as a system that supports a distributed process of gathering data about personal daily experiences and various aspects of the city. Such a process requires the active participation of people using portable devices to voluntarily share contextual information and/or make their sensed data available, i.e., the users manually determine how, when, what, and where to share the sensed data. Thus, through PSNs we can monitor different conditions of cities, as well as the collective behavior of people connected to the Internet in (almost) real time. 5, 70

Q

QoC Any information that describes the quality of the inferred context, which is consistent with our needs. 41, 70

QoD Any information about the technical properties and capabilities of a given device. 41, 70

QoS Any information that describes how well a service operates. 41, 70

R

Resolution Denotes the granularity of the information. 42

S

Selfishness It is the act of benefiting oneself instead of another. 36

sensing layer It represents data, with its attributes, from a particular data source, for example, a particular PSN. 29

T

Trust-worthiness It is similar to the probability of correctness, but it is used to classify the quality of the user thatgenerated the data. 42

tweets Personal updates in texts up to 140 characters shared on Twitter. 9

U

up-to-dateness Describes the age of the data. 42

V

Validity It is defined as a set of rules that can be used to validate the generated data, according to a previous knowledge about the type of the data and the behavioral pattern of the users. 42

Acronyms

A

ARIMA AutoRegressive Integrated Moving Average. 48

C

CCDF Complementary Cumulative Distribution Function. 13

CDF Cumulative Distribution Function. 15

G

GFS Google File System. 52

GPS Global Positioning System. 43

H

HDFS Hadoop Distributed File System. 52

J

JSON JavaScript Object Notation. 21

L

LTFF Long-term traffic flow forecasting. 47

P

P.C. Principal Component. 27

PoI Point of Interest. 32

PSNs Participatory Sensor Networks. 5, *Glossary:* PSN

Q

QoC Quality of Context. 41, *Glossary:* QoC

QoD Quality of Device. 41, *Glossary:* QoD

QoS Quality of Service. 41, *Glossary:* QoS

S

SARIMA Seasonal AutoRegressive Integrated Moving Average. 48

STFF Short-term traffic flow forecasting. 47

V

VANETs Vehicular Networks. 45

W

WSNs Wireless Sensor Networks. 5