Urban Computing Leveraging Location-Based Social Network Data: a Survey

THIAGO H. SILVA, Federal University of Technology - Parana
ALINE CARNEIRO VIANA, Inria
FABRÍCIO BENEVENUTO, Federal University of Minas Gerais
LEANDRO VILLAS, University of Campinas
JULIANA SALLES, Microsoft Research
ANTONIO LOUREIRO, Federal University of Minas Gerais
DANIELE QUERCIA, Bell Labs

Urban computing is an emerging area of investigation in which researchers study cities using digital data. Location-Based Social Networks (LBSNs) generate one specific type of digital data, which offers unprecedented geographic and temporal resolutions. We discuss fundamental concepts of urban computing leveraging LBSN data and present a survey of recent urban computing studies that make use of LBSN data. Besides, we point out the opportunities and challenges that those studies open up.

CCS Concepts: • **Information systems** → **Data mining**; *Data management systems*; • **Human-centered computing** → **Collaborative and social computing**; **Ubiquitous and mobile computing**; • **Applied computing** → *Law, social and behavioral sciences*;

Additional Key Words and Phrases: Urban Computing, Urban Informatics, Location-Based Social Networks, Big Data, Urban Sensing, City Dynamics, Urban Societies

ACM Reference format:
Thiago H. Silva, Aline Carneiro Viana, Fabrício Benevenuto, Leandro Villas, Juliana Salles, Antonio Loureiro, and Daniele Quercia. 2018. Urban Computing Leveraging Location-Based Social Network Data: a Survey. *ACM Comput. Surv.* 0, 0, Article 0 (November 2018), 37 pages.
https://doi.org/0000001.0000001

This work is partially supported by the project CNPq-UrbComp, process number 403260/2016-7, project EMBRACE Inria associated-team, as well as by the authors' individual grants and scholarships from FAPEMIG, CNPq, CAPES, Fundação Araucária. Fabrício is also supported by Alexander Von Humboldt-Foundation.

Authors' addresses: Thiago H Silva (thiagoh@utfpr.edu.br), Informatics, Federal University of Technology - Parana. Av. Sete de Setembro, 3165, Curitiba - PR, 80230-901. Brazil; Aline C Viana (aline.viana@inria.fr), Inria - 1 rue Honore d'Estienne d'Orves. Campus de l'Ecole Polytechnique, 91120 Palaiseau; Fabrício Beneveduto (fabricio@dcc.ufmg.br), Computer Science, Av. Antônio Carlos, 6627 - Prédio do ICEx Pampulha, Belo Horizonte, MG, Brasil; Leandro Villas (leandro@ic.unicamp.br), Computer Science, Av. Albert Einstein, 1251 Cidade Universitária, Campinas, SP, Brasil; Juliana Salles (jsalles@microsoft.com), Microsoft Research, 14820 NE 36th Street, Building 99, Redmond, Washington, 98052, USA; Antonio A. F. Loureiro (loureiro@dcc.ufmg.br), Computer Science, Av. Antônio Carlos, 6627 - Prédio do ICEx Pampulha, Belo Horizonte, MG, Brasil; Daniele Quercia (daniele.quercia@gmail.com), Bell Labs, Broers Building 21 J J Thomson Avenue, Cambridge, CB3 0FA, UK.

0360-0300/2018/11-ART0 $15.00
https://doi.org/0000001.0000001

1 INTRODUCTION

Urban computing is an interdisciplinary area in which urban issues are studied using state-of-the-art computing technologies. This area is at the intersection of a variety of disciplines: sociology, urban planning, civil engineering, computer science, and economics, to name a few [70, 78, 80, 124, 163].

More than half of the world's population today live in cities [94] and, consequently, there is enormous pressure on providing the proper infrastructure to cities, such as transport, housing, water, and energy. To understand and partly tackle these issues, urban computing combines various data sources such as those coming from Internet of Things (IoT) devices [153]; statistical data about cities and its population (e.g., the Census); and data from Location-Based Social Networks (LBSN), sometimes also termed as location-based social media [144, 161, 162].

One fundamental difference between data from LBSNs and data from other sources is that the former offers unprecedented geographic and temporal resolutions: it reflects individual user actions (fine-grained temporal resolution) at the scale of entire world-class cities (global geographic resolution). Never before it has been possible to study urban social behavior and city dynamics at such scale. Consequently, in the last few years, a significant number of efforts have been making use of LBSN data as a source to study aspects of our society in urban settings, opening space for a new avenue of applications in several segments, especially those related to the understanding of urban societies. This article is dedicated to surveying these efforts that explore LBSNs, discussing the related challenges and opportunities for the use of LBSN data to the field of urban computing.

Urban computing with LBSN data has its particularities. For instance, users who share data in Foursquare[1], a popular LBSN, usually have the goal of showing to their friends where they are while also providing personalized recommendations of places they visit. Nevertheless, when correctly analyzed for knowledge extraction, this data can be used to better understand city dynamics and related social, economic, and cultural aspects. To achieve this purpose, new approaches and techniques are commonly needed to explore that data properly. This survey provides an extensive discussion of the related literature, focusing on major findings and applications. Although its richness concerning knowledge provision, LBSN data presents several challenges, requiring extra attention to its manipulation and usability, which drives future research opportunities in the field of urban computing using LBSN data.

It is important to highlight that our work is complementary to two existing surveys in the area of urban computing [70, 163]. Broadly speaking, these efforts cover studies based on data collected from an existing city infrastructure and deployed sensors, usually dedicated to some predefined application (e.g., GPS, traffic, CDR, meteorological, RFID cards, as well as economic data). More specifically, Jiang et al. [70] focus on efforts that explore mobile phone traces, whereas, Zheng et al. [163] surveys a diverse set of techniques and methodologies to gather urban computing data, but only mention briefly few studies that explore LBSN data, neglecting key challenges that revolve around LBSNs. Our work also complements another previous study in the area of urban computing [130]. Silva et al. [130] aim to characterize key properties of participatory sensor networks data, mainly from LBSNs, and present some of the main challenges related to the exploration of this data source, not having the objective of covering a broad range of efforts related to urban computing with LBSN data. We hope that taken together, our effort and these existing ones, provide a broad perspective of urban computing studies and its development through the lens of different data-driven approaches.

The remainder of the study is organized as follows. Section 2 spells out what we mean by "urban computing" and dwells on the main data sources that are typically under study. Section 3 presents

[1] In 2014 Foursquare was divided into two parts, the one called Foursquare Swarm is responsible for letting users perform check-ins in places [16]. The other part, called Foursquare, focuses on the personal, location-based discovery. For simplicity, when we refer to Foursquare, we include the functionalities of Foursquare Swarm.

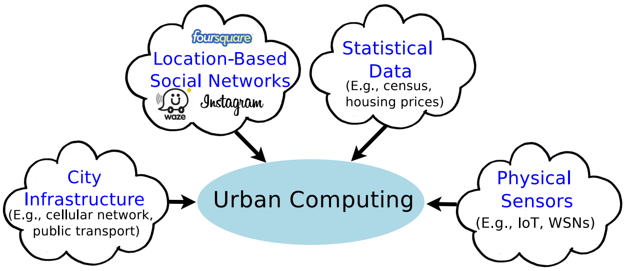


Fig. 1. Typical urban data sources.

some of the main advantages of LBSN data to perform urban computing studies. Section 4 presents a framework for urban computing with LBSN data. Section 5 focuses on recent research trends, while Section 6 focuses on the research questions that are still open before the concluding remarks in Section 7.

2 URBAN COMPUTING IN A NUTSHELL

The term “urban computing” was first introduced by Eric Paulos et al. [40] in the 2004 edition of the UbiComp conference and in his article The Familiar Stranger [111], published in that same year. Recently, Zheng et al. [163] presented a more precise definition for the term defining urban computing as a process of acquiring, integrating and analyzing a large volume of heterogeneous data produced by various sources in urban areas, for instance, vehicles, sensors, and human beings, to help solve various problems that cities face such as traffic congestion and air pollution. Thus, one of the primary objectives of that area is to help improve the quality of life of people living in urban environments.

Urban computing is a computer-mediated mean to understand the aspects of the urban phenomena and also provide estimates about the future of cities. It is an interdisciplinary area resulting from the fusion of computer science with traditional areas such as economics, geography, transportation, and sociology in the context of urban spaces. Within the computer science area, urban computing intersects with, for example, distributed systems, human-computer interaction, computer networks, and data mining.

As urban computing is quite comprehensive, a possible way of classifying various research efforts in this area is through the data considered. Figure 1 illustrates the main data sources used by studies in the area of urban computing. Each of these sources shown in the figure is described below:

- **Physical sensors:** They provide data that is obtained through the installation of sensors dedicated to particular applications, for example, inductive-loop traffic detectors to detect the volume of traffic in streets, sensors for monitoring air quality in various parts of the city, sensors for monitoring noise levels, and sensors in vehicles. Regarding the last example, it is increasingly common for buses, taxis, and private vehicles to have built-in GPS. The location of vehicles contributes, among other things, to the understanding of the city traffic. An et al. [3], for instance, developed an approach for measuring the evolution of recurrent urban congestions through the use of mobility data gathered from a GPS-equipped vehicle. One problem with the physical sensor data source is the difficulty in obtaining the data. In addition, there is a considerable cost for building a sensor network, when it is needed, and, generally, the deployment of sensors in the city demands special authorization from the city hall not easy to obtain. Besides, when it is desired to build a vehicular network, permissions and adaptations of vehicles of users are necessary, which could be troublesome.

- **Infrastructure of cities:** It provides data that is captured by taking advantage of existing city infrastructures that are created for other purposes. That includes cellular telephone networks. Cell phone signals from a large group of people have been used to characterize and predict individual's mobility and, consequently, to improve urban planning [101, 107]. Other examples of city infrastructures able to provide usage data include WiFi service providers or public transportation systems. In particular, in this latter, it is widespread the use of RFID cards to record users' bus and subway usage. Nevertheless, the difficulty here is that, typically, only the city or specific companies have access to this type of data.
- **Statistical data:** It consists of data related to a statistical study on a specific population, e.g., its demography, its health, and its social aspects. Also, data on urban dynamics, such as: economic, e.g., stock prices and housing prices; environment, e.g., flooding occurrences or agriculture details; safety, e.g., crimes committed and prisons made; and energy, e.g., gas consumption and electricity demand. It is possible to find multiple data sources on the Web from this category for some cities, and, typically, these data are open and accessible to obtain. This type of data source is gaining popularity, particularly, after government initiatives related to open data, such as Data.gov and Data.gov.uk. However, these data may not always be available for the location we may intend to study. Another difficulty is the diversity of formats in which the data are available, for instance in tables, maps, graphs, forms, among others [9].
- **Location-based social networks (LBSNs):** They are systems that combine online social networks features and also allow users to share data containing spatiotemporal information. Location-based social networks provide urban data that implicitly have social aspects, such as user's preferences and routines [144, 161, 162]. That is due to the active and voluntary user participation, acting as a sort of social sensor, in a distributed process of sharing personal and also data about various aspects of the city in Web services. This process is also known as participatory sensor network [129, 130], in fact, LBSNs are the most popular examples of it. One key point is that users in these systems can manually determine when, how, where, and what to share.

  LBSNs became quite popular partially due to the increased use of mobile devices, such as smartphones and tablets. These devices typically contain several sensors, e.g., GPS and accelerometer, enabling users to explore them to sense the environment, and, with that, having the opportunity to enrich LBSN data. In addition, users can also use their physiological sensors, e.g., vision, in this sensing process, producing more subjective data. LBSNs provide a new avenue of opportunities to access data on a global scale.

  There are several examples of location-based social networks already deployed on the Internet. For instance, (1) Foursquare, with more than 50 million users monthly using it [44], which allows users to share locations they are visiting with their friends; (2) Waze[2], with 65 million active monthly users [55], which serves to report traffic conditions in real-time; and (3) Instagram[3], a company with 700 million monthly active users in 2017 [59], which allows users to send real-time images to the system. Another example of LBSN is Twitter[4], a system with about 313 million monthly active users in 2016 [145], which allows its users to share personal updates as short text messages with up to 280 characters, known as "tweets".

[2] https://waze.com.

[3] https://instagram.com.

[4] https://twitter.com.

3 ADVANTAGES OF LBSN DATA FOR URBAN COMPUTING

This section highlights some of the main advantages regarding LBSN data to help the study of different phenomena related to urban societies.

Data from LBSN systems allow us to monitor various aspects of cities in near real-time. If we consider, for instance, traffic conditions, people could use their portable devices to share messages containing real-time information about demonstrations or accidents in the city, allowing, for example, unexpected problems to be identified by city authorities, as demonstrated by Pan et al. [108]. The real-time nature of LBSN data also has been demonstrated to be useful to identify earthquakes in (near) real-time [123].

Traditional data collection techniques, such as volunteer recruitment, census population surveys, and GPS track data, are not promptly available on the same scale reached by LBSNs. Taking as an example the human mobility, researchers in different areas, such as computer science and physics, has demonstrated interest in the modeling of mobility patterns [20, 23, 35, 52, 75, 85, 98, 139, 165]. Typically, researchers rely on a GPS track or cell phone usage data (i.e., Call Detail Records) to perform their studies; however, such data do not scale well or suffer from spatiotemporal sparsity. Besides, such data is commonly hard to obtain. Mobility patterns studies, as done by Zheng et al. [166] and Cheng et al. [24] would be hard (or not possible) using other data sources.

It is noteworthy also that LBSN data are distinct from GPS track data or Call Detail Records and have particular characteristics. For instance, photos in a photo-sharing service (e.g., Instagram), or check-ins in a location-sharing service (e.g., Foursquare), bring extra information on a specific location: A photo may convey information on the current situation within its location, while a check-in is usually associated with a location type, e.g., bar. Regarding, for example, mobility patterns investigation, LSBN data enable the study of the semantics of the mobility, i.e., the type of places users visit, as performed by Ferreira et al. [42]. Besides, since the access to the users' social network is typically available on LBSNs, it could also be explored to enrich our knowledge on different urban phenomena, including, mobility: Zhang and Pelechrinis [160] explore that to investigate the causes that provoke homophilous patterns in urban places.

The extra information provided by LBSNs on a geographic location can also enable other types of studies. For instance, it could be used to better understand the semantics of areas of the city, as done by Cranshaw et al. [31] and Noulas et al. [106]. Besides, it can represent valuable opinions that could also be explored to the study of well-being status of urban societies, as performed by De Choudhury et al. [34], and other city issues, as done by Quercia et al. [116].

Finally, LBSN data can be explored to study the social and economic aspects of city dwellers as well. For example, one may argue that a small amount of shared data in one area of the city might suggest that local population does not have proper access to technology, as the use of LBSN applications usually relies on smartphones and data plans that could be expensive in certain countries. In this direction, Venerandi et al. [149] showed evidence that the analysis of LBSN data allows the study of socio-economic issues of a city. Urban behavioral differences worldwide, such as the study performed by Silva et al. [133] and Gonçalves et al. [97] can also be enabled with LBSN data. Note that the same information used in those studies could be obtained using traditional methods such as questionnaires, but this process tends to be much slower and more expensive, which could prevent the observation of dynamic changes in a short period.

4 URBAN COMPUTING FRAMEWORK CONSIDERING LBSN DATA

Urban computing using LBSN data connects advanced management and analytic models of big and heterogeneous data generated by diverse location-based social networks as well as helps service and application improvement in different areas (e.g., urban planning and environmental conditions).

A general framework regrouping these components into layers is thus usually considered in the literature.

Figure 2 shows an overview of this framework, highlighting the three most important components: (i) management (Section 4.1); (ii) analytics (Section 4.2); and (iii) development of services and applications (Section 4.2). Hereafter, we briefly discuss these components and suggest the reading of [163] for a more detailed discussion on some of the techniques frequently used.

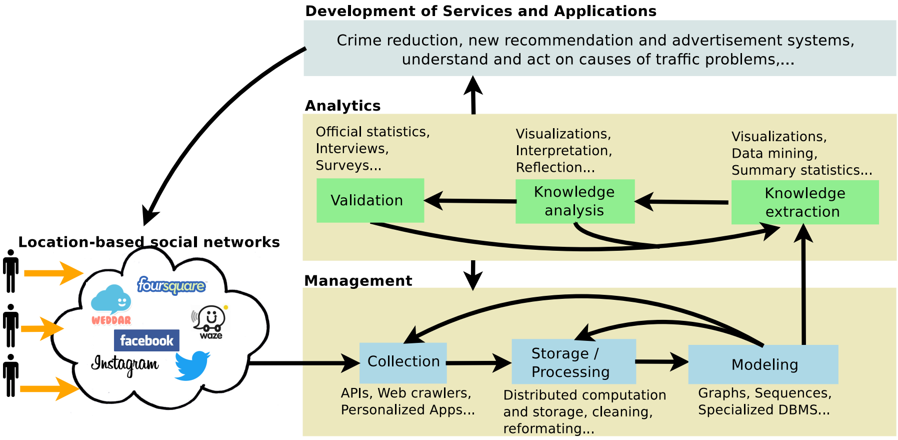

Development of Services and Applications
Crime reduction, new recommendation and advertisement systems, understand and act on causes of traffic problems,...
Analytics
Official statistics, Interviews, Surveys...
Visualizations, Interpretation, Reflection...
Visualizations, Data mining, Summary statistics...
Validation
Knowledge analysis
Knowledge extraction
Location-based social networks
foursquare
WEDDAR
waze
facebook
Instagram
Management
Collection
Storage / Processing
Modeling
APIs, Web crawlers, Personalized Apps...
Distributed computation and storage, cleaning, reformating...
Graphs, Sequences, Specialized DBMS...

Fig. 2. Overview of the urban computing framework with LBSN data.

4.1 Management

As illustrated in Figure 2, the management of LBSN data is composed of some important steps. The first one is the collection of LBSN data that could be obtained from several sources. LBSN data can be gathered, mainly, by APIs and Web crawlers.

There are two different key approaches to access APIs: (1) Based on streaming; (2) Based on requests. A streaming API allows one to gather data in (almost) real-time in which they are published in the system. It does that by keeping a persistent connection that continuously sends updated data to the user until the connection is terminated. Twitter Streaming API[5], for instance, allows one to gather near real-time public tweets. On the other hand, an API based on requests makes data available upon specific requests one might desire. The user makes a single request for data and gets the appropriate data in a single response. After the response is returned to the user, the connection closes only to be re-opened when the user sends another request. It is common to find programming libraries to ease the access to APIs. For example, the python libraries Tweepy[6] and TwitterAPI[7] are examples that ease the use of the Twitter API.

Not all LBSNs provide direct access to their public data through APIs. For this reason, it is necessary to use other strategies to obtain data, such as Web crawlers. Data collection through Web crawler depends on the data source structure and typically demands text mining efforts to parse and extract the desired information. Further discussion about data collection can be found in [130].

The second step refers to data storage and processing. This step might demand techniques to deal with a large volume of data. For this reason, we concentrate our discussion on this topic. Data from

[5] https://developer.twitter.com.
[6] http://www.tweepy.org.
[7] https://github.com/geduldig/TwitterAPI.

location-based social networks might increase quickly, knowing that, storage platforms have to be scalable, distributed, secure, fault-tolerant, and consistent [58]. We can explore available distributed file system technologies, such as Hadoop Distributed File Systems (HDFS), to help with this task. Regarding the processing of these data, one fundamental aspect is how to distribute computation, especially if real-time requirements have to be achieved. MapReduce is one of the first significant contributions on this front [36]. The idea behind this model combined with HDFS forms the Hadoop core[8], which allows the distributed processing of large datasets across clusters of computers. As an alternative, Apache Spark[9] is a general engine for large-scale data processing, and it is commonly used by applications that reuse a working dataset across multiple parallel operations. Examples of such applications are interactive algorithms for data analysis and machine learning [159].

Raw (untreated) LBSN data may not be in a convenient format to perform a particular analysis. Depending on the data type, it is possible to find semantic errors, missing entries or inconsistent formatting. In these cases, they need to be "cleaned" or "completed" before analysis. These tasks tend to be time-consuming and tedious, but they are essential in the production of new knowledge. The task of data cleaning and reformatting can provide insights on the assumptions that can safely be made about the data, on the peculiarities existent in the data gathering process, and on the analysis and models suitable to be applied. Data integration is a related issue at this stage, but it is discussed in Section 6. As another example of problem, LBSN data sometimes do not come with a location, as the case of a tweet, which may not be associated with a particular geolocation. Since linking data with a specific geolocation from where it was created enables a powerful way of modeling geographic aspects, this might be a procedure desired to be applied to the data. Several approaches have been proposed for this purpose, and an evaluation of some of them was performed by Jurgens et al. [73].

Typically, LBSNs provides high-dimensional data, and there are a variety of benefits to reduce dimensionality. In this direction, feature selection has proven to be an effective approach to deal with high-dimensional data for efficient data mining [88]. However, LBSN data can bring extra information that could make this task challenging in some cases. LBSNs can provide information regarding social aspects, such as who share the data, i.e., user-data relations, and who have a social connection to whom, i.e., user-user relations. In this context, it is possible to observe relevant correlations among objects related by social aspects. For instance, data, e.g., tweets, from two related users, e.g., two friends, tend to have higher similarity among topics. Tang and Liu [142], for instance, discuss more details about these situations and propose an approach for feature selection under these circumstances.

The final step refers to data modeling. It is common to assume that LBSN data is a collection of records, each of which consisting of a fixed set of attributes, with no explicit relationship among records or attributes. A data matrix is an example, where objects have the same fixed set of numeric attributes, e.g., timestamp and geospatial coordinate. In this way, objects, e.g., check-ins, can be thought as points in a multi-dimensional space, where each dimension represents a distinct attribute [141]. There are several other data formats to represent LBSN data, being popular the use of graphs [163]. Consider three users sharing data containing their locations in social media sites in different moments in time. This kind of data can be analyzed in many different ways. For instance, one could aggregate them in a directed graph in which nodes represent the user locations where the data has been shared, and edges connect locations that were shared by the same user. Using this graph one can extract mobility patterns of users, which could be useful, for example, to perform more efficient load management in urban infrastructure of mobile networks. Not surprisingly, knowledge discovery with LBSN data goes together with the use of network science theory [39, 102, 103, 158].

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

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

As shown in Section 5, widely known techniques used for graph analysis can be applied directly to the study of graphs derived from data that reflect city conditions.

Data from the mentioned example could also be modeled as a spatial trajectory, i.e., a model to represent data produced by a moving object (e.g., a user) in geospatial areas. This model is typically represented by a series of points in chronological order. Based on our example, consider $p_1\ p_2\ ...p_n$, where each point p represents a geospatial coordinate (i.e., the latitude and longitude of the check-in) and a timestamp when the user shared the data: $p = (lat, long, time)$. With that, we could extract a set of unique movements that share the property of visiting the same sequence of locations with close travel times [49]. We could also enrich this model with, for instance, the categories of the places visited. This could provide a way to give semantics to trajectories [109]. More details regarding this challenge are discussed in Section 6.2. Note also that the notion of trajectory could also be represented as graphs. In this direction, Guo and Liu [54] propose an approach that converts trajectory data to a graph in the context of vehicle trajectories.

LBSN data could also be represented in one geographic data model to be explored in a geographic information system (GIS) [91, 128]. In fact, any object that can be spatially located can be referenced using a GIS. A GIS enables the geographical combination of different unrelated data. This allows the provision of information on the environment highlighted by the LBSN data as well as the data visualization in the form of maps, supporting different analysis. This enables the recognition and analysis of important spatial relationships that might exist between spatial data [91]. For instance, we can infer possible explanations for a high concentration of check-ins in a particular area of the city by looking at the type of buildings in the surrounding areas. Other examples include the analysis performed in [34, 115, 149, 152]. Besides, hybrid models can also be built. Some of the challenges associated with that are discussed in Section 6.2.

4.2 Analytics and Development of Services and Applications

The analytics component is composed of the steps of knowledge extraction, knowledge analysis, and results validation. Knowledge extraction could explore different approaches, depending on the problem we are trying to address. A preliminary investigation of the data to understand their properties can help in the selection of analysis techniques, for this reason, it is a typical procedure performed. Visualization techniques, such as histograms and scatter plots, and summary statistics, such as mean and standard deviation of a set of values, are common methods used for exploring data properties in this preliminary investigation [141].

Visualization of raw data could provide valuable insights about the data, such as important features to be considered in a data mining process. In fact, visual analytics, the act of visually inspecting the data, penetrated many research efforts in the urban computing area. LBSN data brings new challenges in this context, such as large amounts of spatiotemporal data, which most current analysis methods cannot cope with [4, 5].

To illustrate new research efforts to help tackle this sort of problem, Watson [151] describe an approach for visualizing on a single image many events across multiple timescales, without the necessity of any zooming. In Figure 3, the author shows how this technique, called Time Maps, can provide valuable insights into the behavior of Twitter accounts. In that figure, each tweet is colored based on the time of day, and the time axes are shown in logarithmic scale. With this type of visualization, it is possible to see that there is the suggestion of two clusters representing two distinct modes of behavior, namely: “business as usual”, tweets are posted roughly once per hour; and “major events”, tweets occur in burst or rapid succession [151].

The clusters suggestion mentioned above helps to illustrate that performing data exploration with visualization techniques can aid to address some of the questions typically answered by data mining algorithms, which are fundamental for knowledge extraction. For instance, cluster identification

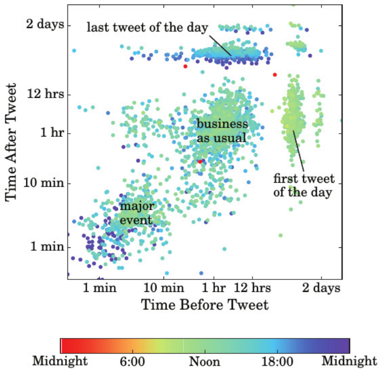


Fig. 3. Time map of tweets written by @BarackObama. The points are color-coded by time of day [151].

is a standard procedure for data mining. Its goal is to divide data into groups (clusters) of similar (or related) objects to one another and different (or unrelated) to other objects in other groups. It is possible to find different notions of a cluster that can be useful in different types of studies; several methods can be found in [12, 56]. Likewise, there are also other data mining tasks, for instance, association analysis, which is useful for discovering important relationships in large datasets, classification, the task of assigning objects to predefined classes, and anomaly detection, which aims to find objects that differ from the majority of other objects. These approaches encompass diverse applications, including urban computing ones [22, 56, 141, 163].

As location-based social networks enable the creation of a large amount of textual content by different users, data mining techniques to deal with such content deserves particular attention in the context of urban computing. Topic modeling is a tool usually used for discovering hidden semantic structures in a text data, and a popular technique is Latent Dirichlet Allocation (LDA) [15]. Another example is sentiment analysis and opinion mining, which aims to automatically extract opinions expressed in textual content shared in LBSNs [51, 53, 119, 127]. With that, subjective aspects expressed in the data might be understood and explored. Other examples of such techniques are available in [2, 13].

The obtained knowledge has to be studied, and several methods could help on that, for instance, visualizations. Visual analytics is a crucial procedure to better interpret the discovered knowledge, especially, complex ones. However, knowledge analysis is not dependent exclusively on visualizations because results could be outputted in different formats. A fundamental aspect in this step is the human capability of interpretation of results.

As shown in Figure 2, we might have to return to the knowledge extraction phase. This repeated *iteration cycle* might happen to gain new insights, and discover and correct mistakes. That is because much of the knowledge extraction is trial and error. We have to reflect on the results, making the comparison between outcomes variants to decide if it is necessary to explore new alternatives.

Another key step is the validation of results. As discussed in Section 6.1, LBSN data may suffer from representativeness or different types of bias. For this reason, when dealing with knowledge extracted from LBSN data it is important to contrast them with a ground truth, for example, data obtained in a traditional way, such as surveys, or official statistics provided by governments, especially when it is desired to use them to draw conclusions from city dynamics or urban societies. This validation step might not be possible or necessary for all types of problems.

Eventually, more experiments have to be repeated, entering in the same iteration cycle mentioned above, until useful information for an individual problem is obtained. It also might be the case of coming back to the data management steps, for example, to collect new data or to adjust the modeling. If this is not the case, we can explore the useful information obtained in new services or applications, such as: new recommendation and advertisement systems; understand the causes of traffic problems and act on them; and increase the safety of cities.

5 RECENT RESEARCH EFFORTS

Several studies in the urban computing literature explore location-based social networks data. We discuss in the following some of the most representative recent research efforts. We divide the discussed studies in six categories, as presented in Figure 4 : (i) Social and Economic Aspects (Section 5.1); (ii) City Semantics (Section 5.2); (iii) City Problems (Section 5.3); (iv) Urban Mobility (Section 5.4); (v) Health and Well-being (Section 5.5); and (vi) Events/Interest Identification and Analysis (Section 5.6). Nevertheless, it is important to mention that a particular study can belong to one or more categories, despite being discussed in a specific one.

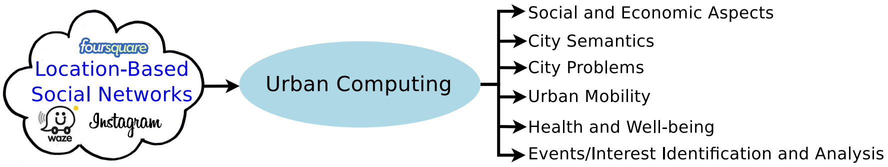


Fig. 4. Taxonomy of recent research efforts in the urban computing area that explore LBSN data.

5.1 Social and Economic Aspects

Aiming to better understand social patterns from LBSN data study, Quercia et al. [114] investigated how virtual communities, observed in the studied system, resemble real-life communities. The authors tested whether established sociological theories of real-life (offline) social networks are valid in these virtual communities. They have found, for instance, that social brokers in Twitter are opinion leaders who venture sharing tweets on different topics. They also discovered that most users have geographically local networks and that the most influential users express not only positive but also negative emotions.

In a similar direction, Joseph et al. [72] studied Foursquare data to identify groups of users in the city by analyzing users by the places they visit. They explore a clustering model inspired by the concept of topic modeling, more specifically Latent Dirichlet Allocation Model, which is, typically, used to study textual documents. In the model instantiation, each user's check-in is viewed as a word from a document representing a user, similar to text documents that can contain many words. Their approach enabled the identification of groups of users who represent spatially close groups and users who seem to have close preferences, confirming that geospatial and social homophily might be, indeed, essential features in clustering users into different communities [31, 62, 95].

Also, when investigating the social behavior in urban areas, an important question that emerges is: how similar/different is one culture from another? In this direction, it is known that eating and drinking preferences are important to describe strong cultural differences. Based on that, Silva et al. [133] proposed a new methodology for the identification of cultural boundaries and similarities between societies, which considers food and drink habits, as described briefly in Section 4.2. This analysis surprisingly tells a lot about the similarities and differences between cultures. The results for neighborhoods, cities, and countries, show how similar cultures are well separated using the methodology. This corroborates with other results in the context of food preferences in the Web, for instance, Wagner et al. [150] showed that dietary patterns observed in an online recipes system reflect well-known habits of the studied countries.

Hochman and Schwartz [63] also studied cultural differences using LBSN data, investigating color preferences in photos shared on Instagram. Hochman and Schwartz uncovered significant differences between images of countries with different cultures. In the same direction, Garcia-Gavilanes et al. [47] and Poblete et al. [113] studied how the usability behavior of Twitter change in different countries and what would be the potential reasons for these differences. In particular, in [47] the authors considered three aspects, widely studied, that vary across countries: Individualism, Power Distance [64], and Pace of Life [84]. They found that cultural differences are also evident in the way users use social media, not being only visible in the real world. Also, Garcia-Gavilanes et al. [46] performed a study of international communication on Twitter, which is a platform that allows users to maintain "weak social ties". The authors found that the best prediction of these ties happens when exploring both spatial distance, as well as socio-economic and cultural factors of the users involved.

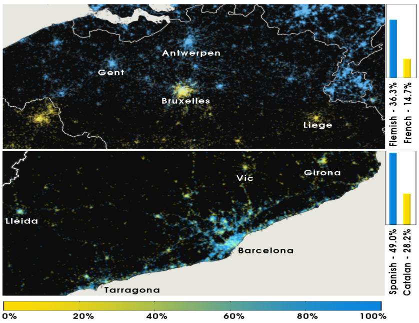


Fig. 5. Language division in Belgium and Catalonia, Spain [97].

In line with those studies, State et al. [140] considered Twitter communications to revisit the theory of changing international alignments proposed by Samuel Huntington [67]. The authors found the persistence of the eight culturally distinct societies postulated by Samuel Huntington, with the divisions being associated with distinctions in religion, spatial distance, economic development, and language. That is opposed to the hypothesis of the world without frontiers of cyberspace. Large-scale micro posts of Twitter are also studied by Gonçalves et al. [97]. The authors showed that the considered data enables the reproduction of the geographic location adoption of languages for different resolution scale, being able, for instance, to identify cultural diversity. As an example of their results, Figure 5 shows language division in Belgium (top) and Catalonia, Spain (down). Note that users use predominantly Flemish in the north part of Belgium, while French is the dominant language in the south of the country. Studying the results for Catalonia, Catalan and Spanish are mixed. The most popular language is Spanish; however, Catalan is also quite significant.

These studies are examples of the potential of empirically exploring large-scale sociocultural distinctions online. The investigation of sociocultural differences between distinct urban areas is important in several fields and can help many services and applications. For instance, since culture is an essential element for economic purposes, identifying similarities between geographically disconnected places might be necessary for enterprises that have business in one country and desire to verify the similarity of preferences across distinct markets [133].

Table 1. Summary of all discussed studies of the class Social and Economic Aspects.

Publication		Dataset				Granularity of analysis	Main technique (s)	Focus
Name	Date	Source	Time	Volume	Coverage			
Quercia et al. [114]	01/06/12	Twitter	Sep to Dec 2010	~258K profiles and ~31M tweets	City	London	Network analysis, sentiment analysis (text)	Study about whether sociological theories of offline social networks is still valid in Twitter.
Joseph et al. [72]	Sep 2012	Foursquare	Sep 2010 to Jan 2011 and June to Dec 2011	~18M check-ins	Worldwide	GPS	Topic modeling (LDA)	Approach to identify groups of people in the city by analyzing users by the places they visit.
Silva et al. [133]	01/06/14	Foursquare	May 2012	~4.7M check-ins	Worldwide	City, Country	Data characterization, clustering (k-means), dimen. reduction (PCA)	Methodology for the identification of cultural boundaries and similarities between societies, considering food and drink habits.
Hochman and Schwartz [63]	01/06/12	Instagram	Jan to Feb 2012	~550K photos	City	City (NYC and Tokyo)	Image processing	Investigate color preferences in shared photos on Instagram.
Garcia-Gavilanes et al. [47]	01/07/13	Twitter	Mar to May 2011	~2.34M users (with associated features)	Worldwide	Country	Statistical analysis	Study how the behavior of Twitter use varies among countries, considering three aspects that vary across countries.
Poblete et al. [113]	Oct 2011	Twitter	2010	~6.2M users and ~5.2M tweets	Worldwide	Country	Network analysis, sentiment analysis (text)	Study of possible differences and similarities in several aspects of the use of Twitter.
Garcia-Gavilanes et al. [46]	01/02/14	Twitter	01/03/11	~13M users (with associated features)	Worldwide	Country	Network analysis, regression (linear regression)	Study international communication on Twitter.
State et al. [140]	May 2015	Twitter	Sep 2009	~51.9M users and ~1.9B follow links	Worldwide	Country	Network analysis	Study the theory of changing international alignments of Samuel Huntington.
Gonçalves et al. [97]	Apr 2013	Twitter	Oct 2010 to May 2012	~400M tweets	Worldwide	Country	Data characterization	Study worldwide linguistic indicators and trends through the analysis of tweets.
Karamshuk et al. [74]	01/06/13	Foursquare	May to Nov 2010	~621K check-ins	City	City (NYC)	Data characterization, regression (various)	Investigate the optimal allocation problem of stores in urban areas.
Lin et al. [87]	01/07/16	Facebook (places)	-	~21K Facebook pages	City	City (Singapore)	Regression (various)	Study the identification of an optimal physical location for a business by looking at Facebook Pages data.
Llorente et al. [89]	May 2015	Twitter	Nov 2012 to Jun 2013	~19.6M tweets	Country	City (several in Spain)	Data characterization, regression (linear regression)	Demonstrate that behavioral features related to unemployment can be recovered from posts of users shared on Twitter.
Hristova et al. [65]	Apr 2016	Foursquare (1) and Twitter (2)	Dec 2010 to Sep 2011	1: ~550K check-ins; 2: ~38K users (with metadata)	City	City (London)	Data characterization, network analysis	Propose a model to capture the relationship between users and the locations they visited.

Related to the economic aspect of cities, Karamshuk et al. [74] studied how to best allocate retail stores in the city. The authors explored data from Foursquare to analyze how the popularity of three international business chains is defined by the number of check-ins in New York City. A set containing several features were evaluated, modeling semantic and spatial information regarding the patterns of users' movements in the vicinity of the studied area. The authors noted that, for example, the existence of locations that naturally attract many users, such as a railway station, is one of the most reliable indicators of popularity. Similarly, Lin et al. [87] also studied the identification of an optimal physical location for business by looking at Facebook Pages data. Among other results, they show that the popularity of neighboring business is a crucial feature in this task.

Llorente et al. [89] demonstrated that behavioral characteristics connected to unemployment could be obtained from the posts of users shared on Twitter. As shown using their analyzed dataset, users in neighborhoods with elevated unemployment rate present distinct social interactions, daily activity, and mobility compared to those in neighborhoods with low unemployment rates. Hristova et al. [65], inspired in multilayer networks, proposed a model to capture the relationship between users and the locations they visited. This model couples the network of places and the social network of users, by connecting users to locations in case they visited them. To exemplify their model, they used check-ins from Foursquare and the users' social network. They found, among other results, that their approach could predict urban area gentrification. Table 1 summarizes the studies discussed in this section, also presenting extra information about the studies not discussed in the text.

5.2 City Semantics

LBSN data can be explored to change our notions of space and perception of physical boundaries, i.e., better understand our perceived physical limits in urban environments, as well as to better understand city dynamics. Some studies in this direction are discussed as follows.

Using Foursquare data, Cranshaw et al. [31] presented a model to identify different regions of a city that reflect current patterns of collective activities. By doing so, they introduce new boundaries for neighborhoods. The main idea is to uncover the nature of local urban areas, which tend to be dynamic, considering the social proximity (obtained from the distribution of users who check-in) and the spatial proximity (obtained from geographical coordinates) of locations. For that, the authors developed a model that groups similar locations considering social and spatial characteristics, according to the considered data from Foursquare. Each cluster represents different geographic boundaries of the neighborhoods. The clustering method used is a variation of the spectral cluster proposed by Ng et al. [104]. Figure 6 shows two clusters, discovered in New York City (numbers 1 and 2 in the figure). Black lines represent the official city limits.

Noulas et al. [106] introduced a method to classify users and areas of a city exploring the types (categories) of places used by Foursquare. The method could be explored to discover communities of users visiting similar type of places. This is useful for comparing urban areas within and between cities or in recommendation systems. More specifically, the authors take into account the activity of Foursquare users in New York. The data considered is illustrated by Figure 7, where the center of a circle represents a location and its radius the popularity concerning the number of check-ins. Each color represents one of the main categories considered by Foursquare, eight in total. For each studied area, the activity performed by the users is calculated based on the visits to places available in the area under study. Thus, the similarity between two areas is estimated among the observed activities.

Silva et al. [135] introduced a technique named *City Image*, which offers a visual summary of the city dynamics exploring users' movements. This approach explores urban transition graphs $G(V, E)$ (also called place networks [65]) to map user movements between city locations. This type of particular graph represents, for example, a set of places V in the city (i.e., vertices) and a set E of pairs of V that represent the movement of users in the city (i.e., edges). Place networks represent an

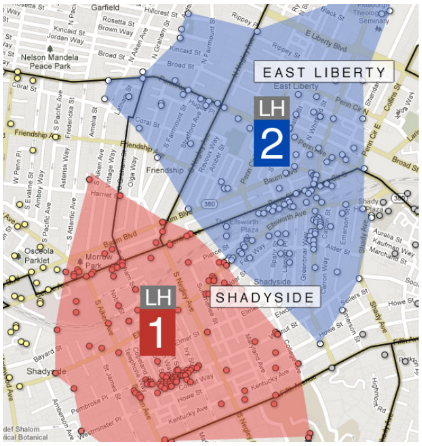


Fig. 6. Clusters found in New York City [31].

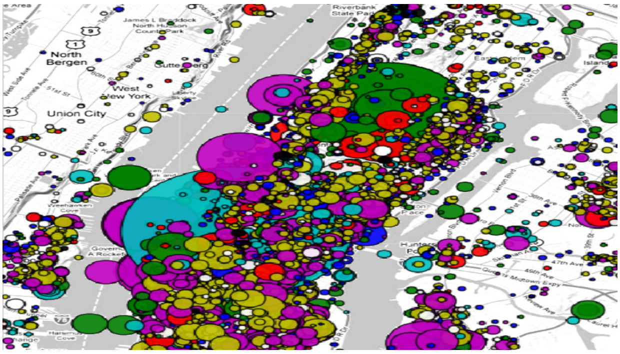
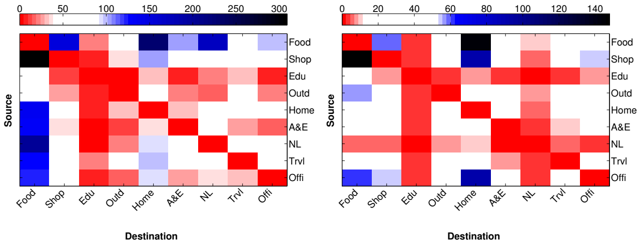


Fig. 7. Foursquare users activity in New York. Categories and the assigned colors: magenta (Nightlife), yellow (Food), red (Arts & Entertainment), cyan (Travel), white (Shops), green (Parks & Outdoors), blue (Home/Work/Other), black (College & Education) [106].

example of an informative model on the dynamics of the city and urban social behavior. City Image considers a place network where a node $v_i \in V$ is the category of a specific location (for example, *Arts & Entertainment*) and a directed edge $(i, j) \in E$ marks a transition between two categories performed by the same user [135].

Two examples of the City Image technique for São Paulo and Kuwait are presented in Figures 8(a) and 8(b), respectively. Each cell in the image represents how favorable is a transition from a particular category in a certain location (vertical axis) to another category (horizontal axis). In the image, blue represents favorability, red indicates rejection, and white indifference. In both cases, the images represent activities performed on the weekend during the night (representative period of free time, i.e., without typical predefined routines). Note, for example, the lack of favorable transitions to *NL* (Nightlife Spot) category in Kuwait. For São Paulo this is not the case, the transition *Food* ⟶ *NL* is very favorable to occur. This indicates that in São Paulo users like to visit venues related to food (*Food*) before visiting nightclubs (*NL*) related venues. Analyzing the case of Kuwait, users, instead, are more likely to make the transitions *Shop* ⟶ *Food* and *Food* ⟶ *Home* in the weekend's evenings [135].



(a) São Paulo (b) Kuwait

Fig. 8. City Images to São Paulo (SP) and Kuwait (KU) during weekends at night. Abbreviations of category of venues (names adopted by Foursquare): Arts & Entertainment (A&E); College & Education (Edu); Great Outdoors (Outd); Nightlife Spot (NL); Shop & Service (Shop); and Travel Spot (Trvl) (images from [129]).

There are several other important studies in this direction. For instance, Long et al. [90] explored a dataset collected from Foursquare to introduce an approach based on topic model to study the intrinsic relations among the different venues in an urban area. Considering a sequence of users' check-ins, they assume that the venues that appear together in several sequences will likely represent geographic topics, for example, indicating coffee shops people typically visit before going to a mall. In their proposal, they employ the Latent Dirichlet Allocation approach to identify the local geographic topics. Similarly, Frias-Martinez et al. [45] explored a Twitter dataset and presented a technique that, by studying tweeting patterns, identify the types of activities that are most common in a city. Their results suggest that geolocated tweets could be an essential data source to describe dynamic urban areas, which tend to be costly using other conventional approaches.

Jiang and Miao [69] demonstrated that LBSN data could serve as a proxy for studying the underlying evolving mechanisms of cities. In their study, instead of using conventional definitions of cities, they use the concept of geographic events clustered spatially, for instance, groups from geographic locations of particular users present in the data, to define what they call "natural cities". Studies in that direction are interesting because data to follow the changes of cities are scarce. Vaca et al. [146] considered the problem of mapping the functional use of city areas. For example, uncover if a particular area of the city is a hotel area. They propose an approach that clusters points based not only on their density, typically used in spatial clustering algorithms, but also on their semantic relatedness. Using the proposed approach, they demonstrated that Foursquare data could help on this task.

Furthermore, we can also mention the following studies. Falcone et al. [41] proposed a methodology to identify venues categories from geolocated tweets. For that, they extract spatiotemporal patterns from tweets and use them to build a framework to infer the category of the visited places. They address the problem as a classification task, achieving promising results in the identification of place semantics based purely on spatiotemporal features from tweets. Naaman et al. [100] study social media activity in different geographic regions. Perform this type of study is not trivial because LBSN data, especially Twitter, the one they used, can be noisy. Besides, content can fluctuate widely in response to events and other breaking news, from Carnival to the news about a tragedy. Since the content can expose a varied set of temporal patterns, they characterize within-day and across-day variability of diurnal patterns in cities. Their study shed some light on possible reasons that could explain the differences between cities regarding the aspects under consideration. Their results could be useful, for instance, in the comparison of cities. Le Falher et al. [82] focused on the study of measures and characteristics that could be explored to quantify how similar city neighborhoods are. In this regard, the authors take into account the activities that take place in certain areas. For example, some users might visit a specific neighborhood mainly for shopping, while others for drinking. Their methodology explores those type of activities that are observed in LBSN data.

These studies show that LBSN data may provide essential characteristics of areas, as well as the behavior that users perform on them. LBSN data enables such type of understanding of the city, a task that would be hard to do using other urban data sources. This section discussed some of the primary studies related to city semantics investigation with LBSN data. However, indeed, other relevant related works in the literature could be mentioned here, such as [26, 33]. Table 2 summarizes the studies grouped in the City Semantics category, providing also extra information about the studies not discussed in the text.

[10] Foursquare-like application closed in 2010.

[11] http://www.untappd.com.

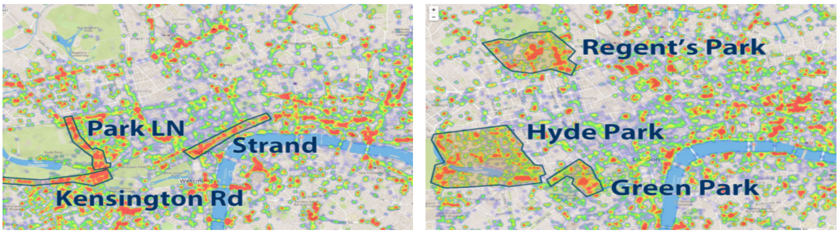
Table 2. Summary of all discussed studies of the class City Semantics.

Publication		Dataset				Granularity of analysis	Main technique (s)	Focus
Name	Date	Source	Time	Volume	Coverage			
Cranshaw et al. [31]	01/06/12	Foursquare	Sep 2010 to Jan 2011 and June to Dec 2011	~18M check-ins	Worldwide	Neighb. (Pittsburgh)	Clustering (spectral clustering)	Model to identify distinct regions of a city that reflect current patterns of collective activities.
Noulas et al. [106]	01/07/11	Foursquare	May to Sep 2010	~12M check-ins	Worldwide	City (NYC and London)	Clustering (spectral clustering)	Strategy to classify users and urban areas exploring the categories of places considered by Foursquare.
Silva et al. [135]	Dec 2014	Foursquare	Apr 2012	~4.7M check-ins	Worldwide	Cities	Network analysis, clustering (hierarchical)	Technique that summarizes visually city dynamics based on people's mobility.
Long et al. [90]	Sep 2012	Foursquare	Feb to May 2012	~800K check-ins	City	City (Pittsburgh)	Topic modeling (LDA)	Approach to investigate relations among distinct venues in an urban area.
Frias-Martinez et al. [45]	Sep 2012	Tweets	Oct to Dec 2010	~24M tweets	Worldwide	City (NYC)	Clustering (k-means, mean-shift), self-organizing map, voronoi tessellation	Strategy to study landmarks and land uses exploring the information provided by geolocated tweets.
Jiang and Miao [69]	01/11/14	Brightkite[10]	Apr 2008 to Oct 2010	~2.7M check-ins	Country	Cities	Network analysis	Show that LBSN data could be used for studying the underlying evolving mechanisms of cities.
Vaca et al. [146]	May 2015	Foursquare	-	~115K venues	City	GPS	Clustering (agglomerative hierarchical, DBSCAN)	Propose a framework for discovering functional areas of cities.
Falcone et al. [41]	May 2014	Twitter	Jun to Nov 2013	~7.4M tweets	City	GPS	Clustering (OPTICS [6]), classification (various methods)	Methodology to identify venues categories from geolocated tweets.
Naaman et al. [100]	01/06/12	Twitter	May 2010 to May 2011	All public tweets	Worldwide	City	Text mining	Study social media activity in different geographic regions.
Le Falher et al. [82]	May 2015	Foursquare	Mar to Jul 2014 and Sep 2010 to Jan 2011	~3M check-ins	City	City	Clustering (k-means, DBSCAN), classification (k-nn)	Focused on the study of measures and features that can be used to express the similarity of neighborhoods.
De Choudhury et al. [33]	01/06/10	Flickr	-	-	City	City	Network analysis	Automatically construct travel itineraries based on Flickr photos.
Chorley et al. [26]	May 2016	Untappd[11]	Aug to Dec 2015	~5.3M check-ins	USA and Europe	City	Data characterization	Characterization of user drinking habits around the world.

5.3 City Problems

Collecting data on problems faced by cities can be facilitated by using Web systems such as Colab.re[12]. This type of system enables users to create, view and share problems of various kinds about the city. Besides general systems such as Colab, there are also specialized applications for monitoring specific

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



(a) Gas emission (b) Nature

Fig. 9. Heatmaps of smell-related tag intensity in London, the more red the higher is the value [116].

issues about the urban environment. For example, NoiseTube is an LBSN that allows users to share noise level in a certain area of the city [93].

Exploring NoiseTube, D'Hondt and Stevens [37] performed a study to map noise levels in Antwerp, Belgium. One of the objectives was to investigate the quality of the noise maps constructed by participatory sensing [21, 129], in comparison to the official noise maps based on simulation. For that, many calibration experiments were carried out, investigating several aspects of noise patterns. The authors were able to construct noise maps with a margin of error comparable with official noise maps based on simulation.

In addition to these initiatives, New York City has made available a system called 311[13] to enable users to complain about problems of the city using a mobile application. Each data (complaint) has a location, time and date, and in some cases, detailed complaint information, such as loud music or building noise (for noise problems). Using the data from 311, and also from Foursquare and Gowalla[14], Zheng et al. [164] infer a noise pollution indicator at different times of the day for regions of New York. By exploring the considered data, it is possible to verify the noise patterns of a given location (e.g., Times Square), and how it changes over time. Noise information not only can facilitate the quality of life of an individual (for instance, help identify a quiet place to live), but also can assist cities in combating noise pollution.

Studying a different problem in a similar direction, Quercia et al. [116] explored the possibility of using shared data in social media to map smells perceived in different regions of the city. The results are promising and show that this may be a new way to classify areas according to their most characteristic smell. To perform this study, the authors considered Instagram, Flickr[15], and Twitter data. They combined photo tags and tweets with the words of an existing "smell dictionary". Then they analyzed these occurrences in the city and show, for instance, that the smell of nature is strongly observed in parks and the smell of gas emission is commonly observed in streets with heavy traffic. Figure 9 illustrates this result, showing that, as one expects, the nature category is present where the gas emissions category is absent, and vice-versa. Focused on the city traffic problem, Ribeiro et al. [118] studied the possibility of using LBSN data as a feature for predicting heavy traffic. The authors noted that data from Instagram and Foursquare are correlated with heavy traffic and, thus, it could be explored in more efficient congestion prediction models.

Gender segregation can also be considered a problem in cities. Traditional ways to investigate differences between distinct gender groups depend on, for instance, questionnaires, which tend to be expensive and do not scale up easily. In addition, data gathered under such circumstances are typically released after long time periods. Thus, these data do not enable the fast identification of changes in

[13] http://www1.nyc.gov/nyc-resources/service/5460/nyc311-mobile-app.
[14] Foursquare-like application closed in 2012.
[15] https://www.flickr.com.

Table 3. Summary of all discussed studies of the class City Problems.

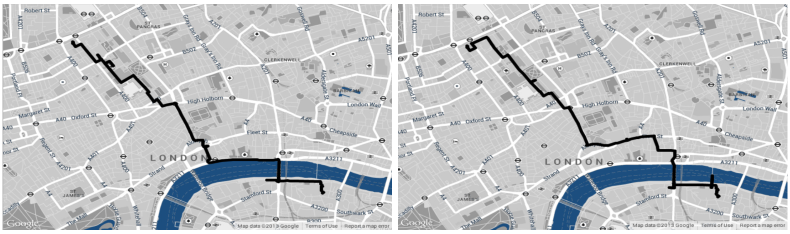
Publication		Dataset				Granularity of analysis	Main technique (s)	Focus
Name	Date	Source	Time	Volume	Coverage			
D'Hondt and Stevens [37]	Sep 2013	NoiseTube	01/11/10	∼85K measurements	City	City (Antwerp)	Statistical analysis	Study the quality of the noise maps constructed by the collaboration of users.
Zheng et al. [164]	Sep 2014	Foursquare (1), Gowalla (2), and 311 (3)	1: May 2008 to Jul 2011; 2: Apr 2009 to Oct 2013; 3: May 2013 to Jan 2014	1: ∼173K check-ins; 2: ∼127K check-ins; 3: ∼67K complaints	City	City (NYC)	Data characterization, tensor decomposition	Infer the situation of noise in different periods for distinct region of NYC.
Quercia et al. [116]	May 2015	Flickr (1), Instagram (2), Twitter (3)	2: Dec 2011 to Dec 2014; 3: year 2010 and Oct 2013 to Feb 2014	1: 17M photos; 2: 154M photos; 3: 5.3M tweets	Worldwide	City (London and Barcelona)	Text mining, clustering (graph-based)	Explored the possibility of using shared data in social media to map smells perceived in different regions of the city.
Ribeiro et al. [118]	Sep 2014	Instagram (1) and Foursquare (2)	Jun to Aug 2013	1: 1M photos; 2: 65K check-ins	City	City (NYC)	Data characterization	Study the possibility of using LBSN data as a feature for predicting heavy traffic.
Muller at el. [99]	May 2017	Foursquare	Apr to May 2014	∼2.9M check-ins	Worldwide	Country, City, GPS	Outlier detection, clustering (k-means)	Approach to obtain and explore data that could help the study of global gender differences study.

the dynamics of societies. Also, the results from studies of gender differences between regions are typically released only for large geographic regions, usually countries. Therefore, although studies based on questionnaires could be performed in small regions, such as a city or a particular place, for example, a restaurant, information regarding gender differences is not typically released on fine spatial granularities. In that sense, Muller at el. [99] reveal another way to obtain and explore similar data that could help the study of global gender differences study. They propose to explore publicly available LBSN data to numerically extract differences between female and male preferences for locations in distinct urban regions around the world at different spatial granularities. Comparing their results with an official gender difference index, they found evidence that their methodology might identify important characteristics of gender differences. This study motivates the investigation of new approaches that use LBSN data in the future construction of indices that express gender differences. Table 3 summarizes the studies discussed in this section. It also contains extra information not discussed in the text.

5.4 Urban Mobility

We present now studies that focus on investigating urban mobility patterns of users with LBSN data. The investigation of user mobility is valuable for several purposes. It helps to understand, for example, how users spend time on distinct tasks. In addition, it could enable the design of new applications to aid traffic engineers to understand better how people move in urban areas.

Quercia et al. [115] proposed a methodology for recommending routes that take into account not only the smallest path but also emotional characteristics, for example, beauty. Not always the shortest way is what we would like to go through. A tourist, for example, could opt for a more beautiful route, even if the distance is a little higher. To quantify how pleasant urban areas are, the authors used data from a crowdsourcing system. After that, they build a graph whose nodes are places and edges on this graph connect geospatial neighbors. This graph allows the discovery of pleasant paths.



(a) Shortest path (b) Most beautiful path

Fig. 10. Maps showing different paths between the same places [115].

Figure 10 shows two paths between the same places in the city of London, where one is the shortest (Figure 10a), and the other is the most beautiful (Figure 10b). The authors also generalized their proposal by showing an approach that predicts the beauty characteristic of an urban area exploring Flickr metadata. Users ascertained the effectiveness of their results, indicating that the proposed approach might be explored in practice in new mapping applications.

Ferreira et al. [42] studied the urban mobility of tourists using check-ins shared in Foursquare, by analyzing at when and where they visit particular locations. In order to accomplish this goal, they build a graph containing temporal attributes. The authors used a directed weighted graph $G = (V, E)$, where the nodes ($v_i \in V$) are particular locations in the studied area at a specific time, and a directed edge (i, j) exists from node v_i to v_j if at a particular time a user gave a check-in at a location v_j right after giving a check-in in v_i. The labeling of the vertices follows the rule: the location's name merged with the hour (integer value) of the check-in. For example, a check-in at Empire State Building at $11 : 00$ a.m. would be "Empire State Building [11]". Edges' weights are incremented when another user performs the same transition, i.e., the weight $w(i, j)$ of an edge is the total number of movements that were observed from node v_i to node v_j. The authors show that their methodology could be valuable, for instance, in a novel recommendation service that would recommend which venue to visit after visiting a particular venue at a particular time.

In a similar direction, Zheng et al. [166] demonstrated that geolocated photos shared on Flickr could provide a useful solution to analyze tourist mobility automatically. They propose an approach to analyzing tourist mobility using regions of attraction and topological features of trip routes followed by distinct tourists. Among other results, they note that despite the variety of trip routes, some tourists' groups do share common routes. That is more evident when they go to regions of attractions that are more similar to each other. Nguyen and Szymanski [105] used check-ins from a location-sharing service to create and test models of human relations and mobility. Nguyen and Szymanski introduced a mobility model exploring users' friendship, considering social ties, envisioning to offer a human mobility model more precise. This model enabled the authors to study the frequency that friends move together. Such type of model could be used to improve the precision of a variety of services, for instance, transportation systems and traffic engineering in communication networks.

Furthermore, Zhang and Pelechrinis [160] study the causes that provoke homophilous patterns noted in visits performed by users in real world places, when exploring check-ins data from a location-sharing service. Besides, they also investigate the levels of social selection and peer influence in the studied service. Social selection is the mechanism that makes users associate with other users who are similar to them concerning the characteristic under study, while peer influence refers to the influence that one user may have to another on decisions related the characteristic under examination. Among their results, they show that peer influence tends to happen while friends are in the proximity, besides, and it depends on the context. Machado et al. [92] studied the impact of mobility of users

Table 4. Summary of all discussed studies of the class Urban Mobility.

Publication		Dataset				Granularity of analysis	Main technique (s)	Focus
Name	Date	Source	Time	Volume	Coverage			
Quercia et al. [115]	Sep 2014	Flickr	-	5M photos	London and Boston	GPS	Network analysis, text mining, regression (linear regression)	Methodology for recommending routes that take into account not only the smallest path but also emotional characteristics.
Ferreira et al. [42]	01/11/15	Foursquare	May 2012	~247K check-ins	London, Rio de Janeiro, NYC, Tokyo	GPS	Data characterization, network analysis	Study of urban mobility of tourists, proposing an approach to identify when sights are popular.
Zheng et al. [166]	May 2012	Flickr	-	~769K	London, Paris, New York City, and San Francisco.	GPS	Clustering (hierarchical, DBSCAN, mean shift), markov model	Approach to analyze tourist movement according to regions of attraction and topological characteristics of travel routes.
Nguyen and Szymanski [105]	Aug 2012	Gowalla	Sep to Oct 2011	~26M check-ins	Worldwide	GPS	Markov model	Mobility model based on friendship envisioning to offer a human mobility model more precise.
Zhang and Pelechrinis [160]	Apr 2014	Gowalla	May to Aug 2010	~10M check-ins	Worldwide	GPS	Clustering (DBSCAN), network analysis	Study the reasons behind the homophilous patterns observed in visits made by users in real-world venues.
Machado et al. [92]	01/06/15	Foursquare	120 days in 2014	-	Six cities in different countries	Cities	Data characterization	Study the impact on the mobility of users according to different weather conditions.
Cheng et al. [24]	01/07/11	Several location sharing services	Sep 2010 to Jan 2011	~22M check-ins	Worldwide	City, country, global	Data characterization, text mining, sentiment analysis (text)	Study mobility patterns provided by check-ins and explore aspects that contribute to the mobility.
Cho et al. [25]	Aug 2014	Gowalla (1) and Brightkite (2)	1: Feb 2009 and Oct 2010 for Gowalla and 2: Apr 2008 to Oct 2010	1: 6.4M and 2: 4.5M check-ins	Worldwide	GPS	Probabilistic modeling	Investigate the relation between user geospatial mobility, its temporal dynamics, and the connections in the user's social network.

observed through Foursquare check-ins according to different weather conditions. The results suggest a behavior change within a specific temperature range for the studied cities. Besides those studies, several others also aim to study user mobility with LBSN data, such as [24, 25]. The studies presented in this section illustrate the growing interest and potential of using LBSN data to study large-scale human mobility patterns. Table 4 summarizes the studies grouped in the category Urban Mobility. This table also provides extra information about the studies not mentioned in the text.

5.5 Health and Well-being

Several studies have shown evidence that LBSN data could also be used to improve our understanding of urban societies regarding its health and well-being status. De Choudhury et al. [34] used Instagram posts to understand food choices in "Food deserts" in the United States. Food deserts are urban areas characterized by inadequate access to affordable and healthy food, known to be connected with diet-related health issues, for instance, obesity. In addition to that study, using Instagram posts together with Foursquare data, Mejova et al. [96] identified obesity patterns based on the content shared by users.

Schwartz et al. [126] studied the geographic variation in well-being using tweets. For that, they mapped tweets from the United States counties where they were shared and correlated the words used on the messages (exploring word topics generated by LDA), with life satisfaction, as captured by surveys answered in these places. The language applied by the users was found to be an essential feature to predict the subjective well-being of users. In the same direction, Paul and Dredze [110] explored Twitter posts to find health-related terms, for instance, symptoms, to show geospatial patterns in syndrome control. More recently, Culotta [32] correlated Twitter activity and found a significant correlation with official health statistics, such as obesity and access to healthy foods. Comparing to models based on demographic features alone, Culotta shows that incrementing models with information derived from Twitter increase the predictive accuracy of these health-related statistics. This suggests that their approach might complement traditional ones.

Kershaw et al. [76] examine tweets to observe the alcohol consumption rate. They applied their approach to visualize changes in drinking patterns throughout different areas in the United Kingdom. The results were validated with a ground truth (official data about alcohol consumption in the United Kingdom). To illustrate another type of effort, we can also mention studies that investigated deprivation in cities. Venerandi et al. [149] propose an approach to compute urban deprivation. Their approach explores LBSN data to discover urban characteristics that exist in a neighborhood, for that they explored data from Foursquare and OpenStreetMap[16]. Among other applications, the proposed method enables the development of "neighborhood profiling". City dwellers could explore it, for instance, as criteria to decide where to buy a house. Table 5 summarizes the studies discussed in this section, also providing extra information not discussed in the text.

5.6 Events/Interest Identification and Analysis

Thanks to the (near) real-time nature of LBSN data, the identification and study of events, anomalous or not, become more feasible. Events can be natural, such as earthquakes, or unnatural, such as changes in the stock market. For instance, Sakaki et al. [123] investigated the real-time sharing of earthquakes messages on Twitter and introduced an approach to detect the occurrence of events in this direction. To demonstrate the efficacy of their approach, they constructed an earthquake warning service in Japan that was able to discover, with significant accuracy, earthquakes announced by the Japan Meteorological Agency. They used a classifier of tweets exploring characteristics, such as the number of words, keywords, and its context. Next, the authors built a probabilistic spatiotemporal model for the event under study, which was able to discover the trajectory and the center of the event site, as illustrated by Figure 11.

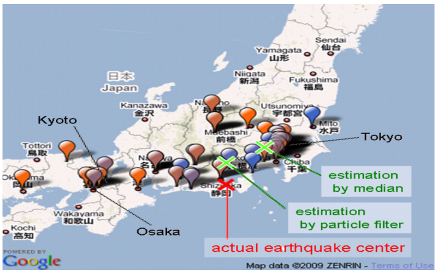


Fig. 11. Estimate of earthquake location [123].

[16] https://www.openstreetmap.org.

Table 5. Summary of all discussed studies of the class Health and Well-being.

Publication		Dataset				Granularity of analysis	Main technique (s)	Focus
Name	Date	Source	Time	Volume	Coverage			
De Choudhury et al. [34]	Feb 2016	Instagram	Jul 2013 and Mar 2015.	14M posts	USA	Regions	Classification (k-nn, SVM), topic modeling (LDA), dimensionality reduction (PCA)	Study of food choices in food deserts areas.
Mejova et al. [96]	May 2015	Foursquare (1) and Instagram (2)	1: Dec 2010 to Sep 2011; 2: Sep to Nov 2014	1: ~195K food places; 2: ~21M posts	USA	GPS	Data characterization	Study of food choices in food places.
Schwartz et al. [126]	01/07/13	Twitter	Nov 2008 and Jan 2010	82M tweets	USA	Counties	Topic modeling (LDA), text mining, regression (linear regression)	Study of geographic variation in well-being.
Paul and Dredze [110]	01/07/11	Twitter	May 2009 to Oct 2010	1.63M tweets	USA	Regions	Topic modeling (LDA), text mining	Investigate a variety of public health data that can be automatically extracted from Twitter.
Culotta [32]	Apr 2014	Twitter	Dec 2012 to Aug 2013	4.3M tweets	USA	Counties	Regression (ridge, two-Stage least squares), text mining	Approach for estimating health statistics.
Kershaw et al. [76]	01/06/14	Twitter	Nov 2013 to Jan 2014	31.6M tweets	UK	Regions	Text mining, statistical analysis	Approach to model alcohol consumption pattern of a population.
Venerandi et al. [149]	01/03/15	Foursquare (1) and OpenStreetMap (2)	1: Mar to Apr 2014; 2: May 2014	1: ~32M check-ins; 2: ~131K point of interest	Three cities of the UK	Neighborhoods	Classification (various)	Alternative method to compute urban deprivation.

Gomide et al. [50] studied how Dengue disease is discussed on Twitter and whether this information can be used to monitor this disease. The authors have shown that tweets can be used to forecast, temporally and spatially, Dengue epidemics. They analyze how Twitter data reflect Dengue looking at four dimensions: location, volume, audience perception, and time. The authors investigate how users talk about Dengue using sentiment analysis techniques [51, 127], and explore the result to concentrate on only messages that express personal experience linked to Dengue.

Related to event detection, Arcaini et al. [7] explore LBSN messages to discover spatiotemporal aperiodic and periodic features of events happening in particular geographic areas. The strategy can potentially help to identify geospatial areas related to a particular event. Studying other types of real-world events, Bollen et al. [17] investigated if the collective mood obtained from tweets are correlated with the value of the Dow Jones stock market over time. Their findings suggest that the accuracy of the standard stock market prediction models is considerably enhanced when particular mood types are considered. Kisilevich et al. [79] proposed a visual analysis environment to detect relevant spatial and temporal patterns in urban areas observed through LBSN data. Sklar et al. [137] used Foursquare data to build an event detection engine that is based on a probabilistic model for measuring how unusually busy a place becomes. Similarly, Georgiev et al. [48] improved the understanding of this problem by investigating event participation of users from the viewpoint of LBSNs, which has implications for event recommender systems.

Becker et al. [10] introduce a method for identifying real-world event content on Twitter. Their approach could be used, for example, to provide better content visualization and to improve the filtering of content extracted on Twitter. Pan et al. [108] present an approach to describe and identify traffic outliers, which could be provoked, for instance, by car accidents or demonstrations. To describe the event, the approach mines terms from tweets of WeiBo, a Twitter-like social website in China, that are correlated geographically and temporally with the anomaly. To illustrate a different direction of efforts, Bakhshi et al. [8] studied the effect of a weather event in online restaurant reviews. They found that exogenous factors, such as rain, exert a significant effect on online restaurant reviews.

In addition to events that tend to happen sporadically, cities, typically, have areas that tend to be more popular among visitors or residents of the city. These areas are called points of interest (POI). Examples of POIs are the sights of cities. However, other places may also be a POI, for example, a popular area of entertainment among residents but unattractive for tourists. The task of identifying POIs is facilitated by the use of LBSN data, since this type of data, e.g., check-ins, may implicitly represent an interest of a user at a given instant.

In this way, when many check-ins are shared in a certain place within a particular time interval, this place might be a POI. That was the premise considered by Silva et al. [134]. In that work, the authors considered photos shared on Instagram to identify POIs. Each photo is associated with a geographic location (latitude and longitude) and to identify POIs, the authors first cluster geographically close photos. After that, they use a null model to exclude clusters that could have been generated by random situations (i.e., random people movements) and therefore, do not reflect relevant points of interest. Also, using datasets from different periods, the authors have shown that LBSNs can automatically capture changes in city dynamics. A famous Soccer Stadium was closed for remodeling during the period covered by one of the datasets, being identified as a POIs only when using the dataset, which covers the period when the stadium reopened.

Crandall et al. [30] considers in their study Flickr photos shared by users and their association with physical places. By exploring the collective behavior of users, they were able to discover landmarks at different granularity levels. For any granularity, they find important places by exploring a mean shift process [28] to identify places with high densities of shared photos. Their results could be used to identify, in an automatic manner, the best places to check while visiting a city, according to the opinion of several LBSN users.

Brilhante et al. [19] explores Flickr photos and also Wikipedia entries to obtain information related to POIs in cities and thus, to recommend itinerary. By using those data, they were able to create a touristic database that includes, among other information, the POIs themselves, their popularity, categories, and visiting patterns. The authors consider their problem as an instantiation of the Generalized Maximum Coverage problem [27]. The technique builds the itinerary that maximizes the user interest over the POIs and at the same time, respects his/her time available.

Levandoski et al. [83] explore location-based ratings (i.e., the evaluation associated to venues that a user checked-in in Foursquare) to develop a recommender system, which considers spatial aspects of evaluations when generating recommendations. To build the recommendations, their proposal also relies on two critical concepts: preference locality and travel locality. Preference locality indicates that preferences of users are influenced by its spatial region. Travel locality suggests that users tend to travel small distances when visiting recommended spatial items, e.g., venues, and this should be taken into account when making recommendations. The authors show that their recommendations can better predict user tastes compared to collaborative filtering.

Table 6. Summary of all discussed studies of the class Events/Interest Identification and Analysis.

Publication		Dataset				Granularity of analysis	Main technique (s)	Focus
Name	Date	Source	Time	Volume	Coverage			
Sakaki et al. [123]	Apr 2010	Twitter	2009	Thousands (different datasets)	Japan	Country	Classification (SVM), probabilistic modeling	Approach to monitor Twitter messages and to identify a target event.
Gomide et al. [50]	01/06/11	Twitter	2006 (start of Twitter) to Jul 2009 and Dec 2010 to Apr 2011	~500K tweets	Brazil	Cities	Clustering (ST-DBSCAN [14]), Regression (linear regression), classification (associative classifier [148]).	Analyze how the Dengue epidemic is announced in Twitter and whether this information could be used to monitor this disease.
Arcaini et al. [7]	May 2016	Twitter	Jul-Sep 2013 and Jun-Jul 2014	~140K	Worldwide	GPS	Clustering (density-based, proposed extension of DBSCAN)	Approach to discover spatiotemporal aperiodic and periodic features of events happening in particular geographic areas.
Bollen et al. [17]	Oct 2010	Twitter	Fev to Dec 2008	~9M tweets	Worldwide	USA	Sentiment analysis (text), regression (linear regression)	Study if the collective mood is linked with the Dow Jones stock market value.
Kisilevich et al. [79]	01/07/10	Flickr (1) and Panoramio (2)	01/06/09	1: ~86M photos; 2: ~11M photos	Worldwide	Cities	Clustering (DBSCAN)	Provide a visual analysis environment to detect spatial and temporal patterns.
Sklar et al. [137]	Sep 2012	Foursquare	20 weeks in 2011	-	City	NYC	Probabilistic modeling	Use Foursquare data to build an event detection engine.
Georgiev et al. [48]	01/07/14	Foursquare	Dec 2010 to Sep 2011	~3.5M check-ins	London, NYC and Chicago	Cities	Regression, network analysis	Investigate event participation from the viewpoint of LBSNs.
Becker et al. [10]	01/07/11	Twitter	Feb 2010	2.6M tweets	NYC	City	Classification (SVM, logistic regression, naive bayes), clustering (online, proposed)	Method for identifying event content on Twitter.
Pan et al. [108]	01/11/13	WeiBo	Mar-May 2011	Thousands (different datasets)	Beijing	GPS	Anomaly detection	Method to detect and describe traffic anomalies.
Bakhshi et al. [8]	Apr 2014	CityGrid (several sources)	2002 to 2011	1.1M restaurant reviews and ratings	Cities in the USA	GPS	Probabilistic modeling, regression	Study the effect of a weather event in online restaurant reviews.
Silva et al. [134]	May 2013	Instagram	Jun and Jul 2012	~2.3M photos	Worldwide	City, GPS	Network analysis, clustering (agglomerative hierarchical)	Characterization of Instagram and technique to discover points of interest.
Yin et al. [157]	Aug 2013	Foursquare and Douban-Event	2012	1,385,223 check-ins and 300,000 events	Several cities	GPS	Probabilistic modeling	Recommender system that provides a set of venues or events considering user's local preference and personal interest.
Crandall et al. [30]	Apr 2009	Flickr	Summer and fall of 2008	~35M photos	Worldwide	City, GPS	Clustering (mean shift), classification (SVM)	Techniques for analyzing geolocated photographs, for instance, to automatically identify places that people find interesting to photograph.
Brilhante et al. [19]	Oct 2013	Flickr	-	~330K photos	Cities in Italy	GPS	Optimization modeling (generalized maximum coverage)	Technique to recommend personalized POIs visitation itineraries.
Levandoski et al. [83]	01/07/12	Foursquare	-	~23K venues ratings	Minnesota (USA)	GPS	Collaborative filtering	Location-aware recommender system.

Yin et al. [157] propose a venue/event recommender system that uses user activity history in LBSNs and data coming from event-based social network services[17], specifically DoubanEvent[18]. Infer user preferences using those data is challenging because users can only visit a limited number of venues and attend a limited number of events. That results in a sparse user-item matrix for most location-based recommenders that explore collaborative filtering methods [120]. Also, when users travel to a new place they do not have an activity history to be explored. The authors propose a probabilistic generative model that quantifies and considers item content and local preference information in the recommendation process. The system presented good performance in recommending venues and events for users especially when they are traveling to new cities.

POIs are dynamic, i.e., a location that is popular today may not be tomorrow anymore. One advantage of using LBSN data to identify points of interest in the city is that we can get robust results to dynamic changes. That is, because LBSNs provide dynamic data, they could automatically capture changes in users' interests over time, helping to quickly identify areas that may become a POI (for example, due to the opening of a new business) or cease to be popular. Table 6 summarizes the studies discussed in this section. It also provides extra information about the studies not discussed in the text.

5.7 Illustration of the Urban computing Framework with LBSN Data

In this section, we consider some of the studies discussed to illustrate the components of the urban computing framework with LBSN data (Section 4), and also provide a more concrete illustration of the framework. First, we consider the study [133], which is one example from the class of studies Social and Economic Aspects. Figure 12 presents an overview of the discussed framework of urban computing for our example. In this considered study, the authors introduce a new approach to identify cultural boundaries between urban societies, taking into account users' preferences for food and drink. To accomplish their goal, they use users' check-ins performed in Foursquare (announced on Twitter) to represent users' preferences about what is eaten and drunk locally, for example, in a particular city. These data were collected using APIs. In possession of those data, they have stored (using MongoDB[19]) and processed them in order to make the desired filtering (only places related to Food and drink), and to create a personalized format.

Next, the authors were ready to extract knowledge from their data. First, they studied the spatial correlation between check-ins data in different types of restaurants for various cities around the world. The authors observed that cities in the same country, where the inhabitants usually have similar culture and eating habits, have the strongest correlations concerning restaurant preferences. In addition to this experiment, they also performed another one considering the time dimension, being able to find differences in the time when users check-in in restaurants. For these analyses, the use of several visualization techniques was fundamental. These efforts enabled the introduction of an approach for the identification of similar cultures, which could be applied in urban regions of varied sizes, for instance, neighborhoods or countries. For that a prototype-based clustering algorithm (*k-means* [57]) is used, as well as the principal component analysis technique (PCA) [71]. In order to validate the results, the authors used the study of Inglehart and Welzel [68], where it is proposed a cultural map of the world using data from the World Values Surveys[20] (WVS), which is one of the biggest cultural studies performed traditionally. When comparing the results with those obtained by Inglehart and Welzel the similarities are very close, suggesting that the proposed technique could be useful in this context.

[17] Virtual platforms for users to create and promote social events that will occur in the physical world.

[18] http://www.douban.com/events.

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

[20] http://www.worldvaluessurvey.org.

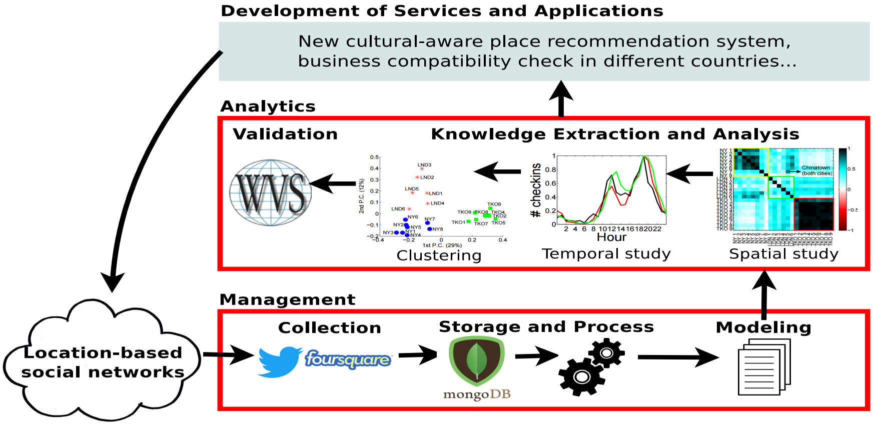

Development of Services and Applications
New cultural-aware place recommendation system, business compatibility check in different countries...
Analytics
Validation
Knowledge Extraction and Analysis
WVS
Clustering
Temporal study
# checkins
Hour
Spatial study
Chinatown (both cities)
Management
Collection
Storage and Process
Modeling
Location-based social networks
foursquare
mongoDB

Fig. 12. Steps for extraction of cultural patterns in Foursquare according to the urban computing framework with LBSN data.

After the analytics stage, the next step is to develop services and applications with the knowledge gained. These applications could be of several types. In our example, instead of exploring traditional approaches to find cultural differences, the introduced approach could represent a cheaper and easier way to find cultural differences in distinct regions of the globe, since it explores data that are voluntarily shared by users on the Web. Also, due to the economic importance of the cultural aspect [47], the introduced approach could help enterprises that have businesses in one area (e.g., country), test the similarity of preferences in distinct markets. A novel place recommendation service exploring a cultural aspect, which could be of interest to tourists and residents, is another application that could explore that methodology.

To provide other examples of the framework of urban computing with LBSN data, we selected randomly one study of each remaining category of studies not exemplified in this section, i.e., City Semantics, City Problems, Urban Mobility, Health and Well-being, and Events/Interest Identification and Analysis. For those studies, we described their key parts according to the framework. Table 7 presents these results.

6 RESEARCH CHALLENGES AND OPPORTUNITIES

Although several research efforts related to urban computing leveraging LBSN data have been performed recently, it is possible to find open issues and opportunities for studying cities and societies using LBSN data. In this context, our extensive overview of the literature presented in previous sections places us in a perfect spot to discuss key challenges and future work for research in the field of urban computing leveraging LBSN. Below, we discuss some of such issues.

6.1 Data Bias

In the previous sections, we presented several examples showing LBSN data does provide solid aggregate information that can help improving understanding of different phenomena related to urban societies. However, it is important to keep in mind possible limitations in LBSN data.

First, it may reflect the behavior of a fraction of consumers. Take for instance popular data sources such as Foursquare, Instagram, and Twitter. Users from those systems are biased towards the citizens who are likely to be young, owners of smartphones, and urban dwellers [18, 38]. Therefore, there could be biases related to the fact that the users of such application might not represent all population

Table 7. Examples of studies described with the urban computing framework with LBSN Data.

	Management			Analytics		
Study	**Collection**	**Storage /Proc.**	**Modeling**	**Knowledge Extract.**	**Knowledge Anal.**	**Validation**
[31]	Foursquare (check-ins from publicly available tweets).	To help in some processing steps it was used Lanczos solver and k-d trees.	Graphs.	Clustering (spectral).	Visualizations of the discovered clusters on a map and the structure of related clusters (metric developed).	Interview with participants.
[99]	Foursquare (check-ins from publicly available tweets).	It was used Mysql to store and manage the data.	Data matrix.	Outlier detection, clustering (k-means).	Visualization and statistical approaches.	Interview with business managers. Comparison of results with official indices.
[166]	Download from Flickr by using its publicly available API.	Filtered non-tourist paths using a proposed mobility entropy-based method to identifying tourist travel paths. Ensured statistical significance of travel paths.	Spatial trajectory.	Clustering (agglomerative hierarchical, DBSCAN, mean shift), markov chain.	Visualizations, topological analysis of travel route.	Compared results with a list of top attractions in Yahoo! Travel.
[110]	Geotagged tweets were downloaded from publicly available Twitter API.	Separation of subgroups of tweets. It was used map reduce pattern to utilize the parallelization power of the Hadoop. It was also performed other preprocessing steps, such as stop word removal and collocation algorithms.	Bag of words.	Text mining.	Various visualizations.	Correlations and other analysis with data from the Health & Social Care Information Centre from UK was used as the ground truth to test their results.
[50]	Tweets were collected using Twitter API. It was collected also locations informed in the Twitter users' profiles.	It was considered only tweets related to "Dengue". It was filtered out invalid locations informed by users and it was inferred locations of users by using a geocoding process.	Data matrix.	Clustering (ST-DBSCAN [14]), linear regression, associative classifier [148].	Visualizations, statistical properties.	It was used an official document containing the summarization of Dengue situation in Brazil to validate their proposal.

of a particular region. With that, areas containing poorer and older residents could provide fewer data and be underrepresented in whatever analysis is made. Besides, users may not share data concerning all of their destinations due to privacy reasons, since it will be made public on Twitter. Thus, our LBSN data might offer a partial view of consumers habits, which needs to be taken with care. Also, LBSN data might represent only a sample of data, perhaps, limited. In other words, only a small part of the activities performed might be represented in the data. Adverse weather conditions, among other external factors, might affect the data gathered representing some venues (especially outdoor ones).

Furthermore, we cannot assume that the data shared in LBSNs are correct or precise. For instance, Twitter is a tool that might enable new types of spam [11, 155]. Costa et al. [29] found evidence that in Apontador, a popular Brazilian LBSN, there are irregular contents shared by some users and this could happen in other types of LBSNs as well. Under these circumstances, data quality, one of the issues discussed in [130], becomes even more serious due to the possibility of the production of false data. That could potentially compromise approaches and methodologies presented in Section 5. In this direction, Hecht et al. [60] identified users that sometimes provide misleading information in their location field. That is particularly important in cases when we want to map informed locations to geographic areas, a useful procedure explored in several studies [42, 86].

Specific characteristics of regions could also be factors for data bias. Thebault-Spieker [143] found evidence that users tend to avoid specific areas of the city with low socioeconomic status to perform paid tasks in a mobile crowdsourcing platform. In the same study, the authors also identified that users also avoid suburbs and rural areas. Particularly about rural areas, Hecht and Stephens [61] have provided evidence that data from Twitter, Flickr, and Foursquare, commonly used sources, tend to be biased to urban aspects and distant from rural aspects. This suggests that research that has been showed to be useful in urban areas, such as those discussed in the previous section, might have to be adapted to be also effective in rural areas when working with LBSN data.

Finally, another source of bias could be associated with the fact data from LBSN come from deliberate actions of the users, i.e., data is to some extent actively generated by the user. Therefore, users might introduce bias in what he/she shares, for instance, amplifying check-ins in trendy venues to impress friends.

6.2 Integration of Multiple Urban Data Sources

The exploration of diverse urban data sources simultaneously could bring several benefits in developing more sophisticated applications [122, 131]. With that in mind, the goal is to design algorithms and data structures that process and combine different data types at different levels of abstraction (for example, text, images, videos, and actions) to extract useful information. To achieve this goal, algorithms are needed to deal with massive data flows, generated by several types of LBSN data, and have operations such as aggregation, filtering and indexing in (near) real-time. Integration is, therefore, a critical phase of this process since information is the foundation on which models and mechanisms of action will be built.

Take, for instance, the exemplified model in Section 4.1 to represent a spatial trajectory produced by a moving user in geospatial areas. This model could be enriched, for example, by providing semantics to the trajectories. One way to do that is to annotate trajectories manually; however, this is a hard task to be done at large scale [121, 154]. Therefore, it is necessary to develop new methods to integrate different data sources to enrich movement data semantically automatically. LBSNs data can be treated as annotations that might offer hints to explain movements. In this direction, Fileto et al. [43] proposed a process to annotate key parts of trajectories with concepts (classes) and objects (instances of concepts) described in ontologies and Linked Open Data (LOD) collections. The authors explored Twitter data to enrich and analyze the displacement of moving objects. This example helps to emphasize the importance of semantic enrichment techniques in the context of urban data integration [1, 156].

The tasks and problems discussed in the data integration step raise several research challenges, some of them are: How to integrate multiple heterogeneous and complex data sources at different levels of abstraction? How to design algorithms that are capable of storing, aggregating, filtering and indexing the collected data efficiently? How to assess the quality of information derived from aggregated data? How to achieve the three previous goals while preserving individuals' privacy?

6.3 LBSN Data Collection

LBSN data collection aims to obtain, continuously and straightforwardly, samples from multiple sources of information ranging from urban societies to existing systems in large cities. Data samples can be obtained from dynamic and heterogeneous sources, as we discussed above. In addition to the continued growth of the Web and the explosion of online social networks, the cheapening and modernization of sensing has led to unprecedented growth in the number of data streams available in real-time. In this scenario, the efficient monitoring of such large volumes of information is an open problem [77].

For that, we need to develop efficient mechanisms for the observation of the physical world as a repository of information subject to continuous changes. Major challenges are tied to this issue, such as: How to design data collection systems that efficiently handle the compromises between the representativeness of the obtained information and the cost in terms of energy, space, latency and financial mechanisms applied to collect it? What mechanisms should be used or developed to collect information from very large, noisy and error-prone data flows, taking into account, security and privacy restrictions? How to allow and encourage users to share information and ensure their privacy, so that representative and unbiased data can be obtained?

In this direction, it is also essential to keep the source of data sustainable. Since users are a central element in location-based social networks, incentive mechanisms play a central role. In this direction, understand which incentive mechanisms work is fundamental because it might guide the design of a new system. Focused on that, Santos et al. [125] assess the performance of incentive mechanisms used by Foursquare to motivate users. Among the results, the authors found evidence that incentives based on mayorship[21], which motivates competition among users to become mayor of someplace, seems to be efficient to keep users motivated, while incentives based on badges[22] do not seem to have the same efficiency, except for some specific types of badges. Still related to incentive mechanisms, most proposals to encourage users to contribute urban data focus on just one strategy. However, as noted by Reddy et al. [117], the use of more than one strategy at the same time may yield better results. The authors conclude that incentives worked best when payments (rewards) were combined with other factors such as user altruism and when there was competition among participants.

6.4 Prediction and Classification

There are opportunities to study areas when we jointly consider time and place where LBSN data are shared. Users have periodic patterns thanks to their routines. That presents a high potential for prediction because it is probable that users will repeat their activities periodically. There are many possibilities for prediction considering people's seasonal patterns, for instance, prediction of crowds. This sort of information is vital in several cases, for example, services to prevent traffic in particular locations and provide alternative routes to drivers. As an example, Hsieh et al. [66] introduced a model that considers the time dimension to suggest routes exploring information of a Foursquare-like system.

In general, LBSN data are little explored in models for traffic prediction. Some studies in this direction are: [118, 136]. Ribeiro et al. [118] showed evidence that a geolocated message, either on Twitter, Foursquare, or Instagram, could be used to improve the understanding of traffic conditions. Besides that, imagine a user that performs a check-in at home and then go to work. When he/she gets into the workplace, for some reason, he/she does another check-in. Regardless of whether he/she is on the LBSN or not, there is intrinsic information regarding the time interval between these check-ins consisting of traffic performance. In case the traffic is congested, the mentioned interval between check-ins will be greater than the travel time not presenting congestion (information easily computed by the maximum speed and distance of the urban paths).

In addition, with the use of LBSN data, it is feasible to classify areas in distinct ways. Some of them have been discussed in Section 5, considering, for instance, smell, noise and visual aspects. That could be valuable for several new services. An example would be a new route suggestion tool

[21] Users become mayor of a place by checking in more than anyone else during the last 30 days.

[22] Users earn badges according to their location, their frequency of check-ins, some specific events, or commemorative dates.

that suggests the smallest route that is also the most olfactory pleasing. People who practice urban running may want to avoid streets with high levels of gas emissions.

6.5 New Applications and Services

There are several opportunities to develop services and application exploring LBSN data. In this section, we present some of them. For instance, considering the place networks, mentioned in Section 5.2, we could study centrality metrics in this network. For example, Figure 13, extracted from [135], shows the betweenness centrality values for the nodes of the network. Each color is related to a location category, and the size of the symbol reflects the proportion of the centrality value. This approach can be explored to help several services, for instance, if an uncommon and frequent flow of users is verified between two distinct shops locations in a particular city, shops owners can explore this information to create business agreements to raise their profits, such as advertising among their companies [135].

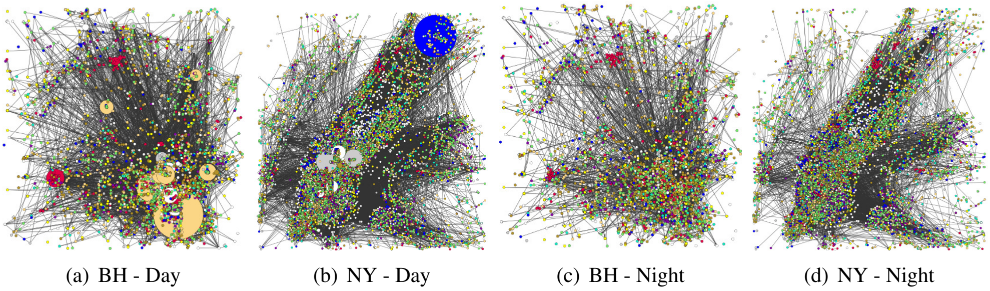

(a) BH - Day (b) NY - Day (c) BH - Night (d) NY - Night

Fig. 13. Betweenness centrality values for the network nodes of the place networks representing New York (NY) and Belo Horizonte (BH) [135].

Furthermore, one could explore in other ways the displacement of users in the city according to the type of places they visit. As data from LBSNs tend to be highly skewed, some of the most popular transitions between types of places, e.g., restaurant or library, could be valuable indicators of the dynamics of the city. Techniques could be developed to measure the similarity between two urban areas, e.g., cities, allowing the comparison and clustering of urban areas that could be explored in different applications.

In another direction, the spatial causes of poverty/ deprivation, including its persistence, is currently a topic of growing interest [65, 81, 112, 138, 147]. Particularly related to the relation between economic marginalization and physical segregation in urban areas, regions that offer few data compared to other regions of the same city may suggest a lack of access to technology by the residents [132]. Similar information can be gathered exploring conventional approaches, for instance, questionnaires, however, this novel approach might enable to obtain this data automatically and cheaply using LBSN data. With this goal, algorithms similar to the one introduced in [31] could be used.

There are opportunities to develop more sophisticated recommendation systems by, for example, exploring some of the studies surveyed. New urban areas recommendation services exploring specific semantic of areas are an example of such systems. For that, one could, for instance, explore novel cultural criteria or the functionality of areas.

Also, we presented several examples of studies suggesting that LBSN data could revolutionize the study of urban societies. Despite the significant advances, there are still work to be done to consolidate the proposed techniques in that direction to enable new services and applications.

6.6 Other Challenges and Opportunities

In the previous sections, we presented some of the main challenges regarding the use of LBSN data to the study of urban societies. However, we did not cover all the challenges and opportunities. For instance, challenges related to the temporal dynamics of LBSNs. Several previous studies model LBSN data as static structures, not taking into account the temporal dynamics. Even though it is an accepted strategy, that representation might result in loss of relevant information in some instances. In addition, another example of challenge is to work with a large number of data that LBSNs can potentially provide. This imposes several challenges related to, for example, processing, storage, and indexing in real-time when using tools of conventional data processing systems and database management. Also, LBSN data exploration may threaten the privacy of users. For example, LBSN data could be explored to deduce users' preferences and particular behavior. With this, users have no guarantee that others will not violate their private life. It is a challenge to ensure people's privacy while relying on data that can be potentially sensitive. A discussion of those and other challenges was presented by Silva et al. [130].

7 CONCLUSIONS

We are facing an unprecedented opportunity for urban (and social) studies, thanks to the significant amount of LBSN data available because of the convergence of social media and geographic information. With that, in this study, we discussed key concepts of urban computing leveraging LBSN data. Also, we surveyed recent efforts available in the literature in the area of urban computing with LBSN data, which is helpful to exemplify research trends and techniques commonly used. In addition, we also presented some of the main challenges and opportunities in the area. We hope this study motivates the development of new initiatives that address challenges related to the improvement of the quality of life of urban societies.

REFERENCES

[1] Fabian Abel, Qi Gao, Geert-Jan Houben, and Ke Tao. 2011. Semantic Enrichment of Twitter Posts for User Profile Construction on the Social Web. In *Proc. of ESWC'11*. Heraklion, Crete, Greece, 375–389.

[2] Charu C Aggarwal and ChengXiang Zhai. 2012. *Mining text data*. Springer Science & Business Media.

[3] Shi An, Haiqiang Yang, Jian Wang, Na Cui, and Jianxun Cui. 2016. Mining urban recurrent congestion evolution patterns from GPS-equipped vehicle mobility data. *Information Sciences* 373 (2016), 515 – 526.

[4] G. Andrienko, N. Andrienko, S. Bremm, T. Schreck, T. von Landesberger, P. Bak, and D. Keim. 2010. Space-in-time and Time-in-space Self-organizing Maps for Exploring Spatiotemporal Patterns. In *Proc. of EuroVis'10*. Bordeaux, France, 913–922. https://doi.org/10.1111/j.1467-8659.2009.01664.x

[5] Gennady Andrienko, Natalia Andrienko, Urska Demsar, Doris Dransch, Jason Dykes, Sara Irina Fabrikant, Mikael Jern, Menno-Jan Kraak, Heidrun Schumann, and Christian Tominski. 2010. Space, Time and Visual Analytics. *Int. J. Geogr. Inf. Sci.* 24, 10 (Oct. 2010), 1577–1600. https://doi.org/10.1080/13658816.2010.508043

[6] Mihael Ankerst, Markus M Breunig, Hans-Peter Kriegel, and Jörg Sander. 1999. OPTICS: ordering points to identify the clustering structure. In *ACM Sigmod record*, Vol. 28. ACM, 49–60.

[7] Paolo Arcaini, Gloria Bordogna, Dino Ienco, and Simone Sterlacchini. 2016. User-driven Geo-temporal Density-based Exploration of Periodic and Not Periodic Events Reported in Social Networks. *Inf. Sci.* 340, C (May 2016), 122–143. https://doi.org/10.1016/j.ins.2016.01.014

[8] Saeideh Bakhshi, Partha Kanuparthy, and Eric Gilbert. 2014. Demographics, Weather and Online Reviews: A Study of Restaurant Recommendations. In *Proc. of WWW'14*. ACM, Seoul, Korea, 443–454.

[9] Luciano Barbosa, Kien Pham, Claudio Silva, Marcos R Vieira, and Juliana Freire. 2014. Structured open urban data: understanding the landscape. *Big data* 2, 3 (2014), 144–154.

[10] Hila Becker, Mor Naaman, and Luis Gravano. 2011. Beyond trending topics: Real-world event identification on twitter. In *Proc. of ICWSM'11*. AAAI, Barcelona, Spain.

[11] Fabrício Benevenuto, Gabriel Magno, Tiago Rodrigues, and Virgílio Almeida. 2010. Detecting spammers on Twitter. In *Proc. of CEAS'10*. Redmond, USA.

[12] P. Berkhin. 2006. *A Survey of Clustering Data Mining Techniques*. Springer Berlin Heidelberg, Berlin, Heidelberg, 25–71. https://doi.org/10.1007/3-540-28349-8_2

[13] Michael W Berry and Malu Castellanos. 2008. *Survey of text mining II*. Vol. 6. Springer.

[14] Derya Birant and Alp Kut. 2007. ST-DBSCAN: An algorithm for clustering spatial-temporal data. *Data & Knowledge Engineering* 60, 1 (2007), 208 – 221. https://doi.org/10.1016/j.datak.2006.01.013 Intelligent Data Mining.

[15] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* 3 (March 2003), 993–1022. http://dl.acm.org/citation.cfm?id=944919.944937

[16] The Foursquare Blog. 2014. *A look into the future of Foursquare, including a new app called Swarm*. The Foursquare Blog. https://goo.gl/BW0QXS.

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

[18] Joanna Brenner and Aaron Smith. 2013. 72% of Online Adults are Social Networking Site Users. http://goo.gl/HTgNy3. (August 2013).

[19] Igo Brilhante, Jose Antonio Macedo, Franco Maria Nardini, Raffaele Perego, and Chiara Renso. 2013. Where Shall We Go Today?: Planning Touristic Tours with Tripbuilder. In *Proc. of CIKM'13*. ACM, San Francisco, California, USA, 757–762. https://doi.org/10.1145/2505515.2505643

[20] Dirk Brockmann, Lars Hufnagel, and Theo Geisel. 2006. The scaling laws of human travel. *Nature* 439, 7075 (2006), 462–465.

[21] J. Burke, D. Estrin, M. Hansen, A. Parker, N. Ramanathan, S. Reddy, and M. B. Srivastava. 2006. Participatory sensing. In *Proc. of World Sensor Web Workshop at ACM Sensys*. Boulder, USA, 117–134.

[22] Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2009. Anomaly Detection: A Survey. *ACM Comput. Surv.* 41, 3, Article 15 (July 2009), 58 pages. https://doi.org/10.1145/1541880.1541882

[23] Guangshuo Chen, Sahar Hoteit, Aline Carneiro Viana, Marco Fiore, and Carlos Sarraute. 2018. Enriching sparse mobility information in Call Detail Records. *Computer Communications* 122 (2018), 44 – 58. https://doi.org/10.1016/j.comcom.2018.03.012

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

[25] Eunjoon Cho, Seth A. Myers, and Jure Leskovec. 2011. Friendship and mobility: user movement in location-based social networks. In *Proc. of KDD'11*. ACM, San Diego, USA, 1082–1090. https://doi.org/10.1145/2020408.2020579

[26] Martin Chorley, Luca Rossi, Gareth Tyson, and Matthew Williams. 2016. Pub crawling at scale: tapping Untappd to explore social drinking. In *Proc. of ICWSM'16*. Cologne, Germany.

[27] Reuven Cohen and Liran Katzir. 2008. The generalized maximum coverage problem. *Inform. Process. Lett.* 108, 1 (2008), 15–22.

[28] D. Comaniciu and P. Meer. 2002. Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24, 5 (May 2002), 603–619. https://doi.org/10.1109/34.1000236

[29] Helen Costa, Luiz H.C. Merschmann, FabrÃ■cio Barth, and FabrÃ■cio Benevenuto. 2014. Pollution, bad-mouthing, and local marketing: The underground of location-based social networks. *Information Sciences* 279 (2014), 123 – 137. https://doi.org/10.1016/j.ins.2014.03.108

[30] David J. Crandall, Lars Backstrom, Daniel Huttenlocher, and Jon Kleinberg. 2009. Mapping the world's photos. In *Proc. of WWW'09*. ACM, Madrid, Spain, 761–770. https://doi.org/10.1145/1526709.1526812

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

[32] Aron Culotta. 2014. Estimating County Health Statistics with Twitter. In *Proc. of CHI '14*. ACM, Toronto, Ontario, Canada, 1335–1344. https://doi.org/10.1145/2556288.2557139

[33] Munmun De Choudhury, Moran Feldman, Sihem Amer-Yahia, Nadav Golbandi, Ronny Lempel, and Cong Yu. 2010. Automatic Construction of Travel Itineraries Using Social Breadcrumbs. In *Proc. of HT'10*. ACM, Toronto, Canada, 35–44. https://doi.org/10.1145/1810617.1810626

[34] Munmun De Choudhury, Sanket Sharma, and Emre Kiciman. 2016. Characterizing Dietary Choices, Nutrition, and Language in Food Deserts via Social Media. In *Proc. of CSCW'16*. ACM, San Francisco, USA, 1157–1170.

[35] Manlio De Domenico, Antonio Lima, and Mirco Musolesi. 2013. Interdependence and predictability of human mobility and social interactions. *Pervasive and Mobile Computing* 9, 6 (2013), 798–807.

[36] Jeffrey Dean and Sanjay Ghemawat. 2008. MapReduce: Simplified Data Processing on Large Clusters. *Commun. ACM* 51, 1 (Jan. 2008), 107–113. https://doi.org/10.1145/1327452.1327492

[37] Ellie D'Hondt, Matthias Stevens, and An Jacobs. 2013. Participatory noise mapping works! An evaluation of participatory sensing as an alternative to standard techniques for environmental monitoring. *Pervasive and Mobile Computing* 9, 5 (2013), 681–694.

[38] Maeve Duggan and Aaron Smith. 2014. Social Media Update 2013. (Jan 2014). http://goo.gl/JhuiOG.

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

[40] Ken Anderson Eric Paulos and Anthony Townsend. 2004. UbiComp in the Urban Frontier. In *Proc. of Ubicomp'04 Workshops*. Nottingham, UK.

[41] Deborah Falcone, Cecilia Mascolo, Carmela Comito, Domenico Talia, and Jon Crowcroft. 2014. What is this place? Inferring place categories through user patterns identification in geo-tagged tweets. In *Proc. of MobiCASE'14*. IEEE, Austin, Texas, USA, 10–19.

[42] A. P. G. Ferreira, T. H. Silva, and A. A. F. Loureiro. 2015. Beyond Sights: Large Scale Study of Tourists' Behavior Using Foursquare Data. In *Proc. of IEEE ICDMW'15 Workshops*. Atlantic City, USA., 1117–1124.

[43] Renato Fileto, Cleto May, Chiara Renso, Nikos Pelekis, Douglas Klein, and Yannis Theodoridis. 2015. The Baquara2 knowledge-based framework for semantic enrichment and analysis of movement data. *Data & Knowledge Engineering* 98 (2015), 104 – 122. https://doi.org/10.1016/j.datak.2015.07.010 Research on conceptual modeling.

[44] Foursquare. 2017. *About Us*. Fourquare. https://foursquare.com/about.

[45] V. Frias-Martinez, V. Soto, H. Hohwald, and E. Frias-Martinez. 2012. Characterizing Urban Landscapes Using Geolocated Tweets. In *Proc. of PASSAT and SocialCom*. Amsterdam, The Netherlands, 239–248. https://doi.org/10.1109/SocialCom-PASSAT.2012.19

[46] Ruth García-Gavilanes, Yelena Mejova, and Daniele Quercia. 2014. Twitter Ain'T Without Frontiers: Economic, Social, and Cultural Boundaries in International Communication. In *Proc. of CSCW'14*. ACM, Baltimore, Maryland, USA, 1511–1522.

[47] Ruth Garcia-Gavilanes, Daniele Quercia, and Alejandro Jaimes. 2013. Cultural Dimensions in Twitter: Time, Individualism and Power. In *Proc. of ICWSM'13*. AAAI, Boston, USA.

[48] Petko Georgiev, Anastasios Noulas, and Cecilia Mascolo. 2014. The Call of the Crowd: Event Participation in Location-based Social Services. In *Proc. of ICWSM'14*. AAAI, Ann Arbor, USA.

[49] Fosca Giannotti, Mirco Nanni, Fabio Pinelli, and Dino Pedreschi. 2007. Trajectory pattern mining. In *Proc. of KDD'07*. ACM, San Jose, USA, 330–339.

[50] J. Gomide, A. Veloso, W. Meira Jr., V. Almeida, F. Benevenuto, F. Ferraz, and M. Teixeira. 2011. Dengue surveillance based on a computational model of spatio-temporal locality of Twitter. In *Proc. of WebSci'11*. Evanston, USA.

[51] Pollyanna Gonçalves, Matheus Araújo, Fabrício Benevenuto, and Meeyoung Cha. 2013. Comparing and Combining Sentiment Analysis Methods. In *Proc. of COSN'13*. Boston, USA, 12.

[52] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. 2008. Understanding individual human mobility patterns. *Nature* 453, 7196 (2008), 779–782.

[53] Pritam Gundecha and Huan Liu. 2012. Mining social media: a brief introduction. In *New Directions in Informatics, Optimization, Logistics, and Production*. Informs, 1–17.

[54] Diansheng Guo, Shufan Liu, and Hai Jin. 2010. A Graph-based Approach to Vehicle Trajectory Analysis. *J. Locat. Based Serv.* 4, 3-4 (Sept. 2010), 183–199. https://doi.org/10.1080/17489725.2010.537449

[55] Kristen Hall-Geisler. 2016. *Waze and Esri make app-to-infrastructure possible*. Tech Crunch. https://goo.gl/HtJxGH.

[56] Jiawei Han, Jian Pei, and Micheline Kamber. 2011. *Data mining: concepts and techniques*. Morgan Kaufmann.

[57] John A Hartigan and Manchek A Wong. 1979. Algorithm AS 136: A k-means clustering algorithm. *Applied statistics* 28 (1979), 100–108. https://doi.org/10.2307/2346830

[58] Ibrahim Abaker Targio Hashem, Ibrar Yaqoob, Nor Badrul Anuar, Salimah Mokhtar, Abdullah Gani, and Samee Ullah Khan. 2015. The rise of "big data" on cloud computing: Review and open research issues. *Information Systems* 47 (2015), 98 – 115. https://doi.org/10.1016/j.is.2014.07.006

[59] Alex Heath. 2017. *Instagram's user base has doubled in the last 2 years to 700 million*. Business Insider. https://goo.gl/PWgLVe.

[60] Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H. Chi. 2011. Tweets from Justin Bieber's Heart: The Dynamics of the Location Field in User Profiles. In *Proc. of CHI '11*. ACM, Vancouver, Canada, 237–246. https://doi.org/10.1145/1978942.1978976

[61] Brent Hecht and Monica Stephens. 2014. A Tale of Cities: Urban Biases in Volunteered Geographic Information. In *Proc. of ICWSM'14*. Ann Arbor, USA.

[62] John R. Hipp, Robert W. Faris, and Adam Boessen. 2012. Measuring 'neighborhood': Constructing network neighborhoods. *Social Networks* 34, 1 (2012), 128 – 140. https://doi.org/10.1016/j.socnet.2011.05.002 Capturing Context: Integrating Spatial and Social Network Analyses.

[63] Nadav Hochman and Raz Schwartz. 2012. Visualizing Instagram: Tracing Cultural Visual Rhythms. In *Proc. of ICWSM'12*. AAAI, Dublin, Ireland, 6–9.

[64] Geert Hofstede, Gert Jan Hofstede, and Michael Minkov. 2010. *Cultures and organizations: Software of the mind. Revised and expanded.* McGraw-Hill, New York.

[65] Desislava Hristova, Matthew J. Williams, Mirco Musolesi, Pietro Panzarasa, and Cecilia Mascolo. 2016. Measuring Urban Social Diversity Using Interconnected Geo-Social Networks. In *Proc. of WWW '16*. Montreal, Canada, 21–30. https://doi.org/10.1145/2872427.2883065

[66] Hsun-Ping Hsieh, Cheng-Te Li, and Shou-De Lin. 2012. Exploiting large-scale check-in data to recommend time-sensitive routes. In *Proc. of UrbComp'12*. ACM, Beijing, China, 55–62.

[67] Samuel P Huntington. 2000. The clash of civilizations? In *Culture and Politics*. Springer, 99–118.

[68] Ronald Inglehart and Christian Welzel. 2010. Changing Mass Priorities: The Link between Modernization and Democracy. *Perspectives on Politics* 8, 02 (2010), 551–567. https://doi.org/10.1017/s1537592710001258

[69] Bin Jiang and Yufan Miao. 2015. The Evolution of Natural Cities from the Perspective of Location-Based Social Media. *The Professional Geographer* 67, 2 (2015), 295–306.

[70] Shan Jiang, Gaston A. Fiore, Yingxiang Yang, Joseph Ferreira, Jr., Emilio Frazzoli, and Marta C. González. 2013. A Review of Urban Computing for Mobile Phone Traces: Current Methods, Challenges and Opportunities. In *Proc. of UrbComp'13*. ACM, Chicago, Illinois, Article 2, 9 pages. https://doi.org/10.1145/2505821.2505828

[71] I. T. Jolliffe. 2002. *Principal Component Analysis* (second ed.). Springer.

[72] Kenneth Joseph, Chun How Tan, and Kathleen M Carley. 2012. Beyond local, categories and friends: clustering foursquare users with latent topics. In *Proc. of Ubicomp'12*. ACM, Pittsburgh, USA, 919–926.

[73] David Jurgens, Tyler Finethy, James McCorriston, Yi Tian Xu, and Derek Ruths. 2015. Geolocation Prediction in Twitter Using Social Networks: A Critical Analysis and Review of Current Practice.. In *Proc. of ICWSM'15*. Oxford, UK, 188–197.

[74] Dmytro Karamshuk, Anastasios Noulas, Salvatore Scellato, Vincenzo Nicosia, and Cecilia Mascolo. 2013. Geo-spotting: Mining Online Location-based Services for Optimal Retail Store Placement. In *Proc. of KDD'13*. ACM, Chicago, Illinois, USA, 793–801. https://doi.org/10.1145/2487575.2487616

[75] P. Katsikouli, A. C. Viana, M. Fiore, and A. Tarable. 2017. On the Sampling Frequency of Human Mobility. In *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*. 1–6.

[76] Daniel Kershaw, Matthew Rowe, and Patrick Stacey. 2014. Towards Tracking and Analysing Regional Alcohol Consumption Patterns in the UK Through the Use of Social Media. In *Proc. of WebSci'14*. ACM, Bloomington, USA, 220–228.

[77] Jaein Kim and Buhyun Hwang. 2016. Real-time stream data mining based on CanTree and Gtree. *Information Sciences* 367-368 (2016), 512 – 528. https://doi.org/10.1016/j.ins.2016.06.018

[78] Tim Kindberg, Matthew Chalmers, and Eric Paulos. 2007. Guest Editors' Introduction: Urban Computing. *IEEE Pervasive Computing* 6, 3 (2007), 18–20. https://doi.org/10.1109/MPRV.2007.57

[79] Slava Kisilevich, Milos Krstajic, Daniel Keim, Natalia Andrienko, and Gennady Andrienko. 2010. Event-Based Analysis of People's Activities and Behavior Using Flickr and Panoramio Geotagged Photo Collections. In *Proc. of International Conference Information Visualisation*. IEEE, London, UK, 289–296.

[80] Vassilis Kostakos and Eamonn O'Neill. 2008. City ware: Urban Computing to Bridge Online. *Handbook of Research on Urban Informatics: The Practice and Promise of the Real-Time City* (2008).

[81] Neal Lathia, Daniele Quercia, and Jon Crowcroft. 2012. The Hidden Image of the City: Sensing Community Well-being from Urban Mobility. In *Proc. of Pervasive'12*. Springer-Verlag, Newcastle, UK, 91–98.

[82] Géraud Le Falher, Aristides Gionis, and Michael Mathioudakis. 2015. Where is the Soho of Rome? Measures and algorithms for finding similar neighborhoods in cities. In *Proc. of ICWSM'15*. Oxford, UK.

[83] J. J. Levandoski, M. Sarwat, A. Eldawy, and M. F. Mokbel. 2012. LARS: A Location-Aware Recommender System. In *Proc. of ICDE'12*. Washington, USA, 450–461. https://doi.org/10.1109/ICDE.2012.54

[84] Robert V. Levine. 2006. *A Geography of Time: The Temporal Misadventures of a Social Psychologist or How Every Culture Keeps Time Just a Little Bit Differently*. University Press.

[85] A Lima, M De Domenico, V Pejovic, and M Musolesi. 2015. Disease containment strategies based on mobility and information dissemination. *Scientific reports* 5, 10650 (2015), –. https://doi.org/10.1038/srep10650

[86] Antonio Lima and Mirco Musolesi. 2012. Spatial Dissemination Metrics for Location-based Social Networks. In *Proc. of UbiComp '12*. ACM, Pittsburgh, USA, 972–979. https://doi.org/10.1145/2370216.2370429

[87] Jovian Lin, Richard Oentaryo, Ee-Peng Lim, Casey Vu, Adrian Vu, and Agus Kwee. 2016. Where is the Goldmine?: Finding Promising Business Locations Through Facebook Data Analytics. In *Proc. of HT '16*. ACM, Halifax, Canada, 93–102. https://doi.org/10.1145/2914586.2914588

[88] Huan Liu and Hiroshi Motoda. 2008. *Computational methods of feature selection*. Chapman and Hall.

[89] Alejandro Llorente, Manuel Garcia-Herranz, Manuel Cebrian, and Esteban Moro. 2015. Social Media Fingerprints of Unemployment. *PLOS ONE* 10, 5 (05 2015), 1–13. https://doi.org/10.1371/journal.pone.0128692

[90] Xuelian Long, Lei Jin, and James Joshi. 2012. Exploring trajectory-driven local geographic topics in foursquare. In *Proc. of UbiComp'12*. ACM, Pittsburgh, USA, 927–934. https://doi.org/10.1145/2370216.2370423

[91] Paul A Longley, Michael F Goodchild, David J Maguire, and David W Rhind. 1999. *Geographical information systems*. Vol. 1. Wiley New York.

[92] Kassio Machado, Thiago H. Silva, Pedro O. S. Vaz de Melo, Eduardo Cerqueira, and Antonio A. F. Loureiro. 2015. Urban Mobility Sensing Analysis Through a Layered Sensing Approach. In *Proc. of IEEE International Conference on*

Mobile Services. IEEE Computer Society, Washington, DC, USA, 306–312. https://doi.org/10.1109/MobServ.2015.50

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

[94] George Martine, Alex Marshall, et al. 2007. State of world population 2007: unleashing the potential of urban growth. In *State of world population 2007: unleashing the potential of urban growth*. UNFPA.

[95] Miller McPherson, Lynn Smith-Lovin, and James M Cook. 2001. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology* 27, 1 (2001), 415–444. https://doi.org/10.1146/annurev.soc.27.1.415

[96] Yelena Mejova, Hamed Haddadi, Anastasios Noulas, and Ingmar Weber. 2015. #FoodPorn: Obesity Patterns in Culinary Interactions. In *Proc. of DH '15*. ACM, Florence, Italy, 51–58. https://doi.org/10.1145/2750511.2750524

[97] Delia Mocanu, Andrea Baronchelli, Nicola Perra, Bruno Gonçalves, Qian Zhang, and Alessandro Vespignani. 2013. The Twitter of Babel: Mapping World Languages through Microblogging Platforms. *PLoS ONE* 8, 4 (04 2013).

[98] Eduardo Mucceli, Aline Carneiro Viana, Carlos Sarraute, Jorge Brea, and José Ignacio Alvarez-Hamelin. 2016. On the Regularity of Human Mobility. *Pervasive and Mobile Computing* (Dec. 2016).

[99] Willi Mueller, Thiago H. Silva, Jussara M. Almeida, and Antonio AF Loureiro. 2017. Gender matters! Analyzing global cultural gender preferences for venues using social sensing. *EPJ Data Science* 6, 1 (19 May 2017), 5. https://doi.org/10.1140/epjds/s13688-017-0101-0

[100] Mor Naaman, Amy X Zhang, Samuel Brody, and Gilad Lotan. 2012. On the Study of Diurnal Urban Routines on Twitter. In *Proc. of ICWSM'12*. Dublin, Ireland.

[101] Diala Naboulsi, Razvan Stanica, and Marco Fiore. 2014. Classifying call profiles in large-scale mobile traffic datasets. In *Proc. of INFOCOM'14*. IEEE, Toronto, Canada, 1806–1814.

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

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

[104] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. 2001. On Spectral Clustering: Analysis and an Algorithm. In *Proc. of NIPS'01*. MIT Press, Cambridge, MA, USA, 849–856. http://dl.acm.org/citation.cfm?id=2980539.2980649

[105] T. Nguyen and B. K. Szymanski. 2012. Using Location-Based Social Networks to Validate Human Mobility and Relationships Models. In *Proc. ASONAM '12*. IEEE Computer Society, Washington, DC, USA, 1215–1221. https://doi.org/10.1109/ASONAM.2012.210

[106] Anastasios Noulas, Salvatore Scellato, Cecilia Mascolo, and Massimiliano Pontil. 2011. Exploiting Semantic Annotations for Clustering Geographic Areas and Users in Location-based Social Networks. In *Proc. of ICWSM'11*. AAAI, Barcelona, Spain.

[107] Eduardo Mucelli Rezende Oliveira, Aline Carneiro Viana, K.P. Naveen, and Carlos Sarraute. 2017. Mobile data traffic modeling: Revealing temporal facets. *Computer Networks* 112 (2017), 176 – 193.

[108] Bei Pan, Yu Zheng, David Wilkie, and Cyrus Shahabi. 2013. Crowd Sensing of Traffic Anomalies Based on Human Mobility and Social Media. In *Proc. of SIGSPATIAL'13*. ACM, Orlando, Florida, 344–353.

[109] Christine Parent, Stefano Spaccapietra, Chiara Renso, Gennady Andrienko, Natalia Andrienko, Vania Bogorny, Maria Luisa Damiani, Aris Gkoulalas-Divanis, Jose Macedo, Nikos Pelekis, et al. 2013. Semantic trajectories modeling and analysis. *ACM Computing Surveys (CSUR)* 45, 4 (2013), 42.

[110] Michael J Paul and Mark Dredze. 2011. You are what you Tweet: Analyzing Twitter for public health.. In *Proc. of ICWSM'11*. Barcelona, Spain, 265–272.

[111] Eric Paulos and Elizabeth Goodman. 2004. The familiar stranger: anxiety, comfort, and play in public places. In *Proc. of CHI'04*. ACM, Vienna, Austria, 223–230.

[112] Jamie Pearce, Tony Blakely, Karen Witten, and Phil Bartie. 2007. Neighborhood deprivation and access to fast-food retailing: a national study. *American journal of preventive medicine* 32, 5 (2007), 375–382.

[113] Barbara Poblete, Ruth Garcia, Marcelo Mendoza, and Alejandro Jaimes. 2011. Do all birds tweet the same?: characterizing twitter around the world. In *Proc. of CIKM'11*. ACM, Glasgow, UK, 1025–1030.

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

[115] Daniele Quercia, Rossano Schifanella, and Luca Maria Aiello. 2014. The Shortest Path to Happiness: Recommending Beautiful, Quiet, and Happy Routes in the City. In *Proc. of HT '14*. ACM, New York, NY, USA, 116–125. https://doi.org/10.1145/2631775.2631799

[116] Daniele Quercia, Rossano Schifanella, Luca Maria Aiello, and Kate McLean. 2015. Smelly Maps: The Digital Life of Urban Smellscapes. In *Proc. of ICWSM'15*. Oxford, UK.

[117] Sasank Reddy, Deborah Estrin, Mark Hansen, and Mani Srivastava. 2010. Examining Micro-payments for Participatory Sensing Data Collections. In *Proc. of Ubicomp'10*. ACM, Copenhagen, Denmark, 33–36. https://doi.org/10.1145/1864349.1864355

[118] Anna Izabel João Tostes Ribeiro, Thiago Henrique Silva, Fátima Duarte-Figueiredo, and Antonio A.F. Loureiro. 2014. Studying Traffic Conditions by Analyzing Foursquare and Instagram Data. In *Proc. of PE-WASUN '14*. ACM, Montreal,

Canada, 17–24. https://doi.org/10.1145/2653481.2653491

[119] Filipe N. Ribeiro, Matheus Araújo, Pollyanna Gonçalves, Marcos André Gonçalves, and Fabrício Benevenuto. 2016. SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science* 5, 1 (2016), 1–29. https://doi.org/10.1140/epjds/s13688-016-0085-1

[120] Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor. 2010. *Recommender Systems Handbook* (1st ed.). Springer-Verlag, Berlin, Heidelberg.

[121] Salvatore Rinzivillo, Fernando de Lucca Siqueira, Lorenzo Gabrielli, Chiara Renso, and Vania Bogorny. 2013. Where Have You Been Today? Annotating Trajectories with DayTag. In *Advances in Spatial and Temporal Databases*. Springer Berlin Heidelberg, Berlin, Heidelberg, 467–471.

[122] Diego Rodrigues, Azzedine Boukerch, Thiago H. Silva, Antonio Loureiro, and Leandro Villas. 2017. SMAFramework: Urban data integration framework for mobility analysis in smart cities. In *Proc. of ACM MSWiM'17*. Miami, USA.

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

[124] Flora Salim and Usman Haque. 2015. Urban computing in the wild: A survey on large scale participation and citizen engagement with ubiquitous computing, cyber physical systems, and Internet of Things. *International Journal of Human-Computer Studies* 81 (2015), 31 – 48. Transdisciplinary Approaches to Urban Computing.

[125] Frances A. Santos, Thiago H. Silva, Torsten Braun, Antonio A. F. Loureiro, and Leandro A. Villas. 2017. Towards a Sustainable People-Centric Sensing. In *Proc. of ICC*. Paris, France.

[126] Hansen Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Richard E Lucas, Megha Agrawal, Gregory J Park, Shrinidhi K Lakshmikanth, Sneha Jha, Martin EP Seligman, et al. 2013. Characterizing Geographic Variation in Well-Being Using Tweets.. In *Proc. of ICWSM'13*. Boston, USA.

[127] Jesus Serrano-Guerrero, Jose A. Olivas, Francisco P. Romero, and Enrique Herrera-Viedma. 2015. Sentiment analysis: A review and comparative analysis of web services. *Information Sciences* 311 (2015), 18 – 38. http://www.sciencedirect.com/science/article/pii/S0020025515002054

[128] Shashi Shekhar, Mark Coyle, Brajesh Goyal, Duen-Ren Liu, and Shyamsundar Sarkar. 1997. Data Models in Geographic Information Systems. *Commun. ACM* 40, 4 (April 1997), 103–111. https://doi.org/10.1145/248448.248465

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

[130] Thiago H. Silva, Clayson S. F. de S. Celes, Joao B. Neto, Vinicius 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, and Antonio A.F. Loureiro. 2016. Users in the Urban Sensing Process: Challenges and Research Opportunities. In *Pervasive Computing: Next Generation Platforms for Intelligent Data Collection*. Academic Press, 45–95.

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

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

[133] Thiago H. Silva, Pedro O. S. Vaz de Melo, Jussara M. Almeida, Mirco Musolesi, and Antonio A. F. Loureiro. 2014. 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.

[134] Thiago H. Silva, Pedro O. S. Vaz de Melo, Jussara M. Almeida, Juliana Salles, and Antonio A. F. Loureiro. 2013. A picture of Instagram is worth more than a thousand words: Workload characterization and application. In *Proc. of DCOSS'13*. Cambridge, USA, 123–132.

[135] Thiago H. Silva, Pedro O. S. Vaz de Melo, Jussara M. Almeida, Juliana Salles, and Antonio A. F. Loureiro. 2014. Revealing the City That We Cannot See. *ACM Trans. Internet Technol.* 14, 4, Article 26 (Dec. 2014), 23 pages. https://doi.org/10.1145/2677208

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

[137] Max Sklar, Blake Shaw, and Andrew Hogue. 2012. Recommending Interesting Events in Real-time with Foursquare Check-ins. In *Proc. of RecSys '12*. ACM, Dublin, Ireland, 311–312. https://doi.org/10.1145/2365952.2366028

[138] Chris Smith, Daniele Quercia, and Licia Capra. 2013. Finger on the Pulse: Identifying Deprivation Using Transit Flow Analysis. In *Proc. of CSCW'13*. ACM, San Antonio, Texas, USA, 683–692. https://doi.org/10.1145/2441776.2441852

[139] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. 2010. Limits of predictability in human mobility. *Science* 327, 5968 (2010), 1018–1021.

[140] Bogdan State, Patrick Park, Ingmar Weber, and Michael Macy. 2015. The Mesh of Civilizations in the Global Network of Digital Communication. *PLOS ONE* 10, 5 (05 2015), 1–9. https://doi.org/10.1371/journal.pone.0122543

[141] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. 2005. *Introduction to Data Mining, (First Edition)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

[142] Jiliang Tang and Huan Liu. 2014. Feature Selection for Social Media Data. *ACM Trans. Knowl. Discov. Data* 8, 4, Article 19 (Oct. 2014), 27 pages. https://doi.org/10.1145/2629587

[143] Jacob Thebault-Spieker, Loren G. Terveen, and Brent Hecht. 2015. Avoiding the South Side and the Suburbs: The Geography of Mobile Crowdsourcing Markets. In *Proc. of CSCW'15*. ACM, Vancouver, Canada, 265–275. https://doi.org/10.1145/2675133.2675278

[144] Declan Traynor and Kevin Curran. 2012. Location-based social networks. *From Government to E-Governance: Public Administration in the Digital Age* (2012), 243.

[145] Twitter. 2017. *It's what's happening*. Twitter.com. https://goo.gl/Mn6R4U.

[146] Carmen Karina Vaca, Daniele Quercia, Francesco Bonchi, and Piero Fraternali. 2015. Taxonomy-based discovery and annotation of functional areas in the city. In *Proc. of ICWSM'15*. Oxford, UK.

[147] Laura Vaughan, David L Chatford Clark, Ozlem Sahbaz, and Mordechai (Muki) Haklay. 2005. Space and exclusion: does urban morphology play a part in social deprivation? *Area* 37, 4 (2005), 402–412. https://doi.org/10.1111/j.1475-4762.2005.00651.x

[148] Adriano Veloso, Wagner Meira Jr., and Mohammed J. Zaki. 2006. Lazy Associative Classification. In *Proc. of ICDM'06*. IEEE Computer Society, Washington, DC, USA, 645–654. https://doi.org/10.1109/ICDM.2006.96

[149] Alessandro Venerandi, Giovanni Quattrone, Licia Capra, Daniele Quercia, and Diego Saez-Trumper. 2015. Measuring Urban Deprivation from User Generated Content. In *Proc. of CSCW'15*. ACM, Vancouver, Canada, 254–264.

[150] Claudia Wagner, Philipp Singer, and Markus Strohmaier. 2014. The nature and evolution of online food preferences. *EPJ Data Science* 3, 1, 38. https://doi.org/10.1140/epjds/s13688-014-0036-7

[151] M. C. Watson. 2015. Time maps: A tool for visualizing many discrete events across multiple timescales. In *Proc. of IEEE Big Data*. 793–800. https://doi.org/10.1109/BigData.2015.7363824

[152] Michael J. Widener and Wenwen Li. 2014. Using geolocated Twitter data to monitor the prevalence of healthy and unhealthy food references across the US. *Applied Geography* 54 (2014), 189 – 197. https://doi.org/10.1016/j.apgeog.2014.07.017

[153] L. D. Xu, W. He, and S. Li. 2014. Internet of Things in Industries: A Survey. *IEEE Transactions on Industrial Informatics* 10, 4 (Nov 2014), 2233–2243. https://doi.org/10.1109/TII.2014.2300753

[154] Zhixian Yan, Dipanjan Chakraborty, Christine Parent, Stefano Spaccapietra, and Karl Aberer. 2013. Semantic Trajectories: Mobility Data Computation and Annotation. *ACM Trans. Intell. Syst. Technol.* 4, 3, Article 49 (July 2013), 38 pages. https://doi.org/10.1145/2483669.2483682

[155] Sarita Yardi, Daniel Romero, Grant Schoenebeck, et al. 2009. Detecting spam in a twitter network. *First Monday* 15, 1 (2009).

[156] Mao Ye, Dong Shou, Wang-Chien Lee, Peifeng Yin, and Krzysztof Janowicz. 2011. On the Semantic Annotation of Places in Location-based Social Networks. In *Proc. of KDD'11*. ACM, San Diego, California, USA, 520–528. https://doi.org/10.1145/2020408.2020491

[157] Hongzhi Yin, Yizhou Sun, Bin Cui, Zhiting Hu, and Ling Chen. 2013. LCARS: a location-content-aware recommender system. In *Proc. of KDD'13*. ACM, Chicago, USA, 221–229.

[158] Reza Zafarani, Mohammad Ali Abbasi, and Huan Liu. 2014. *Social media mining: an introduction*. Cambridge University Press.

[159] Matei Zaharia, Mosharaf Chowdhury, Michael J. Franklin, Scott Shenker, and Ion Stoica. 2010. Spark: Cluster Computing with Working Sets. In *Proc. of HotCloud'10*. Berkeley, USA, 10–10.

[160] Ke Zhang and Konstantinos Pelechrinis. 2014. Understanding Spatial Homophily: The Case of Peer Influence and Social Selection. In *Proc. of WWW'14*. ACM, New York, NY, USA, 271–282. https://doi.org/10.1145/2566486.2567990

[161] Yu Zheng. 2011. Location-Based Social Networks: Users. In *Computing with spatial trajectories. Zheng, Yu and Zhou, Xiaofang*. Springer press.

[162] Yu Zheng. 2012. Tutorial on Location-Based Social Networks. In *Proc. of WWW'12*. Lyon, France.

[163] Yu Zheng, Licia Capra, Ouri Wolfson, and Hai Yang. 2014. Urban computing: concepts, methodologies, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)* 5, 3 (2014), 38.

[164] Yu Zheng, Tong Liu, Yilun Wang, Yanmin Zhu, Yanchi Liu, and Eric Chang. 2014. Diagnosing New York City's Noises with Ubiquitous Data. In *Proc. of UbiComp'14*. ACM, Seattle, Washington, 715–725.

[165] Yu Zheng, Lizhu Zhang, Xing Xie, and Wei-Ying Ma. 2009. Mining interesting locations and travel sequences from GPS trajectories. In *Proc. of WWW'12*. ACM, Madrid, Spain, 791–800.

[166] Yan-Tao Zheng, Zheng-Jun Zha, and Tat-Seng Chua. 2012. Mining Travel Patterns from Geotagged Photos. *ACM Transactions on Intelligent Systems and Technology (TIST)* 3, 3, Article 56 (May 2012), 18 pages.

Received Jul 2017; revised Mar 2018; accepted Nov 2018