On the Social and Spatiotemporal Aspects for Urban Computing

by

Kassio Leonardo da Silva Machado

Thesis submitted in partial fulfillment of the requirements for the
Doctorate in Philosophy degree in Computer Science

School of Information Technology and Engineering
Faculty of Engineering
University of Ottawa

Abstract

In this thesis, we address urban sensing approaches, especially the challenge of providing scalable sensing systems of social, spatial and temporal variables for urban scenarios. We deal with the problem of obtaining adequate information to be used in the design of urban solutions and improve services sensitive to the social and spatiotemporal context. We investigate collective behavior in the spatiotemporal context through Online Social Networks (OSNs) due to their enormous popularity and acceptance among users of mobile personal devices. This acceptance provides significant user's engagement, where substantial amounts of data are shared online daily, resulting in massive repositories of contextual information.

Initially, we investigate the spatiotemporal preferences of users in several cities around the world in a joint analysis with climatic data. The results showed that a subset of cities exhibits a dynamic behavior that concerns the users' spatial preferences, where temperature thresholds can characterize the shift. The preference shift includes places visited by users in the studied cities as well as the transition between regions of the city, such that this phenomenon may be observed in significant portions of the population and places.

In this work, we also investigate the effects of spatiotemporal dynamics on the co-location of users. We propose a network model based on timely meetings to estimate the social graph based on the geographic proximity. The results show the changes in the structural characteristics of the graph over time. Based on these insights, we propose a message forwarding protocol for delay-tolerant applications capable of increasing delivery rate while optimizing message replication and delay.

In these investigations, we include the spatiotemporal dynamics of urban areas regarding content consumption. We jointly investigate the location of users and their content of interest according to the metadata used in this work provided by the OSN. The results show that the studied areas can provide substantial demands for redundant content in small spatiotemporal windows, such that users could cooperate to provide content locally and offload the demand. From these observations, we propose a distributed cache management mechanism able to take advantage of users' social and spatial persistence. Finally, we combine OSN data and official local government data to formulate an urban sensing framework capable of assessing characteristics of urbanism, sociability, and mobility of citizens of a city.

In this work, we propose methodologies of sensing and analysis of social and spatiotemporal characteristics in urban settings. The proposed methods and applications favor the use of publicly available data to provide effective generalization capability. Thus, the applications and contributions of this work can be reproduced in cities not contemplated in this study, and other cities can take advantage of urban sensing.

List of Publications

1. Machado, Kassio, Azzedine Boukerche, Eduardo C Cerqueira, and Antonio Loureiro. "A Data-Centric Approach for Social and Spatiotemporal Sensing in Smart Cities." *IEEE Internet Computing*, 2018.

2. Machado, Kassio, Azzedine Boukerche, Eduardo C Cerqueira, and Antonio AF Loureiro. "A socially-aware in-network caching framework for the next generation of wireless networks." *IEEE Communications Magazine 55*, no. 12, 2017.

3. Machado, Kassio, Azzedine Boukerche, Pedro OS Vaz de Melo, Eduardo C Cerqueira, and Antonio AF Loureiro. "Pervasive forwarding mechanism for mobile social networks." *Elsevier Computer Networks Journal 111*, 2016.

4. Machado, Kassio, Azzedine Boukerche, Eduardo C Cerqueira, and Antonio AF Loureiro. "Long-Term Spatiotemporal Analysis of Social Media for Device-to-Device Networks." *IEEE Global Communications Conference (GLOBECOM)*, pp. 1-6, 2016.

5. Machado, Kassio, Azzedine Boukerche, Pedro OS Vaz de Melo, Eduardo Cerqueira, and Antonio Loureiro. "Exploring seasonal human behavior in opportunistic mobile networks." In *IEEE International Conference on Communications (ICC)*, 2016.

6. Machado, Kassio, Thiago H Silva, Pedro OS Vaz de Melo, Eduardo C Cerqueira, and Antonio AF Loureiro. "Urban mobility sensing analysis through a layered sensing approach." In *IEEE International Conference on Mobile Services (MS)*, pp. 306-312, 2015.

Table of Contents

List of Tables vii

List of Figures viii

1 Introduction 1

1.1 Thesis Statement . . . 2
1.2 Contributions . . . 2
1.3 Thesis Outline . . . 3

2 Social and Spatiotemporal Aspects of Urban Computing 4

2.1 Introduction . . . 4
2.2 Measuring The Urban Scenario . . . 6
2.2.1 Urban Sensing . . . 6
2.2.2 Spatial and Temporal Aspects . . . 7
2.2.3 Social Aspects . . . 9
2.3 Urban Sensing Applications . . . 12
2.3.1 Wireless Networks . . . 12
2.3.2 Transportation and Vehicular Networks . . . 15
2.3.3 Internet of Things . . . 17
2.3.4 Urban Computing . . . 19
2.4 Conclusion . . . 21

3 Urban Sensing Through Data Layers 23

3.1 Introduction . . . 23
3.2 Related Work . . . 25
3.3 Issues and Challenges on Sensing Layers . . . 26

3.4 Sensing Layers . . . 27
3.4.1 Formal Model . . . 27
3.4.2 Layer Characterization . . . 28
3.5 Layer Combination . . . 29
3.6 Discussion . . . 37
3.7 Conclusion . . . 40

4 A Long-Term Analysis of Social and Spatiotemporal Aspects 41
4.1 Introduction . . . 41
4.2 Related Work . . . 43
4.3 Graph Characterization . . . 44
4.4 Spatiotemporal Analysis . . . 45
4.4.1 Proximity Graph . . . 46
4.4.2 Spatial Distribution and Content Similarity . . . 49
4.4.3 Discussion . . . 52
4.5 Conclusion . . . 53

5 Pervasive Forwarding Mechanism for Mobile Social Networks 54
5.1 Introduction . . . 55
5.2 Related Work . . . 56
5.3 Trace-Based Analysis . . . 59
5.3.1 Data Description . . . 59
5.3.2 Data Combination . . . 61
5.4 PervasivePeopleRank . . . 63
5.5 Performance Evaluation . . . 66
5.5.1 Network Model . . . 66
5.5.2 Contact Graph Analysis . . . 67
5.5.3 Network Performance . . . 69
5.6 Conclusion . . . 72

6 Social-Based Distributed Caching For Urban Scenarios **74**
6.1 Introduction . . . 74
6.2 Related Work . . . 76
6.3 Network Model . . . 79
6.3.1 Multigraph Approach . . . 79
6.3.2 Network Model . . . 80
6.4 Popularity-Based Social Caching . . . 81
6.5 Spatiotemporal Analysis . . . 83
6.6 Evaluation . . . 88
6.7 Conclusion . . . 91

7 A Data-Centric Approach for Social and Urban Sensing **93**
7.1 Introduction . . . 93
7.2 Data Sources . . . 94
7.3 Sensing Social Aspects . . . 96
7.4 Sensing Spatiotemporal Aspects . . . 99
7.5 Conclusion . . . 105

8 Conclusion and Future Work **106**
8.1 Summary of the Thesis . . . 106
8.2 Future Research Directions . . . 107

References **109**

List of Tables

3.1 Thermal thresholds of transitions between phases. 32

5.1 Parameters of simulation of the proximity graph. 67

6.1 Online social network dataset description. 82

List of Figures

2.1 Sensing spectrum. 8
2.2 Applications of undirected and directed graphs. 10

3.1 Sensing Layers build using check-ins, traffic, and weather data. 25
3.2 Data layer combination according to formal model of Sensing Layers. . . . 28
3.3 Temperature history along period of the data collection. 30
3.4 Mean popularity of data samples according to the days of week. 31
3.5 Matrix correlation of local temperature and spatial distribution of users. . 33
3.6 Phase transitions of temperature $n = 50$. 34
3.7 Check-in representativity δ and γ subsets. 36
3.8 Intersection representativity of δ. 38
3.9 Graph of difference α - β. 39

4.1 Number of components and edge density of proximity graph. 45
4.2 Mean of nodes during for each time slot T_{th}. 46
4.3 Mean degree of nodes. 47
4.4 Giant component size on weekdays and weekends. 47
4.5 Re-encounter interval probability. 48
4.6 Distributions of encounters and venues. 48
4.7 Spatial distribution of popular spots of encounters. The red circles represent spots that were not among the most popular places in the previous hour. . 50
4.8 Spatial distribution of encounters with similar content. 51
4.9 Complementary cumulative distribution of encounter interval. 51
4.10 Density function of publications with similar content and spatial proximity. 52

5.1 Average of temperature for the selected time series. 60

5.2 Curves of popularity, in terms of the number of users, during the seasons observed. . . . 61
5.3 Principal component analysis of venues' popularity according to temperature. 62
5.4 Popular venues in New York City in different phases. . . . 63
5.5 Entropy average of encounters grouped by phases. . . . 64
5.6 Analysis of the graph of contact. . . . 68
5.7 Distribution of the shortest paths. . . . 69
5.8 Delivery ratio and average cost according to temperature variation. . . . 70
5.9 Average of hops and CCDF of latency. . . . 71
5.10 Delivery ratio considering different C_{time}. . . . 72

6.1 Time series of samples grouped by cities and days. . . . 79
6.2 The cumulative aggregation of content interests, proximity, and social graph. 80
6.3 Pause time considered in the network model. . . . 82
6.4 PopSoC Framework components. . . . 84
6.5 Edge occurrence. . . . 85
6.6 Distance between consecutive encounters. . . . 85
6.7 Interval between consecutive encounters. . . . 86
6.8 Encounter duration 86
6.9 Topology overlap. . . . 87
6.10 Geographic and social persistence of users. . . . 87
6.11 New York encounters during two distinct sports events. . . . 88
6.12 Distribution of content popularity. . . . 89
6.13 The impact of α parameter. . . . 89
6.14 Hit per edge. . . . 90
6.15 Hit per edge with persistence. . . . 90
6.16 Delay of hits. . . . 91

7.1 Characterization of the proximity graph. . . . 97
7.2 Interval of re-encounters. . . . 98
7.3 Edge coverage as a function of the set of venues. . . . 99
7.4 Attributes of neighborhoods. . . . 100
7.5 Treemaps of venues for SoHo and Lower East Side. . . . 101

7.6 Temporal activity registered in the OSN samples. 102
7.7 Classification of temporal activity. 103
7.8 Spatial brokerage. 104
7.9 Spatial homogeneity. 105

Chapter 1

Introduction

The urban setting has developed in many ways, making life in urban centers a convenient and attractive option for more and more people. Estimates say that by 2019, approximately 9% of the world's population will be living in one of the 41 most populous cities on the planet, megacities with more than 10 million inhabitants[1]. This agglomeration of people will create a set of situations and challenges for the local administration of these cities. In order to deal with these challenge, local authorities and governments are regularly applying new technologies to deal with the massive demands for shared services; nevertheless, much of the technology clash in the scalability that makes these approaches impractical [76].

Industries and academia have pursued alternatives to the cities' problems, where several areas of study have contributed to sustainable ways of dealing with various challenges in areas such as pollution [84], urban mobility [93], public safety [24], urbanism [37], and telecommunications [52], among many others. Most of these alternatives rely on sensing systems to monitoring certain environment variables through a conventional sensing system consisting of physical devices with a well-defined purpose. Also, they usually require high implementation and maintenance costs when applied over large coverage areas, making the system unfeasible for most of the cities.

On the other hand, alternatives for urban sensing independently of physical sensors have gained momentum and attention from researchers and government. In this scenario, urban sensing explores virtual sensors, software tools that integrate external data sources, statistical methods and data analyses to provide streams of information on a large number of environmental variables. Especially with complex sensors, virtual sensors, which manipulate two or more data sources, have expanded the sensing spectrum in general, adding the ability to measure variables whenever there are no conventional sensors, such as social and behavioral aspects.

These new sensing capabilities leverage urban computing and allow approaches that operate on variables from different domains and sources. Since urban computing is strongly coupled with human behavior, the close examination of the collective behavior and other social aspects in light of the spatiotemporal conditions represent a new horizon for the understanding of urban areas.

[1] https://www.economist.com/node/21642053

1.1 Thesis Statement

Next, we present the three key elements of this work.

Research Problem: this thesis presents the investigation of the challenges of providing scalable urban sensing in large geographic areas. Since static physical sensors are traditionally the primary technology of urban monitoring, deploying and managing them represent a significant financial and ecological cost. Due to the hardware limitations, there are currently no sensing mechanisms for various social, human behavioral and spatiotemporal variables and adding new variables to the set of monitored characteristics is not a viable task. Therefore, in this thesis, we investigated solutions to overcome the limitations of hardware-based sensing to provide large-scale, multimodal sensing and new sensing capabilities using Online Social Networks.

Key Idea: to explore alternative ways of providing spatial, temporal and social information through complex sensors, a category of virtual sensors that leverage public data and apply statistical analyses to take advantage of citizens' ubiquity and the popularity of personal devices to provide efficient sensing of great capillarity.

Objectives: the main objective of this thesis is to propose a scalable and replicable sensing alternative to support urban computing applications. Specifically, to present mechanisms capable of providing insights and relevant data capable of extending the conventional sensing spectrum of an urban scenario and its citizens. In particular, propose methodologies and frameworks capable of providing information about social, spatial, and temporal aspects, as well as favoring open and popular data sources that allow reproducibility in different cities around the world. Moreover, to use this information to solve particular problems such as a message forwarding protocol and a distributed cache management mechanism.

1.2 Contributions

In the following we summarize the contributions present in this thesis.

- In the context of urban sensing, we have investigated and developed scalable and reproducible sensing mechanisms. The mechanisms and analyses developed in this work explored public data mainly from OSNs to estimate properties of mobility, sociability, urban planning, and others. In this way, user engagement can leverage sensing capabilities, and the massive popularity of these applications can provide reproducibility in different locations around the world [76, 80].

- In the context of computer networks, we investigated how social and spatiotemporal aspects can support wireless networks. Initially, we explored the dynamics of encounters and co-location of users of OSNs to provide efficient Device-to-Device (D2D) communication. We designed a protocol for Mobile Social Networks (MSNs) that is aware of and adaptable to fluctuations in urban and social characteristics [79, 78].

In addition, we explored such dynamics to identify spatiotemporal characteristics of the consumption and publication of online content [77]. From these observations, we developed a content caching framework that supports locally content provisioning on wireless networks through cooperation between users via D2D [80]. The results of these investigations showed gains in the performance of the networks, as well as the relevance and applicability of the social and spatiotemporal aspects that can be considered in the development of network protocols.

- In the context of the social and spatiotemporal aspects, we used the mechanisms of sensing and analysis proposed to extract observations about urban dynamics and collective behavior. Our proposed mechanisms evidenced significant changes in the spatial distributions of users, social graph structure, and the popularity of the points of interest (POI) of the cities. The results indicated a correlation of spatial distribution of users and environmental temperature, such that the climatic variable may be used as a predictor for fluctuations of social and spatiotemporal variables [76]. Furthermore, the collective behavior presents significant regularity, where we identified social persistence through re-encounters of users and spatial persistence through regular visits to the same regions and POI [80]. Spatial and temporal patterns have also been found in online content consumption preferences. The analyses of the content shared by OSN users indicated that, in urban areas, two or more geographically close users may request the same content or service over short periods and this redundant demand has dynamic characteristics across the city regions and daily hours [81].

1.3 Thesis Outline

The remainder of this thesis is organized as follows. Chapter 2 provides an overview of the social and spatiotemporal aspects of urban sensing. Chapter 3 presents a study of the spatiotemporal characteristics of six cities around the world, investigating the relationship between the spatial distribution of their inhabitants and the local temperature. Chapter 4 presents the results of a long-term study of the spatial and temporal characteristics of mobile application users, as well as the content consumed and published by them. Chapter 5 presents an alternative of opportunistic communication centered on the dynamic characteristics of the social and proximity graph. Chapter 6 presents an alternative approach to offload redundant demand in cellular networks by exploiting social and spatiotemporal persistence characteristics of mobile users. Chapter 7 presents a study of the urban scenario considering government and mobile applications data to leverage the capacity of sensing in urban scenarios. Finally, Chapter 8 presents the conclusions of this work and some future directions.

Chapter 2

Social and Spatiotemporal Aspects of Urban Computing

Modern urban settings are environments with an abundance of variables and characteristics that make these places interestingly complex. The crowding of people and the way they behave are critical to the city's essential services such as transportation and communication. In this chapter, we introduce the urban computing and sensing, as well as the dynamics that surround the main entities of the urban environments, which concern people, places, and objects, under the light of space and time. We present recent works that use social, temporal, and spatial aspects or a combination of them to propose solutions to problems encountered in urban computing, especially in large-scale scenarios.

Initially, Section 2.1 presents the introduction that discusses the use of social and spatiotemporal aspects in pervasive applications and their convergence for urban computing applications. Section 2.2 presents the fundamental concepts of urban sensing and analysis of social aspects. Section 2.3, presents recent studies and applications that use social and spatiotemporal data to undertake problems in areas such as wireless networks, urban mobility, transportation, Internet-of-Things (IoT), public security, urban planning, and others. Finally, Section 2.4 presents the conclusion.

2.1 Introduction

Urban areas can be seen beyond a form of social organization: they can also be seen as complex systems consisting of a large number of variables that provide great intrinsic dynamism. Social and economic factors have been determinant in attracting people, services and all kinds of investments to large urban areas, consequently introducing great heterogeneity and additional complexity.

Managing such complex and dynamic systems is a common challenge for technological development, where on one hand we have the governmental authorities with a great interest in solutions for the management of resources and services and on the other hand, we have industries and academia investigating and proposing solutions. From this motivation,

computer-based solutions evolved into urban computing, a concept introduced by Paulos and Goodman [97] and subsequently defined by Kindberg et al. [67] as the integration of computing, sensing, and actuation technologies into urban settings and its inhabitants' lifestyles.

Many authors have explored and contributed to the development of the term and the topic, as Zheng et al. [147] that define the term as a process of collecting, integrating, and analyzing a significant volume of heterogeneous data related to urban areas, and supporting the solutions to the challenges and problems of this type of environment. Recently, Silva et al. [120] defined urban computing as computer-mediated means to understand the aspects of urban phenomena and estimate the future of the cities. The authors also describe it as an interdisciplinary area resulting from the fusion of computer science fields, as human-computer interaction, computer networks, and data mining with traditional areas such as economics, geography, and sociology in the context of urban environments. Therefore, the development of solutions for urban settings should consider the effects of their implementation in the short and long term, especially considering the social, economic, and environmental aspects. Nevertheless, the development process of these solutions can be expensive because of the usual large scale deployment required by urban scenarios.

The subset of technology solutions, especially computer-based ones, has explored alternatives to address the challenges of scalability by using alternative forms of manipulation of the required data. The growing popularity of online content-sharing platforms has had a major impact on how we interact with the web nowadays; in particular, mobile applications have become one of the most popular forms of online content consumption. In addition, smartphones have become the primary device for a significant portion of Internet users.

Thus, solution developers have explored mobile applications as an alternative way of drawing conclusions about the state of the city in various aspects due to their massive acceptance among citizens and the potential for data collection [84, 144, 50]. Such developers have explored social, temporal, and spatial aspects through these applications to get real-time data about the population needs [120]. For instance, the study of computer network worloads has provided significant performance improvements in supporting resource provisioning.

In recent years, the studies of demand provisioning for computational resources have developed substantially, especially network-provided services, evolving from temporal-based to spatiotemporal-based prediction. In the computer networks, Content Distribution Networks have played an important role in this evolution by promoting the displacement of content, traditionally stored in the core of the network, toward the edge of the network.

Spatial properties are an important factor in the performance of these networks by providing indicators of geographic locations suitable for the deployment of cache services, content replicators and mobility management in general [16, 143, 9]. In mobile networks, these cache mechanisms can be part of users' equipment, which makes the users' behavior a critical aspect, where the influence of the behavior can be observed in the temporal, spatial, and social contexts. Therefore, despite the challenges related to the node mobility [], behavioral and social aspects are effective predictors of demand, among other characteristics.

From the point of view of behavioral aspects, the applications have exploited the individual preferences of the users concerning the content consumed and the characteristics that orbit the moments of consumption. An abstract example is a user who prefers a specific content type from a particular provider according to a given environmental condition; in a concrete example, this is a user with a preference for watching videos provided by YouTube when he is at home.

From the perspective of the social aspect, applications have explored social links between users, such as co-location, friendship, and multiple encounters to identify clusters of content and users, as well as manage resources in the infrastructure and user equipment. In addition, social relationships can be imprinted on devices through manufacturers' initiatives and the behavior of their owners.

2.2 Measuring The Urban Scenario

In this section, we introduce fundamental concepts for the study of urban variables through spatial, temporal, and social properties.

2.2.1 Urban Sensing

The task of measuring urban variables represents a multidisciplinary challenge. For this reason, alternative mechanisms have been designed over the years to sense urban areas considering many constraints, in special the economic and spatial challenges.

Urban sensing is one of the prominent alternatives that has been studied to address the limitations of coverage and sparsity of sensing in large geographical areas. As defined by Campbell et al. [21], urban sensing is the action of collecting data about people's immediate surroundings and how people interact and interpret their surroundings. The authors argue that this people-centric perspective explores the premise that people are no longer only data consumers, nonetheless data producers and consumers form two sets with a significant intersection. According to Shin et al. [114], urban sensing is a type of social sensing through mobile devices that provides opportunities to track multiple data points in real-time, and therefore to sample the dynamic behavior and inherent complexity of human activity within the city. In this way, urban sensing goes beyond conventional sensing based on Wireless Sensor Networks and Internet-of-Things and explores personal devices and crowds of users as key players in the data collection process.

Similarly, Jaimes et al. [60] defined Crowd Sensing as mechanism for sensing based on the collection of observations through a large number of individuals. According to this definition, one individual may not provide sufficient data, however the aggregated data from many individuals can provide sampling with significant quality and coverage of the target phenomena. Ma et al. [75] defined the concept of Mobile Crowd Sensing as a sensing paradigm in which individuals with sensing and computing devices collectively share data and extract information to measure and map phenomena of common interest, where applications are categorized into two groups.

- Participatory Sensing: requires the participants to meet the application requests consciously. The user is directly involved in the sensing task. Applications in this category place the user as part of the decision and collection process, where each user acts as an individual sensor node [47].

- Opportunistic sensing: users act passively on the sensing process. The sensing engine runs in the background and collects data opportunistically without active user involvement. It shifts the burden of supporting an application from the custodian to the sensing system [66, 70].

Thus, the abundance of data collected through users, personal devices, conventional sensors throughout the cities' elements, makes it possible to combine data sources that broaden the sensing spectrum through software-based sensing, or virtual sensors. According to Madria et al. [82], a virtual sensor is the emulation of a physical sensor that obtains data from one or more underlying physical sensors, capable of providing a customized view or measurement to the user. Kabadayi et al. [62] defined virtual sensors as sources of indirect measurements and abstract conditions using heterogeneous groups of physical sensors through the combination of their data. In this approach, signals from different physical sensors can be used together to provide the measurement or output signal of a variable for which there is no physical sensor.

In this study, we explore urban sensing techniques as defined by Campbell et al. [21], such that the approaches presented mainly explore the data obtained through users or other distributed methods, as well as mechanisms indirectly dependent on physical sensors, such as the virtual sensors as defined by Kabadayi et al. [62].

2.2.2 Spatial and Temporal Aspects

The valuation of data has caused great impacts on the technologies that we use every day. Industry, academia, and governments have seen data as a valuable asset for scientific, economic, and social development. Data related or indexed by spatial and temporal characteristics have been widely used in real applications within the reach of most citizens of urban areas. This massive data supply has driven the urban sensing initiatives that see in these data opportunities to develop methods and mechanisms able to expand the capabilities of observing urban variables [76].

Since urban sensing naturally requires the sensing of a wide geographic area, traditional sensing has been investigating alternatives to provide the monitoring of urban variables using scalable approaches. Wireless sensing systems and data aggregation mechanisms have been investigated, however these approaches do not overcome the challenges directly related to the use of physical sensors [28]. Sensing based on physical sensors has a fundamental role in the monitoring of the urban scenario, especially the sensing of environmental variables. Nevertheless, in this sensing paradigm, the equipment used is strongly limited by the purpose, i.e., an immutable set of observed variables, also the costs of implementation and maintenance.

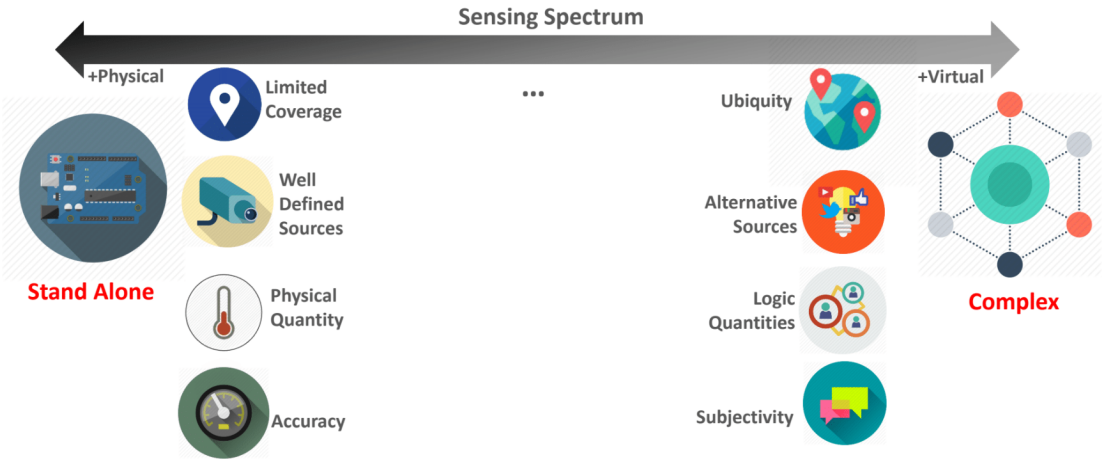


Figure 2.1: Sensing spectrum.

Recently, the urban sensing have faced the challenge of improving and expanding sensing capabilities [145, 66]. Figure 2.1 presents the sensing capabilities through a sensing spectrum divided into two major paradigms. The virtual sensing paradigm is the result of urban sensing studies that explore the use of one or more alternative data sources to provide ubiquitous and efficient sensing, as well as sensing to variables that do not have conventional sensors. The alternative data sources used in these approaches represent the reuse of data collected for a different purpose that require statistical methods, machine learning, and artificial intelligence to extrapolate the usefulness of the data [120], in addition to providing sensing that uses less or dispenses specific equipment [12].

Spatiotemporal characteristics have been used to evaluate the urban scenario for decades; however, the increasing ease of data collection has increased the relevance of these properties as effective ways of indexing the dynamics of urban variables. Schuster et al. [110] argues that the context data can be modeled on a taxonomy organized into space, time, people, and information dimensions. The authors demonstrate that the spatial dimension determines where urban and social dynamics occur and the geographical area of observation determines the scope and types of interactions observed. According to the authors, on a geographic scale, a small spatial scope would cover the events observed at the level of a place; ideally, events limited by a few meters observed in places like a club or a conference. In the medium scope, the spatial characteristics are indexed at the city level, referring to the citizens, regions of the city and neighborhoods as examples, while in a large scope, the events are defined on a global scale.

The data from Online Social Networks (OSN), mobile telephony, and traffic have been used to evaluate the spatiotemporal properties of various cities around the world. From this point, we use the term OSN for any online platform for social interaction capable of capturing users' spatiotemporal properties, such as Twitter[1] and Instagram[2], which indirectly

[1] https://www.twitter.com
[2] https://www.instagram.com

capture and make these properties available through metadata, as well as Foursquare[3] and Swarm[4] that focus and depend directly on this type of data, also known as Location-Based Social Networks (LBSN). Pappalardo et al. [96] explored the dichotomy of spatial behavior that determines two distinct classes of clients, according to the mobility estimated through the records of the calls registered by a cellular operator. The analyses showed that users could be classified among returners, which are users with significant geographical persistence, and explorers, which are users who frequently visit places never visited before. Using mobile data, Horanont et al. [52] investigated users' daily activity preferences and identified the influence of weather conditions. From the spatial information captured by phone call records, the authors estimated the mobility at the city level, as well as user activities and Points-of-Interest (POI).

2.2.3 Social Aspects

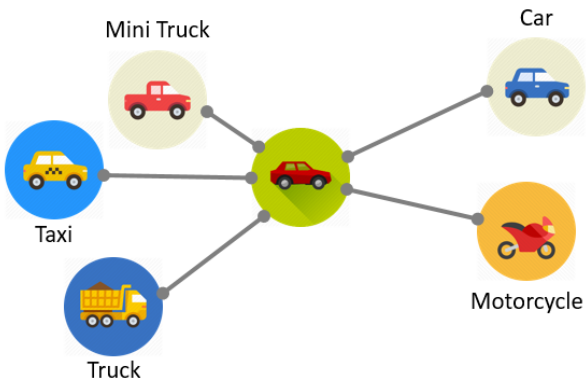
In this section, we present some fundamental characteristics of the evaluation of social aspects through graph theory. Complex networks and graph analysis have provided robust methods that are widely used for evaluating social dynamics and investigating the relationships and interactions among users themselves and with other entities, such as places and objects [15].

The undirected graph is one of the most used representations for social characteristics. Formally, an undirected graph is defined as $G = (V, E)$, where V is a finite set of vertexes representing the network nodes, and E is the finite set of edges that represent the relationships among the nodes; thus, a $e \in E$ is a set $\{u, v\}$, where u and $v \in V$. Directed graphs are an alternative representation of the relationship between a set of vertexes, where the edges represent an unequal relationship. Thus, in a directed graph $G' = (V, E)$, where the set of edges $e \in E$ is formed by ordered pairs (u, v), such that u, and $v \in V$.

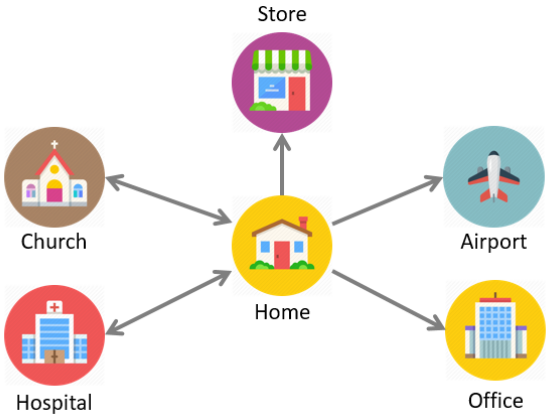
The representation of proximity graphs typically uses undirected graphs to model homogeneous relationships, such as co-localization. Nevertheless, directed graphs are also used to represent social relationships, as following and matching, which are common in OSN applications. Figure 2.2 exemplifies two scenarios with distinct graph instances. Figure 2.2a presents an undirected graph of proximity between vehicles, where the vertexes are the vehicles, and the edges represent the connectivity between them. In this scenario, the connectivity relationship is presented as bidirectional, such that the existence of an edge represents the communication capability between the two involved vehicles. In Figure 2.2b, the graph represents the transitions between a set of places, where the edges represent the origin and destination of the commute. Unlike the non-directed graph, in this example, the edges indicate one-way transitions to the airport and the office and the round-trip, or bidirectional trip, to other places.

The connectivity of a graph is one of the fundamental characteristics of these data structures. An undirected graph is said to be connected if there exists a path between any two vertexes u and $v \in V$, where a path is a sequence of edges $s = \{u_i, v_i\}, \{u_j, v_j\}, \ldots, \{u_n, v_n\}$,

[3] https://www.foursquare.com
[4] https://www.swarmapp.com



(a) Undirected graph of wireless connectivity



(b) Directed graph of transitions among places

Figure 2.2: Applications of undirected and directed graphs.

such that these edges represent a sequence of path fragments from u to v. Similarly, a directed graph is strongly connected if there is a path between any vertices u and $v \in V$. Therefore, in a connected or strongly connected graph, there are no unreachable vertices. Additionally, a graph is biconnected or non-separable when this graph is connected or strongly connected, and it retains this property even when any vertex is removed. This property evidences the characteristic structural resilience of a network or its ability to maintain connectivity, independent of the edges and vertices removed.

The set of dominant vertexes of a graph $G = (V, E)$ is a subset of vertices $D_v \subset V$, where any vertex $v \notin D_v$ is connected through an edge to at least one vertex $u \in D_v$. Similarly, the set of dominant edges of a graph G is the set of edges $D_e \subset E$, where any edge $e \notin E_v$ is adjacent to at least one edge $i \in D_e$. It is important to note that finding the smallest dominant set of a graph is a classic NP-complete decision problem with efficient approximation algorithms.

Assortativity, also known as homophily, is a property widely observed in networks with social characteristics. It is the natural preference of the network's vertexes to connect to other vertexes with similar characteristics. The similarity of the vertexes is calculated using their attributes, often using the correlation of the number of vertex connections, or degree

of the node, as well as other similarity metrics, such as cosine similarity. Real networks usually show non-zero assortativity concerning the degree of the vertexes.

From this point, we will define some of the essential measures of centrality used in graph studies. The measures of centrality support the identification of prominent vertexes of the network considering local or global properties. The measure of centrality based on the degrees of the vertices is one of the fundamental measures that evaluate the connections of the vertex with the network by using the number of incident edges.

In directed graphs, a single vertex has independent degrees that denote the edges directed to the vertex and the edges directed to the neighboring vertexes, respectively named indegree and outdegree. The degree of centrality is a metric that denotes a local property because it considers only the immediate neighborhood of a vertex made of the vertexes directly connected. For this reason, the global evaluation of this property may use the distribution of degree centrality.

The distance between two connected vertices u and $v \in V$ is defined by the total of edges contained in a path between these two vertexes, such that the shortest path is any path that connects the two vertexes through the smallest series of non-repeated edges. The proximity metric of a vertex is a way to evaluate the separation of a vertex in the network; it considers the shortest path from the vertex to every other vertex of the network. Thus, the closeness centrality value of a vertex v is defined by

$$C(v) = \frac{|V| - 1}{\sum_{u \in V} d(v, u)} \tag{2.1}$$

where V is the set of vertexes of the graph, and $d(v, u)$ is the function of the shortest path between the vertexes v and u.

Betweenness centrality is a metric that evaluates vertexes that act as bridges between any other two vertexes of the graph via the smallest path between them. The betweenness of a vertex has great applicability in the study of the structural properties of networks in general, especially in studies of propagation of information due to its evaluation concerning the global relevance of a vertex. The betweenness centrality of a vertex v is defined by

$$B(v) = \sum_{v \neq u \neq w \in V} \frac{\sigma_{u,w}(v)}{\sigma_{u,w}} \tag{2.2}$$

where $\sigma_{u,w}$ represents the total of shortest paths from vertex u to w, and $\sigma_{u,w}(v)$ represents the total of shortest paths that pass through the vertex v.

The clustering coefficient measures the probability of group formation between the vertices of a graph. Group formation is a natural phenomenon commonly found in real networks, especially social networks. Social networks present subsets of vertices with a high density of edges that represent groups of friends, relatives, and co-workers, among others. The clustering coefficient of a vertex v is given as

$$C_c(v) = \frac{2|\{e_{uw} : u, w \in N_v, e_{u,w} \in E\}|}{k_v(k_v - 1)} \tag{2.3}$$

where N_v represents the set of neighbour vertexes of v, E the set of edges, and k_v the cardinality of N_v; therefore, there are $k_v(k_v - 1)$ possible edges among the vertexes in N_v. From this, the clustering coefficient of a graph can be calculated as

$$G_c = \frac{1}{|V|} \sum_{v \in V} C_c(v) \tag{2.4}$$

Studies of aspects and social structures have extensively explored graph theories to understand the relationships and interactions of people with each other and the world around them. Other properties, methods, and metrics not included in this discussion can be found in [129, 102].

2.3 Urban Sensing Applications

In this section, we gather applications related to the spatiotemporal and social aspects for urban sensing or smart cities. The selected works are grouped into four major areas: wireless networks, vehicular networks and transportation, Internet-of-Things, and urban sensing in general.

2.3.1 Wireless Networks

Due to the continuous increase in the popularity of smartphones and other personal devices, in the coming years, the expected growth of cellular data traffic is by dozens of times, causing overload on cellular networks. As promising ways to address this problem, the approaches based on cooperation between geographically close users and offloading techniques are becoming a turning point for wireless network development.

The User Equipment (UE), Radio Access Networks (RAN), and Core Networks (CN) may benefit from observable social aspects and features able to support the management of content caching, redundant content requests, and match-making systems, among others. The Content-Centric paradigm and the direct communication between devices, as foreseen in the D2D model, may provide significant advantages over the traditional model of communication, such as efficient spectrum usage, energy saving, extension of network coverage, and increased throughput.

Chen et al. [27] promoted the offload of network demand and improved the throughput through direct communication between devices that are geographically close using a cooperative approach for relay selection. In the proposed cooperative D2D model, each user has two options for relay selection: social trust or social reciprocity. In the social-trust based selection, the knowledge of preliminary social context is used to evaluate the trustworthiness of the neighbors. Thus, family members and friends are usually trustworthy enough to serve as relays, as these social links often characterize altruistic users with reliable behavior.

According to Chen et al. [27], the personal devices carried by human beings can provide knowledge of human social ties and trustworthiness that can be used to achieve effective relay assistance for cooperative D2D communications. The presented model assumes that two users have social trust toward each other when they present some social link, which can be kinship, friendship, or colleague relationship.

Furthermore, identifying social links between users is a critical task through their devices. Thus, the authors adopted a network-assisted approach, where the devices carry out the identification process exploring common social features in a matching process that may evaluate OSN profiles and contacts book information, such as address and phone number. In the opposite of social trust that requires strong social ties among users, social reciprocity is a mechanism for promoting cooperation among the users in the absence of social trust and still guaranteeing mutual benefits directly or indirectly using a coalitional game. In the direct reciprocity, only the two individuals who share resources and cooperate obtain the benefits, following the principle "I help you, and you help me." On the other hand, in the indirect reciprocity, a third party will reward and cooperate, following a concept of "I will help you, and someone else will help me."

Bao et al. [11] leveraged the observation that cellular networks are typically overloaded in crowded areas, such as peak hours of commutation or during high-density events. In addition, a small number of online services and content providers are extremely popular, concentrating the access and requests of a large number of clients, for this reason, many people located in small areas may request the same popular content. Looking at these scenarios, they proposed a solution exploring the possibility of reply requests from spatially clustered clients sharing content stored locally on their mobile devices, where such D2D communication can be mediated by servers, avoiding many of the known problems of pure ad hoc communication.

In the proposed approach, the cellular operator activates the *DataSpotting* service during high-demand situations, and the clients are instructed to report their locations periodically, enabling the operator to build a map to indicate clusters of clients, called data spots. The server uses the clients' position to estimate the location and radii of the data spots, and the spatial content availability. The clients are periodically updated with information about near data spots.

When a client enters a data spot, it alerts the operator of its content requests, and the operator manages a digest of the content catalog in different clients within the same data spot. Then the operator applies a match-making service and notifies the requesting client, using cellular connectivity as the control channel, to directly connect to the appropriate neighboring device through WiFi to retrieve the content. The authors argue that, in this way, the knowledge of cached content in the clients does not spread among other devices, avoiding privacy concerns.

Furthermore, cell-tower mediated transfers are amenable to accounting and the clients may have the choice to intentionally approach a data spot with interesting content. The author investigated the existence of *DataSpotting* in real scenarios and conducted a real measurement in the Manhattan area in order to understand client density dynamics, geographic distribution, and the typical contact duration for pedestrian users.

Shafiq et al. [112] presented a characterization of the spatiotemporal dynamics of the workload in cellular networks, investigating real traffic datasets of a mobile operator in a wide metropolitan area of the United States. The analysis combines data collected from RAN, capable of providing customer geolocation records according to the cell used by the mobile device, as well as records from the CN that contain flow-level IP traffic information originating from or directed to customers. The analysis presents characteristics of traffic volume organized in four categories: byte volume, number of flows, packets, and users, indexed temporally and spatially.

Despite the thousands of observed applications in the datasets, only 100 more popular applications correspond to more than 95% of the volume of transferred bytes. The results of the temporal characterization of aggregate traffic, that is, the compound of traffic from all observed applications, indicated strong diurnal behavior and two daily peaks in the four traffic categories analyzed. However, the same traffic presents different temporal dynamics when classified by applications. For instance, social media applications presented a growing volume of concurrent users and bytes transferred throughout the day beginning at 8 a.m.; however, after the single peak at 3 p.m., the user volume decreases linearly, whereas the volume of bytes transferred continues to grow for another eight hours.

The geospatial analysis has shown that the geographical distribution of application usage is skewed, such that the customers' interests are heterogeneously distributed and observable by the types of applications used. According to the results, *web browsing* is the most ubiquitous category, where 80% of traffic is originated in only 50% of all monitored cells, and all the demand related to *dating* applications originates in less than 5% of cells.

In this way, geographic areas and neighborhoods present different characteristics according to traffic volumes; however, two or more disjoint regions of the same category, such as suburbs, university regions, and downtown, still show significant similarity to the same characteristics. Such evidence shows that user interests in cellular data networks are highly dependent on human behavior and social aspects, such that the spatiotemporal characteristics of user groups, in particular locations and periods of the day, present great potential in favor of network optimization.

Wang et al. [130] investigated the file-sharing dynamics through the Xender application. The application transfers files independently of cellular networks using WiFi tethering. According to the authors, this is one of the reasons that make the application popular in countries with underdeveloped economy.

The study shows the analyses of approximately 5 million users and 90 million file shares on the platform, where most of the shares are multimedia files, especially videos. The results showed that the traffic presents significant variations according to the temporal characteristics, where the total traffic increases by up to five times on Sundays compared to the days of the week; in addition, ten to forty percent of the traffic load corresponds to redundant content.

2.3.2 Transportation and Vehicular Networks

The modernization of public and private transport systems and vehicles has given positive results in the daily lives of their users, especially in metropolitan areas. Since traffic is one of the critical issues in these environments, many efforts have been directed toward optimizing urban mobility and user experience while making use of these services [17].

Spatial and temporal aspects have been used for many decades to address these problems; however, only recently have social characteristics been added to the new approaches. Social characteristics can be observed collective transportation systems when the routine pushes random users to be together for minutes or hours multiple times throughout the days and weeks. Similarly, users of private or individual transportation can show social characteristics through their pairs of origin and destination, or by sharing a portion of the path to their respective destinations [14]. In this section, we present recent studies that have improved solutions to issues like demand prediction, content provisioning, and traffic, using the social, spatial, and temporal contexts.

Vegni and Loscri [129] described a large number of challenges and applications of social aspects related to vehicle networks, and argue that the human behavior largely impacts Vehicular Network (VANET) characteristics, from the drivers' behavior to the structure of the network. The authors focused on the development of Vehicular Social Networks (VSNs), a special category of network created by the ability of socialization among nearby vehicles.

The concept of sociability in vehicular environments emerges from the assumption that drivers or vehicles may have common interests with their neighbors denoted by the applications used and content or data consumed by these entities. According to the authors, a VSN is defined as a group of vehicles that may have common needs, preferences, or interests considering a single and shared spatiotemporal context. In particular, it is a VANET that includes Vehicle-to-Vehicle (V2V) and Vehicle-to-Infrastructure (V2I) paradigms, and additionally considers human factors such as personal preferences, regular mobility, and selfishness, among others aspects that affect vehicular connectivity and performance.

Therefore, the connectivity between two or more individuals exceeds the exchange of packages and allows the formation of social networks on-the-fly, where vehicle and their neighbors group together opportunistically based on common characteristics, goals, or binding relationships. The presented discussion emphasizes that the VSN concept is not limited to private vehicles; it also includes trucks, buses, motorcycles, and other vehicles private and public.

The authors discuss the VSN as a technology composed of two fundamental parts: (i) a physical layer represented by the vehicular ad hoc network, and (ii) a social network framework running on top of a physical vehicular network. The social framework presented by the authors characterizes the existence of VSNs considering three criteria that also determine the type of the network:

- Content-based is a VSN created from the relevance of the content available among members. The goal is to link users and their content of interest.

- Relationship-based is VSN where members are people who present common individual characteristics. It groups similar users, such as co-workers.
- Position-based is a VSN created when vehicles cross a particular zone of interest and form a network for exchange of geographically relevant information.

Ni et al. [89] examined the prediction of passenger demand in subway systems in the occurrence of social events using OSN data. The authors argue that predicting demands based on human behavior is a complex challenge when it involves non-recurrent events due to their irregular and inconsistent nature, in such a way that the particular characteristics of the investigated regions are relevant to the prediction model because they reflect the characteristics the visitors.

For this reason, the authors have developed a systematic approach to examine the activities of OSN users to assess the social characteristics surrounding subway stations during sporting events in New York City. The study presented by the authors jointly examines passenger demand, provided by local metropolitan public transport authorities, and messages published on Twitter. The results indicated a moderate correlation between the passenger flow and the activities recorded in the OSN during different events recorded during the seven-month evaluation. The proposed approach takes advantage of the metadata and latent features of topics covered by Twitter messages to enhance the generalization capability of the model.

From the results, the authors developed a prediction pipeline composed of a hashtag-based event detection algorithm and an optimization model that combines linear regression and the results of the Seasonal Autoregressive Integrated Moving Average (SARIMA). The approach proposed by the authors is capable of achieving 98.27% accuracy and recall of 87.68% for the sporting events investigated. The authors argue that user-generated content published in OSN has great value for the prediction of the social, spatial, and temporal context of the inhabitants of the city, being able to reflect the interests of users, social interactions, and characteristics of certain regions of the city, such that correlations, as presented by the authors, can be used as alternatives for predicting urban transport demand.

Uppoor et al. [127] presented the challenges of simulating network environments and argued that most of today's network solutions have been precariously evaluated. For this reason, the authors presented a synthetic dataset based on real data for a reliable simulation of these networks. The authors proposed a dataset generation process and discussed the relevance of the simulation of micro- and macroscopic features.

The dataset describes 24 hours of vehicle traffic in a 400 km2 area around the city of Koln, Germany; specifically, the dataset replicates the driver behavior and the traffic flow. The results showed that even simplistic assumptions about micro- and macroscopic dynamics could significantly affect the topological properties of the simulated network. Thus, the authors conjecture that network evaluations that neglect spatiotemporal aspects pose a high risk of skewing the evaluation or producing overoptimistic results.

Nikolaou et al. [90] investigated the cache management formed by individual caches distributed among network clients. The work introduces cache placement strategies that

take advantage of the demand in the service and the social links among the clients to manage popular content stored in the clients' devices. Specifically, the proposed mechanisms manage cached contents considering content consumption statistics and social aspects of the network clients through their social relationships as represented in the social graph. Thus, the mechanisms proactively replicate the cached content that is more likely to be requested based on the characteristics of the relationships. The main goal is proactively storing and positioning the content through replication to strategically provide it locally to other customers.

The authors simulated a Content-Centric and vehicular network scenario; simulation results showed substantial improvement in customer perceived latency at the cost of low bandwidth overhead of approximately two percent. In addition, the relative cost of proactive content replication decreases as the size of the system increases.

2.3.3 Internet of Things

IoT devices have become popular and provided valuable sources of behavioral and environmental data. Such devices are increasingly taking part in the urban scenario, promoting ubiquitous sensing of common spatial and temporal characteristics. However, in addition to wireless sensor networks and other environmental sensing technologies, researchers have explored the latent social characteristics of the devices used in IoT. In this section, we present recent works that discuss the cooperation between IoT devices using mainly social-inspired approaches.

Atzori et al. [5] argued that the main challenges related to the current IoT model are the discovery and composition of service, and attack the problem by introducing the social context proposing a Social-Internet-of-Things (SIoT) based on the social relationship between objects. The authors proposed a distributed architecture to enhance the ability of people and objects to discover, select, and use IoT services and resources, giving the traditional paradigm the necessary structure to ensure network navigability such that discovery of objects and services can be effective and scalable as in social networks.

The proposed architecture evaluates the trustworthiness between objects to leverage the interactions autonomously creating social groups, such as:

- Parental object relationship (POR): defined during the item production among similar objects, built in the same period and implemented by the manufacturer.
- Co-location and co-work object relationship (CLOR and CWOR): objects establish the relationship when they cooperate frequently to provide application services or constantly share the same geographic area, such as homes and offices.
- Ownership object relationship (OOR): this is when two or more objects are associated through the same user in a relationship of property or domain.
- Social object relationship (SOR): this is a link among objects whose association occurs due to sporadic or continuous contact between their owners who also have a social link with a significant trust.

The proposal also establishes categories for social structures that emerge from the relationship between objects through relational models observed in Sociology and Anthropology. In this direction, a set of objects can configure a social structure of *Communal Sharing* when all objects present equivalence, collectivism, and absence of any form of distinctiveness, in contrast to the structures of *Authority Ranking*, where objects present relational asymmetry in hierarchical models that establish precedence. In structures of *Equality Matching* relationships are based on reciprocity and balanced exchange, in order that the cooperation between objects is egalitarian, while in *Market Pricing* structures, relationships are defined via proportionality, where interactions are defined on a shared scale of rate and values.

Lin and Dong [74] argued that the social perspective can support the management of large sets of connected objects in IoT scenarios. They introduce the evaluation of trust among objects to improve the nodes interactions and their perceptions of uncertainty and risk during the execution of tasks cooperatively, such that the trust in the SIoT should be evaluated as a dynamic process.

The authors proposed a relational model based on fundamental elements: the agents trustor and trustee, the goal, the evaluation of trustworthiness, the action, the result, and the context. The trustor relies on its need for the trustee's action to achieve a goal. Thus, if the evaluation of trustworthiness is favorable, the trustor delegates the tasks to the trustee and has the expectation of the result. The expectation is positive if the trustee can return a favorable result and is useful in achieving the goal; otherwise, the expectation is negative, and the result is a threat against the goal. Since both the trustor and the trustee are cognitive, the evaluation of trustworthiness is mutual. Both pre-evaluate each other before the delegation of the task based on the context and past experiences. In this stage, the trustor ranks the potential trustees, while the trustees recognize malicious intents.

After the delegation and action execution, the results are used to perform the post-evaluations in the trustor and the trustee. The evaluation is based not only on the success rate but also on the gain, the damage, the cost, and the environment. The authors argue that trust is context-dependent; therefore, the trustworthiness evaluation of an agent's performed actions must consider the task type and the environment, such that a single agent may have different trust values for distinct actions and in different environments.

Nitti et al. [91] argued that the license-free ISM (Industrial, Scientific, and Medical) band experiences coexistence issues, and as a result, the upload on the Internet will then become more expensive in the next years, making it necessary to find alternatives. In scenarios with shared channels and scarcity of radio resources, the cognitive radio (CR) may represent a feasible alternative. Nevertheless, CR solutions require the design of sensing techniques for continuous monitoring of the status of the channels. The authors proposed the use of the SIoT paradigm to leverage the autonomous cooperation among objects capable of establishing social relationships to provide accurate sensing of radio resources.

Since, the CR systems should be aware of their operational and geographical environment to define policies, internal state, and operational parameters. Thus the cooperation of hundreds of devices, grouped in communities composed of users and devices, is the prin-

ciple behind the required temporal and spatial accuracy. Therefore, the synergy of the SIoT paradigm and the CR technology may address the spectrum scarcity based on the concept that the many are smarter than the few.

The authors demonstrate that the cooperation restricted to the benefited devices may not represent an accurate enough view of the channel status. Accordingly, devices not directly benefited by CR still have a significant part in the sensing procedures, in particular in crowded scenarios. In this way, a distributed approach of cooperation based on SIoT allows the secondary users to have a reliable vision of the spectrum usage, and the CR mechanism can minimize the use of overcrowded bands by allowing devices to transmit in another band, and, in addition, to decouple the functions of sensing the spectrum and transmission.

2.3.4 Urban Computing

The interdisciplinary nature of urban computing is a promising field for the exploration of the socio-spatiotemporal aspects. Many cities around the world have provided data about economic, social, environmental, welfare, and other aspects, such that numerous challenges in different fields have used this data in innovative approaches to provide efficient and intelligent applications and services to citizens and local governments. In this section, we present studies that provide services and analyses oriented to city management.

According to Castells and Himanen [23], the future of cities depends on the integration of the population and local governments to develop mechanisms that empower citizens individually and collectively by leveraging the capacity to improve their lives. The SenseCityVity [106] project is a mobile crowdsourcing platform that encourages citizens of Guanajuato City, Mexico, to investigate, document, and expose the city's problems. The goal of the platform is to operate as a civic reporting system capable of collecting the publications of geolocated multimedia content provided by users. The authors argue that this type of system promotes the engagement of the local population and public authorities and accelerates the process of solving urban problems due to the ubiquity of the citizens.

Yan et al. [137] discussed the need to study the geographical characteristics of online content consumption, especially multimedia content. According to the authors, the content providers need an understanding of spatial characteristics of consumption to manage the content distribution networks and consequently guarantee quality of experience to the user. However, current studies focus on the spatial and behavioral aspects of viewers and do not provide large-scale analyses covering wide metropolitan areas. Based on this motivation, the study investigates the spatial characteristics of video content consumption through data from a network provider in Shanghai, China, using a two-month dataset with the history of content requests directed to six major content providers.

The results showed that the popularity of content and the similarity among requisitions could be observed at different levels throughout the studied area, where the downtown region represents a spatial reference to the consumption characteristics. For instance, the concentration of video popularity in a region is proportional to its distance from the

downtown area, in such a way that the regions around downtown present more similar requests when compared to the regions in the downtown area.

The analysis also showed that the type of content exhibits popularity variations according to the regions of the city. For example, the similarity of video requests is proportional to the size of the region; however, the downtown regions present greater similarity of requests for movie content, while the suburban regions present more similar requests related to shows.

Catlett et al. [24] investigated the spatiotemporal characteristics of crimes, a major social problem in large urban areas. The authors argued that crime data can provide patterns and trends that can be used to support the development of new technologies and policies to cope effectively with crime. The authors proposed an approach based on auto-regression techniques and spatial analysis to predict trends and identify areas of risk.

Thus, the proposal is a spatiotemporal model to forecast crimes based on the distribution and spatial density of crimes and a set of crime predictors capable of estimating the number of crimes in each region of the city. Specifically, the model identifies crime hotspots, high-density crime areas identified using spatial data analysis techniques; so, for each high-risk area identified in the city, a crime prediction model is designed to estimate the amount of crime that will happen.

The experimental results showed considerable accuracy in spatiotemporal crime forecasting in the evaluation of two million real records of crimes collected from the Chicago metropolitan area. It is important to emphasize that the authors also identified different patterns that present peaks, dips, and spatial seasonality of crimes.

Ge et al. [48] argued that modeling the criminal characteristics of a city is essential to assisting police efforts and improving the quality of life of citizens. The authors also investigated Chicago's crime data and argue that the data obtained are the result of a collective effort that applies the crowded sensing paradigm or Human-as-a-Sensor.

According to the authors, the cognitive ability of contributors, users, and administrators was essential to provide this fine-grained data collection. In addition, community crime surveys have traditionally used demographic characteristics such as socioeconomic conditions, race, and poverty level. However, these methods do not reflect the difference between crime categories and temporal aspects. Therefore, the authors modeled crime records throughout the city using a three-dimensional tensor and a decomposition approach, where the three dimensions represent communities, crime categories, and time windows. In this way, the model is able to infer the crime rate for different categories, communities, and months of the year.

Smarzaro et al. [121] discussed that urban planning requires a wide variety of data and is usually collected and provided independently with very particular properties that hinder the analysis process. For this reason, the authors leverage the use of LBSN data as an auxiliary data source to provide useful information in a timely manner. The authors presented a case study where they used data from multiple LBSNs to estimate metrics used in urban planning.

The objective of the study was to simplify the process of data collection and analysis

that supports urban planning through non-governmental alternative data sources that can be used to investigate problems of deprivation, diversity, and availability of local services. The study presents an analysis of multiple data sources, including government census, the geography of the area studied, as well as data about the places and POI collected through Facebook[5], Foursquare[6], Google Places[7], and Yelp[8].

The results present evaluations of spatial inequalities in the supply and access to products and services by the population. The study also presents the limitations of the use of LBSN and emphasizes that the data are naturally skewed by being crowdsourced. According to the authors, the coverage area can be compromised due to the demographic characteristics of the users of these applications, usually young and technology-friendly.

D'Silva et al. [41] used crowdsourced data to explore the popularity of places in the city and infer users' activities. Based on this, the authors proposed a prediction framework that uses the weekly popularity of places as a signature or spatiotemporal identity and group places and similar areas to estimate the popularity of new places. The results showed that the transfer of information from the urban subarea level to a newly opened venue may reduce the error of estimates by 41%.

Daggitt et al. [34] presented growth patterns of urban areas on a global scale, where pairs of cities geographically close are more likely to share similar growth than pairs of remote cities. Intra-city analysis showed the existence of two major classes of places: cooperative places and competitive places. The former category represents places that leverage larger flows of mobility for their own region, while the latter disrupts flows to the region and nearby places. In addition, the authors have identified that the more-than-expected growth of local place density is highly localized, while the below-expected growth is diffuse.

2.4 Conclusion

Cities are complex and dynamic systems that face substantial challenges in the modern world. Therefore, understanding the fundamental aspects of this scenario represents a valuable advantage to overcoming these challenges.

Hence, the temporal aspects have been explored to understand the dynamics of urban variables. Indeed, understanding the fluctuations of variables over time is essential for the development of robust and resilient systems, where predicting demand for resources is one of the main applications. On the other hand, spatial aspects usually help to understand the particularities of the application clients, resulting in services with a high level of adaptation and capable of providing better results when adapted to these particular properties. It is important to mention that the clients of these applications can be users, places, and vehicles to name a few.

[5] https://developers.facebook.com/docs/places/web/search
[6] https://developer.foursquare.com/places
[7] https://cloud.google.com/maps-platform/places
[8] https://www.yelp.ca/developers

The joint use of these features provides a broad view of the urban landscape, especially when supported by multiple combinable data sources, as it has happened in recent years. The advent of the Internet and the popularization of personal mobile devices, such as smartphones and tablets, play important roles in leveraging user participation.

In this direction, OSN, in special with geographic location capabilities, represent a disruptive mechanism for collecting and analyzing data, providing democratization of spatiotemporal data. Furthermore, governments have taken initiatives to provide official data, once inaccessible, and foster scientific research around the data.

For these reasons, we showed in this chapter introductory concepts and applications of urban computing focused on temporal, spatial, and social aspects. We believe that modern cities will provide efficient services for transportation, energy, and security, among others, therefore, the social and spatiotemporal aspects are indispensable for modern applications.

Chapter 3

Urban Sensing Through Data Layers

Efficient sensing is a current open challenge for many services in smart cities. Accordingly, deploying sensing mechanisms capable of providing data from large geographical areas is an economically unfeasible task for today's conventional sensors. On the other side, the new generation of mobile applications has increasingly more opportunities to obtain personal and environmental data, pushing the academic community to develop new sensors and sensing techniques. Alternative forms of sensing have become popular, especially the mechanisms directed to the abundance of publicly available data, operated with heterogeneous data and supported by sophisticated analysis techniques. In this chapter, we proposed a sensing mechanism that exploits public data from online social networks to obtain information on fluctuations in the city's spatial properties and their correlation with climate characteristics. The results showed that the proposed mechanism is capable of detecting variations of the spatial characteristics through the places' popularity and the transitions between them. Also, the results exhibited a dichotomy in visitation preferences capable of highlighting critical temperature values that mark the transition between dichotomy states and quantify the difference in this phenomenon.

Section 3.1 introduces the problem of urban sensing and layer-based sensing addressed in this chapter, in addition to the proposed sensing mechanism. Section 3.2 presents the related work using data layers obtained through personal mobile devices in urban scenario. Section 3.3 presents the challenges and issues of working with multiple and heterogeneous data sources. Section 3.4 presents the formal model of sensing layers used to combine heterogeneous data layers and the layers's characterization. Section 3.5 presents the combination of data layers as a mechanism of sensing. Section 3.6 presents the implications and results of the approach, as well as the conclusion in Section 3.7.

3.1 Introduction

We have been facing the advances in microelectronics and computer networks every day. This involves popular devices such as laptops, smartphones and the design of disruptive technologies [31], such as Vehicular Networks (VANETs) [1], Wireless Sensors Networks

(WSNs) [103] and Internet of Things [99]. These advances have a substantial effect on data collection, transmission and manipulation of information, expanding the observations of an event, phenomenon or behavior with new details.

The popularization of devices with communication and sensing capabilities was a determining factor for the emergence of online services that exploit the environment around the user, specially the geographic location. This brought a disruptive branch of mobile applications for social interaction, search-and-discovery and recommendations in general, able to engage expressive sets of online users daily [144] such as Foursquare, Waze[1], Twitter and Instagram. Similar to conventional sensors capable of measuring physical quantities, these online services represent a data source for domains where there are no conventional sensors, specially about human behavior, such as mobility of individuals [141], daily routine [144], socio-spatial preferences [116, 117], and human sentiment [6].

The abundance of collected data by those services has been investigated and has become an important source of raw information about objects, places, people, phenomena, and events. However, the combination of these different data sources has been neglected [126]. The non-trivial challenge of combining heterogeneous data for useful information extraction is an important step towards a systematic representation of the environment and a valuable opportunity to use the knowledge to improve the quality of experience of many services and the quality of life in general.

A detailed view of an environment may include space, time, people, and additional variables of other domains such as weather, traffic, economy, and news. Many existing studies reduce the scope of analysis, considering only features within a single domain; however, exogenous variables have the potential to influence the dynamics of environmental attributes or even the occurrence of a specific phenomenon. Therefore, different data sources represented as data layers [119] and the appropriate combination of those layers might be able to provide a wide view required for optimization and creation of new services for mobile and pervasive applications.

Figure 3.1 illustrates the basic idea of combined data layers by means of three instances. A layer can store check-ins from Online Social Networks (OSN), traffic incidents, weather conditions, pulse rate, etc., and means spatiotemporal sensing samples representing people interests, city behavior, and individual or collective health. The data can come from sensors, as follows: users of social media, traffic sensors, weather stations, body sensors, and initiatives such as Open Data [149]. The sensing layers can be used as a source of semantic information for pervasive applications based on ubiquitous sensing and smart city services.

In short, in this chapter, we evaluated the spatiotemporal characteristics in the urban scenario and in the light of temperature variability. We jointly investigated the temperature and the users from OSN, and we found trends and correlations between weather and the collective behavior of users. In this way, we applied a layered approach where we combined data of check-ins and weather conditions to provide ubiquitous sensing for smart cities.

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

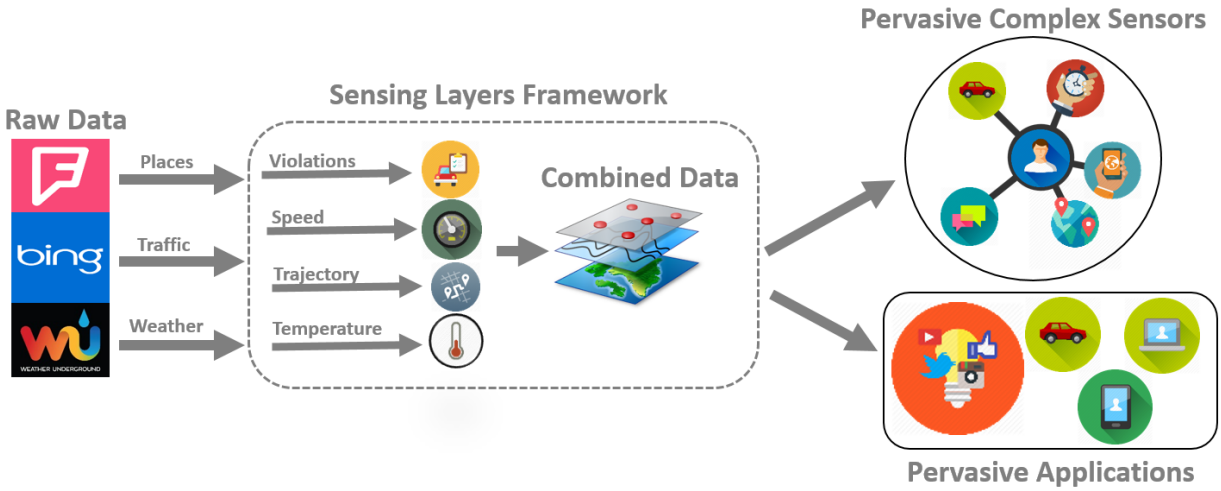


Figure 3.1: Sensing Layers build using check-ins, traffic, and weather data.

3.2 Related Work

Many studies already consider alternative data sources to measure characteristics, such as natural phenomena [108], human behavior [50], and their interactions [56]. Therefore, the state-of-art presents a considerable number of studies based on only one domain or data source. We argue that there is a large number of spatiotemporal data available, and the combination of different data sources could provide useful additional details and semantic meanings about the environment.

Silva et al. [118] presented a characterization of the Waze application from Participatory Sensing Networks (PSN) point of view. The authors presented the properties of coverage area of sensing and distribution of incidents reported according to the temporal aspects and users. Furthermore, they showed the influence of routine users in the pattern of reported incidents. Ribeiro et al. [105] proposed a congestion identification methodology using different data layers. By means of check-ins on Foursquare and Instagram, combined with traffic information collected from Bing Maps, the authors showed that check-ins and congestion were correlated. In addition, the results showed that check-ins may anticipate the traffic conditions up to 36 minutes earlier. Silva et al. [117] presented the cultural boundaries that distinguish cities, using check-ins on Location Based Social Networks (LBSN). The authors investigated habits, the commutation among places for different categories, and regional differences regarding common activities, such as eating and drinking.

Bakhshi et al. [8] constructed a conditional inference tree of social signals from 230K Yelp reviews to study how social signals shape the deviance in review rating from the mean rating. Similarly, Bakhshi et al. [6] analyzed the reviews according to climatic conditions of restaurants, demographics, and local characteristics. In these studies, the authors analyzed factors that influence human feelings and their indirectly impact on the reviews. Schuster et al. [110] highlights four areas that form the context in pervasive mobile applications: space, time, people, and information. The authors emphasize that exploring the four domains in different granularities can reveal important details to mobile services'

performance, innovation in pervasive applications, and characterization of scenarios.

Most studies that investigated the social properties considered isolated data layers but nevertheless showed the opportunity to explore pervasive applications and mobile systems as a source of valuable semantic information. The work that considered multiple information sources detected the relationships and influences of exogenous variables for specific applications. However, urban mobility requires special investigation capable of characterizing their phenomena considering the correlations and potential relationships between variables from different domains.

3.3 Issues and Challenges on Sensing Layers

Combining different data layers requires a specific treatment for each integration process. Even so, some challenges are prevalent to all applications involving multiple sources of data. For this reason, efficient approaches to integration need to be investigated to combine multidomain relationships, entities, and attributes. The understanding of the multidomain relationships potentially provides novel insights into the underlying processes or cause-effect relationships. Below, we list some of the major challenges to the studies of urban sensing based on heterogeneous data.

- Entity-Attribute Issues: characterize a particular situation to infer relationships requires a complete distinction between entities and attributes. Data models are usually heterogeneous when we combine layers of different applications; while a model presents a user as an entity with a set of attributes, another model can represent it through time series. This heterogeneity requires specific remapping and mining mechanisms [140].

- Implicit Relations: relationships between variables are a constantly studied topic, especially the implicit ones due to the nature of difficult detection. The relationships that transcend one domain are less intuitive and involve two or more attributes, entities, or a combination of both [126]. Some relationships are essential for the occurrence of some particular phenomena, and these events can be implicit if the analysis assumes a single-layer information.

- Sensing Issues: consider each data source as a sensing layer that transforms each data-publisher layer in a sensor with particular characteristics, such as coverage area and sensing frequency. Mainly in sensing spatiotemporal data, eventual data gaps caused by different sensor configurations require sophisticated mechanisms to build the appropriated integration of data sources [68].

- Imprecision and Correctness: each data sample has inherent characteristics of correctness and precision dependent on layer, sensor and measured variable. For example, layers can represent the temperature data for an entire city or comments of users on social media about an event occurring across the world. In the first case, accuracy

may be affected if the measured value does not consider variations in different neighborhoods, and, in the second case, the geographic location may indicate that users are not eyewitnesses of the event and therefore may report incorrect information. In addition, the second case may have classification issues due to the nature of the data expressed in natural language [139].

- Generalization: characterization, prediction models, and observations can be expressed according to sets of behavior classes, and summarizing a set of attributes within a single label therefore depends on abstractions of the label entities. For example, the measured value 20°C to the temperature attribute can be classified as hot in a city with a variation of 40°C over the year or normal in a city with a variation of only 15°C (15°C to 30°C). Classifying spatiotemporal data from multiple domains is a challenge for the generalization due to the particularities of the area sensed and temporal fluctuations.

3.4 Sensing Layers

In this section, we present the formal model of layer combination and the characterization of the layers of information used in this study.

3.4.1 Formal Model

The data represented by sensing layers has to come from a source that can be considered a sensor; this means that regardless of the originator entity, the data provided should be able to represent a situation or a fragment in line with the time and space windows.

Let U=$\{u_1, u_2 \ ... \ u_n\}$ a set of sensors, such as smartphones, wireless sensors, and a set of sensing systems P=$\{p_1, p_2 \ ... \ p_n\}$, such as a WSN, PSN or VANET. Each sensor $u_i \in U$ can measure one or more environmental variables, and publish the data on a specific sensing system $p_i \in P$. The j_{th} data sample stored in p_i has the form d_i^{Pi}=(t, m), where t is the timestamp when the sensor u_i reported the data, and m is a n-tuple containing the measure and the metadata, such as the covered area a, device u_i, and the measured value v.

Our research considers a set of data samples D^{P_n} as a set of $n > 0$ data samples d_i^{Pn} according to the sensing settings of $p_n \in P$, i.e. check-ins, photos, temperature measurements, traffic incidents, or comments. It should be able to representing the measures of a feature F^{P_n} spatially, temporally or both. The work plan w is the set of one or more resulting layers from the combination of two or more data sources or even other layers; therefore, it is essential to w to be efficient in the tasks of mapping data samples from different $p_n \in P$ and in query resolution. For instance, a query limited by a moment t and an area a should return $n \leq 0$ data samples from $Q(t, a) = \{d_n^{Pi} \in (D^{P_x} \cup D^{P_y})\}$.

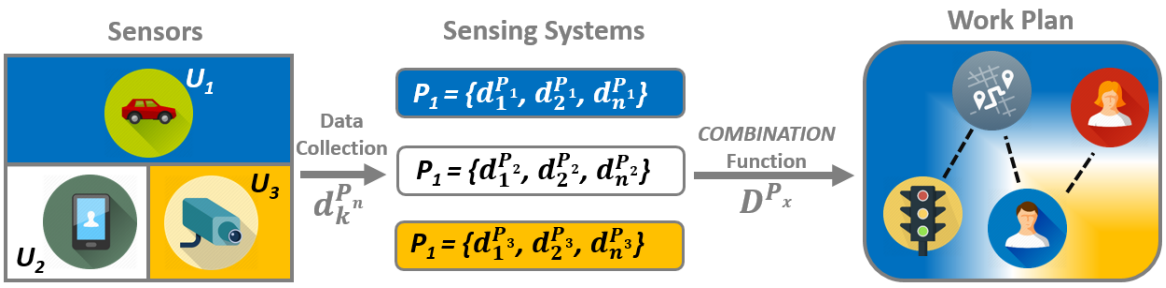


Figure 3.2: Data layer combination according to formal model of Sensing Layers.

3.4.2 Layer Characterization

Our first observations are directed to analyzing individually the temporal characteristics of the data layers used in this chapter. Concerning the OSN data, we focused on the spatial distribution of users and the popularity of Points-of-Interest (POI), due to the effect on citizens' quality of life, communications, transportation and mobility in general. As shown by Zheng et al. [146] and Moosavi and Hovestadt [85], data about these scenarios available through OSN, are urban data streams of sensing able to describe the environment in near-real time. In this way, the observations used in this work were made through public online data collected by the authors or open data initiatives.

The urban mobility involves traffic by public and private transportation, personal routines, and events. Thus, knowing the characteristics of a location and its variation over time and other characteristics is relevant, especially for detection of non-recurring or unusual situations [98]. Nevertheless, recurrent and non-recurring situations may be affected by the events of the same domain or phenomenon that transcend a single domain, for this reason. prediction systems may take advantage of multiple domains and data sources.

The datasets used in this stage are described as follows:

- Personal Check-ins: collected by authors, includes geographic location data, actively informed by the users through the Foursquare application. Based on the sensor information embedded in the user's mobile device, the application recommends places previously registered, indexed by the application and close to the user. Each data sample is a check-in that indicates the local time, latitude, longitude, the unique place identification, and the user. The dataset includes samples from 6 cities in multiple countries and continents over a period of approximately 120 days. From this point, we will refer to this dataset as the check-ins dataset.

- Weather Conditions: collected through a set of weather stations in each city studied and available online through Weather Underground. The dataset contains periodic sensing samples at intervals of up to two hours, with temperature, humidity, and precipitation information. From this point on, we will refer to this dataset as the weather dataset.

The data was collected in two separate time windows and, therefore, do not represent a continuous time series; however, the collection mechanism and methodology were identical.

The analysis was done considering the spatiotemporal intersection between the weather and check-ins datasets. The cities analyzed in this study are New York (NY), Chicago (CH), Los Angeles (LA), Paris (PR), London (LDN), and Sao Paulo (SP). The choices were made considering the importance of the city to the country, continent, tourism, representativeness in OSN, and weather characteristics.

A large number of studies showed the importance of weather conditions for many activities, from restaurant reviews [6] and traffic conditions [94] to the macro mobility of an entire country [10] and mobile data traffic [107]; therefore, it is well known that the weather is an important feature for human behavior and environmental dynamics.

Figure 3.3 presents results of temperature variation over the dataset in the Celsius scale. The values represent the temperature as a daily mean based on the measurements over the day. The period presented in the figure is determined by the intersection between the check-ins and the weather dataset.

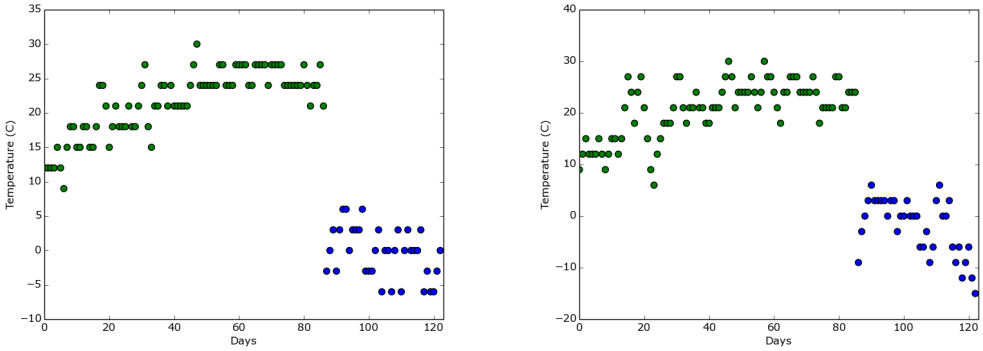
Both figures show scattered charts with green and blue points to distinguish two temporal windows of data collection: 2014 and 2015. Figures 3.3a and 3.3b show a similar temperature spectrum in the cities of NY and CH due to geographic location. However, in Figures 3.3d and 3.3c, the cities of LDN and PR showed different characteristics of temperature compared to North American cities; furthermore, SP and LA presented the smallest variation among the set.

A large number of studies investigated human behavior by considering OSN data and studying limitations and advantages of the application of those data as a source of semantic information [116, 119]. The aspects of context have been investigated by considering different scopes [110], analyzing preferences about food and drink [117], urban lifestyle [141], and the role of gender in online popularity [8]. Most of these studies analyze spatiotemporal characteristics contained in the same layer; therefore, part of our analysis for this specific layer re-evaluates events already reported in these publications, with additional information.

3.5 Layer Combination

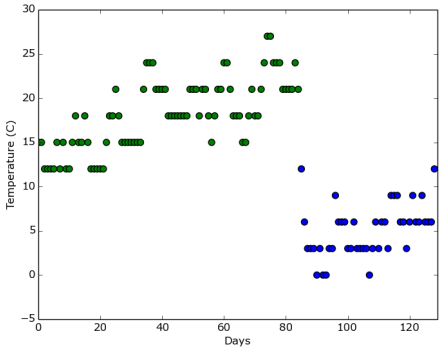
The combination of sensing layers has a potential impact on the acquisition of semantic and contextual information. The use of valuable semantic information resulting from multiple data layers can be an alternative solution to problems such as context disambiguation [63], among others. In particular, the urban scenario treated in this work was investigated considering human behavior as a determinant factor for the mobility dynamics. For this reason, we investigated the variation of visits in the most popular places in the studied cities, using the check-ins dataset.

Considering the weather data layer and its temperature sensing samples, we classified each day of the analyzed period considering the daily average based on n samples from m weather stations. Thus, we established a generalization of temperature value for temporal windows of 24 hours. The spatial window of analysis is the geographic location of a specific

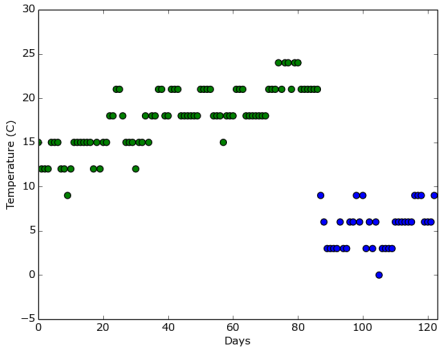


(a) New York

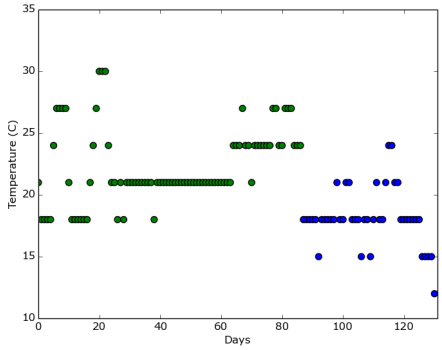
(b) Chicago



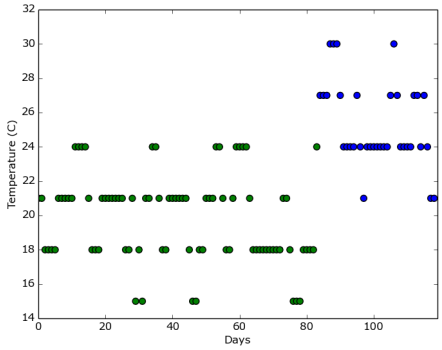
(c) Paris



(d) London



(e) Los Angeles

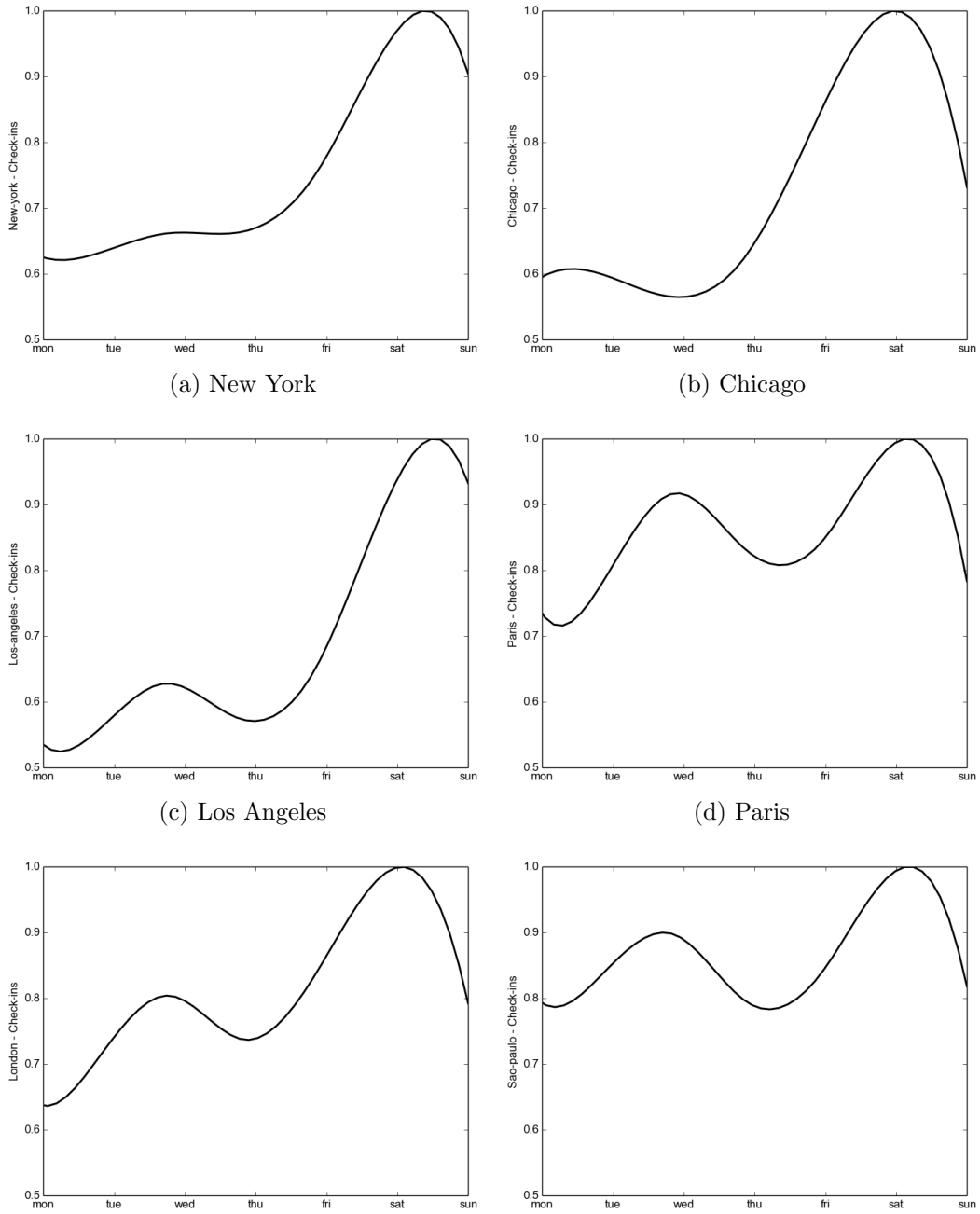


(f) Sao Paulo

Figure 3.3: Temperature history along period of the data collection.

city, built from fragments or spatial sub-windows, represented by places that received at least one check-in during the collection period.

The first combined analysis corresponds to the correlation study of spatial subwindows according to the weather variation. At this point, each mean value of temperature corresponds to a weather class where we calculate the daily average of check-ins for each spatial subwindow. Figure 3.5 shows the correlation matrices calculated using the Pearson cor-



(a) New York (b) Chicago

(c) Los Angeles (d) Paris

(e) London (f) Sao Paulo

Figure 3.4: Mean popularity of data samples according to the days of week.

relation method. Each element of the matrix corresponds to a correlation value obtained comparing the mean daily check-ins to every place in the city according to specific weather class. The objective of this analysis is to compare the entire city in front of different weather conditions through the visited places and quantify the variation of spatial distribution of users within the city.

Figures 3.5a,3.5b and 3.5c show the different results in the correlation study for the

cities of NY, CH and LA, despite the cultural similarity [117]. The NY and LA cities showed high correlation values over all classes of temperature and minimum correlation of 0.77 and 0.72 respectively. However, the two cities have different temperature spectrum, where NY presents variation of 36°C and LA 18°C. CH presented the wider spectrum of temperature and correlation with variation up to 45°C and minimum correlation 0.48. Furthermore, the results visually suggest two possible clusters between 3°C and 6°C. Figures 3.5d, 3.5e, and 3.5f present the cities of PR, LDN, and SP, and similarly to CH, PR visually suggests the presence of clusters despite the minor correlation values.

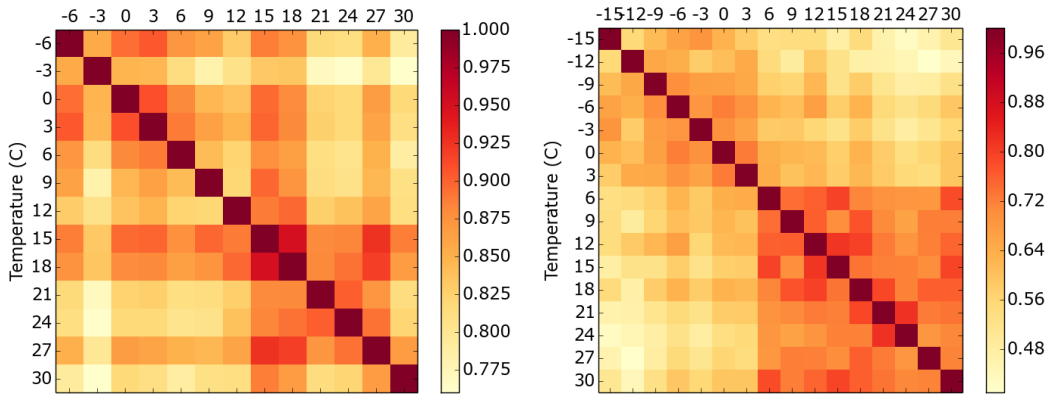
The main observation for these results is the formation of visible clusters in the correlation matrices. The clusters with high correlation values could indicate a similar distribution of check-ins among the places registered; in practical terms, it could represent that most of the people continue frequenting the same places or that the clustered weather classes have a similar subset of popular places. Weather classes that limit the borders of the clusters may represent a critical value in the distribution of check-ins and a hypothetical threshold of transition between two classes of behavior or user preferences.

SP showed high correlation values along the temperature spectrum, suggesting a low spatial sensibility to temperature variations. On the other hand, CH and PR showed a wide correlation spectrum and evident clusters of weather classes. Therefore, we formulated a phase transition hypothesis based on the clustering of weather classes and defined the phase transition thresholds, i.e., critical values of temperature, for each city, as shown in Table 3.1.The values defined for the threshold are weather classes on the borders of visually identified clusters representing an accentuated variation relatively of correlation result.

City	**Threshold τ**
New York	$12 < \tau < 15$
Chicago	$3 < \tau < 6$
Los Angeles	$18 < \tau < 21$
Paris	$9 < \tau < 12$
London	$12 < \tau < 15$
So Paulo	$18 < \tau < 21$

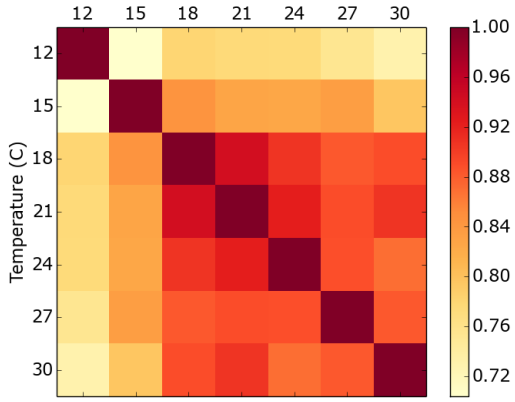
Table 3.1: Thermal thresholds of transitions between phases.

The phase transition thresholds based on correlation can be used as a foundation for measuring the temperature susceptibility for a spatiotemporal window. To evaluate our hypothesis of phase transition according to the established threshold τ, initially we consider S as the set of all places registered by check-ins in the spatiotemporal window and W as the set of weather classes. For each weather class $w \in W$, we selected a subset $s \subset S$ composed by $n\%$ of the most popular places according to the days of a temperature w. At the end of the process, there is a sequence of $|W|$ subsets of most popular places, represented as $P = \{p_1, p_2, ...p_n\}$. Thus, consider $\alpha = \{p_w \in P \mid w > \tau\}$ and $\beta = \{p_w \in P \mid w \leq \tau\}$, respectively, represent subsets of the most visited places after and before the threshold. Using this definition, we can formulate $\delta = \{p \mid p \in \alpha \cap \beta\}$ as the set of the most popular places with low susceptibility to the temperature changes. The low susceptibility discussed in a given place p_i remains among the most popular over the temperature spectrum in

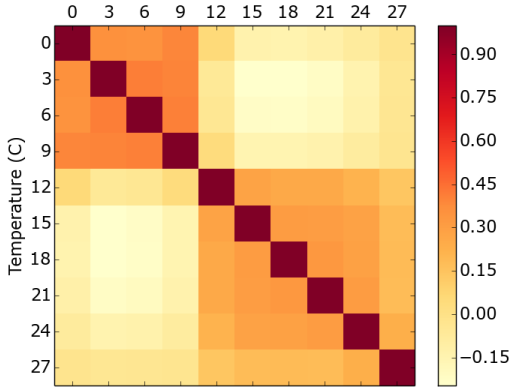


(a) New York

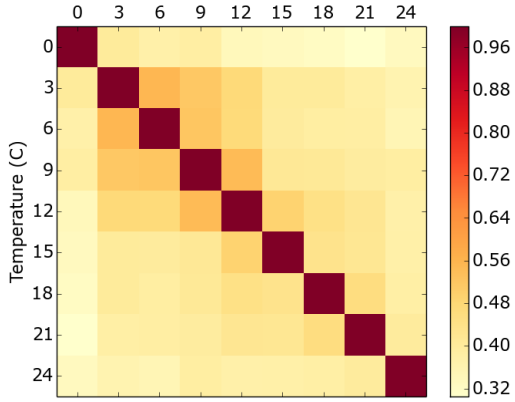
(b) Chicago



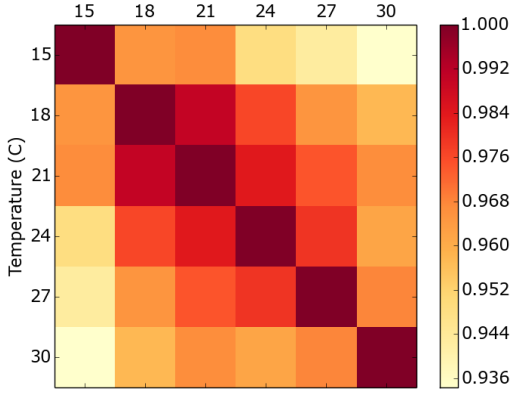
(c) Los Angeles



(d) Paris



(e) London

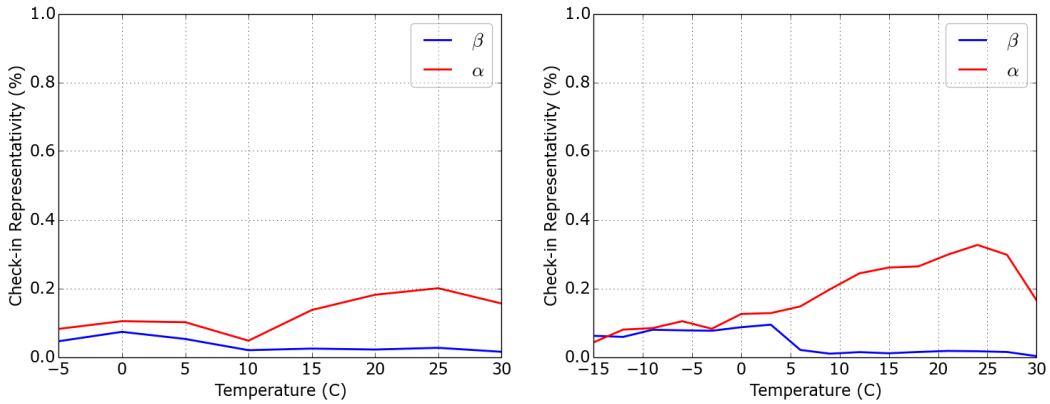


(f) Sao Paulo

Figure 3.5: Matrix correlation of local temperature and spatial distribution of users.

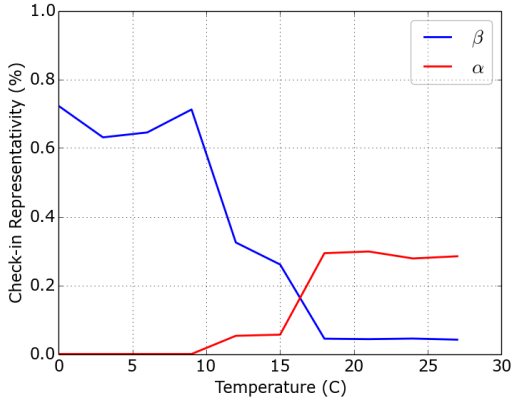
accordance with the selection $n\%$. Therefore, it does not indicate that their check-ins rate is constant along the thermal variation.

The purpose of α and β analysis is fourfold: (1) highlight the most popular places for each temperature class, (2) support the places' classification according to the temperature,

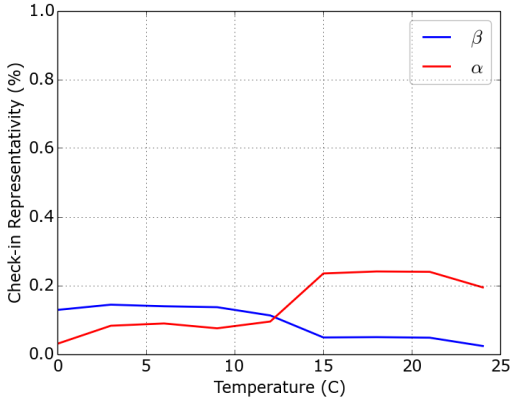


(a) New York

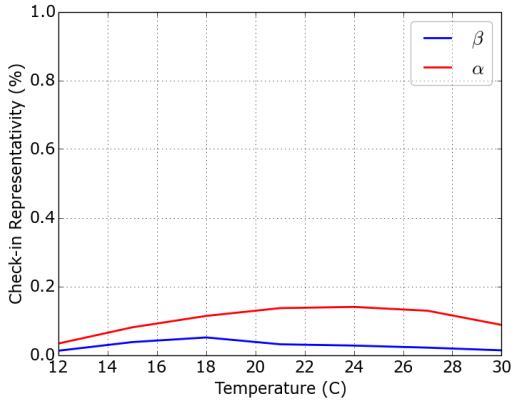
(b) Chicago



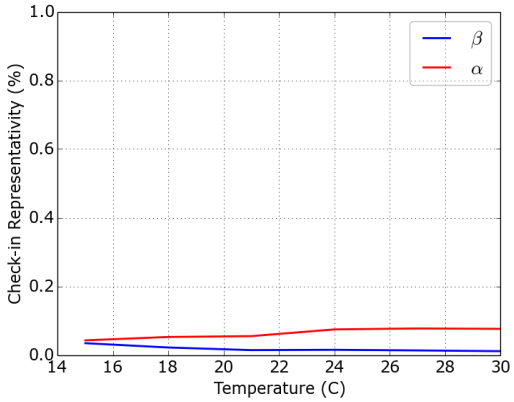
(c) Paris



(d) London



(e) Los Angeles



(f) Sao Paulo

Figure 3.6: Phase transitions of temperature $n = 50$.

(3) quantify the representativity of susceptible places in the local scenario, and (4) reveal the dynamic behavior of the most popular places along the temperature spectrum.

In this way, Figure 3.6 shows the results of α and β representativity over the temperature spectrum. These results do not consider the subset of places δ; moreover, the rep-

resentativity is calculated individually for each weather class, considering only check-ins recorded during the occurrence of the class. One of the main observations of these results is the expected behavior for LA and SP, where both showed no clear transition between phases, supporting the initial hypothesis of less susceptibility to temperature variation. On the other hand, PR and LDN cities showed behavior distinction between the subsets. Especially in PR, the β representativity of check-ins reaches 70% and α 35% approximately. NY and CH showed behavior distinction between phases at high temperatures, while lower temperatures appear with similar behavior, indicating that there was no clear preference of most popular places at lower temperatures.

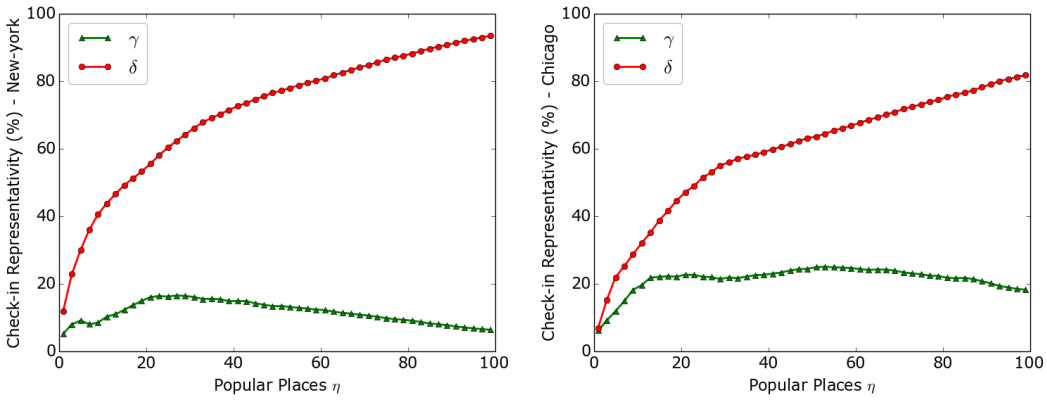
It is important to note that LDN showed clusters that were not clearly evident, although the phase evaluation showed the transition. This occurred because the correlation results evaluated the popularity vectors, the individual popularity of all the places of a city, while in the phase analysis, the popularity was represented as a single value calculated as the sum of the individual popularity of the places belonging to the same phase. For this reason, LDN was able to present low correlation values at the same time as significant representativity of check-ins, an indication that a small subset of very popular places can still be characterized by temperature.

The different behaviors detected in these results may be influenced by the particular characteristics of each city. The geographical location can assign coastal features to cities with beaches and balnearies, attracting people during the holidays and vacation or at high temperatures. Moreover, strong tourist appeal can attract people during specific periods, such as winter and summer, to activities conditioned by the season and, over the year, to non-seasonal activities .

The observations of α and β subsets are evidence of the occurrence of phase transition phenomena on the dynamics of the most popular places. However, the representativity of check-ins is calculated considering only check-ins occurring during the specific weather class; therefore, it is important to quantify this feature to the susceptible and non-susceptible places in front of the check-ins universe. The purpose of analyzing the overall representativity is to estimate the potential impact of the phase transition on city dwellers.

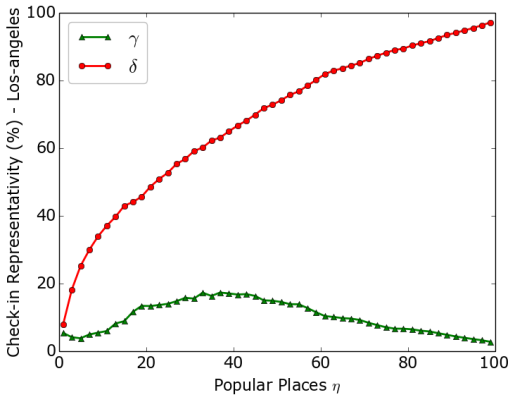
At this point, we define formally $\gamma = \{p \mid p \notin \alpha \cap \beta\}$ as the subset of most popular places susceptible to temperature variation. The curves presented in the results quantify the representativity of the subsets as a function of the value of the parameter n. Figure 3.7 presents the results of global check-in representativity of the δ and γ subsets. The main observation in these results are the different weather susceptibility profiles of the cities, where especially PR presents crescent representativity of check-ins in the set of susceptible places. There was similar behavior with values of approximately 20% in LDN and CH.

The presence of the thermal phase transition phenomenon is an important feature of the environment, and an opportunity for the optimization of services affected by mutable behavior. In the case of the dynamics of urban mobility, optimization gains can be magnified or attenuated according to the arrangement of popular places along the affected geographic area. Subsets of places grouped as susceptible can be crowded in a small area or scattered over the city, identifying a possible region of interest for a given temperature

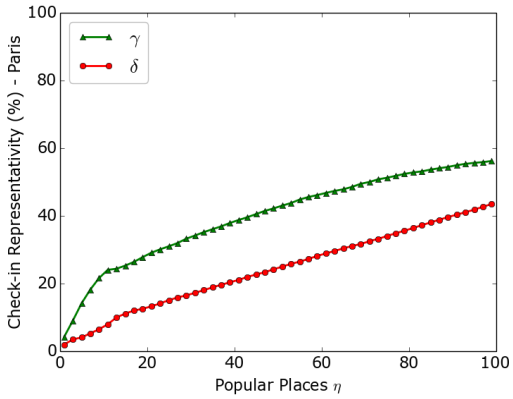


(a) New York

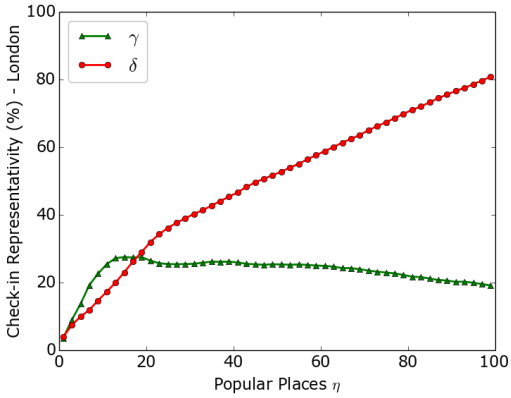
(b) Chicago



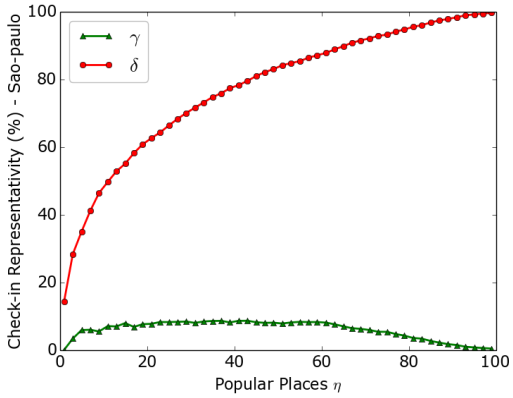
(c) Los Angeles



(d) Paris



(e) London



(f) Sao Paulo

Figure 3.7: Check-in representativity δ and γ subsets.

or range. The representation of sensitive places and their geographical disposition are essential to explore the usefulness of the phase transition phenomenon.

Figure 3.8 shows the results of the proportion of δ subset according to the universe of places for each city and $\alpha \cup \beta$. According to these results, in cities where we observed the

phase transition, as in PR and LDN, the intersection between popular places can reach up to 25% and 50% respectively, indicating a portion smaller or equal to half of the places registered as temperature-insensitive places. This particular feature could be one factor to justify the replacement behavior of the most popular places during the phase transition. In other cities where there was no substitution of most popular places, the intersection reaches values between 86% and 97%, such as in LA and SP.

At this point, it is important to emphasize that even in cities that do not clearly indicate the phase transition, as in LA and SP, at specific temperatures, the most popular places contained in γ reached more than 10% of check-ins, and for this reason, the arrangements of sensible places may still partially affect the city.

To analyze the urban mobility from the geographical point of view, we divided each city into a grid of 3x3 regions to group registered places in spatial macro-windows. The frequency of transition between two regions is computed for every two consecutive check-ins on the same day from the same user and represents a daily mean. The regions identifiers are the same for all cities; however, naturally, every city has different patterns of popularity and transition between regions according to specific geographic properties.

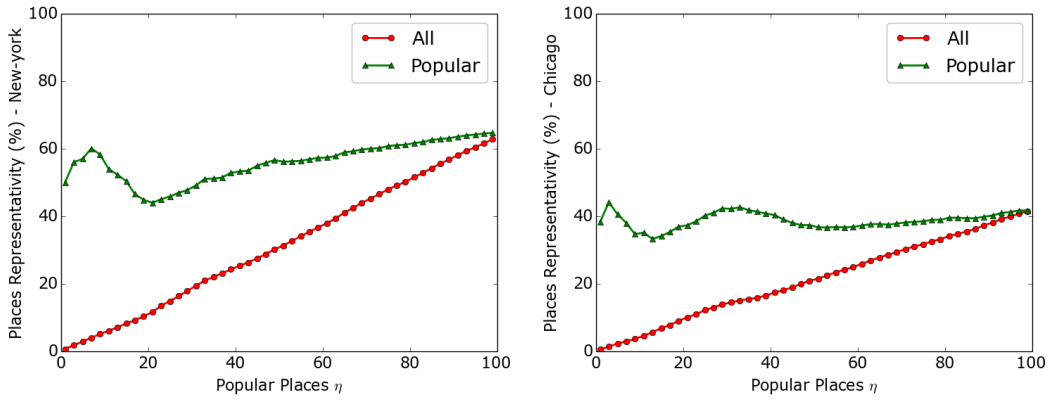
Initially we built the G_β graph grouping all the transitions that occurred in weather classes $w < \tau$, and similarly, we built the graph G_α grouping the classes $w > \tau$. Figure 3.9 shows the resulting graph of the difference $G_{diff} = G_\alpha$ - G_β. The edge width represents the result of the difference and the vertex size represents the difference of popularity registered. The colors represent the phases, where β blue, α red and gray in case of no significant difference.

The first observations in those results are the different concentrations of check-ins in the zones of cities. Gray vertices represent the most popular area in every city that remained with the highest daily average check-ins in both phases. The size of the vertices is defined by the difference in popularity between phases. Thus, even in cities where the phase transition was not identified, the phenomenon may exist on a smaller scale. Furthermore, the representation of transitions between zones shows that the transitions are more common during phase β.

Some cities exhibited a geographical partition, such as LDN and CH, which indicates distinct agglomerations of places and people according to thermal characteristics, especially CH, where the size of the vertices indicates a significant difference in popularity between the phases. Although other cities, such as LA and NY, do not exhibit clear geographic partitions, the difference in popularity of the regions is evidence of spatial preference.

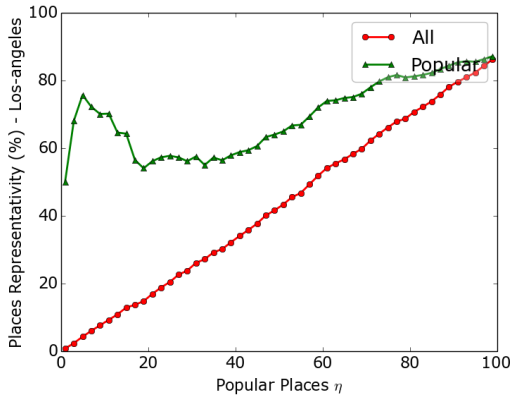
3.6 Discussion

The phenomenon of phase transition has been discussed in physics [39], biology [22, 40], and social sciences [136, 56], among other fields, due to the importance of the characteristics of phases for dynamic environments and complex organisms. In Computer Science and Urban Computing, little attention has been given to the characterization of this type of phenomenon despite its potential for significant environmental impact.

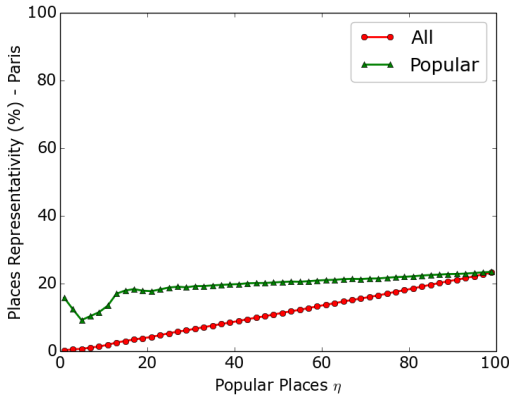


(a) New York

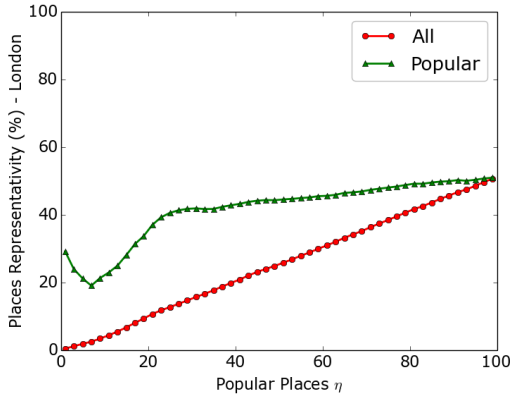
(b) Chicago



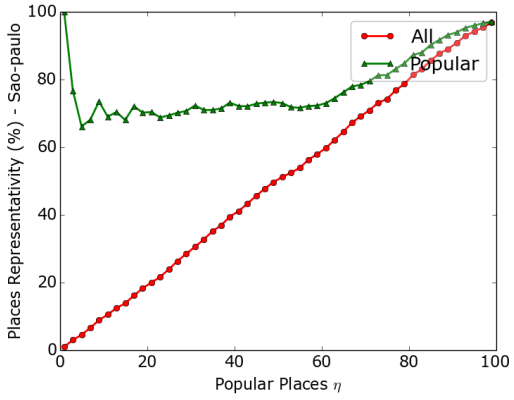
(c) Los Angeles



(d) Paris



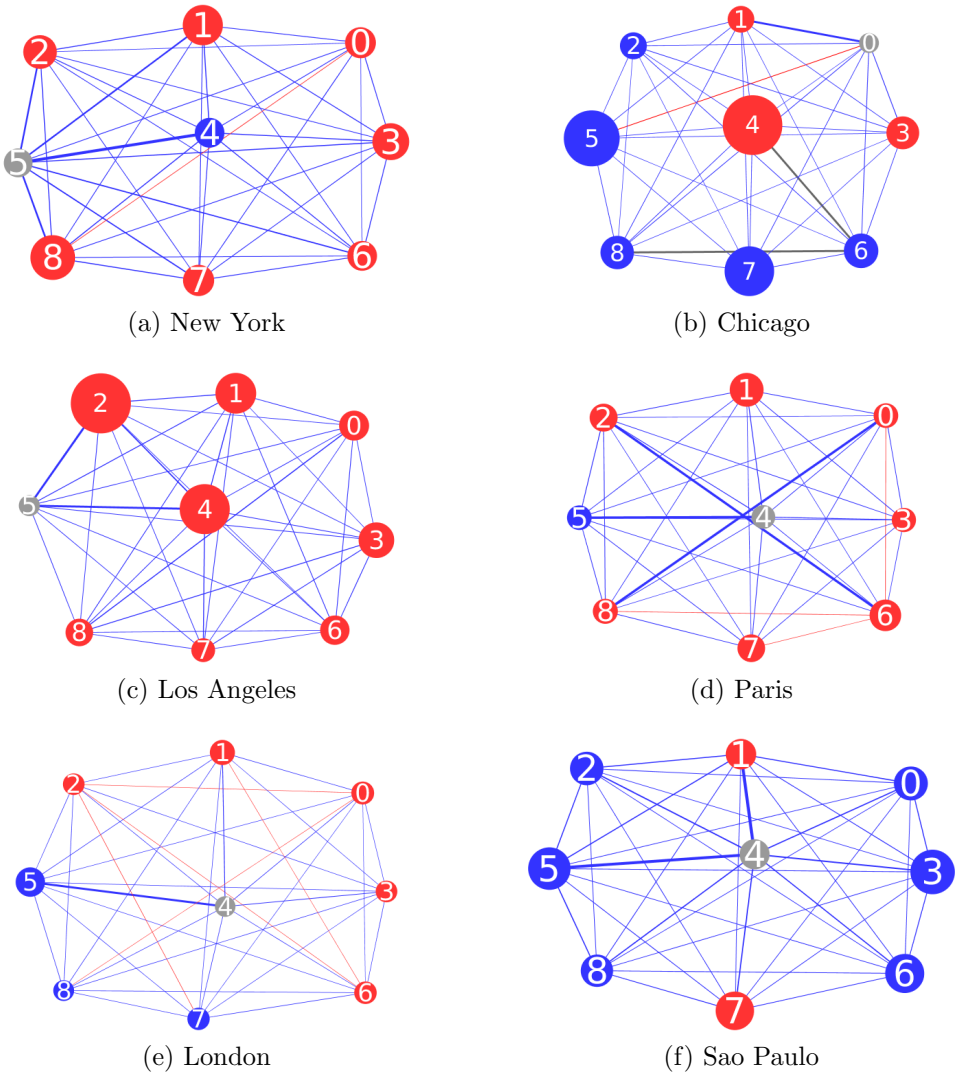
(e) London



(f) Sao Paulo

Figure 3.8: Intersection representativity of δ.

Scheffer et al. [109] characterized common generic properties of phase transitions in different dynamic systems. Although it is a difficult task to predict the critical points that define the shift of behavior, studies in different scientific fields suggest the existence of generic early-warning signals that may be indicators of a wide class of systems if a critical threshold is approaching. For example, Altman et al. [2] explored the modeling of phase



(a) New York (b) Chicago

(c) Los Angeles (d) Paris

(e) London (f) Sao Paulo

Figure 3.9: Graph of difference α - β.

transition in P2P networks. Using two models of file dissemination, epidemic and random, the authors detected the existence of phase transitions: a small change in the parameters caused a substantial change in the network behavior, influencing the online availability of files and time extinction of files.

The methodology of combining layers presented in our study was able to demonstrate characteristics that suggest a phase transition behavior. The results showed the fluctuation of collective preferences observable through the places' popularity for a subset of the cities studied. These preference changes can materialize as a shift of a portion of check-ins to a particular subset of places, configuring a spatial distribution that may result in spatial zones or clusters of interest according to the thermal characteristics.

This phenomenon can catalyze significant changes in the short term as modifications in the proximity graph as well as long-term effects in the social graph. In one of the cities

where there was no detection of the phenomenon, commutations were detected toward areas with evidence of thermal preference, suggesting a revaluation of the phenomenon in a smaller scale considering a zone or a neighborhood instead of the city.

These results also elucidate the feasibility of the sensing-layer approach. Using the framework model, we could implement a technique to combine check-ins and weather data in a data plan for the classification of places and spatial-thermal preferences and demonstrate characteristics of a phase transition. The phenomena detection should also be evaluated considering other weather variables such as precipitation, humidity, and rainfall. Cities like Los Angeles and Sao Paulo, which have a lower temperature range compared to other cities, should consider these variables, quantifying the spatial preferences in the face of these events and identify their transition parameters if they exist.

3.7 Conclusion

In this chapter, we presented the different spatiotemporal and thermal profiles of a set of cities. The results showed the detection of preference shifts in a subset of studied cities. We showed that the transition between phases may be modeled using a temperature threshold. The explicit specification of this threshold can characterize the thermal susceptibility of places and regions of the cities and support the investigation of the dynamics of spatial preferences and related studies in areas such as telecommunications, economics, and transportation.

We found distinct results concerning the phase transition phenomenon. Paris and London showed a clear shift of popularity among the set of places visited by users; Chicago and New York did not present a clear distinction of phase change; moreover, Los Angeles and Sao Paulo showed no significant distinction of behavior.

The initial correlation study of places' popularity assisted the definition of the τ value as a hypothesis of the transition threshold between phases, and the knowledge of this value is fundamental to the classification of places according to their temperature susceptibility. The previous knowledge of the transition threshold allowed us to estimate the macro mobility in the cities, allowing us to investigate the relationship of the thermal condition and the clustering of users in the regions of the city.

These results indicate, further to the thermal preferences, the potential of using alternative data sources for urban sensing. The massive popularity of Online Social Networks can leverage participatory sensing, placing citizens of smart cities as an active and significant part of urban sensing.

Chapter 4

A Long-Term Analysis of Social and Spatiotemporal Aspects

Social and spatiotemporal aspects present intrinsic correlations when investigated from the urban scenario's perspective. Knowledge about the collective behavior observed in these environments has a potential effect on the development of public services and policies. Most of the studies on these environments are restricted to temporal windows of a few days and hundreds of users due to the physical and economic limitations, and for this reason, do not exhibit dynamics observable only in long-term studies. In this chapter, we tackle this challenge using data from online social networks over fifty weeks and hundreds of thousands of users. The results showed the seasonality of the social graph estimated from the geographic proximity between users, as well as the properties of those meetings, such as interval, probability of repetition, and the spatial recurrence. Finally, we have identified spatial properties in content consumption and publishing that may leverage the benefits of caching mechanisms.

Section 4.1 presents the introduction to the topic and the questions addressed in this chapter. Section 4.2 presents an overview of the related studies about opportunistic encounters between users and the local cooperation of their personal devices. Section 4.3 describes the real data and the method used to evaluate the dynamics of the proximity graph. Section 4.4 presents the analysis of the proximity graph according to spatiotemporal features and a Device-To-Device perspective. Finally, Section 4.5 presents the conclusion.

4.1 Introduction

The growth of people agglomerating on large metropolises has been investigated to model and predict the dynamics of urban scenarios. Many researchers from different areas have paid attention to urban dynamics due to their multidisciplinary problems. The people movement represents a critical aspect, with a significant impact on essential services, such as transport and communication. People can move according to their interests, social relationships [29], routine [83], and environment factors [76], among others. In addition, events

such as climactic changes [78], concerts, and sports matches allow the citizens to concentrate in specific geographic areas. These features justify the trajectories and encounters between hundreds and thousands of people.

In the computer systems domains, these events can bring peaks of network usage, increasing the competition for cellular network resources and affecting the quality of service. To cope with these problems, cellular operators are currently relying on additional spectrum and hardware, such as mobile cell towers; however, these solutions increase the costs and may not meet the local demand in time. Meanwhile, the next generation of wireless networks and applications for smart cities envision direct communication between personal devices. The recent results of Device-to-Device (D2D) communication studies presented great potential to offload the local demand on cellular networks [72, 4].

D2D are overlay networks built opportunistically taking advantage of the spatial proximity; therefore, they may be formed of smartphones or other personal devices from mobile users during shared timely co-location situations. It means that these networks are closely related to spatiotemporal features, e.g., the day of the week, time of day, and geographic area. From a pervasive, D2D networks are affected by user preferences and their social aspects, making the investigation of social and physical proximity graphs a relevant issue to improve the performance of this class of network.

Applications and services based on spatial proximity have a disruptive potential to which industry and academia have devoted efforts for their development. The 3GPP collaboration groups are also investigating D2D communications, their feasibility, and use-cases in LTE as Proximity Services (ProSe) scenarios [73]. Importantly, most of the network traffic on personal devices goes beyond proximity-based applications, reaching social media and entertainment content, as investigated by Das et al. [35]. Popular content from Netflix[1], YouTube[2], Facebook, and other applications are downloaded millions of times daily on these devices. In addition, diffusion of content with high local relevance, such as multimedia streaming and traffic conditions may take advantage of D2D cooperation for distributed multimedia decoding, content retrieval, and delay reduction [115].

In this chapter, we investigated the proximity graph formed by users in an urban scenario. Our study evaluated features of mobility and encounters of users as well as the venues where they occur and the content shared. The study addresses the following questions:

- What are the observable dynamics of the proximity graph in an urban environment?
- How can the patterns of encounters be used to improve D2D communication?
- What are the spatiotemporal features of popular content in these scenarios?

In this direction, we used approximately one year of real data collected from social media applications to explore the real human behavior and urban dynamics in large scale.

[1] https://www.netflix.com
[2] https://www.youtube.com

4.2 Related Work

Considering the research questions defined, we addressed related works that evaluated urban spatiotemporal dynamics, especially studies that explored long time series and scenarios with great geographic density. Nevertheless, we also included other relevant studies, which have explored the D2D challenges and social aspects for network cooperation.

Asadi et al. [4] showed that D2D communication has advantages and may improve spectral efficiency and reduce communication delay. The authors argue that D2D introduces additional complexity in terms of interference control and overhead; for this reason, the design of efficient protocols remains an open research issue. The authors discuss the lack of a standard for D2D communications and the role of a central entity to manage opportunistic cooperation in cellular networks. According to them, cell towers orchestrating D2D serve as the fundamental difference between D2D and usual Mobile Ad hoc Networks (MANET). Moreover, the availability of a supervisor central entity, as in scenarios of orchestrated D2D, resolves many of the existing challenges of MANET.

Bao et al. [11] used real experiments to estimate the potential for offloading in cellular networks via D2D content transfer. The authors investigated how cellular networks are most overloaded during high-density events when many people are located in a small area, and these scenarios also present a high potential of redundant consumption of contentin other words, large groups of co-located users interacting with the same online services at the same time. According to the authors' vision, the cellular operators can track the location of personal devices and build maps that indicate dense clusters of users, appointed by the authors as data spots. They conducted experiments by wardriving (using bike rides) in the Manhattan area with Bluetooth scanning and GPS logging. The experiments included simulations of content distribution according to data measured, and the results indicated an improvement of performance for Video-On-Demand and Publish-Subscribe applications.

Wang et al. [133] proposed a framework for traffic offloading supported by social media applications in opportunistic social networks. The goal is to offload social media applications traffic by user-to-user sharing. The framework explores the proximity graphs and social graphs by means of users' profiles on social media applications. The experiments used trace-based simulations and demonstrated that the proposal could reduce up to 86.5% of the cellular traffic and satisfy the access delay requirements of users.

Chen et al. [27] exploited D2D communication in a pervasive approach with a strong social focus. The authors investigated the social ties in human social networks to enhance the cooperation of personal devices. The authors showed two fundamental social phenomena, namely, social trust and social reciprocity. Their work presented a coalitional game-theoretic framework for social-tie-based cooperation strategies and a network-assisted relay selection mechanism. The results of the trace-based simulations showed up to 122% performance gain over the cases without D2D cooperation.

Le et al. [71] proposed a content retrieval scheme for Disruption Tolerant Networks (DTNs) to support cooperative caching. The cooperative caching scheme, based on the social relationship of nodes, explores cached data on nodes with preeminent social levels,

adapts to unstable network topologies in DTNs, and enables sharing and coordination of cached data among multiple nodes to reduce data access latency.

Many recent studies have explored the potential of social aspects in respect of D2D communication improvements. However, the state-of-the-art remains open about the human interactions observed in long-term studies. Our study enlarges the usually simulated scenarios to cover an entire city for an extended period, investigating real spatiotemporal data and content sharing. These characteristics are essential to evaluate the potential of cooperation between nodes. The spatiotemporal characterization of content may identify peaks of consumption, opportunities of cooperation, data spots, and the feasibility of D2D for offloading and estimate the potential for power saving and spectrum usage in challenging scenarios especially found in large metropolises.

4.3 Graph Characterization

In this section, we describe the methods used to estimate and evaluate the spatiotemporal features of contact and proximity graph. The large popularity of social media applications has supported a wide variety of studies about human behavior. Specifically, Location-Based Social Networks (LBSN), such as Twitter and Foursquare, became popular in academia due to their real-time streams of public data capable of mapping people to venues by means of status, check-ins, and photos shared online.

For spatiotemporal evaluation, we used the data from Foursquare and Instagram applications due to their high popularity and acceptance among people, institutions, and venues. The data collected provides real geographic coordinates of users' spatial distribution in the Manhattan borough in New York City (NYC) from January to December of 2015. During the 50 weeks of collection, we observed 196 thousand unique users and obtained 1.5 million data samples.

The data samples represent spatial points on the region of interest that define the user location for a specific date and time t. Formally, we define a data sample as a 3-tuple $d_n = \langle u, p, t \rangle$, where u represents a user $u_i \in U$, t is the timestamp of the sample, and p is the u_i's position defined by latitude and longitude coordinates. According to this, we estimate the encounters between users, hence the proximity graph. An encounter event is identified by analyzing any two data samples d_i and d_j, when they satisfy the following criteria:

- $u_i \neq u_j$;
- the geographic distance between p_i and p_j is less than D_{th};
- t_i and t_j are in the same i_{th} time slot.

The distance threshold (D_{th}) is defined at 50 meters to simulate the current technologies for direct connection between personal devices, such as Bluetooth and WiFi Direct. A time slot is a period, or a time window, of observations on the proximity graph. The time slots

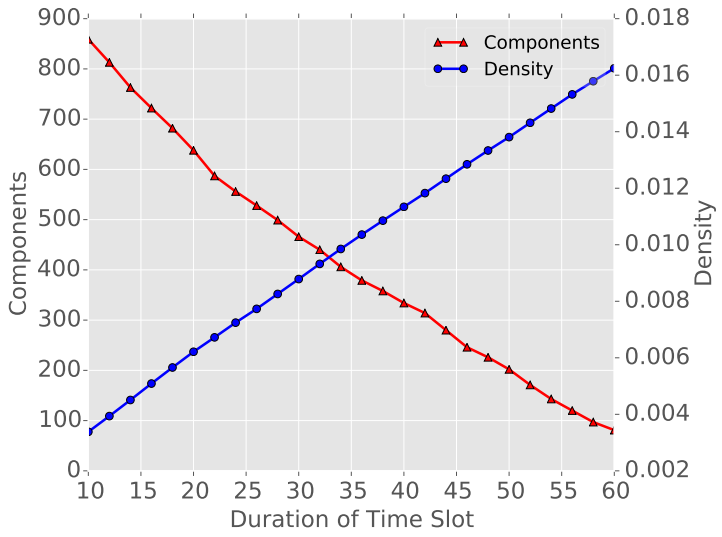


Figure 4.1: Number of components and edge density of proximity graph.

capture in a single and undirected graph $G(t) = (V, E)$ all the encounters that occur in the t time window [53], where $v \in V$ are users and $e_{ij} \in E$ the edges resulting from encounters. Thus, each new data sample potentially represents a new state of the network that is able to change its structural properties. For this reason, the proper definition of the time slot size represents a critical issue to model the dynamics of the proximity graph. To address this challenge, we analyze the number of components and the edge density properties using different time slot sizes, as shown in Figure 4.1.

The result shows that short time slots lead to a low-density graph and a fragmented network, presenting many nodes without any connections. On the other hand, larger slots capture higher density and a small number of components, which may cause over-simplification and obfuscate the accurate structure of the network. The combined analyses of metrics showed a convergence using 33 minutes as time slot size. Thus, we divide the period of observations into discrete time slots of duration $T_{th} = 33$ minutes. It is important to highlight that the proximity graphs considered only nodes with at least one encounter registered on the dataset.

4.4 Spatiotemporal Analysis

This section presents the results of the spatiotemporal analysis of encounters between users. The results show the proximity graph features and the spatial distribution of encounters across the studied area.

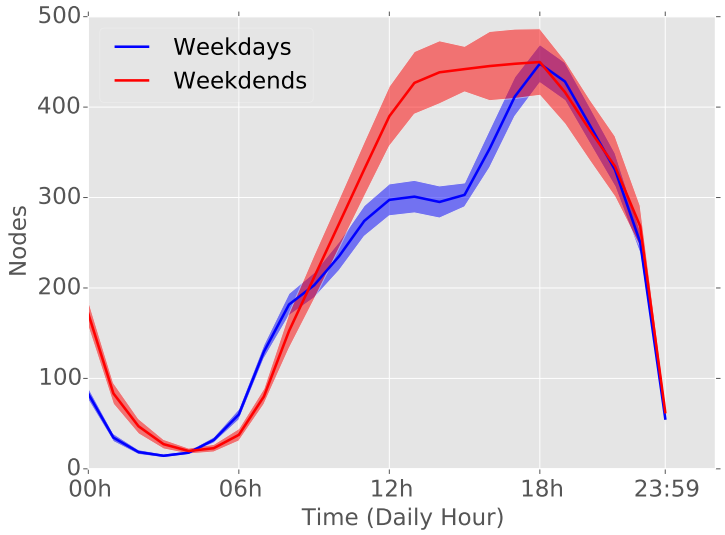


Figure 4.2: Mean of nodes during for each time slot T_{th}.

4.4.1 Proximity Graph

Figure 4.2 shows the mean of the nodes of proximity graphs during weekdays and weekends. The observations of results show the expected variation of nodes according to business hours (08 a.m. to 18 p.m. approximately). However, it is important to emphasize three observations: (1) the proximity graphs showed a higher number of nodes before 5 a.m. on weekends; (2) the number of nodes started to increase at 5 a.m.; however, during the weekends, at 9 a.m., it increased faster despite the late start; and finally, (3) after 6 p.m., both presented similar declinations.

These observations highlight the different features of weekdays and weekends. The weekends have more appeal for recreational activities, and people have more free time to use OSN applications and share their data. It can be observed in the rankings of popular places during the weekends compared to weekdays, as investigated by Bannur and Alonso [10]. In addition, the activities played and the venues visited have an essential influence on content shared on these applications.

Figure 4.3 shows the mean degree of nodes observed on proximity graphs. The observations for weekends present a slow start at 6 a.m.; meanwhile, weekday observations show rapid growth. The degree observed reflects the routine effect of weekdays, especially before 12 a.m. On the weekends, after 3 p.m., the degree of nodes exceeds that of the weekdays, which reinforces the impact of routine and the preference for other activities showing behavior less susceptible to business hours.

Figure 4.4 shows the giant component size compared to $|U|$. The results show a similar trend for both, with peaks at 12 p.m. and 6 a.m. This occurs because these hours are associated with activities related to food, leisure, and commutation in restaurants, bars, and bus terminals to name a few. Venues classified according to these features are frequently more popular, comprising a small set of places during these hours; in addition,

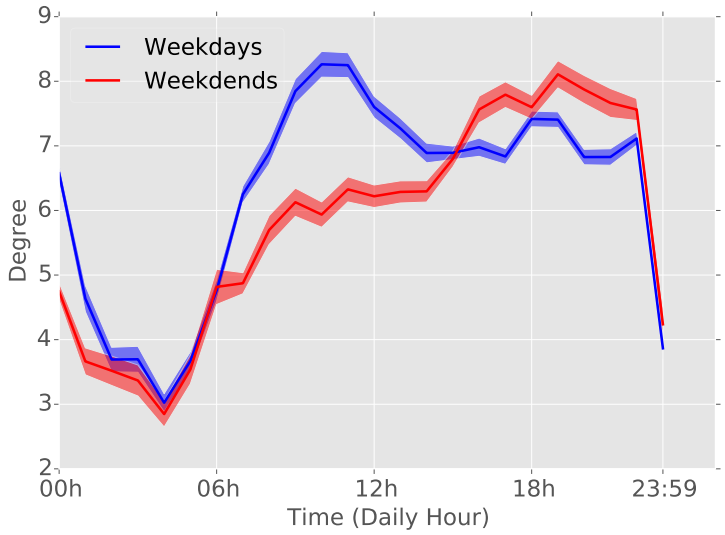


Figure 4.3: Mean degree of nodes.

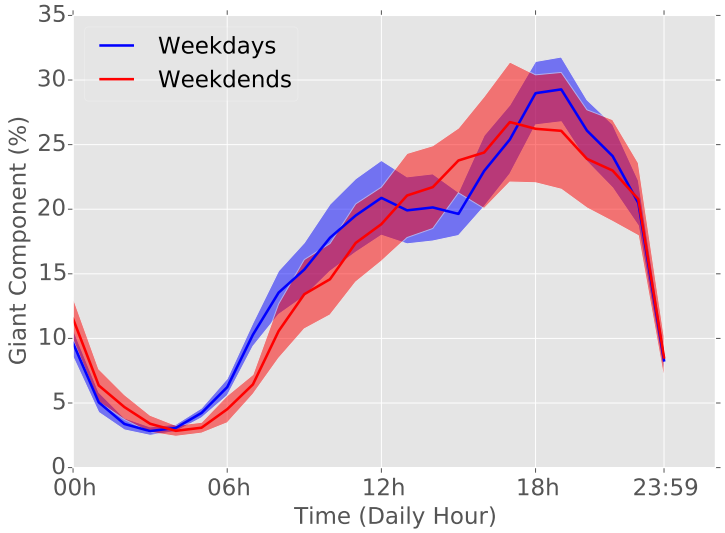


Figure 4.4: Giant component size on weekdays and weekends.

the activities carried out during these hours represent common interests for the majority of users. The higher value registered approximately 28% of node coverage, which shows that despite the T_{th} value, the graph remains fragmented most of the daily time. The number of encounters and their related properties reflect the number of data samples observed. At dawn, a reduced number of encounters occurs due to the natural sparsity of data, which portrays the low usage of OSN applications, low mobility of most users, and few Points-of-Interest (POI) at these hours.

According to our observations, the mean time of re-encounters between two random users $u, i \in U$ is 50 hours. From a periodicity point of view, most of the re-encounters occurred after 12 p.m. of weekdays and represented only 18% of all edges observed, and

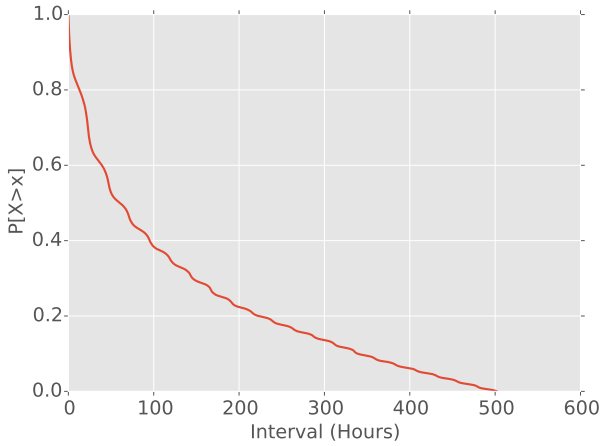


Figure 4.5: Re-encounter interval probability.

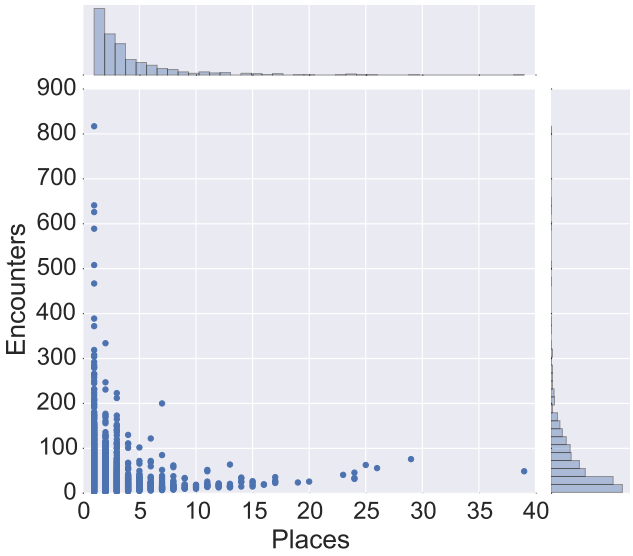


Figure 4.6: Distributions of encounters and venues.

from a spatial point of view, the re-encounters usually happened at the same venues or small sets of places. The points presented in Figure 4.6 represent the total of encounters and the total number of venues where they occurred. The combined analyses of these results indicates a strong influence of routine on encounters with more than one occurrence. In other words, the re-encounters happened in venues previously visited within a period of a few days. The high degree of nodes during business hours on weekdays reinforces these observations.

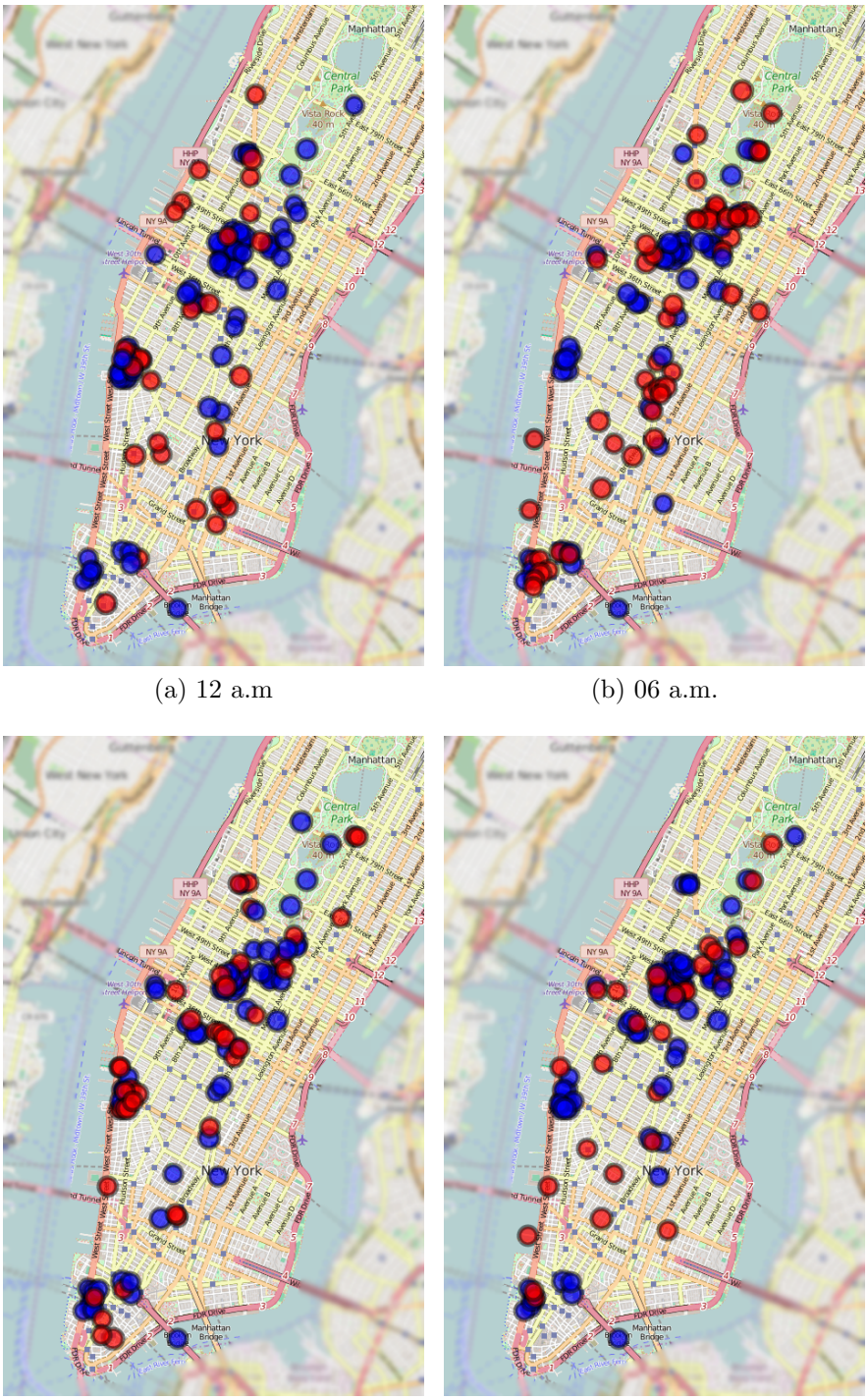
4.4.2 Spatial Distribution and Content Similarity

The encounters observed on the proximity graph are a consequence of the spatial distribution of users on the area of interest. Venues with high popularity receive more visits, and that, consequently, proportionately leads to more encounters. Figure 4.7 shows the spatial distribution of popular venues observed on the dataset. For better presentation, the figure shows only the 100 most popular POIs for selected time windows: 12 a.m., 6 a.m., 12 p.m., and 6 p.m. The circles represent the popular venues on the map, specifically, the red circles represent new POIs that have not figured between the most popular on the previous time window. For example, the red circles in Figure 4.7d are new popular POIs at 6 p.m. not observed at 12 p.m., shown on Figure 4.7cc. The popularity of POI may be influenced by their specific features, such as the class of place and their activities. However, multiple geographically close POIs may indicate a neighborhood, street, or area with particular seasonal interest, such as spots of transport commutation or shopping malls.

The results show the dynamics of venues' popularity during the day, where we can observe venues becoming more popular in specific hours. It is expected that venues like Times Square become one the most popular spots during all day; however, other spots can attract variable demand of users due to the features of venue and environment, characterizing seasonal spots.

Beyond the social links, the mutual interest in specific content also represents a relevant feature to encourage D2D communication for cooperation purposes. We used data from Twitter to also evaluate the spatiotemporal content similarity, in other words, the similarity of content published during the encounters. Twitter is an application for sharing textual and multimedia content and is more proper to this evaluation; therefore, we only use a subset of the collected data. The methodology and period of data collection are the same for all social media used in this work. The goal of the evaluation is to estimate the opportunities for cooperation between users, based on their content interests. The interests $c \in C$, where C is the content pool that includes all contents registered on the dataset, are URLs for websites, photos or videos, and hashtags explicitly defined in the body of messages shared. We previously cleaned C, removing generic hashtags or hashtags coupled with the venue, such as *#nyc*, *#selfie*, *#trip*, etc.

Figure 4.8 shows the heat map of encounters with similar content interests. The encounters are identified according to the same values of D_{th} and T_{th} defined in Section 4.3. The points represent the data samples and the heat map of the occurrences of encounters of two or more users with mutual interests. The results show spatiotemporal proximity and similarity of content across the entire area of interest. Furthermore, the results expose encounters outside of the usual POIs and crowded venues. This observation suggests opportunities for D2D collaboration for purposes beyond the network performance. It is important to emphasize the encounters registered in the central region of the city concentrate a large set of distinct content, which indicates that the cache mechanism used in this area should be able to manage efficiently massive pools of content, 72.8% larger compared to other areas on a daily mean. On the other hand, low-density areas, usually outside the center, presented a lower hit ratio and a daily mean of 43.6% less hit success, despite the usual smaller content pools.



(a) 12 a.m

(b) 06 a.m.

(c) 12 p.m.

(d) 06 p.m.

Figure 4.7: Spatial distribution of popular spots of encounters. The red circles represent spots that were not among the most popular places in the previous hour.

To generalize the findings, we investigated the interval between the publication of data samples with similar content, disregarding the T_{th}. Figure 4.9 shows the interval of samples

Figure 4.8: Spatial distribution of encounters with similar content.

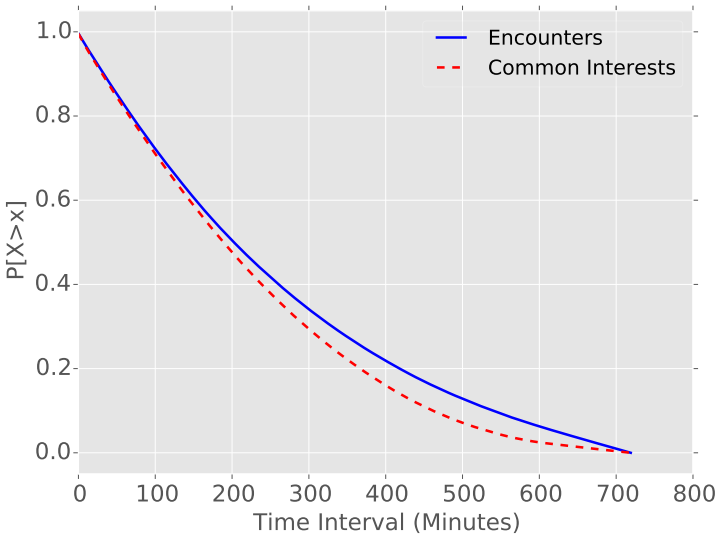


Figure 4.9: Complementary cumulative distribution of encounter interval.

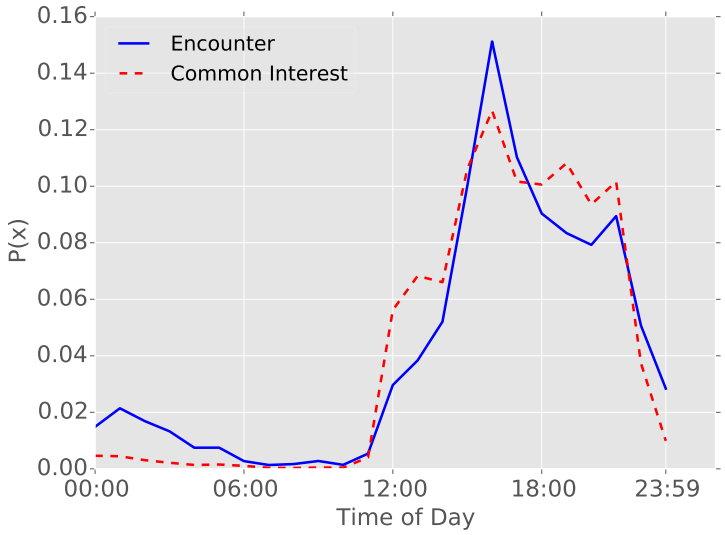


Figure 4.10: Density function of publications with similar content and spatial proximity.

with spatial proximity less than D_{th}. The results show the average interval of approximately 200 minutes, i.e., the usual interval between two publications with similar content, whose authors are geographically close. The data used for this evaluation considered only data samples with the mention of at least one $c \in C$. Figure 4.10 shows the density function of common encounters with similar content. The main observation is the trend of growing near noon, a well-known time of peak usage of OSN applications with many online interactions on these applications.

4.4.3 Discussion

The evaluation presented relevant features of social context for D2D communications, where users with multiple encounters may reveal social links and potential for common interests and cooperation. The patterns observed in New York City may be an example of predictable urban dynamics useful for the next generation of wireless technologies. The re-encounter patterns and their respective frequencies are essential features to the design of policies for D2D-based applications and the adjustment of parameters.

From the users' perspective, the insights may indicate when and where users can cooperate even without social links, only considering mutual interests. In general, the similarity of content usually is related to everyday applications, such as YouTube videos, live broadcasts of events, viral content, or common interests like traffic information and weather conditions, as observed on the data collected. Meanwhile, users with social links tend to cooperate even altruistically, donating resources to other users, motivated by friendship. In this case, users can cooperate to obtain content that represents common interests and obtain gains from a cellular operator or save energy [27]. From the network side, the D2D collaboration can support the critical moments of resource availability and provide the requested content in a low-cost way just in time.

The adoption of the D2D paradigm can foment the popularity and development of applications for social caches, crowd sensing, and public safety, among others [45]. The spatial distribution of encounters, especially with common interests, revealed that not only popular spots and spatial spots of interests could take benefits from D2D. Regions far from the town also presented spatiotemporal proximity among users and similarity of content shared. However, popular and not popular spots presented distinct relevant features that should be considered on the design cache mechanisms and cellular offloading, such as the distinct properties of the content pool and the hit ratio, considering the central and far neighborhoods of the city. In this way, the spatiotemporal features explored in this chapter elucidated the feasibility, in addition to the challenges of D2D, of proximity-based services and human-intensive mobile systems, highlighting the potential of user interaction for distributed processing, storage, and, eventually, formation of mobile clouds [38]

4.5 Conclusion

In this chapter, we used data from Online Social Network (OSN) applications to investigate the encounters between users across a year in New York City. Using real data, we simulated the encounters based on the geographic distance of data samples provided by users and estimated the proximity graph in a long-term analysis. The evaluation explored the seasonality aspects, such as daily hours of weekdays and weekends.

The results provided insights about the encounters with relevant spatiotemporal features, such that the mean interval of re-encounters usually requires a few days and occurred on small sets of venues. These observations characterized the effects of routine and its importance to regular encounters.

Together with these insights, we also investigated the similarity of content shared by users. Results showed peaks of similar content shared by users geographically close enough to take advantage of D2D communication. The evaluated data showed an average of approximately 200 minutes of interval, between two similar contents shared in a close region. These features show the potential of Device-to-Device (D2D) cooperation, the relevance of social-inspired approaches, and the potential of OSN applications as an alternative way of assessing human behavior in the spatiotemporal context.

Chapter 5

Pervasive Forwarding Mechanism for Mobile Social Networks

Modern urban environments must provide communication mechanisms able to overcome the lack of network infrastructure. Given these challenge scenarios and conditions, researchers have developed the Mobile Social Networks (MSNs), a communication alternative for delay-tolerant applications that explores human behavior, social dynamics, and opportunistic encounters among users. This networking paradigm offers unique benefits for information dissemination in urban computing applications. However, the proposed protocols for this type of network usually require sensitive information regarding users' privacy, in such a way the individual behavior is the main factor of decision-making mechanisms. In this chapter, we present an analysis of urban collective behavior, focusing on observable changes on social and spatiotemporal characteristics in a large urban scenario. Based on the observations, we proposed a message forwarding protocol for delay-tolerant MSN applications, capable of adapting to the changes in collective behavior, respecting the user's privacy, and increasing the efficiency of message delivery.

Section 5.1 presents the fundamental characteristics of wireless networks based on opportunistic meetings and lack of infrastructure, as well as the influence of human and social behavior in these scenarios. Section 5.2 presents an overview of the related work of message forwarding mechanisms for MSNs, including flooding-based and socially-aware protocols. We also present investigations of fluctuations in human behavior characterized by environmental features. Section 5.3 presents the simulation model, the data used, and the combination of weather and social features in our simulations. Section 5.4 presents the *PervasivePeopleRank* algorithm, our proposal for forwarding messages in MSN-based applications. Section 5.5 presents the simulation results, analysis, and findings of environmental effects on opportunistic social communications. Finally, Section 5.6 presents the conclusions.

5.1 Introduction

The future of computer networks comprises a large variety of applications, composed of different devices and scenarios with many particular features and challenges. Among the new technologies, opportunistic wireless networks is an emergent paradigm focused on direct communication between devices for scenarios independent of infrastructure. Both industry and academia endorse the benefits of opportunistic communication for Delay-Tolerant Networks (DTN) [135], Vehicular Networks (VANET) [1], Participatory Sensing Networks (PSN) [119], and Mobile Social Networks (MSN) [138], and also, reinforce the challenges of the area. In these scenarios, regular nodes are mobile and have limited resources; the communication occurs based on spatial proximity between peers due to friendship, routine, mobility, or simply by chance. These characteristics provide time-sensitive scenarios with frequent topology changes and lack of end-to-end paths most of the time. For this reason, traditional network protocols are neither efficient nor feasible, since they were not designed to deal with intermittent connections and network partitions.

The current ubiquity of portable wireless devices and increasing enhancement of hardware capabilities contribute to the growing interest in applications using this network class. The popularity of personal devices, such as smartphones, has led to the significant development of online services focused on user content. Location-Based Social Networks (LBSNs), such as Facebook, Instagram, and Twitter, capture a significant amount of spatiotemporal data about environments and human behavior, turning those applications into valuable repositories of social data, especially when the samples are indexed temporally and spatially. These online services capture user preferences and urban dynamics [116] and provide highly contextualized data using real-time streams of observations regarding large sets of features [85].

The large volume of records describes interactions in social media applications, people boarding public transportation systems, and cell phone calls, among others. When chronologically grouped, these observations represent the time-series of urban reality and its implicit attributes. Mining these data sources to formulate mobility models and peer encounters has become an important issue in mobile network scenarios. Moreover, insights about human behavior and its fluctuations have been shown to be relevant aspects for opportunistic networks [138]. Many proposals have studied opportunistic networks as complex systems sensitive to social and spatiotemporal aspects using real data [54, 57, 13]. They explore pervasive social context [110], such as social network contacts, personal interests, and previously visited venues, in addition to complex network metrics, such as node centrality, betweenness, assortativity, and network density.

In MSN scenarios, users are individuals carrying handheld devices with direct connection capabilities such as Bluetooth and WiFi Direct in a device-to-device manner. Due to the relevance of human behavior in MSN applications, social features have been explored to identify communities and nodes with high centrality as a critical issue for improving network performance, since social aspects usually have long-term characteristics. In this direction, many forwarding algorithms have been proposed, but only a few consider the temporal changes of these features [86]. For this reason, they are inefficient in front of

variations in user mobility and network density, which are common in urban scenarios, due to the different characteristics of days of the week and time of the day.

This variability in scenarios represents a challenge to the communication method used in opportunistic networks. The Store-Carry-Forward method requires efficient mechanisms for choosing the best nodes and the best time to forward or replicate messages, a non-trivial procedure, considering device constraints such as buffer size, energy consumption, and overhead. Usually, the proposed socially-inspired protocols select the relay nodes considering endogenous variables related to social aspects, and disregard environmental variables with potential influence on human behavior, and failing to incorporate mechanisms to adapt to fluctuations. Thus, in this chapter, we investigated the following:

- The spatiotemporal variations in urban scenarios, according to several parameters including social characteristics, the day of the week, month, and seasonal weather
- The effects of these variations on the performance of MSNs

The contributions of this chapter are threefold: first, we show that the levels of venues' popularity and their visit patterns present distinct behaviors according to seasonal and weather conditions. These findings suggest that environmental variables can support the design of pervasive protocols (spatiotemporal and socially-aware), especially in urban environments. Second, we have designed a simulation of opportunistic communications based on real data from social media applications, incorporating different settings of months, seasons, weather, and mobility in New York City.

The results present variations in network metrics capable of being characterized according to thermal conditions, which evidences the relationship between environmental variables and human mobility, and their effects on the performance of MSN protocols. Finally, we propose a message-forwarding mechanism based on environmental features and node mobility, which applies the insights gained from observing fluctuations in human behavior.

5.2 Related Work

One of the most significant challenges of communication in opportunistic networks is the design protocol for optimized routing mechanisms. The protocols require sophisticated decision mechanisms to cache, replace, and forward messages through the network, using one or more instances of them (replicas). These proposals investigate the use of the personal device capabilities of computing, sensing, communication, and data storage in order to monitor, predict, and model entities and events that exist in the physical world, such as the cyber-physical Systems [104]. Therefore, the message forwarding mechanisms should be able to select the best nodes to forward messages and improve main performance indexes, such as delivery ratio and end-to-end delay, taking into account the overhead caused by multiple replicas, hops, and energy.

The Spray-and-Wait [123] (S&W) is one of the most popular algorithms for forwarding messages, and it uses a flooding-based architecture divided into two steps. The split approach enables rapid diffusion of replicas on the network during the first step, in addition to using a customizable utility function for managing the replicas during the second step. Initially, each created message has λ replicas to spread on the network during the *spray* step. A relay node can be any node in the network that meets other nodes with $n > 1$ copies of the original message. As defined by a utility function, the relay node receives $c < n$ copies forwarded by the source or another relay node. When a node has only one replica of the message, it initiates the *wait* step. During this stage, it will not deliver the last replica until it meets the destination node.

Different mechanisms have been proposed for the *spray* and *wait* steps which extend the original algorithm, including Spray-and-Focus [124], which changes the *wait*. The new *focus* step determines that messages with one local replica will be forwarded to their destinations or other relay nodes, based on an evaluation of the time interval since the last two meetings between nodes. The main advantage of this approach is the controlled number of replicas in the network; this is defined by λ, which represents an upper bound to the overhead.

Recent studies have investigated MSNs, considering the nodes as users of personal devices such as smartphones, to take advantage of social aspects [57]. These proposals have explored social aspects, such as the node popularity [88], social group labeling [58], expected delay and the number of encounters [25], explicit mutual interests [32], and a combination of communities and node centrality [59]. In this direction, Moreira and Mendes [87] investigated the impact of human behavior on opportunistic social networks. They studied the use of social aspects and data similarity to develop opportunistic forwarding systems for essential services in extreme networking conditions and dense networking scenarios. Furthermore, their work shows suitable types of opportunistic forwarding schemes according to the network density. Their experiments used simulations based on real and synthetic mobility traces, and their findings point to the investigation of self-awareness mechanisms and adaptable forwarding schemes based on network features and the dynamism of user behavior.

Chen and Lou [25] proposed a forwarding scheme that considered information from node encounters and Time-To-Live (TTL) message property. The authors proposed a routing protocol for delay-tolerant applications that distributed multiple replicas between nodes in proportion to their expected encounter ratio. They proposed the Expected Encounter-based Routing protocol (EER) that uses the following metrics: Expected Encounter Value (EEV), which is the expected number of encounters for each node, and the Expected Meeting Delay (EMD), which is the minimum waiting time for the meeting of the current node and the destination node of a message. Similar to the Spray-and-Focus approach, the messages are created with λ replicas and spread on the network in proportion to the EEV. Thus, when the number of replicas of a held message is reduced to 1, the single replica is forwarded only to the destination node or a relay node with a lower EMD. The experiments used the vehicle-based mobility model, which is part of the Opportunistic Network Environment (ONE) simulator [65].

Mtibaa et al. [88] proposed a forwarding mechanism based on node popularity derived from the *PageRank* algorithm [19]. The *PeopleRank* proposal explores the popularity of nodes using a distributed approach, forwarding new copies of the original message to nodes ranking higher than the current node. The messages are duplicated on demand and without a specific limit of replicas. The performance evaluation presented results using six datasets of real data, with 27 up to 414 nodes.

Ciobanu et al. [30] explored the social graph from social media applications to provide additional information and support the message forwarding mechanism. The proposed algorithm, *OpportuNistic Socially-aware, and Interest-based DissEmination* (ONSIDE), takes users' interests and contact history into consideration to decrease the congestion and required bandwidth, taking into account the overall network's hit rate and the delivery latency.

Similarly, Socievole et al. [122] introduced the multi-layer social network model, which combines social networks based on proximity and online social networks. The authors investigated the relationship between different social network layers regarding node centrality, community structure, link strength, and prediction. Both works discuss the advantages of using social aspects to improve opportunistic dissemination and the benefits of using online social media applications to obtain the social graph. Nevertheless, these proposals assume an eventual connection to the Internet or remote servers of social media applications. These assumptions make it difficult to use these proposals in scenarios without infrastructure.

Environmental features can change the social and network variables used by these proposals when a contextual variable (e.g., weather, traffic conditions, the day of the year) reaches a critical value, causing changes in the variable of interest (e.g., connection duration, distance traveled, node degree, clustering coefficient). These contextual tipping points, according to the definition of Lamberson et al. [69], can represent symptoms of change in environmental characteristics. Bakhshi et al. [7] discussed how weather conditions could influence people's mood, retail sales, the stock market, among others. The authors argued that many of the effects seen in online communities could be explained using offline theories from experimental psychology. Results showed, that during visits to restaurants, user experiences varied according to weather conditions, which also influenced customers' online reviews.

Similarly, Bannur and Alonso [10] studied social media check-in data from the user's perspective, investigating seasonal polarity of check-ins in different regions of the United States. Results showed the seasonal behavior of check-ins for specific categories of venues during the 12 months of 2013 by quantifying the popularity of movies, restaurants, shopping locations, among other venues, in different days of the week, and different months. In addition to seasonal variation in visits, the results showed that ranking of the most popular venues varied during the year. Considering an urban scenario and social media traces, Cho et al. [29] showed that humans experience a combination of strong, short-range spatially and temporally periodic movement, which is not impacted by the social network structure. Their work showed that, by investigating the Brightkite and Gowalla LBSNs, social relationships could explain about 10% to 30% of all human movement, while periodic

behavior can explain 50% to 70%.

The state of the art of both topics, social-based protocols, and social media data mining classified urban scenarios as dynamic systems and pointed to the influence of social aspects and exogenous variables. Most of the performance evaluations carried out by recent studies considered real mobility traces, but the data analyzed represented only a few hundred users, small sets of communities or limited geographic areas, such as universities and conferences centers [111, 100, 42]. Moreover, existing social-based studies have implemented mechanisms based on the history of encounters regardless of their fluctuations and characteristics, which have the potential to deteriorate communication network performance. For this reason, the design of social-based forwarding mechanisms with the ability to adapt to different network configurations is a recent challenge, in which prediction of critical points of change can support the pervasive mechanisms in improving the performance in MSNs.

5.3 Trace-Based Analysis

In this section, we describe the real data used in simulations, as well as the methodology used to combine weather and social media data. Many papers have explored social media applications to simulate large urban scenarios and investigate their dynamics [54, 118, 105]. On the face of it, we reinforce the use of real data in our experiments, because environmental conditions are complex to simulate, and their effects on the behavior of users are better observed in situ [101].

5.3.1 Data Description

Many geolocalized data samples about daily life in urban environments are available through urban streams [85] and can be combined as layers of information [119]. Each geolocalized record represents an event limited by a temporal and spatial window, such as sensing samples of mobility, content interest, and venue popularity. We used public data sources in a combined approach to analyze the spatial distribution of users, and encounters between them, in different environmental configurations.

The data collected comprises geolocated data samples of weather conditions and human mobility limited to Manhattan in New York City (NYC) from February to August 2015. The traces of human mobility were built using data from social media applications, specifically, geolocalized photos on Instagram and check-ins on Foursquare, resulting in a dataset of 1.3 million samples.

By using social media applications as data sources, we obtained real data about venues, users, and encounter routines. Thus, in this work, our simulations consider commutes between real locations, a large number of users with distinct behaviors, and areas with time-sensitive agglomerations. According to the public data collected from those sources, we defined a data sample from social media as a 3-tuple $s_m = \langle u, p, t \rangle$, where u represents a user $u_i \in U$, t is the timestamp of the sample, and p is u_i's position defined by latitude

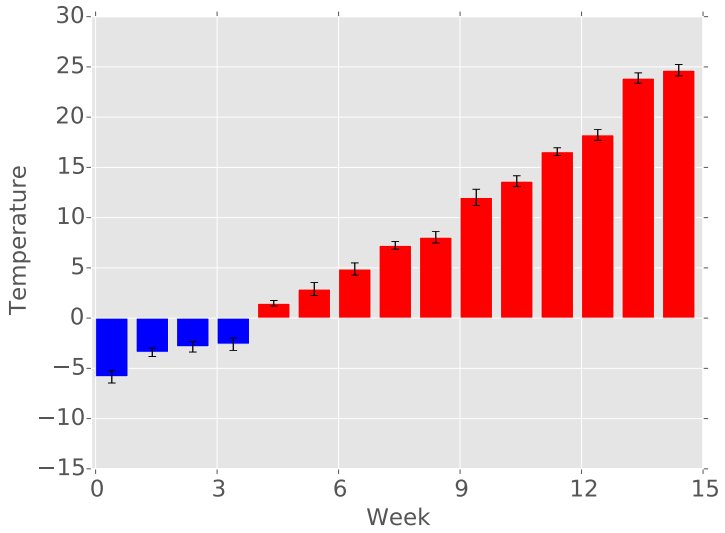


Figure 5.1: Average of temperature for the selected time series.

and longitude coordinates. In addition, we defined the path traveled by u_i within a time window as $u_i^{ts} = \{s_{m1}, s_{m2}, \ldots, s_{mk}\}$.

Data on weather conditions were collected from the National Weather Service (NWS) and public stations via the Weather Underground[1] service. The service provides data about weather variables with a sensing frequency of up to 60 minutes of interval, obtained from 54 weather stations in the area of interest. Weather data samples are defined as a 3-tuple $w = \langle t_p, p, t \rangle$, where t_p is the temperature measured in degrees Celsius, p is the position of the weather station, and t is the sample timestamp. The weather conditions of a simulated time series are summarized according to the average of all temperature measures during the selected time window and classified according to variance.

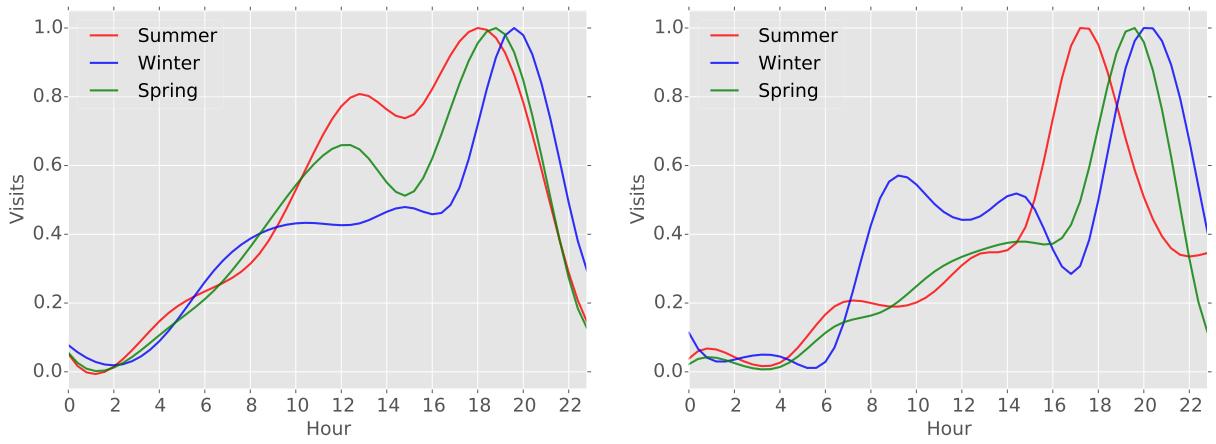
Using this model, we defined each trace as $T = \langle U_t, t_p, \Delta t \rangle$, where U_t is a set of u_i^{ts}, t_p is the average of temperature measures, and Δt is the time window of analysis. The set of traces comprises 15 independent time series grouped into seven days, starting on Monday and ending on Sunday, which are subsets of collected data and selected according to the absence of holidays and low variance of temperature. By using this methodology, we defined classes of temperature grouped by intervals of 5° Celsius, as shown in Figure 5.1.

The collected data refers to the period previously mentioned and is limited by the bounding box of Manhattan, defined by geographic coordinates[2]. The social media data samples were collected using the Twitter Stream API[3], and represent data samples obtained at the moment of its online publication, and originally published by mentioned

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

[2] The guidelines for data collection, as well as tools used and their parameters, are available on http://homepages.dcc.ufmg.br/~kassiolsm/comnet

[3] Application Programming Interface available online on https://dev.twitter.com/streaming/overview



(a) Central Park

(b) Times Square

Figure 5.2: Curves of popularity, in terms of the number of users, during the seasons observed.

applications; in other words, the samples are collected in real-time and limited to the Foursquare and Instagram applications. The weather data samples are limited according to the geographic position of the weather stations and are obtained using public API of Weather Underground, which provides queries based on geolocation and date.

5.3.2 Data Combination

Figure 5.2 shows the time series of visits for two Points-of-Interest (POI) in NYC: Central Park (CP) and Times Square (TS). The data represents the normalized average number of visits normalized by the max of individual time series during daily hours in different seasons and weather conditions. Both places present similar peaks of popularity during the night, but more than one peak occurs in the summer season, specifically at CP, where two similar peaks were registered and did not happen with the same intensity during winter and spring. The difference seen in these time series illustrates how thermal and temporal variations can characterize the visiting patterns. Note that even popular venues, which can attract crowds any day of the year (such as in well-known POIs), present fluctuations characterized by environmental variables and seasonality.

To verify whether there are significant differences in the activities done in NYC when the weather changes, we created a $m \times n$ matrix M that represents the places that people visit in NYC at different temperatures. Each row $i \in \{1, 2, \ldots, m\}$ of M is a 5°C temperature range, and each column $j \in \{1, 2, \ldots, n\}$ is the mean amount of data users in place p_j when the temperature was in the range defined by row i. Thus, Figure 5.3 shows the Principal Component Analysis (PCA) for matrix M; that is, each point in the graph is a 5°C temperature range, and the horizontal and vertical axes represent the first and second principal components of M according to PCA, respectively.

The first two components can explain 74% of the variance seen in the data. The results

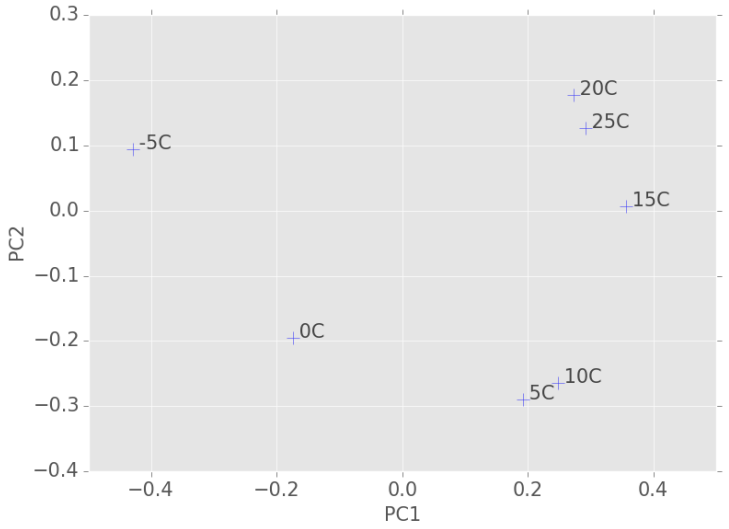


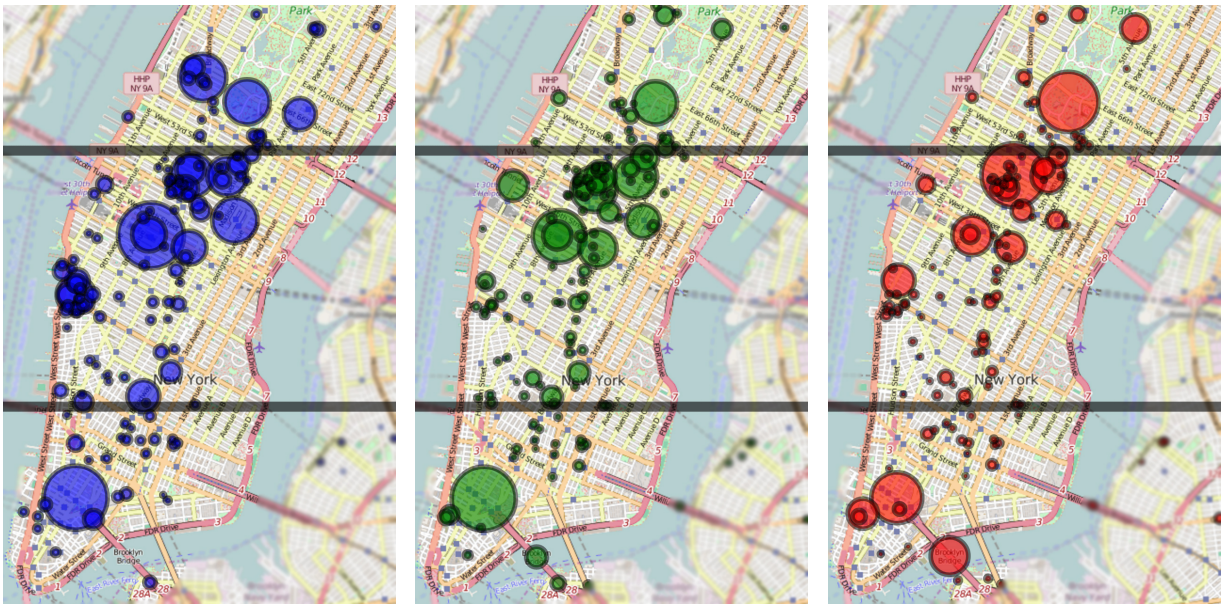
Figure 5.3: Principal component analysis of venues' popularity according to temperature.

presented distinct values for the set of temperatures observed; that is, venue popularity in NYC varies according to the local temperature. The first component, on the horizontal axis, shows the difference between cold and hot temperatures, while the second component, on the vertical axis, measures extremes temperature. Based on these observations, we modeled the popularity of venues in three phases: *negative*, *transition*, and *positive*. The *negative* phase comprises time series with average temperatures lower than 0°C; the *transition* phase includes time series with average temperatures between 0°C and 10°C; and the *positive* phase comprises time series with average temperatures greater than 10°C.

Figure 5.4 shows the analysis of geolocalized data samples according to the three phases defined in the PCA. The circles represent popular venues in the area of interest, and the size of the circle represents popularity according to the average daily number of visits (for better visualization, we maintained a limit of only 150 of the most popular venues). The results show a variation in popularity, with new venues observed only in specific phases. For example, during the phase *negative*, three POIs with similar levels of popularity close to Central Park are observable in the North, but their popularity changes during the *transition* and *positive* phases. A similar situation was registered with the Brooklyn Bridge on the South, where at least three POIs were observable in the *positive* phase.

Figure 5.5 presents the entropy matrices, grouped according to the phases defined in PCA. Each element of the matrix represents the entropy calculated using $i \in \{1, 2, \ldots, n\}$ that represents the number of data samples at a place p_i observed in intervals of two hours, and according to the days of the week. Entropy values are related to the total number of check-ins observed, where low values indicate few opportunities for encounters between users due to sparse check-ins and their spatial distribution. Hours with lower entropy values occur in periods outside regular business hours in the *transition* and *positive* phases.

Entropy begins increasing at 8 a.m. and decreasing at 0 a.m. during the weekdays, a consequence of the usual behavior of the citizens of NYC. The entropy values show



(a) Phase negative (b) Phase of transition (c) Phase positive

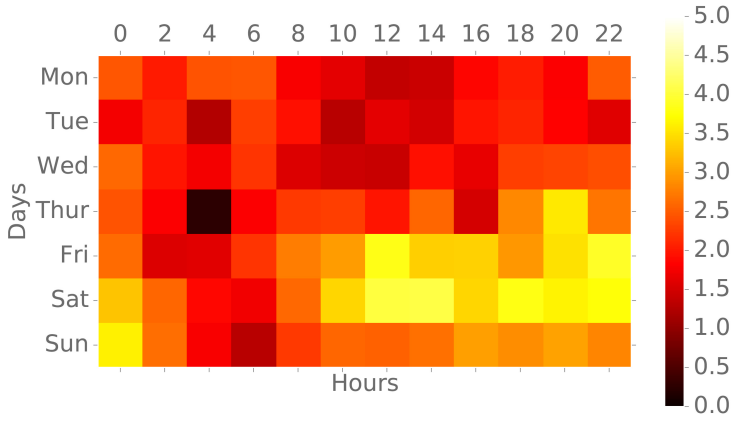
Figure 5.4: Popular venues in New York City in different phases.

critical hours; they are time windows with low mobility, capable of negatively impacting opportunistic communication performance. The phases emphasize the distinct patterns of critical hours, showing the fluctuation of spatial distribution and mutable characteristics of the critical hours set. Few users keep moving according to their particular features; therefore, forwarding mechanisms should pay attention to nodes with high mobility, or nodes with high potential to connect disjoint communities, to improving network performance in critical hours. It is important to note that several particular situations and variables can influence the spatial distribution of people, such as holidays, musical events, traffic jams, and weather conditions. In particular, weather conditions such as snow, rain, or severe temperatures can influence personal preferences and urban mobility in the form of traffic conditions, an inclination to indoor places, and increased demands on public transportation.

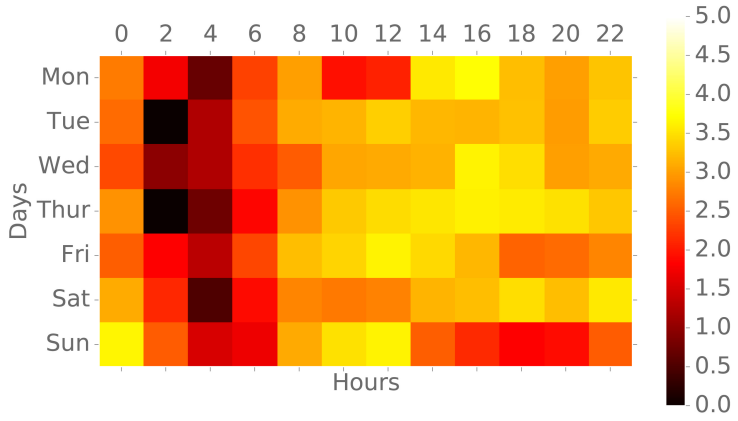
5.4 PervasivePeopleRank

In this section, we present the *PervasivePeopleRank* (PPR), an algorithm designed for forwarding messages in MSN applications, which selects relay nodes based on information about users and the environment.

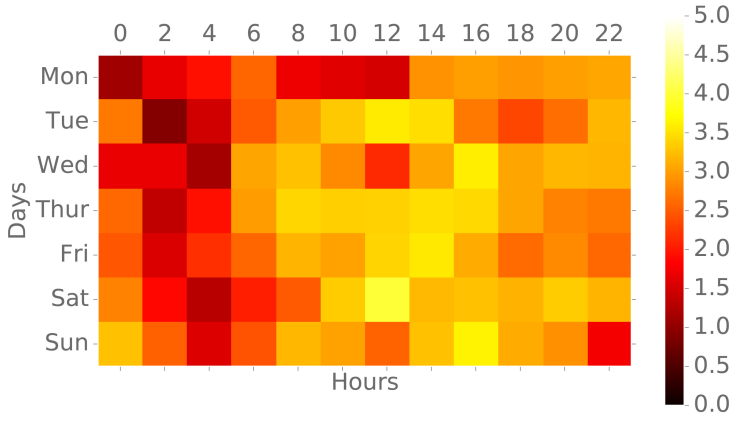
The PPR extends the previous protocol *PeopleRank* (PeR) proposed by Mtibaa et al. [88], which ranks the nodes according to their social links. When an encounter between two nodes N_i and N_j occurs, the algorithm calculates the individual PeR value using the following equation:



(a) Phase negative



(b) Phase of transition



(c) Phase positive

Figure 5.5: Entropy average of encounters grouped by phases.

$$PeR(N_i) = (1-d) + d \sum_{N_x \in F_i} \frac{PeR(N_x)}{|F_x|} \tag{5.1}$$

Equation 5.1 describes the PeR computation performed on both nodes, where F_i is the set of neighbors connecting to N_i (social links) and d is a damping factor defined as the probability, at any encounter, that the social link between nodes improves the rank of the nodes involved. The damping factor ($0 < d \leq 1$) controls the weight given to the social links on the forwarding decision. The PeR value is the metric used for replicating and sending messages towards the central nodes of the network, which have a higher probability of knowing the destination node.

Originally, the social links used in the metric are collected from social media applications. Therefore, the metric eventually requires a connection to the Internet or a server capable of providing the users' social graph. Meanwhile, we adapted the protocol to compute the social links using nearby devices close enough to connect directly. The PeR protocol represents a feasible alternative to large scenarios, with a lack of infrastructure and susceptibility to variable features. PeR provides customization of the impact of social links using the damping factor, which provides the adaptability to work in scenarios without additional resources (remote servers and Internet) and the low complexity to compute the main metric PeR in a distributed form.

The PPR considers seasonal and thermal aspects due to their effects on mobility preferences and node connectivity, taking into account the date, hour, and temperature. Algorithm 1 shows the PPR forward decision, in which nodes N_i and N_j share their PeR values and the size of their respective sets of social links. The two nodes then update their PeR values and replicate messages, if N_j has a greater PeR_j value than PeR_i or the node destination is known by N_j.

The case of $PeR_j < PeR_i$, PPR applies a time-dependent mechanism that evaluates two features:

- environmental: PPR evaluates whether the current hour is a critical hour of encounters employing the entropy matrices. In our experiments, we defined a critical hour as one that demonstrates lower entropy than the daily average.
- node mobility: the algorithm also evaluates the ΔM_i, which is the daily average of time intervals between mobility events of the node N_i.

We assume the nodes are capable of storing the entropy matrices and the social links locally. The data can be stored in key-value data structures indexed by phases, the day of the weeks, and the time of the day in the case of the matrices and by the ID of the user in case of social links. To mitigate the storage cost of social links and the impact of encounters with a single occurrence, we assume that each social link has a lifetime of τ hours. The τ defines the maximum interval between two consecutive meetings of two random users; if the encounter does not happen again before the deadline, the social link is removed. Otherwise, the deadline is renewed.

The environmental and node mobility features are evaluated to cope with hours of the low ratio of encounters. Thereby, we assume that, in addition to the capability of knowing the day and hour, all nodes are equipped with sensors or other resources for measuring temperature and mobility events. Obtaining information about time and calendar are

trivial tasks for modern personal devices. Additionally, these devices have sensors for temperature, luminosity, pedometer, accelerometer, and compass, capable of acquiring data about the environment and users' activities, such as weather conditions, walking, and cycling. Therefore, we point out that mobility events can be obtained using alternatives to Global Positioning System (GPS). Thus, the PPR does not enable forwarding based on geographic location; it mitigates privacy issues using the size of the social links set, not the identity of social links, and the time registered for mobility events, instead of the users' geographic coordinates.

Urban scenarios can provide a large number of users with different patterns of mobility. The PPR exploits this feature during critical hours, creating *ephemeral* copies of messages, a kind of replica forwarded to nodes with lower PeR and higher mobility ($\Delta M_j > \Delta M_i$). Messages flagged as *ephemeral* are forwarded normally, with TTL=min (T_m, H_n), where T_m is the original TTL of the message and H_n is the end of the critical hour.

Algorithm 1: PervasivePeopleRank Algorithm

```
1  PeR_i ← PeR(N_i);
2  PeR_j ← send(PeR_i);
3  F_j ← send(F_i);
4  PeR_i ← update(PeR_j, F_j);
5  for m ∈ buffer(i) do
6      if PeR_j ≥ PeR_i OR destination(m) ∈ F_j then
7          forward(N_j, m);
8      else
9          ΔM_j ← send(ΔM_i);
10         if critical(hour) AND ΔM_j ≤ ΔM_i then
11             forward-ephemeral(N_j, m);
```

5.5 Performance Evaluation

In this section, we present the network model used for simulating the opportunistic communications, the connectivity graph, and the network performance of the *PervasivePeopleRank* algorithm.

5.5.1 Network Model

The node mobility is determined according to the definition given in Section 5.3.1. Therefore, given two data samples s_{mi} and $s_{mj} \in T$, the settings of opportunistic communication experiments consider an encounter and network connection event between users u_i and u_j when:

- the distance $dt \leq DT_{range}$ between positions p_i and p_j;

Table 5.1: Parameters of simulation of the proximity graph.

	Parameter	Value
Network	Contact Interval (C_{time})	5, 30 and 60 minutes
	Communication Range	50 meters
	Area	25.15 x 24.01 km
	# of Nodes	$12854 \leq n \leq 18315$
	Message creation	Each data sample and random $n \leq C_{time}$ minutes
Spray and Wait (S&W)	Replicas (λ)	1000
Expected Encounter Routing (EER)	Replicas (λ)	1000
	Re-encounter Time Frame	48 hours
PeopleRank (PeR)	Damping Factor (d)	0.8
PervasivePeoplerank (PPR)	Damping Factor (d)	0.8
	Social Link Lifetime (τ)	48 hours
dLife	Re-encounter Time Frame	48 hours

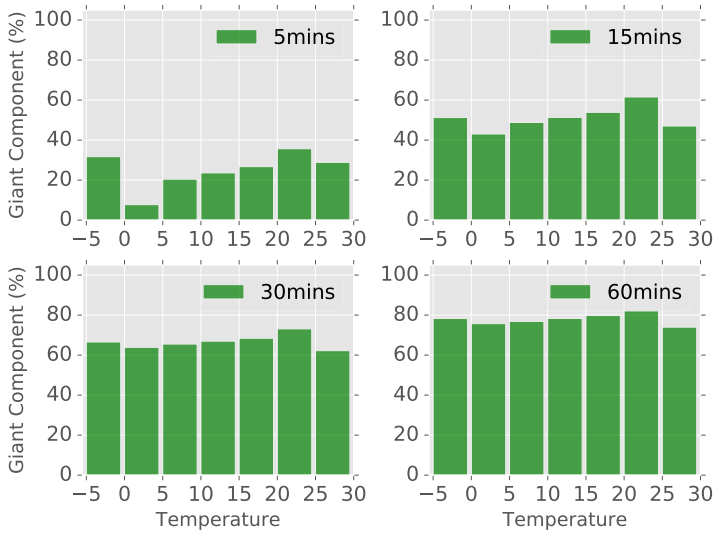
- the contact interval $c_{ij} \leq C_{time}$ between time stamps t_i and t_j;

where the DT_{range} is the distance threshold, defined as 50 meters, usually reached by Bluetooth or WiFi Direct technologies, and the interval C_{time} was experienced as a parameter that varied between five minutes and one hour. The encounters are formally described as a network contact graph $G(V, E)$, in which the stochastic process of encounter between two nodes $i, j \in V$ is modeled as an edge $e(i, j) \in E$. We assume that the network contact graph is undirected, therefore node i contacts j whenever j contacts i.

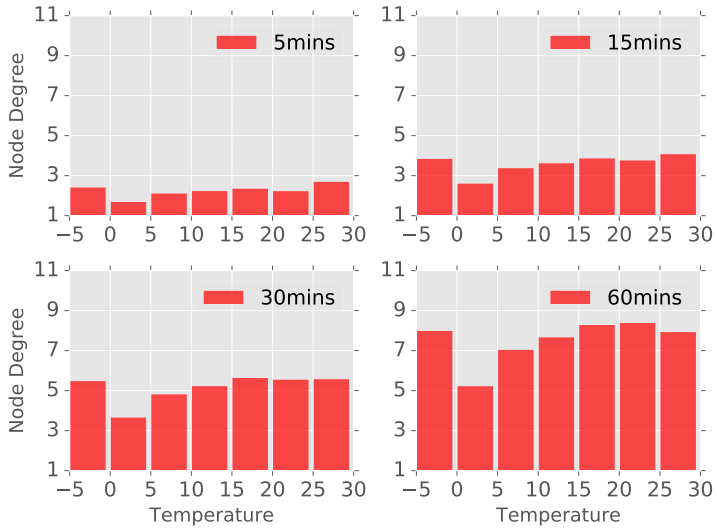
The parameters of simulations are described in Table 5.1. The fixed number of replicas used in EER and S&W simulations is enough to compare with related work, as shown in Section refsec:network. The damping factor used by PeR and PPR are defined as shown in [88, 122] to provide significant relevance to social links. The lifetime of social links defined by τ and the re-encounter time frame was defined considering time series used in simulations composed of seven days.

5.5.2 Contact Graph Analysis

The network analysis takes into account the contact graphs G_{ct} formed during the trace-based simulations, and the observed environmental temperatures. The graphs are grouped into four configurations of C_{time}. Figure 5.6a shows the size of the giant component of the contact graph for simulations of different durations and temperatures. Observe that the size of the giant component, when temperatures are inside the *transition* phase, is reduced by up to 19.1% when compared to other phases. The differences in size are noticeable in simulations with C_{time} of five and fifteen minutes, which represent 10.1% and 28.3% of all observed encounters in the dataset, respectively. Additionally, results show that C_{time} equal to fifteen minutes is enough to connect more than half of the nodes in the giant component for most scenarios.



(a) Giant component size



(b) Average degree of nodes

Figure 5.6: Analysis of the graph of contact.

Figure 5.6b shows the average degree of nodes, according to contact graphs and C_{time} configurations. The results showed that the temperature shifts from -5°C to 0°C signals the most significant changes in the network structure, where the degree of nodes decreases by an average of 32.2%. Figure 5.7 shows the Complementary Cumulative Distribution (CCDF) of the shortest path between any i and $j \in G_{ct}$ using C_{time} as 60 minutes. The changes in graph structure are characterized by the specific range of temperatures defined in the *transition* phase. The metrics showed the *positive* and *negative* phases as well connected, which provide efficient communication; however, the temperatures of the *transition* phase indicated sparse connectivity and longer paths. Thus, adaptive approaches to forwarding mechanisms are required to deal with the variations of the network structure. Also, the

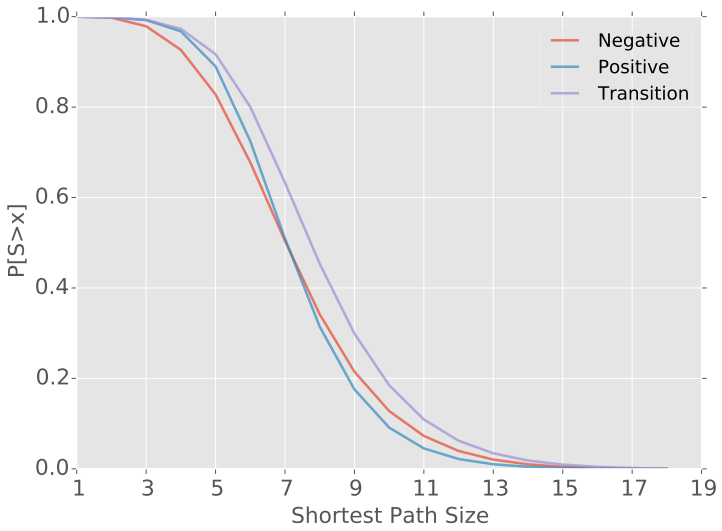


Figure 5.7: Distribution of the shortest paths.

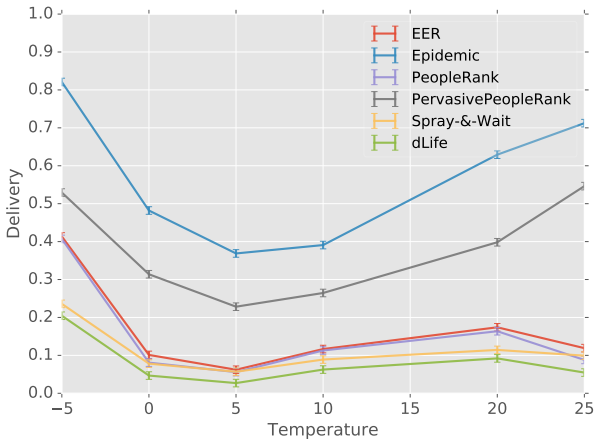
environment can characterize the changes and provide early-warning signals [109].

5.5.3 Network Performance

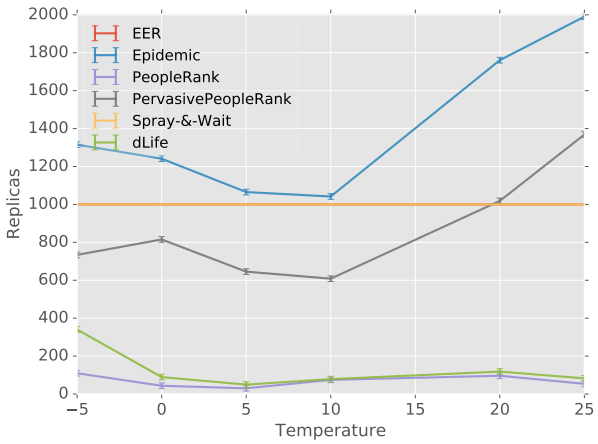
To evaluate the effects of environment and human behavior on MSN applications and on the proposed forwarding mechanism, we compared *PervasivePeopleRank* (PPR) with five other mechanisms: EER [25], PeopleRank [88] (PeR), Spray-and-Wait (S&W) [123], dLife [87], and Epidemic. The performance evaluation of the opportunistic communication is presented in terms of delivery probability, obtained as the ratio between the number of delivered messages and the number of messages that should be delivered; delay, calculated as the time elapsed between the message creation and the delivery; cost, which is the amount of replicas available in the network at the moment of delivery; and hops, as the number of nodes in the message's delivery path. The network traffic is generated based on time and mobility, where the messages are created for random destinations in two moments: when a node publishes a new data sample reporting its geographic position, and after random n minutes since the last published data sample, such that $n \leq C_{time}$.

In terms of node buffers, the default TTL of messages is 72 hours to attend the usual sparse nature of opportunistic networks. Also, we defined messages as generic packets independent of content to focus on message diffusion. Each message represents a unit on the buffer, with a capacity for 1000 unique messages. Figures 5.8 and 5.9 present the simulation results using C_{time} as 60 minutes and λ as 1000 replicas. The delivery results in Figure 5.8a show decreasing performance in temperatures within the *transition* phase.

Nevertheless, the PPR algorithm delivered at least 57.8% more messages than the remaining related protocols for the same phase. In the simulations with temperatures corresponding to the *positive* phase, the improvement is 69%. Messages delivered during critical hours of encounters increased by 48.2% using PPR. The average number of replicas



(a) Delivery ratio

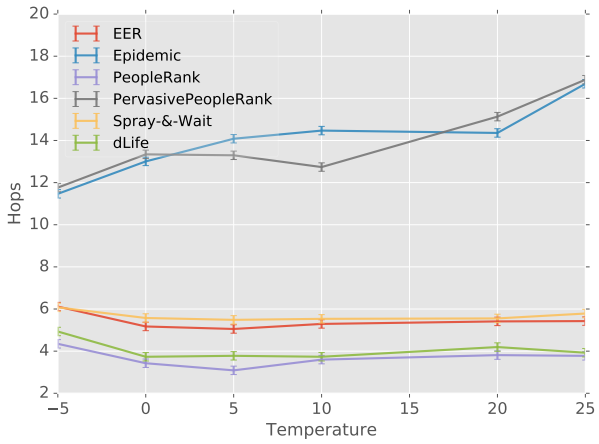


(b) Average of replicas

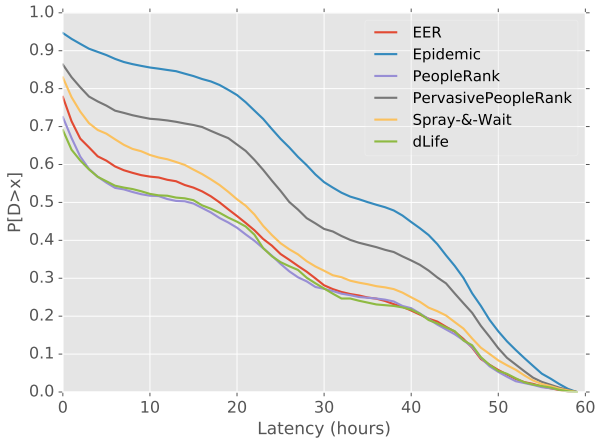
Figure 5.8: Delivery ratio and average cost according to temperature variation.

presented in Figure 5.8b shows the constant value for protocols EER and S&W, which are based on the replica limit λ. The increased number of replicas at higher temperatures using PPR occurs as a result of the higher number of contacts provided through mobility. The average interval between mobility events ΔM_n decreases by 11.7% in these temperatures.

Figure 5.9 shows the average number of hops and the CCDF of latency. Concerning these results, it is worth emphasizing that simulations of urban areas, such as NYC, can provide a large number of single encounters (in other words, encounters with just one occurrence). Moreover, these application scenarios provide subsets of nodes with few connections or low mobility, that is, nodes walking in small sub-areas or visiting unpopular places. Nodes with these features are accessible mainly through long paths or specific nodes, such as bridge nodes, which are responsible for connecting different communities and areas [87]. For this reason, Epidemic with the simple flood technique provides the best performance of delivery ratio and high average of hops. Indeed, the related protocols select relay nodes primarily considering centrality and social aspects, in an attempt to use



(a) Average of Hops

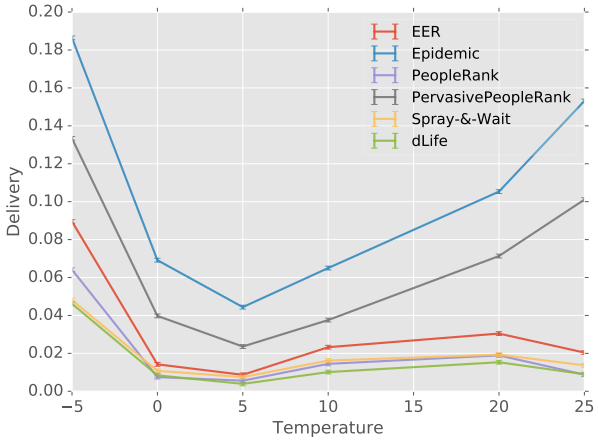


(b) Latency

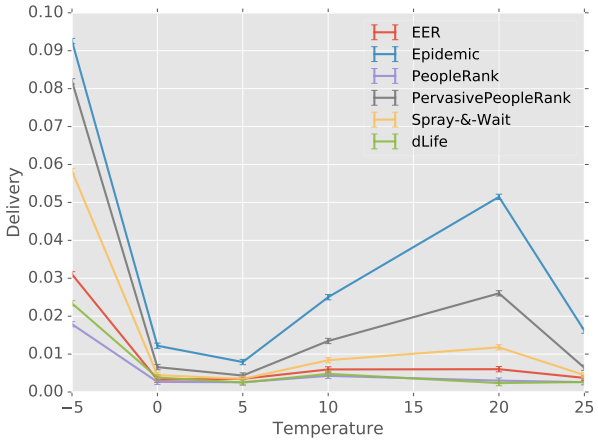
Figure 5.9: Average of hops and CCDF of latency.

short paths and lower delay. However, in large geographic areas, these approaches limit the number of feasible encounters to message transfer to a set of low-frequency events, and negatively affect delivery. That is, the related protocols quickly reach the well-connected nodes (Figure 5.9a); nonetheless, the messages are replicated or forwarded to another node with higher centrality, another node that had previously met the destination, or directly to the destination. In case of few connected destinations or low relay node mobility, more time may be required before a more suitable candidate for relaying the node is encountered, or a node from the destination social group is found.

dLife and PeR use 85.8% fewer replicas because the messages are usually forwarded to the high centrality nodes, but the infrequent encounters with feasible relay nodes, according to their respective decision mechanisms, stops the diffusion. Hence, the delivery rate is 76.4% less than PPR. The greedy approach of PPR reaches distant nodes and improves the delivery ratio, but naturally increases the overall number of hops. Nevertheless, PPR delivers 20.1% more messages using 15% fewer replicas than EER and S&W.



(a) 30 minutes



(b) 5 minutes

Figure 5.10: Delivery ratio considering different C_{time}.

Figure 5.10 presents the delivery results, using C_{time} as thirty and five minutes, and the performance is proportionately similar. Observe that the delivery rate in these scenarios decreases as temperatures fall in the *transition* phase. Nevertheless, the delivery rate using the proposed protocol is 54.4% and 47.9% better than the related proposals in these scenarios, respectively. Considering all the scenarios, the delivery rate is improved by at least 54.1% and 61.4%.

5.6 Conclusion

In this chapter, we investigated the seasonal patterns of urban mobility and their features facing thermal variation. Our observations indicated some effects of spatiotemporal features in human mobility and encounters in an MSN scenario. The social media data used in our investigation presented a fluctuation in venue popularity and of probable encounters between peers. Results showed that temperature could explain 74% of the variance in the

popularity of venues. Moreover, we showed that three ranges of temperatures could characterize distinct patterns of encounters. The changes in environmental variables provided the identification of distinguished behaviors observable by the spatial distribution of users, an essential feature for the design of message-forwarding mechanisms for people-centric approaches and large geographic areas.

In addition, we used the spatiotemporal insights to propose the *PervasivePeopleRank*, a cyber-physical message-forwarding mechanism for Mobile Social Networks. The mechanism improves delivery by an average of 57.8% by distributing multiple replicas of messages according to node centrality, mobility, and seasonal aspects.

Finally, our results indicate that environmental factors can characterize the state of the network, providing insights about the dynamism of urban scenarios. In particular, the temperature was shown to be a relevant feature in assisting the forwarding decision process for networks based on physical proximity and susceptible to human behavior.

Chapter 6

Social-Based Distributed Caching For Urban Scenarios

Cache mechanisms have been employed in computing systems for many decades, but recently they have retaken the attention of the academic community of next-generation wireless networks. One reason for such studies is the need to overcome performance bottlenecks in cellular networks, where massive temporary demands for content in regions with limited resources can use storage capabilities in the network itself. For this reason, in-network caching is a scientific challenge due to its distributed nature; therefore, in scenarios with cooperation among users, we can address the social and spatiotemporal characteristics to deliver content locally and effectively. In this chapter, we analyze the dynamics of the social graph of four metropolises around the world, where we address the spatiotemporal characteristics of encounters between their inhabitants and the content consumed and published by them. The results indicated social and geographic persistence capable of leveraging the use of caching mechanisms to provide content through local resources. Thus, we proposed a caching mechanism that assesses social and mobility characteristics to define the cached content of user devices.

Section 6.1 presents an introductory discussion on the relevance of social aspects and caching mechanisms for next-generation wireless networks. Section 6.2 presents a review of the recent studies about Device-to-Device cooperation in wireless networks. Section 6.3, presents the network model and the data used in the quantitative analysis. Section 6.4 presents the *PopSoC*, a mechanism for in-network caching. Section 6.5, presents the numerical results of mobility and sociability features that support *PopSoC*, whereas Section 6.6 discusses and analyzes its performance evaluation. Finally, Section 6.7 presents the conclusions.

6.1 Introduction

Urban centers are moving towards modernization to improve the quality of life and the experiences of their citizens. A large number of sophisticated mechanisms have been pro-

posed to optimize essential services such as traffic, energy distribution, and communications. In this way, the smart city has been widely discussed by industries, academia, and governments as a set of efficient services with positive effects on inhabitants' lifestyles, the economy, the environment, and plenty of other fields. Since the popularity of smartphones, tablets, and other personal devices has grown rapidly in recent years, many services that compose the urban scenario have converged for online and mobile platforms. This phenomenon allows us to capture data from myriad variables continually and emphasizes the fundamental role of wireless networks for smart cities.

Citizens and their personal devices are part of the landscape of large urban centers nowadays. According to the study [51], more than half of all web traffic comes from smartphones and tablets, and thirty percent of all mobile queries on search engines are related to the user's location. The large data traffic generated by those devices has motivated researchers to propose new mechanisms to support the network infrastructure to handle the demand. In this context, Information-Centric Networking (ICN) [44] has been presented as a promising technology in which the data transfers are no longer host-oriented but content-oriented. This approach has gained attention due to its benefits, such as improvement of spectrum efficiency, multicast transmissions, and in-network caching. Specifically, mechanisms for in-network caching have been extensively investigated due to their advantages for content dissemination, offload, and energy efficiency.

The next generation of wireless networks envisions the active participation of users to compose communication services in smart cities and support the offload of the network demand through user-to-user cooperation. The current personal devices are equipped with sensors and multiple network interfaces, which allow modern mobile applications to consider the social and environmental data captured through devices. For this reason, the Device-to-Device (D2D) paradigm and social characteristics will play a significant role in the efficiency of network resources; therefore, many researchers have made dedicated efforts to understanding the properties of Mobile Social Networks (MSN) [125]. Since the paradigm depends directly on users' cooperation, understanding the dynamics of the proximity graph, the graph from casual encounters of users based on physical proximity, is an essential element in the viability of D2D-based services.

Recently, the advent of Online Social Network (OSN) applications brought the opportunity to investigate the human dynamics by using data provided by users during many daily activities, such as exercises, content sharing, chats, etc. Studies of content recommendation, personal preferences, and analysis of social networks and communities have widely used OSN data and Location-Based Social Networks (LBSN), a subclass of OSN where users can interact by using geotagged content or enjoying unique features based on geolocation. These applications have been used to investigate spatiotemporal properties of content engagement, features of venues, variations of cultural aspects, among others. However, they have little explored the dynamics of MSN in terms of encounters among users. In the context of the next generation of wireless networks, data from OSN and LBSN may be used by protocol designers to explore social aspects and support mechanisms for resource sharing, cooperative computing, and opportunistic communication.

Social properties, such as the formation of communities and the emergence of hubs

and influencers may provide intersections of interests among friends, co-workers and set of users. Also, social properties contribute to content popularity, typically characterized by heavy-tailed distributions. Scenarios with these characteristics make in-network caching a suitable solution for providing content in overload situations. However, the mobility of users can negatively affect the spatial distribution of content supply and demand. Deployment of cache servers in the Core Network (CN) does not prevent the performance deterioration caused by backhaul bottleneck, and dedicated hardware for content caching on the Radio Acces Network (RAN) represents spatially static resources and additional cost [11]. On the other hand, users can provide distributed in-network caching through their personal devices, taking advantage of the ubiquity of these devices, associated human behavior, and distributed storage capacity.

In this chapter, we investigate the temporal, spatial, and social properties of human mobility to propose an in-network caching framework. The contributions of this chapter are as follows:

- We present a network model to simulate encounters between citizens in an urban scenario and estimate the proximity graph and its properties. The simulations use real data collected from a social media application for several continuous months.
- We analyzed the spatiotemporal properties of the simulated encounters and the dissemination of content, where the results indicated persistent behavior for geographic and social characteristics.
- Lastly, we proposed a framework for in-network caching based on content popularity and social context. The framework explores the behavior of users to manage the content stored. The performance evaluation indicated an increase in hit probability with fewer replicas per content.

6.2 Related Work

The design of mechanisms for optimized content caching and distribution is one of the most critical challenges of distributed caching in wireless networks. The problem requires sophisticated decision mechanisms to select the relevant content to be cached and forward it via opportunistic encounters among users, considering the content replication if necessary. Thus, the proposed solutions to the problem explore the computing, sensing, communication and data storage of personal devices to model entities and events around users and contents. Therefore, the mechanisms for content caching should take advantage of this information to be able to select the best nodes to store and serve contents, improving the main performance indexes, such as cache hit and delay, taking into account the overhead caused by multiple replicas and hops.

In this direction, Wang et al. [133] proposed a framework for traffic offloading assisted by OSN services via opportunistic networks. Firstly, the proposed framework selects a subset of users to receive the same content as initial seeds, depending on their content,

spreading impacts in OSNs, and their mobility patterns collected previously. After that, the seed users store the content until they share with neighbor users using opportunistic encounters. The results indicated reduction up to 86.5% of cellular traffic, considering the delay requirements. The proposal presented by the authors considers, that the integration between OSN services and the mobile operator in the decision process, of what content to store and which user should do it. In addition, the proposal assumes that the mobile operator can push content to the users in the initial seed selection step using a global view of the network, and independent of the user's interest in the content pushed.

Wang et al. [132] proposed a framework for information-centric virtualized cellular networks using D2D communications. The framework considers the local storage of contents in the users' devices and incorporates the content caching strategies in a resource allocation optimization problem to maximize the total utility of Mobile Virtual Network Operators (MVNOs). The work discusses the need for Network Function Virtualization (NFV) to deal with the typical scenario of multiple mobile operators, where users from different operators are unable to directly communicate and collaborate due to the policies and economic factors. Hence, the virtualization-based approaches may use general-purpose hardware infrastructure and ensure the modularity of the components [131, 49]

Sheng et al. [113] designed a multilayer architecture for content delivery, which explores caching in the personal devices and the edges of the network. The authors discussed the need for caching along the path from content servers to personal devices and advocates the importance of infrastructure coordination in the cooperation of users-to-users and users-to-infrastructure. The results presented the benefits of hybrid caching mechanisms, where the improvement concerning latency can be approximately 40% better compared to traditional approaches and 60% for the Region Hit Ratio (RHT), which indicates the ratio of locally retrieved contents without visiting remote servers and the total of requests.

Chen et al. [26] proposed a relay selection mechanism for D2D applications based on social-trust and social- reciprocity. The identified trusted relationships may relay data among their users, and social without trust, form a coalition with social reciprocity to relay for those in the same coalition. Bao et al. [11] evaluated the potential for data offloading via D2D using real experiments. According to the experiments, the cellular networks are mostly overloaded during high-density events, when many people located in small geographic areas, where they usually consume similar content or use a single application massively. The authors proposed that the mobile operators track the geolocation of clients and build maps to indicate dense clusters of users and content, defined by authors as data-spots. They conducted experiments using bike rides in New York with Bluetooth scanning nearby devices. Furthermore, the experiments included simulations of content, and the results indicated performance improvement for multimedia and publisher-subscriber applications.

Recently, social networks have been widely studied to improve wireless networks in different paradigms, especially Delay-Tolerant Networks (DTN) [148, 135] and vehicular networks [90, 129]. Since most of those networks employ store-carrying-forward mechanisms and the data is transferred mainly through short-range wireless interfaces, those similar paradigms can exploit their users' social properties similarly. In DTN scenarios,

the proposals have explored social aspects, such as nodes' popularity [88], community labeling [58], expected number of encounters and delay [25], explicit mutual interests [32], and a combination of centrality and communities [59].

In this direction, Moreira and Mendes [87] investigated the effects of human behavior in DTN applications with social aspects. They evaluated the social properties and similarity of interests among users to develop opportunistic message forwarding systems, focusing on services for extreme networking conditions, and dense networking scenarios. The experiments used simulations based on real and synthetic mobility traces, and the findings indicated the need for adaptable forwarding and self-awareness mechanisms based on the dynamism in behavior.

Chen and Lou [25] proposed a forwarding scheme based on node encounters and the time-to-live (TTL) property of messages. The authors proposed a routing protocol for delay-tolerant applications that distributed multiple replicas between nodes, in proportion to their expected encounter ratio. Their work presented the Expected Encounter-based Routing protocol (EER), using the metrics Expected Encounter Value (EEV) of each node and the minimum Expected Meeting Delay (EMD) between the current node and the destination. The messages are created with a predefined number of replicas and spread on the network proportionally to EEV. Thus, when the number of replicas of a held message is reduced to one, the single replica is forwarded only to the destination node or a relay node with lower EMD. The experiments used the vehicle-based mobility model, which is part of the Opportunistic Network Environment (ONE) simulator [65].

Mtibaa et al. [88] proposed a forwarding mechanism based on node popularity, derived from the *PageRank* algorithm [19]. The *PeopleRank* proposal explores the popularity of nodes with a distributed approach, forwarding new copies of the original message to nodes that rank higher than the current node. The messages are duplicated on demand and without a specific limit of replicas. The performance evaluation presented results by using six datasets of real data with 27 up to 414 nodes.

Ciobanu et al. [30] explored the social graph from social media applications to provide additional information and support the message-forwarding mechanism. The proposed algorithm takes users' interests and contact history into consideration to decrease the congestion and required bandwidth, taking into account the overall network's hit rate and the delivery latency. Similarly, Socievole et al. [122] introduced the multi-layer social network model, which combines social networks based on proximity and online social networks.

The authors investigated the relationship between different social network layers regarding node centrality, community structure, link strength, and prediction. Both works discuss the advantages of using social aspects to improve opportunistic dissemination and the benefits of using online social media applications to obtain the social graph. Nevertheless, these proposals assume an eventual connection to the Internet or to remote servers of social media applications. These assumptions make it difficult to use these proposals in scenarios without infrastructure.

The state of the art indicated the social properties as dynamic and complex systems and pointed to their influence in D2D based systems. Most of the performance evaluations carried out by recent studies considered real traces of human mobility, although the data

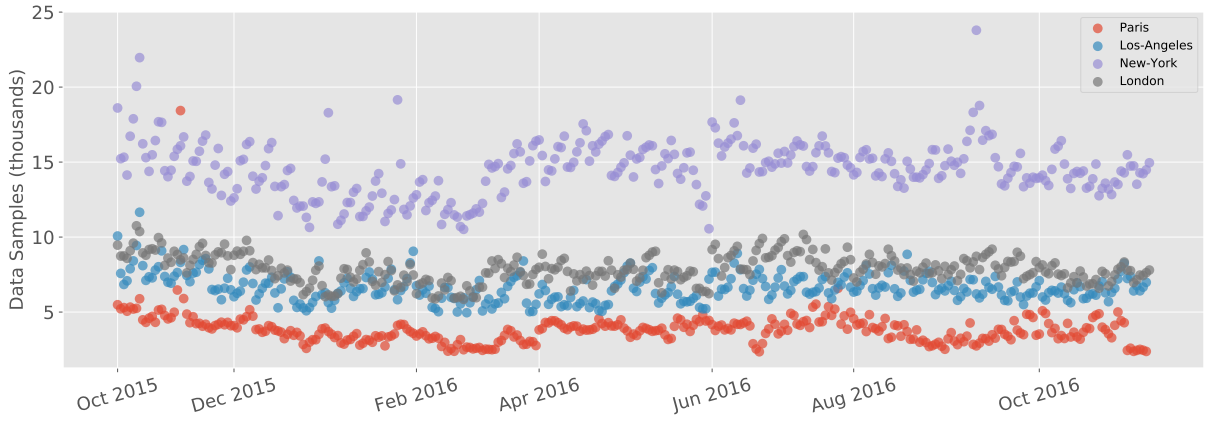


Figure 6.1: Time series of samples grouped by cities and days.

analyzed represent a few tens or hundreds of users and communities, such as universities or conference centers [111, 100, 42]. Moreover, the majority of socially-aware studies have proposed mechanisms considering scenarios in experimentation scale, which may not capture the temporal variations of personal preferences, typically observed in long-term data. For this reason, the design of socially-aware content caching mechanisms with D2D should consider clients' inclination to not disclose their private data. Furthermore, personal preferences and social properties must be investigated as dynamic features of urban scenarios, taking into account the effects of their fluctuations.

6.3 Network Model

In this section, we describe the multigraph approach, the dataset built from social media application, as well as the network model and traffic used to simulate the encounters among users and content dissemination.

6.3.1 Multigraph Approach

The graph presented in Figure 6.2 shows some distinct classes of links among users. The multiple edges among the same pair of nodes represent different layers of data combined, where a pair of nodes may be connected by more than one edge or connected in different layers. In other words, the graph describes the interactions of the same pair of nodes in a multitude of domains cumulatively combining two or more graphs.

In this hypothetical network, the social relationship layer indicates the existence of association among users, friendship or some other human relationship; the layer of D2D links indicates sufficient geographic proximity for a direct connection of the users' devices; and the edges in the content-based layer indicate the request of identical content for both nodes. The edge (u_1, u_4) represents the most common edge class for large sets of users dis-

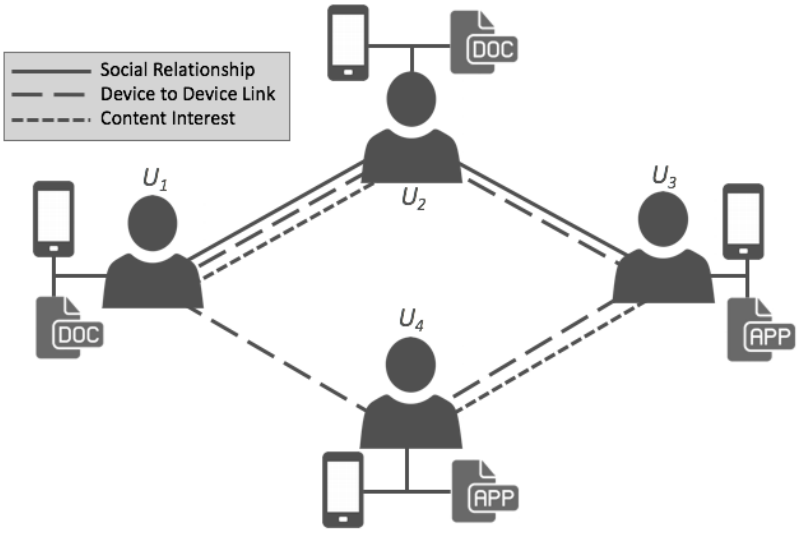


Figure 6.2: The cumulative aggregation of content interests, proximity, and social graph.

tributed in broad geographic areas. Those edges are a result of casualness, a spatiotemporal coincidence of two random users with low chances of happening repeatedly.

The edge (u_1, u_2) similarly represents an opportunistic encounter; nonetheless, the social factor can group users with interest in similar content due to homophily, a phenomenon observed in social networks that defines the natural grouping of nodes with similar attributes [134]. The intersection of interests may also result from spatiotemporal factors capable of forming crowds momentarily and independently of social relationships (edge (u_3, u_4)), such as the encounter of random users in musical concerts, sports events, and other public occasions.

The physical proximity layer and content requests are visible to the base station and potentially useful for the caching mechanism. Since the social information layer is rarely available due to privacy constraints, cache mechanisms must infer the social attributes from alternative layers, investigating frequently visited places and mobility patterns, for example.

6.3.2 Network Model

Urban data streams are services for data dissemination that provides structured data of variables from urban scenario, such as transportation, traffic, environmental conditions, and health [85, 64, 18]. Some local governments are officially providing those streams, while other general-purpose applications, such as OSN, may also be alternatively used. In OSN streams, a sample of data may register an event indexed in time and space, enabling multidimensional analyses capable of representing mobility and personal preferences, among others.

To investigate real urban areas, we used OSNs as data sources to analyze the spatial distribution of people and estimate the encounters between them. We collected data samples

from Twitter, the popular application where users can share text and multimedia messages indexed geographically. Third-party applications can collect the data publicly available, in addition to the metadata with the user identification, date of publication, latitude and longitude coordinates, terms, hashtags, and URLs, among others.

Therefore, we collected data from October 2015 to November 2016 and prepared a dataset, which represents the cities London, Paris, New York, and Los Angeles. The choice of the set of cities follows their representativeness in the dataset that sums 11.4 million of samples from 998,000 users, as described in Table 6.1.

In this way, we formally define a data sample as a 3-tuple $d =< u, p, t >$, where u represents a user $u_i \in U$, t is the time index of the sample defined by local time, usually the local time, and p is the u_i's spatial index defined by latitude and longitude coordinates. Therefore, a set of data samples classified by city and chronologically sorted represents a time series, as presented in Figure 6.1.

Given that the samples do not provide the pause-time, defined as t_p, which represent the period that a user remains in p, we assume the temporal threshold $T_{th} = 60$ minutes as the maximum natural interval between two chronologically ordered samples of a $u_i \in U$. The simulations presented in the next sections comprise time series of months uninterruptedly; therefore, the pause-time t_p should consider different situations of mobility, such as from high or moderate mobility during business hours or routine to low mobility at home during the night. To cope with it, any natural interval between two consecutive samples that exceed the temporal threshold is replaced by a random $0 < t_p \leq$ 60 minutes following the Poisson process, as shown in Figure 6.3.

The distance threshold D_{th} defines the maximum geographic distance for an encounter based on physical proximity between two data samples. To simulate the D2D capabilities, we defined $D_{th} = 100$ meters. Accordingly, an opportunistic encounter event able to establish a D2D link is identified when any two data samples d_i and d_j satisfy the following criteria:

- $u_i \neq u_j$;
- $d_g \leq D_{th}$ where d_g is the geographic distance between p_i and p_j;
- $t_s < min(t_u + t_{pu},\ t_j + t_{pj})$ where t_s is the current simulation time.

Naturally, user mobility provides a temporal graph $G = (V, E)$, where E is the set of edges that represent the geographical proximity among users and V is the set of users available at time t_s. In this analysis, we evaluated the properties of the graph using the $G_t = \{g_1, g_2, \ldots, g_n\}$ time series, where g_i is a snapshot of the graph at time t_i. We also evaluated the aggregated graph $G_c = \{g_1 \cup g_2 \cup \cdots \cup g_n\}$.

6.4 Popularity-Based Social Caching

In this section, we present the framework Popularity-based Social Caching (PopSoC) for in-network distributed caching. The framework consists of a set of services implemented

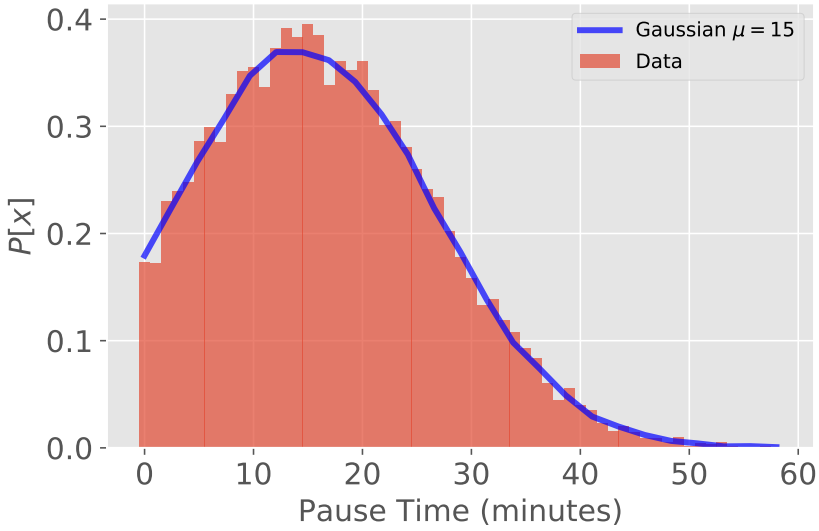


Figure 6.3: Pause time considered in the network model.

Table 6.1: Online social network dataset description.

	Samples	Users	Period
Paris	1324051	156823	October 25th 2015 to November 7th 2016
London	2733343	269283	
Los Angeles	2313217	209907	
New York	5057454	362752	

on the client and server sides. The client side provides two essential services to capture the user data and answer the queries of the base station:

- Mobility Manager: captures and stores the mobility of the client from multiple sources and replies to the queries from the base station.
- Cache Manager: stores and provides the content indicated by the base station and defines the cache replacement policies.

In the Cache Manager service, the client side provides a portion of the local storage for caching the content indicated by the base station. Since the framework exploits content popularity and homophily for content reuse, a client should be able to store content consumed by itself as well as potentially relevant content indicated by the base station. Therefore, the local storage of the personal device reserved for the cache is divided into social and popularity-based, where the cache allocated for social caching is proportional to the social persistence P_s value of the client.

The Mobility Management module captures the user mobility when using multiple sources available, such as sensors and the network infrastructure. Spatial indexing or positioning can be obtained through the base station while the client remains associated, and

geographic position can occasionally be obtained during the use of user-level applications for navigation, web browsing, and others.

The server side defines the base station tasks according to the following services:

- Peer-Discovery Service: continuously updates the sets of associated clients and the content objects available through them;
- Content Manager: decision mechanism that indicates whether the content should be cached and whether the content is required by the client;
- Content Match Making: service for content pull and push, handles content requests and the correspondent popularity. It examines the requests and indicates the clients to serve with the cached content.

The base station needs to be aware of the content available on the associated clients; thus, during the client association process, an extra step is included to update the content table at the base station with the cached content at the client's device, as widely considered in related proposals [90]. We assume that the base stations are capable of providing the cooperation among clients from different mobile operators through Network Function Virtualization (NFV) [132].

The Content Management module monitors the content objects available at the cell range through the Content Availability Table (CAT), formed by the union of the cache of all clients associated to the base station, where each entry maps the content identifiers to their sets of client providers. Furthermore, the module also monitors the popularity of the contents individually using the Content Popularity Table (CPT), where each request increases the popularity of a particular content indexed in the table.

When a client sends a content request to the base station, it indicates the required content and its social persistence *sper* obtained through the Mobility Manager. The base station evaluates the CAT and identifies the set of clients associated with the same base station that is capable of providing the content through D2D communication, characterizing a hit event. If the content is not available, i.e., a miss event, the base station obtains the content from the remote server and forwards it to the requesting client. In both cases, the base station indicates the corresponding popularity of the content using the CPT, and the requesting client stores the content in cache, overwriting a less popular content object previously stored. In hit events, the client stores the content requested in the social cache; however, in miss events the base station also pushes content with high popularity to the requesting client through a neighbor client using D2D.

6.5 Spatiotemporal Analysis

The understanding of the properties of the proximity graph has a vital role in the design of D2D and social-based applications. Suitable predictions of node characteristics

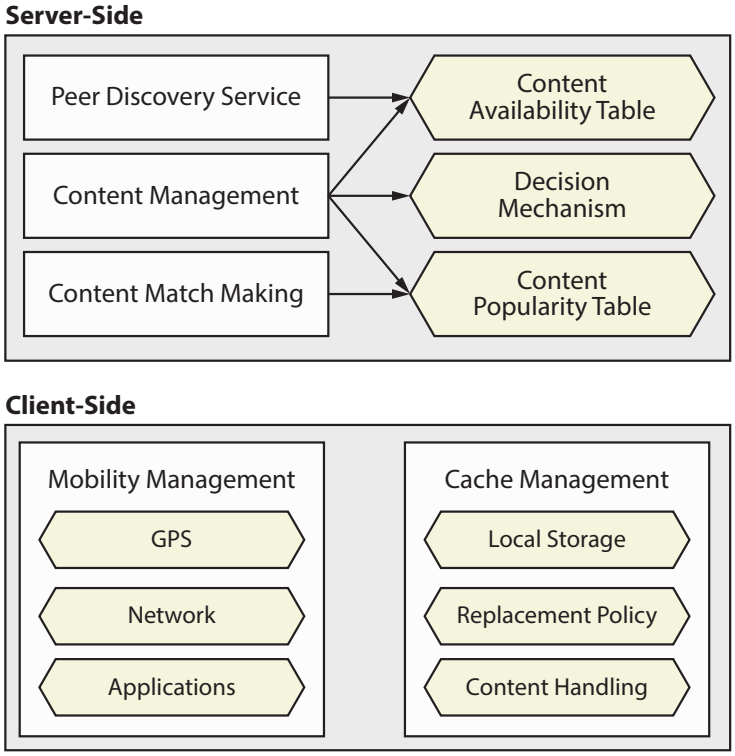


Figure 6.4: PopSoC Framework components.

and structural properties of the network are critical tasks for decision mechanisms in distributed systems. Therefore, in this section, we investigate the spatiotemporal properties of simulated opportunistic encounters according to the network model presented in the last section.

Figure 6.5 shows the complementary cumulative distribution function (CCDF) of the edge occurrences recorded in the simulations. The occurrences indicate the repetition of the encounters for an approximately small portion of the edges observed. Naturally, most of the encounters are ephemeral and happen just once, which is an expected result considering the vast urban areas studied. Figure 6.6 shows the geographic distance between two consecutive encounters of the same pair of nodes. The result shows the distribution of distances between i_{th} and i_{th-1} occurrences of edges. The combination of results indicates that edges with at least two occurrences are likely to occur in small geographic areas, configuring edges with some regularity as regionally limited.

Additionally, the inter-contact time or interval between two occurrences of the same edge, shown in Figure 6.7, shows a mean time of approximately 11 minutes. The contact time, or the duration of these edges, shows a mean of approximately 2 minutes, as shown in Figure 6.8. We emphasize that inter-contact time results contain samples with natural and synthetically defined pause-time. Therefore, encounters registered more than once characterize that edges with some regularity experience quick encounters of a few minutes in limited geographic areas, usually by a few hundred meters, in addition to intervals

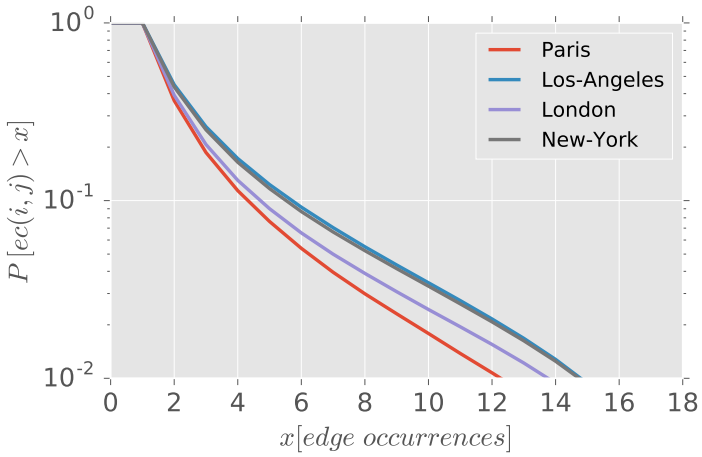


Figure 6.5: Edge occurrence.

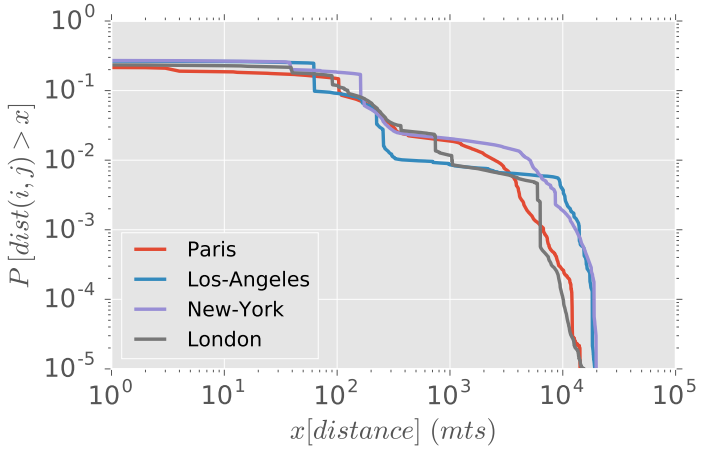


Figure 6.6: Distance between consecutive encounters.

that indicate re-encounters in the same day. The cities investigated presented similar trends that indicate analogous social behavior, as observed in [92], where different cities present a universal behavior of human mobility, independent of organizational and cultural differences.

The geographic persistence P_g of a user u is calculated as the probability of u visiting a place previously visited. Formally, let V_u the set of places visited by u, then the geographic persistence, is the ratio of the total visits represented by V_u and the subset $R_u = \{p \mid p \in V_u\ ,\ \lambda(p) > 1\}$, where $\lambda(p)$ is a function that quantifies the total of visits to the place p. Similarly, we define the user's social persistence P_s based on the percentage of repeated encounters or the probability of occurrence of edges already registered.

Figure 6.9 shows the results of topology overlap analysis, a metric that calculates the ratio of total shared neighbors between two nodes. In this analysis, we calculated the topology overlap values for all pairs of neighbors in the proximity graph, considering the

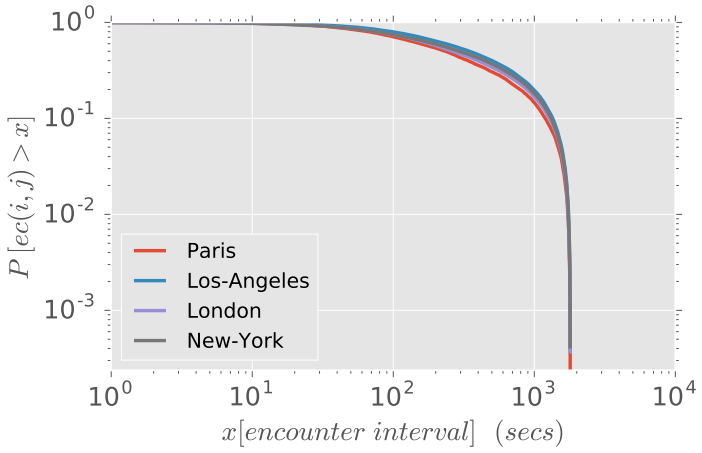


Figure 6.7: Interval between consecutive encounters.

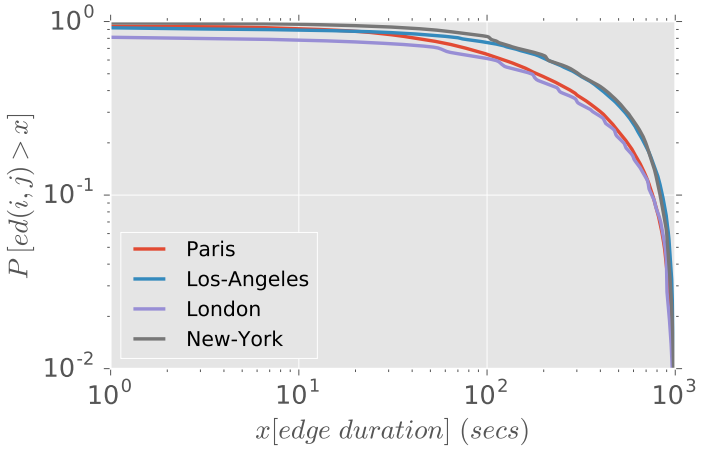


Figure 6.8: Encounter duration

cumulative social graph of all the encounters over the dataset. The result shows, for all analyzed cities, that more than half of the observed users have some topology overlap with one or more neighboring users. The curves, although similar, show greater disparity between New York and Paris. However, in general, the result indicates that these cities present consistent social characteristics that result in proximity graphs with relevant properties of navigability.

Figure 6.10 presents mean values of the results of social persistence correspondent to the geographic persistence. The result shows peaks on P_s for extreme values of P_g. Users with great P_g and P_s represent individuals with a high probability of visiting the same places and meeting the same people, characterizing the behavior of consistency in both aspects. Meanwhile, users who presented high values of P_s despite the low P_g are users who configure strong and stable edges regardless of location. Users with significant social persistence values indicate opportunities to cache relevant content for their neighboring

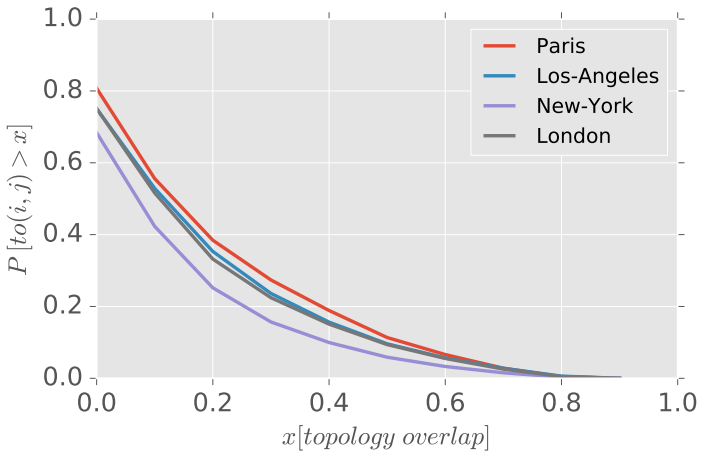
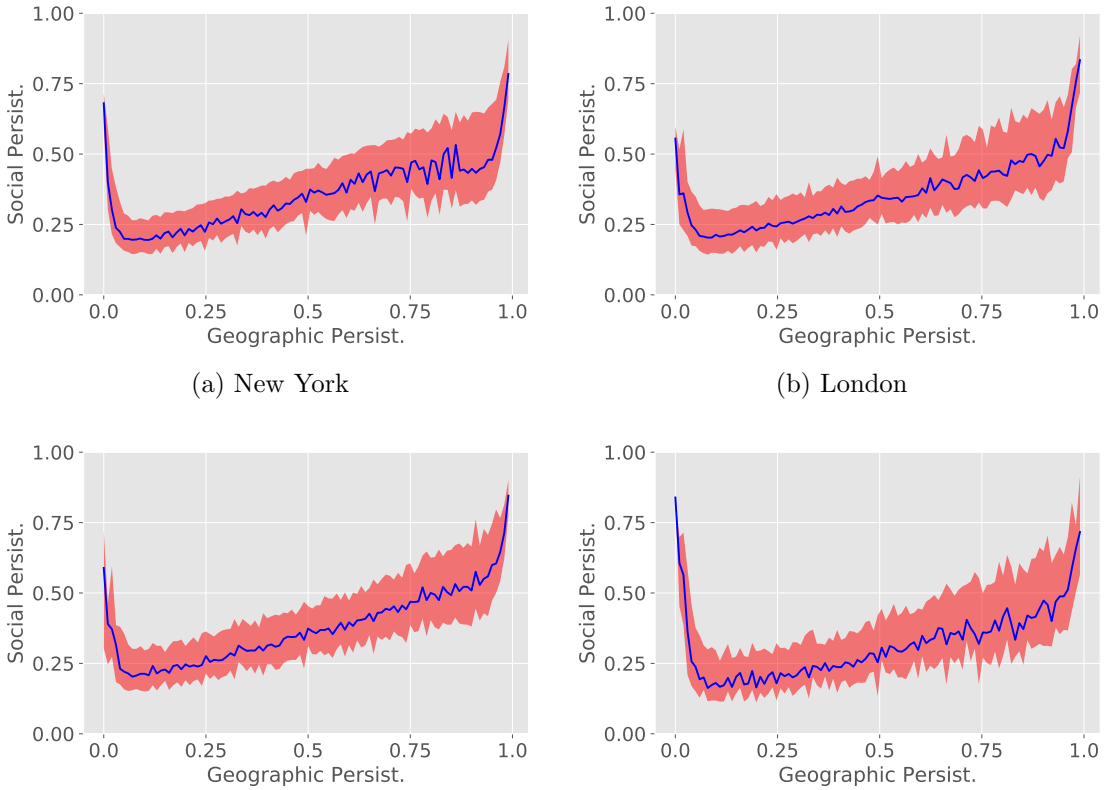


Figure 6.9: Topology overlap.



(a) New York

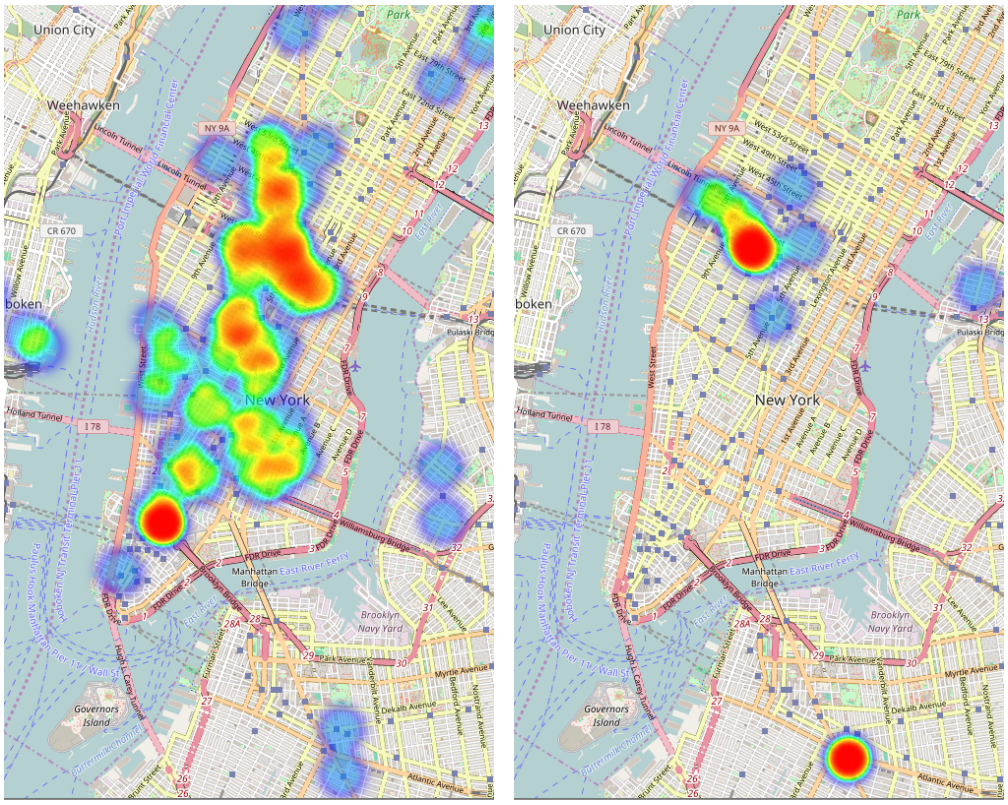
(b) London

(c) Los Angeles

(d) Paris

Figure 6.10: Geographic and social persistence of users.

nodes and their communities; thus, these users can store contents of their interest for reuse of their future contacts rather than content recommended and pushed by the base station. Thus, geographic persistence can be used as a predictor for social persistence and a tool for decision mechanisms.



(a) NFL Super Bowl (b) NBA Match

Figure 6.11: New York encounters during two distinct sports events.

6.6 Evaluation

In this subsection, we present the performance evaluation of the framework proposed, as well as the parameters of the opportunistic caching simulation.

The dataset studied captures the dynamic preferences of users in broad real-world scenarios; therefore, it is necessary to use traffic with corresponding spatiotemporal characteristics to evaluate cache policies properly. The diffusion and consumption of content in large and crowded areas present spatiotemporal characteristics that are difficult to replicate synthetically. As shown in [11], data spots arise naturally from the agglomeration of users in the same region with excessive demand; for this reason, synthetic models may not provide the particularities of these scenarios and instead provide oversimplified or generic network traffic.

Figure 6.11 presents the encounters estimated on the days of two popular sports events in New York. The heat map cumulatively indicates the encounters among users who published messages in the OSN that contain the same term or keyword related to the event. Despite the difference in engagement of the two events, as determined by the number of users involved, both events show similarities of interest and spatial proximity; however, the spatial distribution is significantly unequal. The spatial features evidenced in this

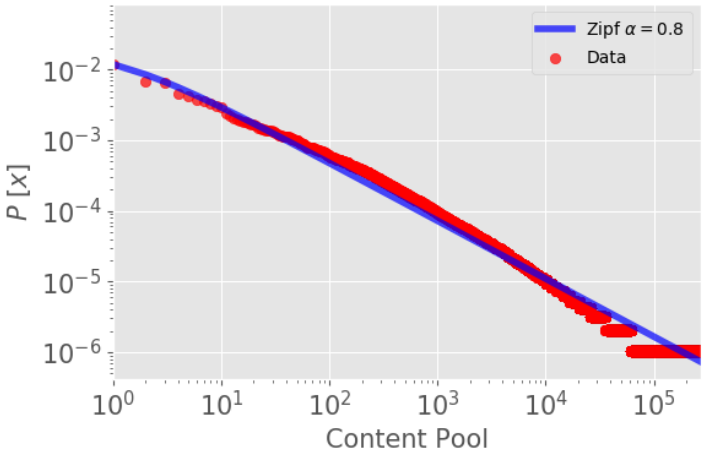


Figure 6.12: Distribution of content popularity.

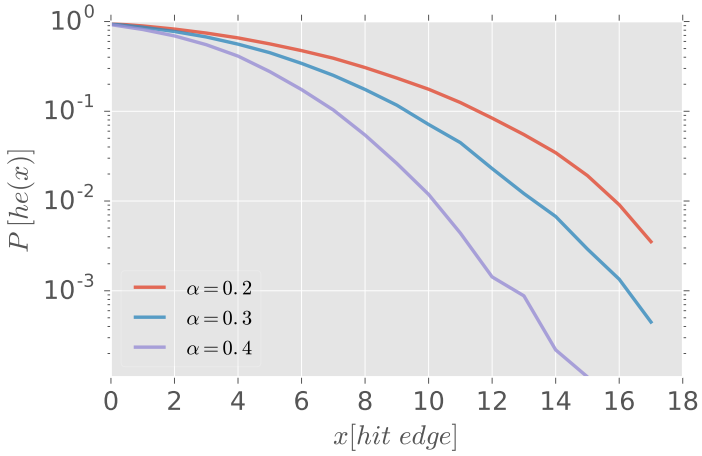


Figure 6.13: The impact of α parameter.

result show the heterogeneity of content interest, and evidence that the massive demand for content has spatiotemporal properties that make it challenging to deploy resources spatially static, for example, the use of additional hardware for content caching on the edge of the network.

According to this, the network model used to evaluate the caching policies is derived from the published messages captured in the dataset. In addition to the metadata, we use semantic parsing of published messages to isolate elements, such as URLs and hashtags. The elements of the messages represent the contents objects of the content catalog, where a sample represents a request to one or more objects in the catalog, and the total number of requests of the same object represents the corresponding popularity. The empirical content catalog of is presented on Figure 6.12 modeled using a Zipf distribution, where $\alpha = 0.8$.

Figure 6.13 shows CCDF of the number of hits per edge, considering different values of parameter α. Since the parameter is dependent on the social persistence threshold, high

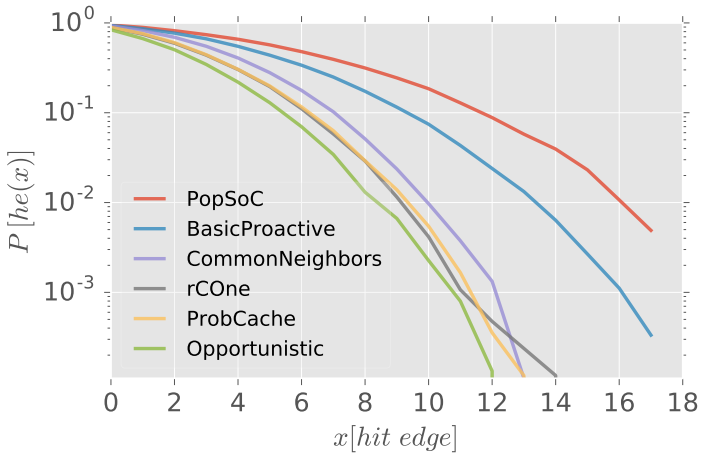


Figure 6.14: Hit per edge.

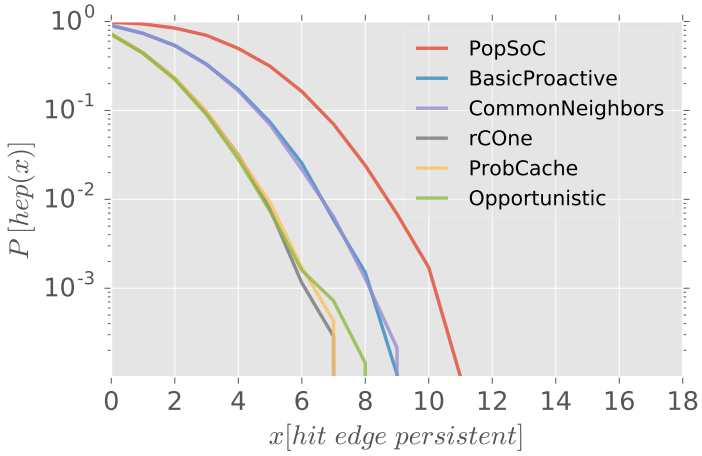


Figure 6.15: Hit per edge with persistence.

values provide significant portions of the device's local storage for social caching. Naturally, high values provide smaller subsets of users selected for social-based caching and lessen the impact of the social factor of the proposal. However, the same values are responsible for the decrease in base station content push operations.

Figure 6.14 shows the hit CCDF considering the edges in the proximity graph resulting from the simulated traces. The results show that PopSoC was able to provide more cached content than the baselines used in the comparison. The BP and CNP mechanisms showed inferior performance due to competition for cache space resulting from a proactive approach. The caching approach divided into popular content and consumer consumption limits the aggressive replication of content, using up to 37.5% fewer content replicas and increasing the overall hit up to 20.2%.

Since the intersection of content interest among clients is a determining factor for caching policies in D2D environments, the challenge of providing the required content

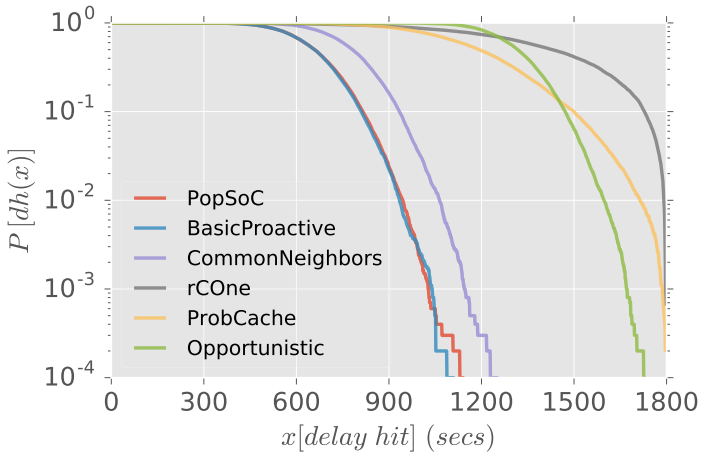


Figure 6.16: Delay of hits.

through caching goes beyond storing highly popular content. Effective in-network caching in D2D scenarios may also require proper selection of nodes and relevant content considering the spatiotemporal and social context.

In this way, cache management according to the social persistence used in PopSoC provides advantage when clients proportionally prioritize the consumed content for social caching. This approach allows the client to store popular content according to their own consumption and content that is potentially relevant in the social context. Therefore, it reduces the total push operations of content not required by the client. Consequently, the positive effect is noticeable on the probability of hits on edges, which present two or more occurrences, as shown in Figure 6.15.

The delay measured in the simulation model used considers late hit cache events in a delay-tolerant scenario. In other words, it is the waiting time of a request for a content object not available in the cache but later provided by a newly arrived neighbor node. PopSoC presented an average delay of approximately 11 minutes, a result similar to the basic proactive protocol that represents an approximate improvement of two times compared to the lower performance of the baseline (Figure 6.16).

6.7 Conclusion

In this chapter, we explored the users' spatiotemporal preferences to provide in-network caching for wireless networks in urban scenarios. We used data collected from Online Social Networks (OSNs) to investigate social and spatiotemporal characteristics of users through the dynamics of mobility and the opportunistic encounters. We used an opportunistic network model with Device-to-Device (D2D) capabilities to estimate the temporal properties of the users' proximity graph and the dissemination of content in those scenarios.

The simulations included real data from four different cities collected during several months and indicated that human mobility has strong social characteristics capable of

influencing the encounters among citizens, evidencing the temporal properties that make the distinct urban proximity graph. In this way, we observed the predictability of human behavior through the persistence of edges regarding time, space, and social aspects.

Using those observations, we proposed the PopSoC framework for distributed cache management in a content-based network architecture. Our proposal explores the social and geographical persistence of users without requiring sensitive personal data from users or external data sources. Numerical results showed an increment in performance, increasing hit events in the cache using a reduced number of replicas per content. This shows the feasibility of using the social context to improve the quality of service of wireless networks in smart cities.

Chapter 7

A Data-Centric Approach for Social and Urban Sensing

The management of services that serve urban areas is fundamentally dependent on data from the areas affected by them. For this reason, efficiently collecting data is a critical task for the development of effective urban policies and services. However, some applications still depend on manual data collection that demands a significant amount of time, such as the collection of urban characteristics and census, while others do not have large-scale sensing mechanisms, such as the social graph of the inhabitants, the graph mobility among places among others. In this chapter, we use data made publicly available through online social networks and the city hall of a major metropolis to provide a sensing mechanism. The results demonstrated effectiveness in urban sensing and insights into city regions and user relationships, places and the combination of these elements.

Section 7.1 presents the big data-based sensing paradigm and the importance of public and open data for the development of new sensing mechanisms. Section 7.2 presents the data sources used in this chapter and its main characteristics. Section 7.3 presents the results of the investigation of social aspects, as well as Section 7.4 presents the results of spatiotemporal aspects. Finally, Section 7.5 presents the conclusion.

7.1 Introduction

The modernization of the urban landscape has transformed the metropolises, accelerating the process of urbanization, and making the urban lifestyle attractive to a large number of people. This agglomeration introduces numerous new challenges for city management, requiring sophisticated mechanisms for observing the urban variables. Understanding the urban characteristics from observations is a fundamental step towards the development of smart cities through services for people, mobility, environment, living, governance, and economy [95]. Industry and academia have investigated different sensing solutions to improve the monitoring in extensive geographic areas, making the Internet of Things (IoT) a fundamental part of modern urban scenarios and a widely disseminated approach [3].

Nonetheless, many of the urban sensing initiatives, based on conventional sensing or Wireless Sensor Networks (WSNs), depend on specific resources that may require significant investment, additional hardware, and sensor management, making the financial factor a challenge for implementing urban sensing in cities that are geographically extensive or have limited budgets.

Meanwhile, the vast popularity and reach of mobile computing and wireless networks in recent years have had a significant impact on social interactions, media consumption, business, education, and many other fields [57, 128]. These advances have consolidated a vibrant ecosystem of online applications and services that compose the urban lifestyle, where a wide variety of data is generated at high speed by users, devices, companies, and transactions among these entities, making the urban scene more measurable.

The observations registered by these applications are capable of portraying a broad set of urban variables through urban data streams [18, 43]. The data may reveal cities as complex systems in which the human factor has a significant impact on their characteristics. In special, Online Social Networks (OSNs) applications have gained the attention of researchers who have investigated the urban dynamics through users' characteristics [142] using a new sensing paradigm capable of creating complex virtual sensors [110].

The virtual sensors are software tools that provide analytics-based sensing while exploring alternative data sources; process large sets of variables; characterize and recognize multiple phenomena as well as different entities (physical or not); perform the tasks of capturing, representing and analyzing urban data as a scalable alternative to overcome the costs of physical and dedicated sensors. Nevertheless, virtual sensors that operate with multiple sources demand great efforts of integration due to the lack of standardization and the sources' particularities that affect the frequency of the sample collection, anonymity, spatial coverage, among other aspects. These challenges hinder the development of more abstract, replicable, and scalable urban sensing services.

To address these challenges, we propose a data-centric urban sensing framework exploring OSNs, placing users and their mobile devices as fundamental tools for measuring the urban reality. The services shed light on the social and spatiotemporal characteristics of the urban panorama and provide insights through a data analytics pipeline, offering new observations at a city scale using a scalable and reproducible methodology. In this chapter, we present sensing approaches to characterize social, spatiotemporal, and living aspects, exploring encounters among users, venue distribution, and visitation preferences, among others.

7.2 Data Sources

The public data available through OSNs represent a new opportunity for urban sensing for two main reasons. First, OSN applications are a modern phenomenon of enormous popularity and rely on sets of users in the order of millions, including large and diverse sociodemographic groups. Second, these applications have platforms for developing and integrating them with third-party applications that leverage the democratization of users'

data. Therefore, popular OSN applications with high user engagement can provide real data about people, places, events, preferences, and collective behavior through contextual information and metadata.

We promote this potential by exploring the collected data from different applications as follows:

- Twitter[1] is an OSN primarily for textual content sharing, although it provides support for multimedia content formats, where users can post content publicly or to their followers exclusively. We collected samples of public content from users whose metadata indicates the users' precise location at the time of the publication through geographic coordinates using the Twitter Stream[2].
- Instagram[3] is an OSN focused on sharing photos and videos of short duration that motivate interaction between its users through private conversations, public comments, and the concept of like. The published content can index the user's location through the name of the venue or event. The collected data adds contextual information to the user's location registered in the samples, indicating the venue corresponding to the stated geolocation.
- Facebook[4] is a popular OSN with a significant number of daily active users. It connects people, interests, and venues through user profiles, pages, and posts with textual and multimedia content. The collected data correspond to queries for nearby places based on latitude and longitude coordinates within the studied area.

All collected samples used in this study are restricted by the official boundaries of the Manhattan region, the most densely populated borough in New York City (NYC), as well as their official districts as determined in the dataset compiled and publicly distributed by the local city hall at http://on.nyc.gov/2CmX3pN.

Accordingly, we compiled two data sources. First, the mobility dataset that comprises samples collected from Twitter and Instagram and define the geographic coordinates of users along the observation window. Second is the venues dataset, which semantically defines the set of places indicated by users in the first dataset, is designed on the basis of the synthesis of metadata from Instagram samples and Facebook's geographic database.

Using these data, we investigated 3.5 million mobility samples shared by 256,000 users collected from May 2016 to July 2017 to shed light on the mobility dynamics of users and their preferences as well as the spatial characteristics of the city.

[1] https://developer.twitter.com
[2] https://developer.twitter.com/en/docs/tutorials/consuming-streaming-data.html
[3] https://www.instagram.com/developer
[4] https://developers.facebook.com

7.3 Sensing Social Aspects

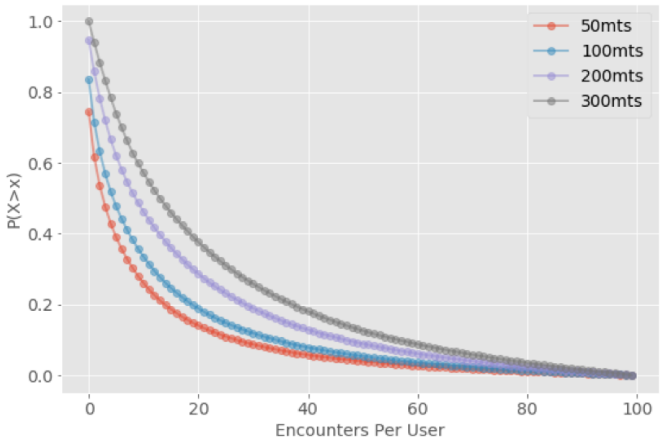
Monitoring social aspects of user groups with thousands of individuals can be a challenging task directly related to the monitored variables. Monitoring variables related to the social context in the physical environment typically requires specific approaches to measure the proximity between users and estimate their encounters and interactions. In this way, researchers have used custom hardware or mobile applications to this task, however, the capillarity and reach of the experiment are affected by these approaches. In mobile applications, the additional power consumption to scan users around via Bluetooth or WiFi is discouraging for most users and significantly decreases the number of participants, while in hardware-based approaches, the cost is a critical factor that limits large-scale implementations.

Considering those challenges, we used OSN data and a spatiotemporal window model to estimate users' encounters in a trace-based analysis [54]. Each sample geographically indexed from the mobility dataset represents an event limited by a temporal window and spatial area. According to the public data collected, we defined a data sample from social media as a 3-tuple $S_m = (u, p, t)$, where u represents a user $u_i \in U$, t is the timestamp of the sample, and p is the u_i's position defined by latitude and longitude coordinates.

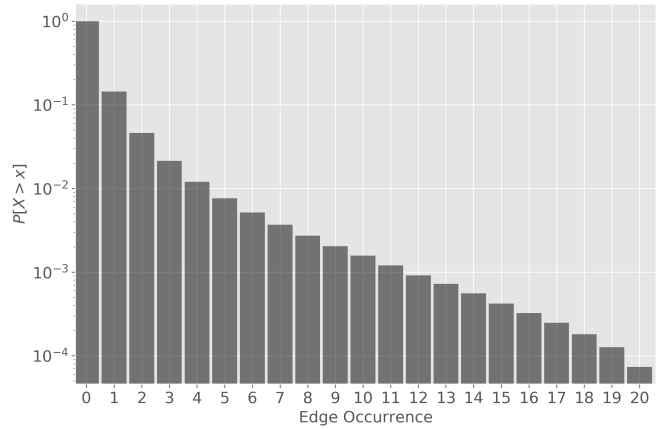
The pause-time t_p, is a piece of information not present in the dataset that represents the period a specific user remains at p location. Considering that the compiled dataset comprises an uninterrupted time series of weeks, the t_p should contemplate different mobility scenarios, from moderate to high mobility during business hours to low mobility outside business hours, at home or at other places of recreation. Our preliminary study showed a mean interval of fifteen minutes between two mobility samples from the same user, and distances of a few meters. In addition, consecutive samples with semantically distinct places, or with a distance greater than 100 meters, are predominantly reported with intervals of less than one hour (93rd percentile).

Since the samples do not provide the pause-time, we assume a temporal threshold T_{th} =60 minutes as the maximum natural interval between two chronologically ordered samples. Therefore, any natural interval between two consecutive samples that exceed the temporal threshold is replaced by $0< t_p \leq 60$ minutes following the Poisson process modeled from the empirical pause time distribution observed in the dataset. The multiple services based on geographical position and physical proximity are common elements of IoT and smart city applications. For this reason, we defined the distance threshold D_{th} to 100 meters in order to simulate the Bluetooth capabilities and assisted Device-to-Device (D2D) communication [46]. The D_{th} defines the maximum geographic distance for an encounter based on physical proximity between two data samples. Accordingly, an opportunistic encounter event is identified when any two data samples s_i and s_j satisfy the following criteria:

- $u_i \neq u_j$;
- $d_g \leq D_{th}$ where d_g is the geographic distance between p_i and p_j;



(a) Encounters per user



(b) Edge occurrences

Figure 7.1: Characterization of the proximity graph.

- t_s $min(t_i + t_{pi}, t_j + t_{pj})$ where t_s is the current simulation time.

The individual mobility of users and their neighbors estimated by means of encounters can provide the proximity graph [33] $G(V, E)$, a temporal graph where E is the set of edges that represent the encounter by geographical proximity among users and V is the set of users online at time t_s. In this analysis, we evaluated the properties of the graph using the T_g time series, where g_i is a snapshot of the graph G at time i.

Figure 7.1a presents the complementary cumulative distribution function of the daily number of encounters per user considering the pause-time defined previously and the variations of D_{th}. The results of the communication range variation did not show changes in behavior concerning the temporal aspects; the curves did not present changes in shape, such that the seasonality and trends are not affected, only the natural variation in the number of encounters per user.

Figure 7.1b presents the results of the investigation of the occurrence of edges in the proximity graph. Edges with more than one recorded occurrence represent re-encounters

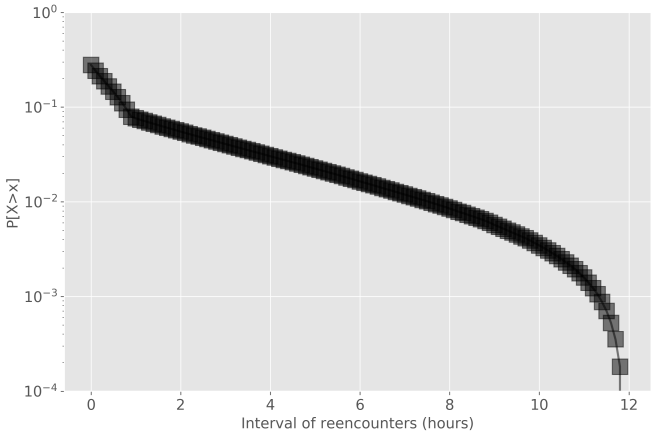


Figure 7.2: Interval of re-encounters.

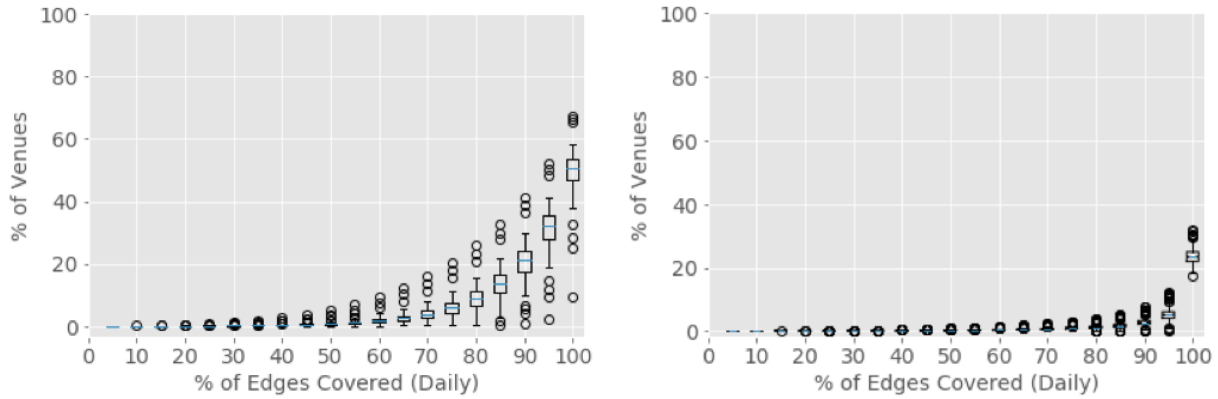
and may characterize the social relationship between nodes or similarity of interests and preferences [36]. Unique encounters among random users represent most edges, which depict the ephemeral characteristics of connections in urban scenarios. However, re-encounters can still be observed.

Figure 7.2 exhibits the distribution of the interval between two consecutive encounters of the same pair of nodes. The joint analyses of these results indicates that, as expected, edges with more than one occurrence represent a small subset of the set of observed edges. Nonetheless, they are characterized by intervals of a few hours between their representations. Short intervals are related to the spatial distribution of these re-encounters, such that these edges, mostly cover one or two directly connected neighborhoods. Also, re-encounters usually occur in small geographic regions or sets of places previously visited by one of the edge's nodes. Such characteristics indicate that users' spatiotemporal preferences show observable regularity for individuals and groups of people.

Notably, when analyzing the number of nodes grouped per hour, the results show curves that follow the trend of use of OSNs, such that most interactions occur in the second half of the day, especially at night, an expected behavior, since the use of OSN is intensified during hours of leisure and recreation.

Furthermore, a complementary analysis of the number of nodes and connected components in the proximity graph indicates, that the set of nodes are distributed in components, usually an order of magnitude smaller. In this way, the observed users are commonly clustered in a small set of spatiotemporal communities. The proper identification of places with a significant regularity of meetings, as well as the spatiotemporal indexing of communities, are critical tasks for the development of user-centered services in smart cities. In this way, this alternative model of urban sensing may add latent or implicit information to the context of geographic areas and other entities of urban scenarios. In order to spatially contextualize the estimated encounters, we evaluated the location associated to the edges of the proximity graph; thus, we group the meetings in semantic places as indicated by the users and present in the metadata of samples.

Thereby, Figure 7.3a shows the coverage of edges as a function of the size of the subset



(a) Edges covered considering the venues.

(b) Edges covered considering the edges of venues' visitors.

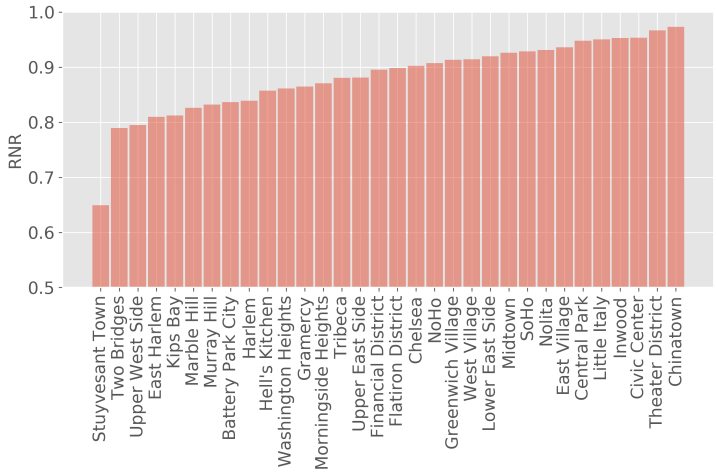
Figure 7.3: Edge coverage as a function of the set of venues.

of observed locations. According to this result, approximately half of the venues can cover the entire set of daily edges observed in the proximity graph. Similarly, we evaluate the edge coverage based on locations using the edges of their visitors, such that Figure 7.3b shows the coverage of all edges using a quarter of the observed places. These results adduce the navigability of the proximity graph guided by the spatiotemporal characteristics, as well as the potential for exploration of the natural spatial agglomeration of users.

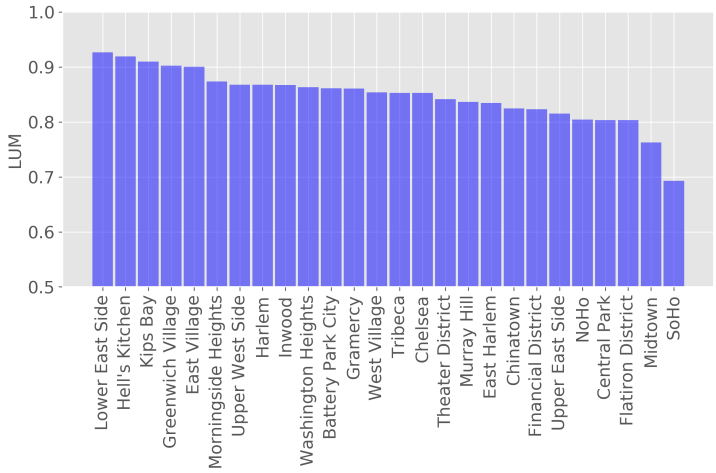
7.4 Sensing Spatiotemporal Aspects

Studies indicate that a city's vitality is related to the activities of its inhabitants throughout the day, and can be compromised by events and characteristics that especially affect pedestrians [37]. One of the fundamental characteristics of the quality of urban life is the diversity offered to the inhabitants regarding activities, services, housing, and others. Accordingly, the spatial distribution of venues and Points-of-Interest (POI) represent another essential characteristic of the dynamics of the urban scenario, as well as social well-being. Jane [61] discussed a set of evaluation techniques to quantify the spatial diversity of urban subareas and argued that the lack of diversity has a potentially adverse influence on the inhabitants' quality of life. For this reason, we investigated the spatial distribution of the observed places and their categories, as well as the effects of the unbalanced distribution. Firstly, we calculated the Residential/Non-Residential balance (RNR), a metric used to estimate the area occupied by constructions classified into two categories: residential and non-residential. The RNR is calculated using the samples recorded in the venues dataset as follows:

$$RNR = 1 - \left| \frac{R_i - NR_i}{R_i + NR_i} \right| \tag{7.1}$$



(a) Residential Non-Residential Ratio (RNR).

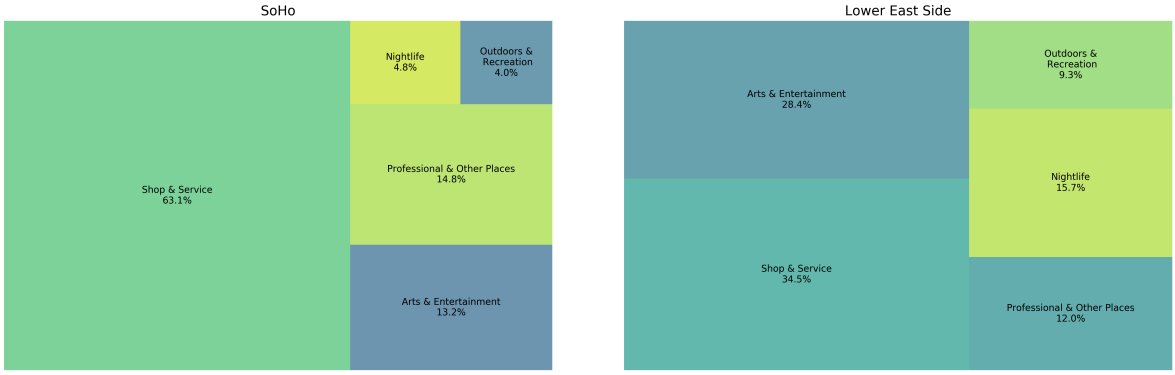


(b) Land Use Mix (LUM).

Figure 7.4: Attributes of neighborhoods.

where i represents the neighborhood and R_i and NR_i are the corresponding portions of places used for residential and non-residential purposes within the neighborhood i, respectively. Thus, low RNR values close to zero represent unbalanced neighborhoods. Figure 7.4a shows the RNR results for the analyzed neighborhoods. Although OSNs are appealing to points of collective interest and places associated with social status, mapping the venues through the content published by users can reflect the zoning characteristics of the city.

A zoning district is a residential, commercial, or manufacturing area of the city based on the government and local authorities' definition of the regulations for land use. The NYC zoning regulations are available online at http://on.nyc.gov/2rBGnmx. For instance, according to the official zoning, the Stuyvesant Town (RNR=0.65) is a mostly residential area, composed mainly of districts for medium-density apartment houses and small commercial areas for local retail needs. However, approximately 19% of its area is intended



(a) SoHo

(b) Lower East Side

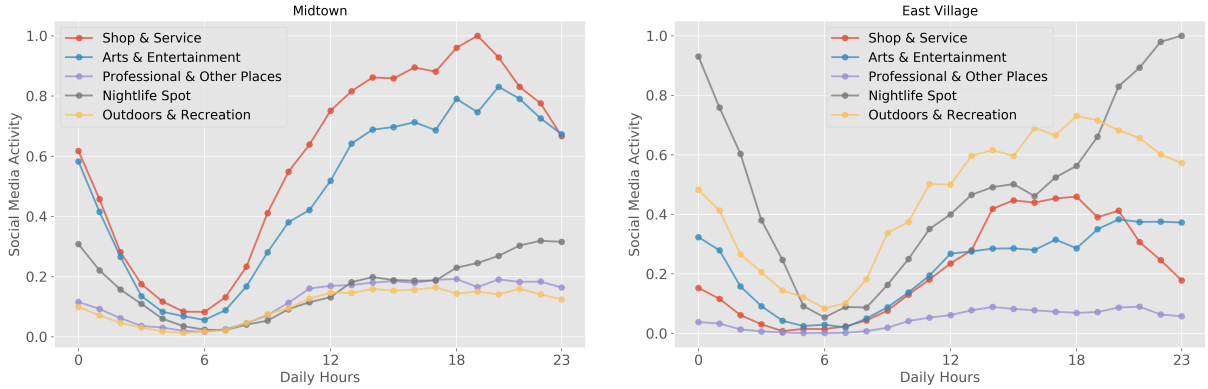
Figure 7.5: Treemaps of venues for SoHo and Lower East Side.

to districts related to industrial and manufacturing activities with heavy industries that generate noise, traffic, or pollutants. Meanwhile, Chinatown (RNR=0.97) is a neighborhood composed mostly of districts for mixed buildings with residential and commercial purposes, in addition to districts for green areas, parks, and landmarks. Thus, the RNR results calculated from the OSN data are consistent with the city's official zoning; as a result, the data may complement the analysis of the districts, adding social and temporal characteristics and potential evaluation of zoning plans.

The Land Use Mix (LUM) is a metric to estimate the diversity of a geographic area through the purposes of use. LUM values close to one represent areas with a significant diversity of use equally distributed, such as areas ablethat provide schools, public squares, offices, shopping malls, among others. Values close to zero represent areas of a specific use, such as industrial districts. The LUM value of a neighborhood is calculated as:

$$LUM = -\sum_{j=1}^{n} \frac{P_{ij} \log(P_{ij})}{\log(n)} \tag{7.2}$$

where P_{ij} is the percentage of use for the purpose j in district i, and n is the number of possible purposes. Figure 7.4b presents the LUM values for the analyzed neighborhoods considering n =5, such that n represents the categories corresponding to professional activities, commerce, services, and leisure. According to the official land use data (publicly available in http://on.nyc.gov/2M3YOCY), the Lower East Side (LUM=0.92) neighborhood is made up of high and medium-density residential districts designated for the construction of apartment buildings, as well as portions destined for regional centers with stores, theaters, offices, and light manufacturing. In contrast, the Midtown (LUM=0.76) neighborhood has many districts for commercial purposes to serve the entire metropolitan region, and SoHo (LUM=0.69) is formed mostly by districts for light and heavy industry. Figures 7.5a and 7.5b represent the proportion of venues, grouped by category, within the neighborhoods with higher and lower LUM values, respectively. The LUM evaluation can overcome the abstraction of venues in only two categories as evaluated in RNR.



(a) Midtown OSN activities

(b) East Village OSN activities

Figure 7.6: Temporal activity registered in the OSN samples.

Researchers and city planners have questioned the elements and the size of the set of categories used in the LUM calculation for different evaluations, making this metric adjustable for different scenarios. Thus, with appropriate generalization, OSN data can contribute to the evaluation of the use of urban land, with updated information and low-cost evaluation. Since the spatial distribution of venues can characterize the offer of services and activities within a geographic area, we used the users' location to evaluate the interests and activities of inhabitants as represented by the visited venues. Hence, we define the popularity of a venue as the number of samples corresponding to the venue during a time window t, and similarly, the predominance of a category of venues is given by the aggregated popularity of their corresponding places during the same period.

Regarding the daily activities, we evaluated the interests of inhabitants for each neighborhood considering the hours of the day and categories of venues as their interests or activities. The results showed distinct registers of daily activities, although consistent with official zoning information. For instance, the Midtown neighborhood shown in Figure 7.6a has among the lowest LUM results, indicating a limited set of uses of the region such that the daily activities of visitors are mostly related to the Shop & Service and Arts & Entertainment categories. In contrast, the East Village neighborhood (LUM=0.9; Figure 7.6b) exhibits a greater variety of popular activities, and recreational venues related to Outdoors & Recreation and Nightlife Spot are more popular.

The results show the balance of activities throughout the day, an indication of the beneficial and proportional distribution of venues. Figure 7.7 shows the hierarchical grouping of categories computed using the Unweighted Pair Group Method with Arithmetic Mean (UPGMA), such that each feature represents the daily mean of the category's popularity, considering t =60 minutes. The dendrogram presents two significant clusters that define a dichotomy of venues with residential and non-residential characteristics. This dichotomy also emphasizes venues and areas more likely to register repeated edges, in such a way that venues and districts with residential characteristics presented daily rates of re-encounters greater than non-residential (p =0.001).

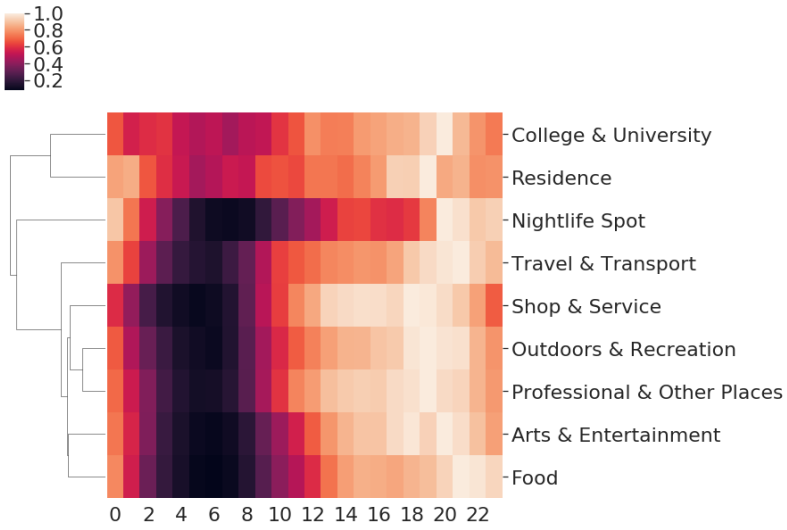


Figure 7.7: Classification of temporal activity.

Additionally, we explored the taxis' trip records, publicly provided by the local government available on http://on.nyc.gov/1EjFCfd. A taxi trip sample describes an individual commutation of an anonymous passenger that includes the pair origin and destination with corresponding timestamps, latitude, and longitude coordinates. Thus, it is possible to compute the temporal profile of a neighborhood and compare it to the mobility dataset. In this way, we analyzed the daily temporal profile of each neighborhood through the hourly popularity of taxi trips and considered the OSN point of view. The obtained correlations are strong for most of the neighborhoods. In particular, during business hours, correlations are stronger, with a mean of p =0.78. It is important to note that a portion of the neighborhoods presented a negative correlation. This observation comprises regions with the predominance of districts intended for industry.

The brokerage [20] is an essential feature in complex networks since it can measure the ability of a node to connect other disconnected nodes. In the proximity graph, the brokerage determines the potential of a user to connect its neighbors not connected. In this way, the brokerage measures the reach of the non-redundant portion of the neighbor's graph. Hristova et al. [55] extended the use of metrics in the context of geospatial networks in which the brokerage can estimate the potential social brokering of venues. According to the authors, the potential brokerage of a venue is calculated by its ego network formed by the union of the social networks of its visitors. Thus, the social brokerage of a place p can be expressed in this way:

$$Brk(p) = |S_n(p)| - \frac{\sum_{u,v \in s_n(p)} e_{u,v}}{S_n(p)} \tag{7.3}$$

where $S_n(p)$ is the social network of the visitors of p. In this work, we assume that the social network of a node is a subgraph of the proximity graph formed by all the neighboring nodes up to the distance of two hops, i.e., we included the friend-of-a-friend relationship. Figure 7.8 shows the brokerage results grouped by the categories of the analyzed venues

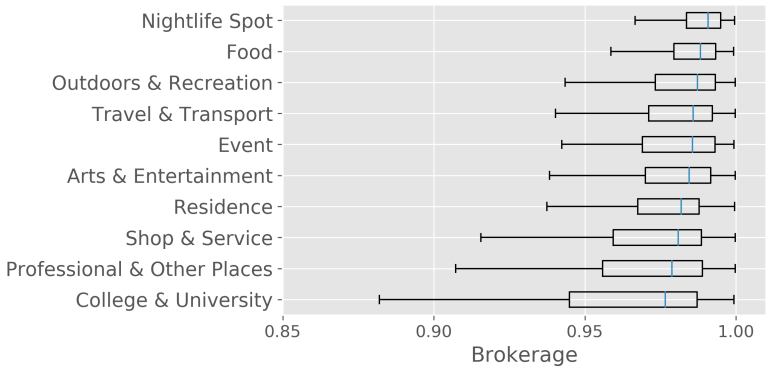


Figure 7.8: Spatial brokerage.

such that values close to zero represent venues with low brokerage potential. The mobility dataset draws an abundant scenario formed by a great variety of users and meetings between them. For this reason, a large portion of the venues presented high brokerage potential. However, the venues corresponding to the categories *College & University* and *Professional & Other Places* presented greater variation and lower values.

The visitors of these corresponding classes showed consistent behavior with multiple visits throughout the dataset, which contributed to the formation of communities of users with well-defined spatial regularity. The evaluation of the social graph of the corresponding visitors showed the gradual densification of the set of edges resulting from the regularity of visits, such that the size of the social network of visitors of the venues in these categories is on an average less than, or equal to half of the other categories. At this point, we measure the social diversity of venues, estimating the homogeneity of their visitors concerning spatial preferences. For each user, we calculate a frequency vector that quantifies the visits of the user, grouped by their respective categories. Then, we calculate the cosine of similarity of a place p using the following equation:

$$Hmg(p) = \frac{\sum_{u,v \in S_n(p)} sim(u,v)}{|S_n(p)(|S_n(p)| - 1)|} \tag{7.4}$$

where $sim(u,v)$ is the cosine of similarity for all pairs of visitors u and v. Figure 7.9 shows the homogeneity results grouped by category, such that values close to zero represent no homogeneity. The *Residence*, *Shop & Service* and *Professional & Other Places* categories, showed that more than half of their pairs have some similarity in visitation preferences. Indeed, these categories usually bring people together daily, such as co-workers and family members. Naturally, the visitors of these places represent more regular edges in the social graph with a high number of re-encounters. Thus, the significant absence of homogeneity in the other categories is indicative of the ephemerality of the meetings in these places, despite their popularity.

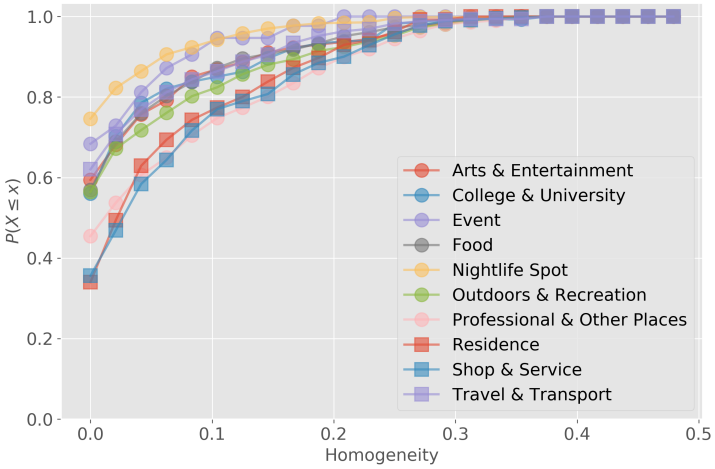


Figure 7.9: Spatial homogeneity.

7.5 Conclusion

In this chapter, we used data analytics to leverage applications and sensing services for Smart Cities exploring mainly Online Social Networks (OSNs). Such methods rely on a pipeline of data analysis that takes advantage of the availability of data from OSNs and governmental data and promote an alternative approach of urban sensing. We evaluated New York City in a case study and investigated spatial, temporal, and social characteristics through the spatiotemporal preferences of OSN users and the geographic properties of the city. From a social perspective, we evaluated a social graph based on the physical proximity of users capable of characterizing meetings and investigated its structural properties. Additionally, we evaluated spatial and temporal characteristics of the city through the places visited by OSN users.

Our study included the spatial distribution of places considering their categories, the official zoning of the city, and calculated city planning metrics such as the proportion of land use for residential, non-residential, mixed, and other purposes. Finally, urban sensing based on OSNs showed significant limitations, which were mainly caused by the sparsity and noise usually observed in users' content available online. Hence, the quality of the results depends on the engagement observed in the OSN, and the data required appropriate treatment for standardization.

Since this paradigm essentially reuses public data from sources that are globally popular, such alternative is scalable, extensible and replicable in different locations, with low implementation costs. The recent advances in machine learning techniques and the increasing popularity of governmental and private open data initiatives, coupled with the ubiquity of OSNs make analytics-based approaches a feasible or complementary alternative to conventional sensing, capable of covering large geographic areas and broadening the sensing spectrum.

Chapter 8

Conclusion and Future Work

8.1 Summary of the Thesis

In this thesis, we present data analyses and mechanisms for urban sensing based on virtual sensing. Using this paradigm, we explored alternative data sources and general-purpose data to provide monitoring of social and spatiotemporal characteristics of urban environments. We present agnostic methodologies, as well as public and open data sources, capable of keeping results and procedures accessible to the scientific community and others.

We primarily explored Online Social Networks (OSNs), which, in turn, showed positive aspects, such as massive data availability, broad capillarity, and variance of sociodemographic groups. However, the use of data provided directly by the user introduces noise, in addition to the limitations related to the nature of the data, such as natural language and unstructured data. Since data are collected for a purpose other than urban sensing, these data need to be adequately addressed previously. On the other hand, local governments are recognizing the importance of data transparency and gradually increasing the supply of official data with information about city management. These actions are fundamental for the development of urban sensing based on complex sensors, and these data are part of the baselines and support the calibration and validation of the proposed systems.

The analyses presented in this study evaluated the dynamics of the characteristics of several cities. From a spatiotemporal point of view, we evaluated the distribution of users across locations in each city, and the results showed seasonality as well as critical temporal variations to the performance of network protocols in applications based on direct communication between devices. These characteristics comprise the sociability dynamics of the inhabitants of the studied cities; thus, we have proposed a message-forwarding protocol for delay-tolerant scenarios. We simulated opportunistic network scenarios and the collective behavior of thousands of users over weeks using real data collected to evaluate the performance of the proposal. The results showed that collective behavior fluctuations could significantly deteriorate the performance of these networks, such that spatiotemporal characteristics can provide relevant contextual information for decision making in these scenarios.

From the point of view of individual behavior, we observed that OSN users are significantly prone to show regularity concerning social, spatial or both aspects. Based on this information, we have proposed an in-network mechanism for caching popular content. The results showed that observing users' regular behaviors is an effective approach for the development of collaborative policies in opportunistic networks. From the collective behavior perspective, we could observe mainly the dynamism of the preferences of groups of users.

The analysis of collective preferences assisted the estimation of urban planning characteristics in different granularities. We analyzed the users' historical location data to identify the activities and services offered in the neighborhoods of the city, and thus estimate properties such as the land use, heterogeneity concerning the places' visitors, and the ratio of residential and non-residential buildings. Additionally, we identified preference shifts in significant portions of inhabitants concerning mobility and spatial distribution. These changes can be characterized and anticipated according to the critical values of temperature.

It is important to emphasize that user engagement and sharing of personal information represent a critical aspect for the development of this sensing paradigm. However, the approaches presented in this thesis do not assume the persistence of these data associated with the identities of individuals, in order that the analyses favor collective behavior rather than the individual one.

8.2 Future Research Directions

Possible future studies and analyses should address the relationships between local governments and citizens, in particular, mechanisms that reinforce their participation considering aspects of privacy and engagement. Thus, we list the following topics as future work.

- Incentive mechanisms. Popular OSNs have significant ubiquity and great user retention potential. These features provide a trusted environment for users to share information and interact with other users and objects. However, users may exhibit altruistic or greedy behavior regarding information and resources shared with third parties. It is necessary to investigate and develop effective methods to empower users regarding their data and the responsibility to share it while keeping them involved. In the context of urban computing, local government plays an active role. Also, sensing systems must be developed to keep citizens motivated to cooperate and contribute to be effective without requiring sensitive information or compromising users' privacy. Therefore, mechanisms for rewards and data transparency, in addition to understanding the boundaries of collaboration between citizens and sensing systems, still need to be investigated.

- Generalization and abstraction mechanisms. Urban sensing should be essentially scalable and replicable. However, there are currently no efficient or widely accepted generalization mechanisms capable of defining abstract representations that facilitate the integration of data sources or the processing of data for the purpose of urban

sensing. Effective abstraction methods can significantly affect how data-driven systems can consume and publish data while ensuring the generalization necessary for continuous integration. Standard and abstract data representations can support the process of data democratization and the research; thus, the alternative data sources from different areassuch as finance, transport, and the environmentcan integrate sensing with a low implementation cost.

References

[1] Saif Al-Sultan, Moath M Al-Doori, Ali H Al-Bayatti, and Hussien Zedan. A comprehensive survey on vehicular ad hoc network. *Journal of network and computer applications*, 37:380–392, 2014.

[2] Eitan Altman, Philippe Nain, Adam Shwartz, and Yuedong Xu. Predicting the impact of measures against p2p networks: transient behavior and phase transition. *IEEE/ACM Transactions on Networking (ToN)*, 21(3):935–949, 2013.

[3] Li-Minn Ang, Kah Phooi Seng, Adamu Murtala Zungeru, and Gerald K Ijemaru. Big sensor data systems for smart cities. *IEEE Internet of Things Journal*, 4(5): 1259–1271, 2017.

[4] Arash Asadi, Qing Wang, and Vincenzo Mancuso. A survey on device-to-device communication in cellular networks. *IEEE Communications Surveys & Tutorials*, 16 (4):1801–1819, 2014.

[5] Luigi Atzori, Antonio Iera, Giacomo Morabito, and Michele Nitti. The social internet of things (siot)–when social networks meet the internet of things: Concept, architecture and network characterization. *Computer networks*, 56(16):3594–3608, 2012.

[6] Saeideh Bakhshi, Partha Kanuparthy, and Eric Gilbert. Demographics, weather and online reviews: A study of restaurant recommendations. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 443–454, 2014. doi: 10.1145/2566486.2568021.

[7] Saeideh Bakhshi, Partha Kanuparthy, and Eric Gilbert. Demographics, weather and online reviews: A study of restaurant recommendations. In *Proceedings of the 23rd international conference on World wide web*, pages 443–454. ACM, 2014.

[8] Saeideh Bakhshi, Partha Kanuparthy, and David A Shamma. If it is funny, it is mean: Understanding social perceptions of yelp online reviews. In *Proceedings of the 18th International Conference on Supporting Group Work*, pages 46–52. ACM, 2014.

[9] Athanasios Bamis, Azzedine Boukerche, Ioannis Chatzigiannakis, and Sotiris Nikoletseas. A mobility aware protocol synthesis for efficient routing in ad hoc mobile networks. *Computer Networks*, 52(1):130–154, 2008.

[10] Sushma Bannur and Omar Alonso. Analyzing temporal characteristics of check-in data. In *Proceedings of the companion publication of the 23rd international conference on World wide web companion*, pages 827–832. International World Wide Web Conferences Steering Committee, 2014.

[11] Xuan Bao, Yin Lin, Uichin Lee, Ivica Rimac, and Romit Roy Choudhury. Dataspotting: Exploiting naturally clustered mobile devices to offload cellular traffic. In *INFOCOM, 2013 Proceedings IEEE*, pages 420–424. IEEE, 2013.

[12] Arianna Bassoli, Johanna Brewer, Karen Martin, Paul Dourish, and Scott Mainwaring. Underground aesthetics: Rethinking urban computing. *IEEE Pervasive Computing*, 6(3), 2007.

[13] Paolo Bellavista, Rebecca Montanari, and Sajal K Das. Mobile social networking middleware: A survey. *Pervasive and Mobile Computing*, 9(4):437–453, 2013.

[14] Azzedine Boukerche and Steve Rogers. Gps query optimization in mobile and wireless networks. In *Proceedings. Sixth IEEE Symposium on Computers and Communications*, pages 198–203. IEEE, 2001.

[15] Azzedine Boukerche and Carl Tropper. A distributed graph algorithm for the detection of local cycles and knots. *IEEE Transactions on Parallel and Distributed Systems*, 9(8):748–757, 1998.

[16] Azzedine Boukerche, Sungbum Hong, and Tom Jacob. An efficient synchronization scheme of multimedia streams in wireless and mobile systems. *IEEE transactions on Parallel and Distributed Systems*, 13(9):911–923, 2002.

[17] Azzedine Boukerche, Cristiano Rezende, and Richard W Pazzi. Improving neighbor localization in vehicular ad hoc networks to avoid overhead from periodic messages. In *GLOBECOM 2009-2009 IEEE Global Telecommunications Conference*, pages 1–6. IEEE, 2009.

[18] David E Boyle, David C Yates, and Eric M Yeatman. Urban sensor data streams: London 2013. *Internet Computing, IEEE*, 17(6):12–20, 2013.

[19] Sergey Brin and Lawrence Page. Reprint of: The anatomy of a large-scale hypertextual web search engine. *Computer networks*, 56(18):3825–3833, 2012.

[20] Ronald S Burt. *Structural holes: The social structure of competition*. Harvard university press, 2009.

[21] Andrew T Campbell, Shane B Eisenman, Nicholas D Lane, Emiliano Miluzzo, and Ronald A Peterson. People-centric urban sensing. In *Proceedings of the 2nd annual international workshop on Wireless internet*, page 18. ACM, 2006.

[22] Mauricio Canals and Francisco Bozinovic. Huddling behavior as critical phase transition triggered by low temperatures. *Complexity*, 17(1):35–43, 2011.

[23] Manuel Castells and Pekka Himanen. *Reconceptualizing development in the global information age*. OUP Oxford, 2014.

[24] Charlie Catlett, Eugenio Cesario, Domenico Talia, and Andrea Vinci. A data-driven approach for spatio-temporal crime predictions in smart cities. In *2018 IEEE International Conference on Smart Computing (SMARTCOMP)*, pages 17–24. IEEE, 2018.

[25] Honglong Chen and Wei Lou. Contact expectation based routing for delay tolerant networks. *Ad Hoc Networks*, 36:244–257, 2016.

[26] Xu Chen, Brian Proulx, Xiaowen Gong, and Junshan Zhang. Social trust and social reciprocity based cooperative d2d communications. In *Proceedings of the fourteenth ACM international symposium on Mobile ad hoc networking and computing*, pages 187–196. ACM, 2013.

[27] Xu Chen, Brian Proulx, Xiaowen Gong, and Junshan Zhang. Exploiting social ties for cooperative d2d communications: A mobile social networking case. *IEEE/ACM Transactions on Networking*, 23(5):1471–1484, 2015.

[28] Zeqiang Chen, Nengcheng Chen, Liping Di, and Jianya Gong. A flexible data and sensor planning service for virtual sensors based on web service. *IEEE Sensors Journal*, 11(6):1429–1439, 2011.

[29] Eunjoon Cho, Seth A Myers, and Jure Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1082–1090. ACM, 2011.

[30] R. I. Ciobanu, R. C. Marin, C. Dobre, V. Cristea, and C. X. Mavromoustakis. Onside: Socially-aware and interest-based dissemination in opportunistic networks. In *2014 Proceedings IEEE Network Operations and Management Symposium*, pages 1–6, 2014.

[31] Marcelo Nogueira Cortimiglia, Alejandro Germán Frank, and Liziane Seben. Tablets: The next disruptive computing technology? *IT Professional*, 15(3):18–25, 2013.

[32] Paolo Costa, Cecilia Mascolo, Mirco Musolesi, and Gian Pietro Picco. Socially-aware routing for publish-subscribe in delay-tolerant mobile ad hoc networks. *Selected Areas in Communications, IEEE Journal on*, 26(5):748–760, 2008.

[33] Felipe D Cunha, Davidysson A Alvarenga, Aline C Viana, Raquel AF Mini, and Antonio AF Loureiro. Understanding interactions in vehicular networks through taxi mobility. In *Proceedings of the 12th ACM Symposium on Performance Evaluation of Wireless Ad Hoc, Sensor, & Ubiquitous Networks*, pages 17–24. ACM, 2015.

[34] Matthew L Daggitt, Anastasios Noulas, Blake Shaw, and Cecilia Mascolo. Tracking urban activity growth globally with big location data. *Royal Society open science*, 3 (4):150688, 2016.

[35] Aveek K Das, Parth H Pathak, Chen-Nee Chuah, and Prasant Mohapatra. Characterization of wireless multi-device users. In *Sensing, Communication, and Networking (SECON), 2015 12th Annual IEEE International Conference on*, pages 327–335. IEEE, 2015.

[36] Pedro OS Vaz de Melo, Aline Carneiro Viana, Marco Fiore, Katia Jaffrès-Runser, Frédéric Le Mouël, Antonio AF Loureiro, Lavanya Addepalli, and Chen Guangshuo. Recast: Telling apart social and random relationships in dynamic networks. *Performance Evaluation*, 87:19–36, 2015.

[37] Marco De Nadai, Jacopo Staiano, Roberto Larcher, Nicu Sebe, Daniele Quercia, and Bruno Lepri. The death and life of great italian cities: a mobile phone data perspective. In *Proceedings of the 25th international conference on world wide web*, pages 413–423. International World Wide Web Conferences Steering Committee, 2016.

[38] Hoang T Dinh, Chonho Lee, Dusit Niyato, and Ping Wang. A survey of mobile cloud computing: architecture, applications, and approaches. *Wireless communications and mobile computing*, 13(18):1587–1611, 2013.

[39] Florian Dörfler and Francesco Bullo. Synchronization in complex networks of phase oscillators: A survey. *Automatica*, 50(6):1539–1564, 2014.

[40] John M Drake and Blaine D Griffen. Early warning signals of extinction in deteriorating environments. *Nature*, 467(7314):456, 2010.

[41] Krittika D'Silva, Anastasios Noulas, Mirco Musolesi, Cecilia Mascolo, and Max Sklar. If i build it, will they come?: Predicting new venue visitation patterns through mobility data. In *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, page 54. ACM, 2017.

[42] Nathan Eagle and Alex (Sandy) Pentland. CRAWDAD dataset mit/reality (v. 2005-07-01). Downloaded from http://crawdad.org/mit/reality/20050701, July 2005.

[43] Roberto Espinosa, Larisa Garriga, Jose Jacobo Zubcoff, and Jose-Norberto Mazón. Linked open data mining for democratization of big data. In *2014 IEEE International Conference on Big Data (Big Data)*, pages 17–19. IEEE, 2014.

[44] Chao Fang, Haipeng Yao, Zhuwei Wang, Wenjun Wu, Xiaoning Jin, and F Richard Yu. A survey of mobile information-centric networking: Research issues and challenges. *IEEE Communications Surveys & Tutorials*, 20(3):2353–2371, 2018.

[45] Gábor Fodor, Stefan Parkvall, Stefano Sorrentino, Pontus Wallentin, Qianxi Lu, and Nadia Brahmi. Device-to-device communications for national security and public safety. *Access, IEEE*, 2:1510–1520, 2014.

[46] Pimmy Gandotra and Rakesh Kumar Jha. Device-to-device communication in cellular networks: A survey. *Journal of Network and Computer Applications*, 71:99–117, 2016.

[47] Raghu K Ganti, Fan Ye, and Hui Lei. Mobile crowdsensing: current state and future challenges. *IEEE Communications Magazine*, 49(11):32–39, 2011.

[48] Liang Ge, Junling Liu, Aoli Zhou, and Hang Li. Crime rate inference using tensor decomposition. In *2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, pages 713–717. IEEE, 2018.

[49] Andre Gomes, Torsten Braun, and Edmundo Monteiro. Enhanced caching strategies at the edge of lte mobile networks. In *IFIP Networking Conference (IFIP Networking) and Workshops, 2016*, pages 341–349. IEEE, 2016.

[50] Janaína Gomide, Adriano Veloso, Wagner Meira Jr, Virgílio Almeida, Fabrício Benevenuto, Fernanda Ferraz, and Mauro Teixeira. Dengue surveillance based on a computational model of spatio-temporal locality of twitter. In *Proceedings of the 3rd International Web Science Conference*, page 3. ACM, 2011.

[51] Google. How mobile search connects consumers to stores, 2016. URL https://goo.gl/63kVcz;Accessedon26-August-2017.

[52] Teerayut Horanont, Santi Phithakkitnukoon, Tuck W Leong, Yoshihide Sekimoto, and Ryosuke Shibasaki. Weather effects on the patterns of people's everyday activities: a study using gps traces of mobile phone users. *PloS one*, 8(12):e81153, 2013.

[53] Theus Hossmann, Thrasyvoulos Spyropoulos, and Franck Legendre. Know thy neighbor: Towards optimal mapping of contacts to social graphs for dtn routing. In *INFOCOM, 2010 Proceedings IEEE*, pages 1–9. IEEE, 2010.

[54] Theus Hossmann, Thrasyvoulos Spyropoulos, and Franck Legendre. Putting contacts into context: Mobility modeling beyond inter-contact times. In *Proceedings of the Twelfth ACM International Symposium on Mobile Ad Hoc Networking and Computing*. ACM, 2011.

[55] Desislava Hristova, Matthew J Williams, Mirco Musolesi, Pietro Panzarasa, and Cecilia Mascolo. Measuring urban social diversity using interconnected geo-social networks. In *Proceedings of the 25th International Conference on World Wide Web*, pages 21–30. International World Wide Web Conferences Steering Committee, 2016.

[56] Solomon M Hsiang, Kyle C Meng, and Mark A Cane. Civil conflicts are associated with the global climate. *Nature*, 476(7361):438–441, 2011.

[57] Xiping Hu, Terry HS Chu, Victor CM Leung, Edith C-H Ngai, Philippe Kruchten, and Henry CB Chan. A survey on mobile social networks: Applications, platforms, system architectures, and future research directions. *IEEE Communications Surveys & Tutorials*, 17(3):1557–1581, 2015.

[58] Pan Hui and Jon Crowcroft. How small labels create big improvements. In *Pervasive Computing and Communications Workshops, 2007. PerCom Workshops' 07. Fifth Annual IEEE International Conference on*, pages 65–70. IEEE, 2007.

[59] Pan Hui, Jon Crowcroft, and Eiko Yoneki. Bubble rap: Social-based forwarding in delay-tolerant networks. *Mobile Computing, IEEE Transactions on*, 10(11):1576–1589, 2011.

[60] Luis G Jaimes, Idalides J Vergara-Laurens, and Andrew Raij. A survey of incentive techniques for mobile crowd sensing. *IEEE Internet of Things Journal*, 2(5):370–380, 2015.

[61] Jacobs Jane. The death and life of great american cities. *New-York, NY: Vintage*, 1961.

[62] Sanem Kabadayi, Adam Pridgen, and Christine Julien. Virtual sensors: Abstracting data from physical sensors. In *Proceedings of the 2006 International Symposium on on World of Wireless, Mobile and Multimedia Networks*, pages 587–592. IEEE Computer Society, 2006.

[63] Nikos Kalatzis, Nicolas Liampotis, Ioanna Roussaki, Pavlos Kosmides, Ioannis Papaioannou, and Stavros Xynogalas. Community context management research and challenges in pervasive and social computing. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2013 IEEE International Conference on*, pages 109–114. IEEE, 2013.

[64] Karthik Kambatla, Giorgos Kollias, Vipin Kumar, and Ananth Grama. Trends in big data analytics. *Journal of Parallel and Distributed Computing*, 74(7):2561–2573, 2014.

[65] Ari Keränen, Jörg Ott, and Teemu Kärkkäinen. The one simulator for dtn protocol evaluation. In *Proceedings of the 2nd international conference on simulation tools and techniques*, page 55. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2009.

[66] Wazir Zada Khan, Yang Xiang, Mohammed Y Aalsalem, and Quratulain Arshad. Mobile phone sensing systems: A survey. *IEEE Communications Surveys & Tutorials*, 15(1):402–427, 2013.

[67] Tim Kindberg, Matthew Chalmers, and Eric Paulos. Guest editors' introduction: Urban computing. *IEEE Pervasive Computing*, 6(3):18–20, 2007.

[68] Hisashi Kurasawa, Hiroshi Sato, Atsushi Yamamoto, Hitoshi Kawasaki, Motonori Nakamura, Yohei Fujii, and Hajime Matsumura. Missing sensor value estimation method for participatory sensing environment. In *Pervasive Computing and Communications (PerCom), 2014 IEEE International Conference on*, pages 103–111. IEEE, 2014.

[69] PJ Lamberson, Scott E Page, et al. Tipping points. *Quarterly Journal of Political Science*, 7(2):175–208, 2012.

[70] Nicholas D Lane, Shane B Eisenman, Mirco Musolesi, Emiliano Miluzzo, and Andrew T Campbell. Urban sensing systems: opportunistic or participatory? In *Proceedings of the 9th workshop on Mobile computing systems and applications*, pages 11–16. ACM, 2008.

[71] Tuan Le, You Lu, and Mario Gerla. Social caching and content retrieval in disruption tolerant networks (dtns). In *Computing, Networking and Communications (ICNC), 2015 International Conference on*, pages 905–910. IEEE, 2015.

[72] Kyunghan Lee, Joohyun Lee, Yung Yi, Injong Rhee, and Song Chong. Mobile data offloading: How much can wifi deliver? *IEEE/ACM Transactions on Networking (ToN)*, 21(2):536–550, 2013.

[73] Xingqin Lin, Jeffrey Andrews, Amitabha Ghosh, and Rapeepat Ratasuk. An overview of 3gpp device-to-device proximity services. *IEEE Communications Magazine*, 52(4):40–48, 2014.

[74] Zhiting Lin and Liang Dong. Clarifying trust in social internet of things. *IEEE Transactions on Knowledge and Data Engineering*, 30(2):234–248, 2018.

[75] Huadong Ma, Dong Zhao, and Peiyan Yuan. Opportunities in mobile crowd sensing. *IEEE Communications Magazine*, 52(8):29–35, 2014.

[76] Kassio Machado, Thiago H Silva, Pedro OS Melo, Eduardo Cerqueira, and Antonio AF Loureiro. Urban mobility sensing analysis through a layered sensing approach. In *Mobile Services (MS), 2015 IEEE International Conference on*, pages 306–312. IEEE, 2015.

[77] Kassio Machado, Azzedine Boukerche, Eduardo Cerqueira, and Antonio AF Loureiro. Long-term spatiotemporal analysis of social media for device-to-device networks. In *2016 IEEE Global Communications Conference (GLOBECOM)*, pages 1–6. IEEE, 2016.

[78] Kassio Machado, Azzedine Boukerche, Pedro O.S. Vaz de Melo, Eduardo Cerqueira, and Antonio A.F. Loureiro. Pervasive forwarding mechanism for mobile social networks. *Computer Networks*, 2016. ISSN 1389-1286.

[79] Kassio Machado, Azzedine Boukerche, Pedro OS Vaz de Melo, Eduardo Cerqueira, and Antonio AF Loureiro. Exploring seasonal human behavior in opportunistic mobile networks. In *2016 IEEE International Conference on Communications (ICC)*, pages 1–6. IEEE, 2016.

[80] Kassio Machado, Azzedine Boukerche, Eduardo Cerqueira, and Antonio AF Loureiro. A socially-aware in-network caching framework for the next generation of wireless networks. *IEEE Communications Magazine*, 55(12):38–43, 2017.

[81] Kássio Machado, Azzedine Boukerche, Eduardo Cerqueira, and Antonio Loureiro. A data-centric approach for social and spatiotemporal sensing in smart cities. *IEEE Internet Computing*, 2018.

[82] Sanjay Madria, Vimal Kumar, and Rashmi Dalvi. Sensor cloud: A cloud of virtual sensors. *IEEE software*, 31(2):70–77, 2013.

[83] James McInerney, Sebastian Stein, Alex Rogers, and Nicholas R Jennings. Breaking the habit: Measuring and predicting departures from routine in individual human mobility. *Pervasive Mob. Comput.*, 9(6):808–822, 2013.

[84] Shike Mei, Han Li, Jing Fan, Xiaojin Zhu, and Charles R Dyer. Inferring air pollution by sniffing social media. In *Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 534–539. IEEE Press, 2014.

[85] Vahid Moosavi and Ludger Hovestadt. Modeling urban traffic dynamics in coexistence with urban data streams. In *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*, page 10. ACM, 2013.

[86] Waldir Moreira and Paulo Mendes. Social-aware opportunistic routing: the new trend. In *Routing in Opportunistic Networks*, pages 27–68. Springer, 2013.

[87] Waldir Moreira and Paulo Mendes. Impact of human behavior on social opportunistic forwarding. *Ad Hoc Networks*, 25:293–302, 2015.

[88] Abderrahmen Mtibaa, Martin May, M Ammar, and C Diot. Peoplerank: Combining social and contact information for opportunistic forwarding. *Proceedings of INFOCOM, San Diego, USA (March 2010)*, 2012.

[89] Ming Ni, Qing He, and Jing Gao. Forecasting the subway passenger flow under event occurrences with social media. *IEEE Transactions on Intelligent Transportation Systems*, 18(6):1623–1632, 2017.

[90] Stavros Nikolaou, Robbert Van Renesse, and Nicolas Schiper. Proactive cache placement on cooperative client caches for online social networks. *IEEE Transactions on Parallel and Distributed Systems*, 27(4):1174–1186, 2016.

[91] Michele Nitti, Maurizio Murroni, Mauro Fadda, and Luigi Atzori. Exploiting social internet of things features in cognitive radio. *IEEE Access*, 4:9204–9212, 2016.

[92] Anastasios Noulas, Salvatore Scellato, Renaud Lambiotte, Massimiliano Pontil, and Cecilia Mascolo. A tale of many cities: universal patterns in human urban mobility. *PloS one*, 7(5):e37027, 2012.

[93] Bei Pan, Yu Zheng, David Wilkie, and Cyrus Shahabi. Crowd sensing of traffic anomalies based on human mobility and social media. In *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 344–353. ACM, 2013.

[94] Bei Pan, Ugur Demiryurek, Chetan Gupta, and Cyrus Shahabi. Forecasting spatiotemporal impact of traffic incidents for next-generation navigation systems. *Knowledge and Information Systems*, pages 1–30, 2014.

[95] Gang Pan, Guande Qi, Wangsheng Zhang, Shijian Li, Zhaohui Wu, and Laurence Tianruo Yang. Trace analysis and mining for smart cities: issues, methods, and applications. *IEEE Communications Magazine*, 51(6):120–126, 2013.

[96] Luca Pappalardo, Filippo Simini, Salvatore Rinzivillo, Dino Pedreschi, Fosca Giannotti, and Albert-László Barabási. Returners and explorers dichotomy in human mobility. *Nature communications*, 6:8166, 2015.

[97] Eric Paulos and Elizabeth Goodman. The familiar stranger: anxiety, comfort, and play in public places. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 223–230. ACM, 2004.

[98] Francisco C Pereira, Ana LC Bazzan, and Moshe E Ben-Akiva. The role of context in transport prediction. *IEEE Intelligent Systems*, 29(1):76–80, 2014.

[99] Charith Perera, Arkady Zaslavsky, Peter Christen, and Dimitrios Georgakopoulos. Context aware computing for the internet of things: A survey. *Communications Surveys & Tutorials, IEEE*, 16(1):414–454, 2014.

[100] Anna-Kaisa Pietilainen and Christophe Diot. CRAWDAD dataset thlab/sigcomm2009 (v. 2012-07-15). Downloaded from http://crawdad.org/thlab/sigcomm2009/20120715, July 2012.

[101] Anna-Kaisa Pietiläinen, Earl Oliver, Jason LeBrun, George Varghese, and Christophe Diot. Mobiclique: middleware for mobile social networking. In *Proceedings of the 2nd ACM workshop on Online social networks*, pages 49–54. ACM, 2009.

[102] Tie Qiu, Baochao Chen, Arun Kumar Sangaiah, Jianhua Ma, and Runhe Huang. A survey of mobile social networks: applications, social characteristics, and challenges. *IEEE Systems Journal*, (99):1–16, 2017.

[103] Priyanka Rawat, Kamal Deep Singh, Hakima Chaouchi, and Jean Marie Bonnin. Wireless sensor networks: a survey on recent developments and potential synergies. *The Journal of supercomputing*, 68(1):1–48, 2014.

[104] Ricky Henry Rawung and Aji Gautama Putrada. Cyber physical system: Paper survey. In *ICT For Smart Society (ICISS), 2014 International Conference on*, pages 273–278. IEEE, 2014.

[105] Anna Izabel João Tostes Ribeiro, Thiago Henrique Silva, Fátima Duarte-Figueiredo, and Antonio AF Loureiro. Studying traffic conditions by analyzing foursquare and instagram data. In *Proceedings of the 11th ACM symposium on Performance evaluation of wireless ad hoc, sensor, & ubiquitous networks*, pages 17–24. ACM, 2014.

[106] Salvador Ruiz-Correa, Darshan Santani, Beatriz Ramirez-Salazar, Itzia Ruiz-Correa, Fatima Alba Rendon-Huerta, Carlo Olmos-Carrillo, Brisa Carmina Sandoval-Mexicano, Angel Humberto Arcos-Garcia, Rogelio Hasimoto-Beltrán, and Daniel Gatica-Perez. Sensecityvity: Mobile crowdsourcing, urban awareness, and collective action in mexico. *IEEE Pervasive Computing*, 16(2):44–53, 2017.

[107] Günther Sagl, Thomas Blaschke, Euro Beinat, and Bernd Resch. Ubiquitous geo-sensing for context-aware analysis: Exploring relationships between environmental and human dynamics. *Sensors*, 12(7):9800–9822, 2012.

[108] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM, 2010.

[109] Marten Scheffer, Jordi Bascompte, William A Brock, Victor Brovkin, Stephen R Carpenter, Vasilis Dakos, Hermann Held, Egbert H Van Nes, Max Rietkerk, and George Sugihara. Early-warning signals for critical transitions. *Nature*, 461(7260): 53–59, 2009.

[110] Daniel Schuster, Alberto Rosi, Marco Mamei, Thomas Springer, Markus Endler, and Franco Zambonelli. Pervasive social context: taxonomy and survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2013.

[111] James Scott, Richard Gass, Jon Crowcroft, Pan Hui, Christophe Diot, and Augustin Chaintreau. CRAWDAD dataset cambridge/haggle (v. 2009-05-29). Downloaded from http://crawdad.org/cambridge/haggle/20090529, May 2009.

[112] M Zubair Shafiq, Lusheng Ji, Alex X Liu, Jeffrey Pang, and Jia Wang. Geospatial and temporal dynamics of application usage in cellular data networks. *IEEE Transactions on Mobile Computing*, 14(7):1369–1381, 2015.

[113] Min Sheng, Chao Xu, Junyu Liu, Jiongjiong Song, Xiao Ma, and Jiandong Li. Enhancement for content delivery with proximity communications in caching enabled wireless networks: architecture and challenges. *IEEE Communications Magazine*, 54 (8):70–76, 2016.

[114] Dongyoun Shin, Daniel Aliaga, Bige Tunçer, Stefan Müller Arisona, Sungah Kim, Dani Zünd, and Gerhard Schmitt. Urban sensing: Using smartphones for transportation mode classification. *Computers, Environment and Urban Systems*, 53: 76–86, 2015.

[115] Fabrício A Silva, Azzedine Boukerche, Thais RM Silva, Linnyer B Ruiz, Eduardo Cerqueira, and Antonio AF Loureiro. Vehicular networks: A new challenge for content-delivery-based applications. *ACM Computing Surveys (CSUR)*, 49(1):11, 2016.

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

[117] Thiago Silva, Pedro Vaz de Melo, Jussara Almeida, Mirco Musolesi, and Antonio Loureiro. You are what you eat (and drink): Identifying cultural boundaries by analyzing food and drink habits in foursquare. In *Proc. of ICWSM*, Ann Arbor, USA, 2014.

[118] Thiago H Silva, Pedro OS Vaz de Melo, Aline Carneiro Viana, Jussara M Almeida, Juliana Salles, and Antonio AF Loureiro. Traffic condition is more than colored lines on a map: Characterization of waze alerts. In *Social Informatics*. Springer, 2013.

[119] Thiago H. Silva, Pedro V. Melo, Jussara Almeida, Aline Viana, Juliana Salles, and Antonio Loureiro. Participatory sensor networks as sensing layers. In *Proceedings of the IEEE Conf. on Social Computing and Networking (SocialCom'14)*, 2014.

[120] Thiago H Silva, Aline Carneiro Viana, Fabrício Benevenuto, Leandro Villas, Juliana Salles, Antonio AF Loureiro, and Daniele Quercia. Urban computing leveraging location-based social network data: a survey. *ACM Computing Surveys*, 2019.

[121] Rodrigo Smarzaro, Tiago França de Melo Lima, and Clodoveu A Davis Jr. Could data from location-based social networks be used to support urban planning? In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 1463–1468. International World Wide Web Conferences Steering Committee, 2017.

[122] Annalisa Socievole, Eiko Yoneki, Floriano De Rango, and Jon Crowcroft. Ml-sor: Message routing using multi-layer social networks in opportunistic communications. *Computer Networks*, 81:201–219, 2015.

[123] Thrasyvoulos Spyropoulos, Konstantinos Psounis, and Cauligi S Raghavendra. Spray and wait: an efficient routing scheme for intermittently connected mobile networks. In *Proceedings of the 2005 ACM SIGCOMM workshop on Delay-tolerant networking*, pages 252–259. ACM, 2005.

[124] Thrasyvoulos Spyropoulos, Konstantinos Psounis, and Cauligi S Raghavendra. Spray and focus: Efficient mobility-assisted routing for heterogeneous and correlated mobility. In *Pervasive Computing and Communications Workshops, 2007. PerCom Workshops' 07. Fifth Annual IEEE International Conference on*, pages 79–85. IEEE, 2007.

[125] Zhou Su and Qichao Xu. Content distribution over content centric mobile social networks in 5g. *IEEE Communications Magazine*, 53(6):66–72, 2015.

[126] Yizhou Sun, Brandon Norick, Jiawei Han, Xifeng Yan, Philip S Yu, and Xiao Yu. Integrating meta-path selection with user-guided object clustering in heterogeneous information networks. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1348–1356. ACM, 2012.

[127] Sandesh Uppoor, Oscar Trullols-Cruces, Marco Fiore, and Jose M Barcelo-Ordinas. Generation and analysis of a large-scale urban vehicular mobility dataset. *IEEE Transactions on Mobile Computing*, 13(5):1061–1075, 2014.

[128] Nikolaos Vastardis and Kun Yang. Mobile social networks: Architectures, social properties, and key research challenges. *IEEE Communications Surveys & Tutorials*, 15(3):1355–1371, 2013.

[129] Anna Maria Vegni and Valeria Loscri. A survey on vehicular social networks. *IEEE Communications Surveys & Tutorials*, 17(4):2397–2419, 2015.

[130] Hui Wang, Xiaofei Wang, Keqiu Li, Jianji Ren, Xiaohong Zhang, and Tianpeng Jiang. A measurement study of device-to-device sharing in mobile social networks based on spark. *Concurrency and Computation: Practice and Experience*, 29(16): e4021, 2017.

[131] Jianquan Wang, Zhaobiao Lv, Zhangchao Ma, Lei Sun, and Yu Sheng. i-net: new network architecture for 5g networks. *IEEE Communications Magazine*, 53(6):44–51, 2015.

[132] Kan Wang, F Richard Yu, Hongyan Li, and Zhengquan Li. Information-centric wireless networks with virtualization and d2d communications. *IEEE Wireless Communications*, 2017.

[133] Xiaofei Wang, Min Chen, Zhu Han, Dapeng Oliver Wu, and Ted Taekyoung Kwon. Toss: Traffic offloading by social network service-based opportunistic sharing in mobile social networks. In *INFOCOM, 2014 Proceedings IEEE*, pages 2346–2354. IEEE, 2014.

[134] Yunsheng Wang, Jie Wu, and Wei-Shih Yang. Cloud-based multicasting with feedback in mobile social networks. *IEEE Transactions on Wireless Communications*, 12 (12):6043–6053, 2013.

[135] Kaimin Wei, Xiao Liang, and Ke Xu. A survey of social-aware routing protocols in delay tolerant networks: Applications, taxonomy and design-related issues. *Communications Surveys & Tutorials, IEEE*, 16(1):556–578, 2014.

[136] Eleanor P Wolf. The tipping-point in racially changing neighborhoods. *Journal of the American Institute of Planners*, 29(3):217–222, 1963.

[137] Huan Yan, Jiaqiang Liu, Yong Li, Depeng Jin, and Sheng Chen. Spatial popularity and similarity of watching videos in large-scale urban environment. *IEEE Transactions on Network and Service Management*, 2018.

[138] Kun Yang, Xueqi Cheng, Liang Hu, and Jianming Zhang. Mobile social networks: state-of-the-art and a new vision. *International Journal of Communication Systems*, 25(10):1245–1259, 2012.

[139] Surender Reddy Yerva, Hoyoung Jeung, and Karl Aberer. Cloud based social and sensor data fusion. In *Information Fusion (FUSION), 2012 15th International Conference on*, pages 2494–2501. IEEE, 2012.

[140] Xiao Yu, Xiang Ren, Yizhou Sun, Quanquan Gu, Bradley Sturt, Urvashi Khandelwal, Brandon Norick, and Jiawei Han. Personalized entity recommendation: a heterogeneous information network approach. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 283–292. ACM, 2014.

[141] Nicholas Jing Yuan, Fuzheng Zhang, Defu Lian, Kai Zheng, Siyu Yu, and Xing Xie. We know how you live: exploring the spectrum of urban lifestyles. In *Proceedings of the first ACM conference on Online social networks*, pages 3–14. ACM, 2013.

[142] Andrea Zanella, Nicola Bui, Angelo Castellani, Lorenzo Vangelista, and Michele Zorzi. Internet of things for smart cities. *IEEE Internet of Things journal*, 1(1): 22–32, 2014.

[143] Zhenxia Zhang, Richard W Pazzi, and Azzedine Boukerche. A mobility management scheme for wireless mesh networks based on a hybrid routing protocol. *Computer Networks*, 54(4):558–572, 2010.

[144] Jiangchuan Zheng, Siyuan Liu, and Lionel M Ni. Effective routine behavior pattern discovery from sparse mobile phone data via collaborative filtering. In *Pervasive Computing and Communications (PerCom), 2013 IEEE International Conference on*, pages 29–37. IEEE, 2013.

[145] Yixian Zheng, Wenchao Wu, Yuanzhe Chen, Huamin Qu, and Lionel M Ni. Visual analytics in urban computing: An overview. *IEEE Transactions on Big Data*, 2(3): 276–296, 2016.

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

[147] Yu Zheng, Tong Liu, Yilun Wang, Yanmin Zhu, Yanchi Liu, and Eric Chang. Diagnosing new york city's noises with ubiquitous data. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 715–725. ACM, 2014.

[148] Ying Zhu, Bin Xu, Xinghua Shi, and Yu Wang. A survey of social-based routing in delay tolerant networks: Positive and negative social effects. *IEEE Communications Surveys & Tutorials*, 15(1):387–401, 2013.

[149] Anneke Zuiderwijk and Marijn Janssen. Open data policies, their implementation and impact: A framework for comparison. *Government Information Quarterly*, 31 (1):17–29, 2014.