

**ESTUDO EM LARGA ESCALA DA DINÂMICA DE
CIDADES E DO COMPORTAMENTO SOCIAL URBANO
USANDO REDES DE SENSORES PARTICIPATIVOS**

THIAGO HENRIQUE SILVA

**ESTUDO EM LARGA ESCALA DA DINÂMICA DE
CIDADES E DO COMPORTAMENTO SOCIAL URBANO
USANDO REDES DE SENSORES PARTICIPATIVOS**

Tese apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Doutor em Ciência da Computação.

ORIENTADOR: ANTONIO ALFREDO FERREIRA LOUREIRO
COORIENTADOR: JUSSARA MARQUES DE ALMEIDA,
PEDRO OLMO STANCIOLI VAZ DE MELO

Belo Horizonte

Maio de 2014

THIAGO HENRIQUE SILVA

**LARGE SCALE STUDY OF CITY DYNAMICS AND
URBAN SOCIAL BEHAVIOR USING PARTICIPATORY
SENSOR NETWORKS**

Thesis presented to the Graduate Program in
Computer Science of the Universidade Fed-
eral de Minas Gerais in partial fulfillment of
the requirements for the degree of Doctor in
Computer Science.

ADVISOR: ANTONIO ALFREDO FERREIRA LOUREIRO
CO-ADVISOR: JUSSARA MARQUES DE ALMEIDA,
PEDRO OLMO STANCIOLI VAZ DE MELO

Belo Horizonte

May 2014

© 2014, Thiago Henrique Silva.
Todos os direitos reservados.

Silva, Thiago Henrique

S586l Large scale study of city dynamics and urban social behavior using participatory sensor networks / Thiago Henrique Silva. — Belo Horizonte, 2014
xxx, 166 f. : il. ; 29cm

Tese (doutorado) — Universidade Federal de Minas Gerais

Orientador: Antonio Alfredo Ferreira Loureiro

1. Computação — Teses. 2. Redes de computadores — Teses. 3. Redes complexas — Teses. I. Orientador. II. Coorientadora. III. Coorientador VI. Título.

CDU 519.6*22



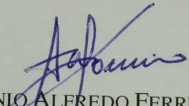
UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

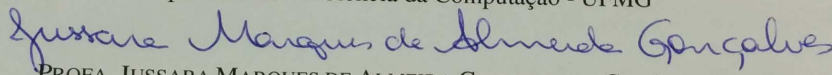
FOLHA DE APROVAÇÃO

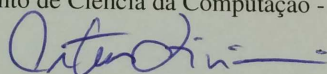
Large scale study of city dynamics and urban social behavior using participatory
sensor networks

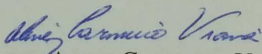
THIAGO HENRIQUE SILVA

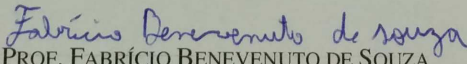
Tese defendida e aprovada pela banca examinadora constituída pelos Senhores:

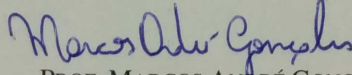

PROF. ANTONIO ALFREDO FERREIRA LOUREIRO - Orientador
Departamento de Ciência da Computação - UFMG

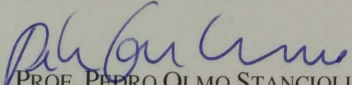

PROFA. JUSSARA MARQUES DE ALMEIDA GONÇALVES - Coorientadora
Departamento de Ciência da Computação - UFMG


PROF. ARTUR ZIVIANI
Laboratório Nacional de Computação Científica - CNPq


PROFA. ALINE CARNEIRO VIANA
INRIA


PROF. FABRÍCIO BENEVENUTO DE SOUZA
Departamento de Ciência da Computação - UFMG


PROF. MARCOS ANDRÉ GONÇALVES
Departamento de Ciência da Computação - UFMG


PROF. PEDRO OLMO STANCIOLI VAZ DE MELO - Coorientador
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 26 de maio de 2014.

To my mother, Aparecida.

Acknowledgments

First of all I thank my family for educating me in the best way possible. Without a strong base I certainly would not be able to get where I got. I dedicate a special thanks to my parents, Aparecida and Hiran, mainly because their great ability in the difficult task of educating, as well as for always supporting me in all my choices. I thank my uncles Juliate (Cabeça) and Altamir, and aunts Maria de Lourdes (Udy) and Simone (Mone) for the support, joy, sympathy, not necessarily in that order. I thank my grandparents, Maria das Dores and Otaviano, which my affection and admiration are inexplicable. I also thank my brothers, Lucas and Gustavo, and all other close members of my family for all the fundamental encouragement in the right moments.

My sincere thanks to my advisors: Prof. Antonio Loureiro, Prof. Jussara Gonçalves, and Prof. Pedro Vaz de Melo. Those professors, besides the extreme competence, were essential in different ways. Loureiro by his singular guidance through the most diverse situations, leading me not just towards a PhD, but to become a full scientist and a better person. I will never forget the priceless support whenever I needed, and all the situations that I saw his true desire to see my growth. Jussara for her compromise of doing good science and her objectiveness in all moments that we had to make crucial decisions. Jussara's ability for dealing with problems is so big that whenever I talked to her I got the impression that everything was solvable and simple. And last, but not least, I thank Pedro, by the fundamental support in the research topic I chose for my thesis. Pedro's openness for new ideas, creativity, and friendship will always be remembered. Together, we all built a great team! Thanks also to other professors of the Department of Computer Science for the high quality classes. Special thanks for Prof. Dorgival, Prof. José Marcos, and Prof. Virgílio. They were important motivators to continue in academic life.

I almost don't have words to express my gratitude to all my friends. Friends from the Department of Computer Science at UFMG (sorted alphabetically): Douglas, Everthon, Fernando, Flávio, Guidoni, Guilherme, Heitor, Leandro, Marcelo, Rafaels (Colares and Santin), Rodolfo, Sumika, and Vinícius (Sapulha). Friends from INRIA/PARIS: Dudu, Felipe, Gra, Indayara, Luiz, and Paula. Friends from the University of Birmingham: Ana, Andrea (Big

Dog), Antonio, Ashley, Francesca, Manlio, Raul, Veljko, and Yusuke. Friends I made during the Telecom Italia internship: Felippo, Juliana, Kiara, Giusy, and Pepe. I cannot forget to thank my old school buddies: Bibinha, Brenudo, Cláudio, Diego, Ighor, Leo, Lilian, Mangões, Monoh, Nádia, Pedro Gabriel, and Xisto. All other friends must be considered mentioned, sorry if I didn't mention your name, but please consider yourself mentioned. Thank you all for the valuable support and friendship. I hope we keep drinking good beers together for many other times.

Finally, I would like to thank Twinings for making an awesome tea, specially Earl Grey, Cabral for keeping his temple for philosophical discussions, Stad Jever for all the good moments, A Obra for all the good songs, CNPq for the essential financial support, and all the other people who were not “tcheba” and contributed in some way to my personal and professional development.

*“You take delight not in a city’s seven or seventy wonders,
but in the answer it gives to a question of yours.”*
(Italo Calvino)

Resumo

A disponibilidade ubíqua de tecnologia computacional, como smartphones, tablets e outros dispositivos facilmente portáteis, bem como a adoção mundial de sites de redes sociais tornam cada vez mais possível estar conectado e compartilhar dados de forma contínua para o processo de publicação de informações massivamente distribuído chamado rede de sensor participativo (RSP). Neste cenário, as pessoas agem como sensores sociais, voluntariamente fornecendo dados que capturam as suas experiências de vida diária, e oferecendo diversas observações, tanto no mundo físico (por exemplo, localização) quanto no mundo on-line (por exemplo, eventos). Esta grande quantidade de dados sociais podem fornecer novas formas de informações valiosas que não estão disponíveis no momento, a esta escala, utilizando métodos de coleta de dados tradicionais, e podem ser usadas para melhorar os processos de tomada de decisão. Nesta tese, mostramos que RSPs, por exemplo as derivadas do Instagram, Foursquare, e Waze podem atuar como valiosas fontes de sensoriamento em larga escala, proporcionando acesso a características importantes do comportamento social urbano de forma mais rápida do que os métodos tradicionais.

O objetivo desta tese é a compreensão das propriedades de RSP, e mostrar como elas podem ser usadas para o estudo da dinâmica de cidades e do comportamento social urbano. Nós estudamos redes de sensor participativos derivados de diferentes sistemas, e demonstramos como modelar e extrair conhecimento a partir delas, de forma individual e simultaneamente. Nossos resultados mostram que PSNs têm o potencial de tornarem-se ferramentas fundamentais para apoiar análises em larga grande escala e (quase) tempo real dos diferentes aspectos da dinâmica de cidades e do comportamento social urbana.

Palavras-chave: Sensoriamento participativo, mídia social, big data, redes sociais móveis, dinâmica de cidades, comportamento social urbano.

Abstract

The ubiquitous availability of computing technology such as smartphones, tablets and other easily portable devices, and the worldwide adoption of social networking sites make it increasingly possible for one to be connected and continuously share data to this massively distributed information publishing process called participatory sensor network (PSN). In this scenario, people act as social sensors, voluntarily providing data that capture their daily life experiences, and offering diverse observations on both the physical world (e.g., location) and the online world (e.g., events). This large amount of social data can provide new forms of valuable information that are currently not available, at this scale, by any traditional data collection methods, and can be used to enhance decision making processes. In this thesis we show that PSNs, for instance those derived from Instagram, Foursquare, and Waze can act as valuable sources of large scale sensing, providing access to important characteristics of urban social behavior much more quickly than traditional methods.

The goal of this thesis are the understanding of properties of PSN, and show how they can be used to the study of city dynamics and urban social behavior. We study participatory sensor networks derived from different systems, and demonstrate how to model and extract knowledge from them, individually and concurrently. Our results show that PSNs have the potential to become fundamental tools to support large scale and near real time analyses of different aspects of dynamics of cities and urban social behavior.

Palavras-chave: participatory sensing, social media, big data, mobile social networks, city dynamics, urban social behavior.

List of Figures

| | | |
|------|--|----|
| 2.1 | Representativeness of authors and institutions. | 9 |
| 3.1 | Participatory sensor network illustration. | 25 |
| 3.2 | Overview of participatory sensor network components. | 26 |
| 4.1 | All sensed locations. The number of locations n per pixel is given by the value of ϕ displayed in the colormap, where $n = 2^\phi - 1$ | 31 |
| 4.2 | Temporal variations in the number of check-ins per continent. | 32 |
| 4.3 | [Best viewed in color]. All sensed locations in six international cities (Foursquare datasets). The number of check-ins in each area is represented by a heatmap. The color range from yellow to red (high intensity). | 33 |
| 4.4 | The complementary cumulative distribution function of the number of check-ins per venue. | 34 |
| 4.5 | The number of locations that were active in a given day. | 35 |
| 4.6 | The distribution of the inter-event times between consecutive check-ins of one popular venue of each dataset. | 36 |
| 4.7 | Weekly location sharing patterns. | 37 |
| 4.8 | Weekdays and weekend location sharing patterns. | 37 |
| 4.9 | All photos shared. Number of photos n per pixel obtained from the value of ϕ shown in the figure, where $n = 2^\phi - 1$ | 39 |
| 4.10 | Temporal variation of the number of photos shared by continent. | 39 |
| 4.11 | Spatial coverage of Instagram in eight cities for all shared photos. The number of pictures in each area is represented by a heat map, where the scale varies from yellow to red (more intense activity). | 40 |
| 4.12 | Example of identification of a quadrant. | 40 |
| 4.13 | Distribution of the number of photos in quadrants. | 41 |
| 4.14 | Temporal variation in the number of sensed quadrants. | 42 |
| 4.15 | Distribution of the time interval between shared photos in a popular quadrant. | 42 |

| | | |
|------|--|----|
| 4.16 | Mean probability of obtain a picture in the next 1-minute, 15-minutes, 30-minutes, and 60-minutes, for eight popular areas during the dawn, morning, afternoon, and night. | 44 |
| 4.17 | Temporal photo sharing pattern. | 45 |
| 4.18 | Photo sharing throughout the day in Rio de Janeiro, Sao Paulo, Osaka, Tokyo, Barcelona, Madrid, Chicago and New York City. | 47 |
| 4.19 | Distribution of the number of photos shared by people. | 47 |
| 4.20 | Distribution of the geographical distance between consecutive pictures of the same person. | 47 |
| 4.21 | Contribution of nodes, distance traveled, and coverage. | 48 |
| 4.22 | Overview of reported alerts. | 50 |
| 4.23 | All sensed locations. The number of locations n per pixel is given by the value of ϕ displayed in the colormap, where $n = 2^\phi - 1$ | 50 |
| 4.24 | Spatial coverage of Waze in Rio de Janeiro. | 51 |
| 4.25 | Distribution of the number of alerts. | 51 |
| 4.26 | Time intervals between consecutive alerts, not necessarily done by the same user. | 53 |
| 4.27 | General temporal sharing pattern (all locations). | 54 |
| 4.28 | Alerts sharing throughout the day in different cities around the world. | 54 |
| 4.29 | CCDF of the number of shared alerts (same user). | 55 |
| 4.30 | Distribution of the geographical distance between consecutive data of the same person. | 56 |
| 4.31 | All sensed locations in three populous cities. The number of check-ins in each area is represented by a heat map. The color range from yellow to red (high intensity). | 57 |
| 4.32 | Analysis of classes of users. | 59 |
| 4.33 | Grids for the areas of New York, Sao Paulo, and Tokyo. | 60 |
| 4.34 | Correlation of popularity of sectors inside cities. | 61 |
| 4.35 | Spearman correlation of popularity between cities. | 61 |
| 4.36 | Temporal sharing pattern for Instagram and Foursquare – new and old datasets. | 62 |
| 4.37 | Cross-correlation between Instagram-New and Foursquare-New datasets, during weekday and weekend. | 62 |
| 4.38 | Temporal sharing pattern of Instagram and Foursquare for New York, Sao Paulo, and Tokyo during weekdays. | 63 |
| 4.39 | Transition graphs – New York. | 64 |
| 4.40 | Transition graphs – Sao Paulo. | 65 |
| 4.41 | Transition graphs – Tokyo. | 65 |

| | | |
|------|---|----|
| 5.1 | Observed transitions occurrences sorted in a descending order for NY city. Periods: weekday and weekend during the day and night. | 71 |
| 5.2 | Observed transitions occurrences sorted in a descending order for Tokyo. Periods: weekday and weekend during the day and night. | 71 |
| 5.3 | Histogram of random generated transitions for NY with a Normal fitting. | 73 |
| 5.4 | Histogram of random generated transitions for Tokyo with a Normal fitting. | 73 |
| 5.5 | The City Image of London for different periods. | 74 |
| 5.6 | The City Image of Kuwait for different periods. | 74 |
| 5.7 | The City Image of Belo Horizonte for different periods. | 75 |
| 5.8 | The City Image of Chicago for different periods. | 75 |
| 5.9 | The City Image of Surabaya for different periods. | 75 |
| 5.10 | The City Image of New York for different periods. | 76 |
| 5.11 | The City Image of Sydney for different periods. | 76 |
| 5.12 | The City Image of Tokyo for different periods. | 76 |
| 5.13 | The City Image of cities in different regions of the world during the day on weekdays. | 77 |
| 5.14 | Heatmap of the number of check-ins, where the color range from yellow to red (high intensity). | 78 |
| 5.15 | Node degree - For two cities in different countries. Each node color represents an specific category of places. Blue=Arts& Entertainment; Red = College & Education; Light Green = Food; Yellow = Home; Green Moss = Office; Purple = Nightlife Spot; White = Great Outdoors; Beige = Shop & Service; Grey = Travel spot; Cyan = no category. | 81 |
| 5.16 | Node Betweenness - For two cities in different countries. Colors legend: see caption of Figure 5.15. | 82 |
| 5.17 | Node Closeness - For two cities in different countries. Colors legend: see caption of Figure 5.15. | 82 |
| 5.18 | Top 50 edge weights and node degrees (hubs); stars represent hubs, black arrows edges, and black circles self-loops. Featured places (nodes) and transitions (edges). | 83 |
| 5.19 | Points of interest of Belo Horizonte. | 87 |
| 5.20 | Sights identified in different datasets. | 88 |
| 5.21 | The temporal photo sharing pattern for different types of POIs. | 90 |
| 5.22 | Examples of possible area classifications. | 91 |
| 5.23 | Check-ins estimation for different times and type of places. | 92 |
| 5.24 | Frequency of check-ins at all subcategories of the three analyzed classes. The names of some places are abbreviated but the semantics of the names is preserved. | 97 |
| 5.25 | Correlation of preferences between countries. | 98 |

| | | |
|------|--|-----|
| 5.26 | Correlation of preferences between cities. | 99 |
| 5.27 | Areas of cities taken into consideration: London/England; New York/USA; and Tokyo/Japan. | 100 |
| 5.28 | Correlation of preferences in regions of London, NYC and Tokyo. | 101 |
| 5.29 | Number of check-ins throughout the hours of the day in different countries (WD = weekday; WE = weekend). | 102 |
| 5.30 | Number of check-ins throughout the hours of the day in different American cities (WD = weekday; WE = weekend). | 102 |
| 5.31 | Clustering results for countries, cities, and regions inside cities. | 105 |
| 5.32 | Clustering results for cities on weekend, considering only the Drink class. | 106 |
| 5.33 | The cultural map of the World given by the World Values Survey [Inglehart and Welzel, 2010]. | 106 |
| | | |
| 6.1 | Sensing layers for a city. Each layer gives information about a specific aspect of the city. | 113 |
| 6.2 | Overview of participatory sensor networks with the concept of sensing layers. | 113 |
| 6.3 | Illustration of sharing data in three PSNs throughout the time, resulting in layers. | 118 |
| 6.4 | Combination by location. | 118 |
| 6.5 | Combination by users. | 118 |
| 6.6 | Illustration of flow graph creation from one single layer, and also from multiple layers. | 126 |
| 6.7 | Illustration of new layers creation from the picture of places layer. | 127 |
| 6.8 | All identified sights with Foursquare and Instagram datasets. | 129 |
| 6.9 | Examples of New York City census tracts. | 131 |
| | | |
| A.1 | The general City Image, which does not consider different periods separately of all cities. | 157 |
| A.2 | The general City Image, which does not consider different periods separately of all cities. | 158 |
| | | |
| B.1 | Dendrogram plots for the binary cluster tree of 30 different cities, in two different time periods. | 160 |
| | | |
| C.1 | General metrics for all similarity networks. | 164 |

List of Tables

| | | |
|-----|---|-----|
| 4.1 | Dataset information. | 30 |
| 4.2 | Dataset information. Note that Foursquare-Crawled was already analyzed in Section 4.1, and Instagram-OLD in Section 4.2. | 57 |
| 5.1 | Distribution of check-ins across the selected cities. | 74 |
| 5.2 | Centrality metrics for NY during the day and night. | 79 |
| 5.3 | Percentage of centrality metrics for all categories of places for BH (day and night). D=degree, B=betweenness, C=closeness. | 84 |
| 5.4 | Percentage of centrality metrics for all categories of places for NY (day and night). D=degree, B=betweenness, C=closeness. | 84 |
| 5.5 | Percentage of centrality metrics for all categories of places for Tokyo (day and night). D=degree, B=betweenness, C=closeness. | 84 |
| 5.6 | The Spearman’s rank correlation coefficient ρ (and its respective p-value) between the rank of similar countries generated from WVS and by our approach. | 107 |
| 5.7 | Summary of the approaches applied to verify our results. | 109 |
| 6.1 | Data stream describing users activity in three different PSNs: Foursquare, Waze, and Instagram. | 117 |
| 6.2 | General sentiment per groups of tracts | 131 |
| B.1 | Clustering results for weekday during the day. | 161 |
| B.2 | Clustering results for weekend during the night. | 161 |

List of Acronyms

| | |
|----------------|--|
| ANEW | Affective Norms for English Words |
| CCDF | Complementary Cumulative Distribution Function |
| CDF | Cumulative Distribution Function |
| DJIA | Down Jones Industrial Average |
| FMM | Friendship-based Mobility Model |
| GPOM | Google-Profile Mood State |
| JMA | Japan Meteorological Agency |
| LBSN | Location-Based Social Network |
| LDA | Latent Dirichlet Allocation |
| MAPE | Mean Absolute Percentage Error |
| OF | OpinionFinder |
| OR | Odds Ratio |
| PCA | Principal Component Analysis |
| POI | Point of Interest |
| PSN | Participatory Sensor Network |
| PSS | Participatory Sensing System |
| SNA | Social Network Analysis |
| ubicomp | Ubiquitous Computing |
| WD | Weekday |

WE Weekend

Contents

| | |
|---|--------------|
| Acknowledgments | xi |
| Resumo | xv |
| Abstract | xvii |
| List of Figures | xix |
| List of Tables | xxiii |
| List of Acronyms | xxv |
| 1 Introduction | 1 |
| 1.1 Motivation | 1 |
| 1.2 Objective | 3 |
| 1.3 Contributions | 3 |
| 1.4 Work Organization | 5 |
| 2 Background | 7 |
| 2.1 Ubiquitous Computing | 7 |
| 2.1.1 Mark Weiser’s Visions | 7 |
| 2.1.2 Ubicomp Today | 9 |
| 2.1.3 Context-Aware Computing | 11 |
| 2.2 City Dynamics and Urban Social Behavior | 11 |
| 2.2.1 Mobility Patterns | 12 |
| 2.2.2 Understanding Cities | 14 |
| 2.2.3 Social Patterns | 16 |
| 2.2.4 Event Detection | 18 |
| 2.2.5 Human Behavior | 20 |
| 2.3 Discussion | 21 |

| | | |
|----------|--|-----------|
| 3 | Humans in the Sensing Process | 23 |
| 3.1 | Participatory Sensor Networks | 23 |
| 3.2 | Challenges | 26 |
| 4 | Properties of PSNs | 29 |
| 4.1 | PSN from Location Sharing Services | 29 |
| 4.1.1 | Data Description | 29 |
| 4.1.2 | Network Coverage | 30 |
| 4.1.3 | Sensing Interval | 34 |
| 4.1.4 | Seasonality | 36 |
| 4.2 | PSN from Photo Sharing Services | 38 |
| 4.2.1 | Data Description | 38 |
| 4.2.2 | Network Coverage | 38 |
| 4.2.3 | Sensing Interval | 42 |
| 4.2.4 | Seasonality | 45 |
| 4.2.5 | Node Behavior | 46 |
| 4.3 | PSN from Traffic Alert Services | 49 |
| 4.3.1 | Data Description | 49 |
| 4.3.2 | Network Coverage | 50 |
| 4.3.3 | Sensing Interval | 52 |
| 4.3.4 | Seasonality | 53 |
| 4.3.5 | Node Behavior | 55 |
| 4.4 | Comparing PSNs from different systems | 56 |
| 4.4.1 | User Behavior | 58 |
| 4.4.2 | Popularity of Areas | 58 |
| 4.4.3 | Routines and the Data Sharing | 60 |
| 4.4.4 | Mapping Transitions | 63 |
| 4.5 | Final Considerations | 66 |
| 4.5.1 | Data Limitations and Bias | 66 |
| 4.5.2 | Results Discussion | 67 |
| 5 | Understanding city dynamics and urban social behavior | 69 |
| 5.1 | Visualizing the Invisible Image of Cities | 70 |
| 5.1.1 | Transition Graph | 70 |
| 5.1.2 | Preferred and Rejected City Transitions | 71 |
| 5.1.3 | Building the City Images | 73 |
| 5.2 | Insights into People Movement Patterns | 77 |

| | | |
|----------|---|------------|
| 5.2.1 | Using Centrality Metrics | 78 |
| 5.2.2 | Network Visualization | 81 |
| 5.2.3 | Information Summarization | 83 |
| 5.3 | Points of Interest | 85 |
| 5.3.1 | POI Identification Algorithm | 85 |
| 5.3.2 | The Vibe of POIs | 89 |
| 5.4 | Socio-Economic Aspects | 90 |
| 5.5 | Cultural Differences | 93 |
| 5.5.1 | Extracting Cultural Preferences | 94 |
| 5.5.2 | Extraction of Cultural Signatures | 98 |
| 5.5.3 | Identifying Cultural Boundaries | 104 |
| 5.6 | Discussion | 108 |
| 6 | Participatory Sensor Networks as Sensing Layers | 111 |
| 6.1 | Sensing Layers | 111 |
| 6.1.1 | Basic Concepts | 112 |
| 6.1.2 | Usefulness of Sensing Layers | 115 |
| 6.1.3 | A Formal Model for Sensing Layers | 116 |
| 6.1.4 | Issues of Data from Multiple Layers | 119 |
| 6.1.5 | Discussion | 120 |
| 6.2 | Processing Sensing Layers | 121 |
| 6.2.1 | Operations | 121 |
| 6.2.2 | Processing Strategies | 126 |
| 6.3 | Applications Using the Sensing Layers Framework | 128 |
| 6.3.1 | Identification of Sights | 128 |
| 6.3.2 | Economic-Cultural Analysis of Regions | 129 |
| 6.3.3 | Discussion | 133 |
| 7 | Conclusions and Future Work | 135 |
| 7.1 | Conclusions | 135 |
| 7.2 | Future Work | 137 |
| 7.3 | Comments on Publications | 138 |
| 7.3.1 | Contributions from the Thesis | 138 |
| 7.3.2 | Other Publications | 140 |
| | Bibliography | 143 |
| | Appendix A General City Images | 157 |

Appendix B Quantitative Comparison of Cities 159

Appendix C Cultural Analysis of Individuals 163

Chapter 1

Introduction

1.1 Motivation

At the beginning, there were mainframes, shared by a lot of people. Then came the personal computing era, when a person and a machine have a close relationship with each other. Nowadays we are witnessing the beginning of the ubiquitous computing (ubicomp) era, when technology recedes into the background of our lives [Weiser and Brown, 1996; Krumm, 2009].

Mark Weiser, in his classical article entitled “The computer for the 21st century” [Weiser, 1991], popularized the concept of ubiquitous computing, which envisions the availability of a computing environment for anyone, anywhere, and at any time. It may involve many wirelessly interconnected devices, not just traditional computers, such as desktops or laptops, but may also include all sorts of objects and entities such as pens, mugs, phones, shoes, and many others. Although this is not the reality yet, much has been done in this direction in the past 20 years after the publication of Weiser’s seminal paper, and the key ingredients are evolving in a favorable direction for it. Observe, for example, the increasing number and popularization of numerous types of portable devices.

A fundamental step to achieve Weiser’s vision is to sense the environment. The research in wireless sensor networks (WSNs) has provided several tools, techniques and algorithms to solve the problem of sensing in limited size areas, such as forests or factories [Yick et al., 2008; Akyildiz et al., 2002]. However, traditional WSNs have their limitations, for example the high costs related to achieve very large coverage spaces, such as metropolises size areas. Consider, for instance, the challenges to build and maintain such networks.

Mobile phones play a fundamental role in today’s technologically-advanced community allowing people to communicate (almost) anywhere in the world and share all kinds of

contextual information (e.g., location and opinion). Modern mobile phones, namely smartphones, are the new frontier for accessing the Internet and the World Wide Web. They are being manufactured with an increasing number of powerful embedded sensors of different categories (e.g., GPS, accelerometer, microphone, camera, gyroscope), enabling a variety of new applications and services. Indeed, smartphones are being used for many personal sensing applications, such as for monitoring physical exercises, and for wide participatory sensing systems, which are not limited to a particular individual (e.g., traffic conditions and noise pollution) [Lane et al., 2010].

Participatory sensing systems (PSSs), such as Instagram¹, Foursquare², Waze³, and Weddar⁴, combine features of online social networks with location-based services. This type of system has started to create new virtual environments that integrate the user interactions and, probably because of that, are becoming very popular. For example, in 2013 Foursquare reported 40 million users [Foursquare, 2013], Instagram 150 million users [Instagram, 2013], and Waze 50 million users [Goel, 2013].

PSSs have been driven by one important aspect: the information the users share, in particular location-related information [Smith et al., 2005]. From a participatory sensing system, it is possible to derive a participatory sensor network (PSN) [Burke et al., 2006]. In this type of network, the users' mobile device plays an important role. Individuals carrying these devices are able to sense the environment and share relevant observations. Thus, each node in a PSN consists of a user with a mobile device. Each PSN provides access to data related to certain aspects of a pre-defined geographic region. For instance, in a PSN derived from Waze, users report traffic conditions, in the one derived from Foursquare, users can share their actual location associated with a specific category of place (e.g., restaurant).

Participatory sensor networks enables the observations of the actions of hundreds millions of people in large scale urban areas in (near) real time and over extended periods of time. This opens an unprecedented opportunity to revolutionize the way social science is done. Unlike traditional methods that rely on survey data, new techniques can be designed to exploit participatory data, which is much cheaper, more dynamic as it reflects current situations in (near) real time, and, more important, can easily reach planetary scale. Moreover, as we argue here, such participatory sensing applications may have the potential to be a fundamental tool to better understand human urban interaction in the future, leveraging our awareness to different aspects of our lives in urban scenarios.

¹<http://www.instagram.com>.

²<http://www.foursquare.com>.

³<http://www.waze.com>.

⁴<http://www.weddar.com>.

1.2 Objective

The main objective of this thesis is to answer the question: Can we use PSNs to perform large scale and near real time study of city dynamics and urban social behavior? To that end, a fundamental step is understand the properties of participatory sensor networks. We aim to analyze PSNs derived from different systems. After that, our goal is to show how to model and extract knowledge from PSNs, individually and concurrently. Thus, we tackle the main objective of this thesis answering three different questions:

- **What are the properties of PSN?** Despite the concept of participatory sensor network be relatively old, coined in 2006 [Burke et al., 2006], very few properties of this type of network are known. With that, our goal is to investigate the properties of PSNs in order to understand its challenges and usefulness;
- **How can we use PSNs?** Our goal here is use the properties we extract from the analysis process to the design of techniques and methodologies to the study of city dynamics and urban social behavior. First, we want a model that enables the knowledge extraction from a PSN individually. Based on this model, the aim is to propose techniques and methodologies to demonstrate the usefulness of PSNs to the study of city dynamics and urban social behavior;
- **Can we combine data from different PSNs to infer new information?** Data from different PSNs can be considered as “sensing layers”, providing data on various aspects of a predefined geographic region. Given that, our main objective is to show the usefulness of using multiple PSNs to the extraction of new information. For that, we aim to define the concept of sensing layers. Next, we envision the proposition of a framework that enables the analysis and exploration of multiple layers simultaneously. Finally, we aim to present applications that use the proposed framework.

1.3 Contributions

Our main contributions can be summarized in:

1. **Characterization and analysis of participatory sensor networks properties:** We have characterized and analyzed properties of three different types of PSNs: (1) photo sharing services, particularly Instagram; (2) location sharing services, particularly Gowalla, Brightkite, and Foursquare; (3) and traffic alert services, particularly Waze. Among the results, we showed the planetary scale of those networks, as well as

the highly unequal frequency of data sharing, both spatially and temporally, which is highly correlated with the typical routine of people. Such characterization provided us with a deeper understanding of the properties of those PSNs, and revealed their great potential to support studies on city dynamics and urban social behavior, motivating then the proposition of techniques in this direction;

2. **Applicability of single PSNs:** From the results obtained in the characterization stage, we propose different methods and techniques that capture several aspects of urban areas, such as people's routines, cultural traits, points of interest, economical particularities, etc. These proposed methods and techniques illustrate how PSNs can be exploited to enable large scale and near real time analyses of city dynamics and urban social behavior;
3. **Definition, modeling, and application of PSNs as sensing layers:** We define the concept of sensing layers, which represent data from different PSNs, each one enabling the access of data related to a certain aspect of the city. A range of fruitful opportunities may emerge from this idea, because as each layer represents a partial view of the city, their aggregation can provide a deeper understanding of it. With this in mind, we propose a framework for integrating multiple sensing layers, which can be applied to more sophisticated services than services based on a single layer. Finally, we present applications that illustrate the use of the proposed framework and the potential of using multiple sensing layers.

The results for the Contribution 1 were reported in the following publications:

- In [Silva et al., 2012b], we perform the first analysis of PSN properties. The PSNs analyzed were derived from location sharing services (Gowalla and Brightkite);
- In [Silva et al., 2013b], we extended the work [Silva et al., 2012b] analyzing also two different PSNs derived from Foursquare, a popular location sharing service;
- In [Silva et al., 2013c] (**2nd best paper award**) and [Silva et al., 2013d], we investigate properties of a PSN derived from Instagram, a photo sharing service;
- In [Silva et al., 2013f], we study properties of a PSN derived from Waze, a popular traffic alert system.

The results for the Contribution 2 were reported in the following publications:

- In [Silva et al., 2012d] (**Best paper award**), we propose a technique that provides a visual summary of the city dynamics based on the movements of individuals. An extended version, [Silva et al., 2014c], got accepted in the ACM Transactions on Internet Technology (to be published in the second semester of 2014);
- In [Silva et al., 2013d], we propose a technique for point of interest (POI) identification, which is also able to extract sights out of the identified POIs;
- In [Silva et al., 2013a], we survey models and approaches applied in PSNs to support different applications and techniques;
- In [Silva et al., 2014b], we propose a new methodology for identifying cultural boundaries across populations using self-reported cultural preferences recorded in PSSs.

The results for the Contribution 3 were reported in the following publications:

- In [Silva et al., 2013e], we perform a comparative study of different PSNs derived from Instagram and Foursquare, and verified if they can complement each other;
- In [Silva et al., 2014a], we, among other things, introduce the concept of sensing layers used in this work;
- In [Silva et al., 2014d], we formalize the concept of sensing layers, presents a framework for working with multiple sensing layers, and also illustrates the potential of the joint use of multiply sensing layers through two applications. An extension of this work is under revision in the ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems, where we present more details about the proposed framework and the applications illustrated.

1.4 Work Organization

The rest of this document is organized as follows. Chapter 2 presents in Section 2.1 the concept of ubiquitous computing, showing its definition (Section 2.1.1), discussing its current state (Section 2.1.2) and also presenting the concept of context aware computing (Section 2.1.3), which is a central piece of ubicomp. This chapter presents also, in Section 2.2, related studies, discussing the approaches and models used to extract and generate context information from PSNs data in order to study city dynamics and urban social behavior.

Chapter 3 discusses the participation of humans in the sensing process, covering particularities of participatory sensor networks (Section 3.1). Besides that, this chapter, in Section 3.2, also discuss the challenges that emerge when dealing with PSNs. Chapter 4 studies

the properties of participatory sensor networks derived from location sharing services, photo sharing services, and traffic alert services in Sections 4.1, 4.2, and 4.3, respectively. Section 4.4 characterizes two distinct PSNs derived from Foursquare and two from Instagram. It compares these PSNs to investigate whether they can complement each other, or if they are compatible to study the dynamics of cities and urban social behavior. Finally, Section 4.5 discusses the chapter's results.

Chapter 5 discusses proposed techniques and applications that exploit PSNs to foster a deeper understanding of relevant aspects related to city dynamics and urban social behavior. Section 5.1 presents a technique called City Image, which provides a visual summary of the city dynamics based on people movements. Section 5.2 discusses other possibilities to better understand city dynamics through people movements. Section 5.3 presents a technique to extract points of interest in the city. Section 5.4 discuss possibilities to use PSNs to the analysis of social and economic aspects of city's inhabitants. Section 5.5 motivates the use of participatory sensing systems to the study of cultural differences. Section 5.6 discusses the chapter's results. methodology to identify cultural boundaries.

Chapter 6 is dedicated to the discussion of the concept of sensing layers. Section 6.1 defines the concept of sensing layers and proposes a framework for working with sensing layers. Section 6.2 discusses how to process sensing layers, defining examples of operations that can be applied to sensing layers, as well as strategies of processing using the proposed operations. Section 6.3 presents some proposed applications that illustrate the potential of using sensing layers. Finally, Chapter 7 presents the conclusions and future work.

Chapter 2

Background

2.1 Ubiquitous Computing

Modern computing can be divided into three eras. The first is characterized by one single computer (mainframe) owned by an organization and used by many people concurrently. In the second era, a personal computer (PC) is usually owned and used by a single person. In the third era, ubiquitous computing (ubicomp), each person owns and uses many computers, especially small networked portable devices such as smart phones and tablets [Weiser and Brown, 1996; Krumm, 2009].

Ubiquitous computing is related to mobile computing, although they are not the same thing, neither a superset nor a subset of each other [Weiser, 1996]. Mobile computing devices are not mere personal organizers. They are devices (computers with processing power) that contemplate a new paradigm: mobility. Mobility has some constraints, such as finite energy sources. This paradigm is changing the way we work, communicate, have fun, study and do other activities while we are moving [Satyanarayanan, 1996]. The fact is that ubiquitous computing must support mobility, since motion is an integral part of everyday life. Hence, ubiquitous computing relay on mobile computing, but goes much further.

2.1.1 Mark Weiser's Visions

To talk about ubiquitous computing we have first to mention Mark Weiser, which has been recognized as the “father” of ubiquitous computing. Weiser, called by many “Visionary”, was head of the Computer Science Laboratory at Xerox Palo Alto Research Center (PARC) when he coined the term ubiquitous computing in 1988. When the ubiquitous computing program emerged at PARC, it was at first envisioned only as an answer to what was wrong with personal computing, because they were too complex, too demanding of attention, among

others things [Weiser et al., 1999]. During the implementation of the first ubicomp system, Weiser's group realized they were, in fact, starting a post-PC era, in other words, ubicomp was emerging [Weiser et al., 1999].

Mark's visions influenced a countless number of researchers. Almost one quarter of all the papers published in the Ubicomp conference between 2001 and 2005 cite Weiser's "foundational articles" [Bell and Dourish, 2007]. Among the Weiser's foundational papers of ubiquitous computing, perhaps the most impacting work is the one entitled "The Computer in the 21st Century", published in *Scientific American* in 1991. In this paper, Weiser describes the ideal ubicomp future, its purposes, concerns and analogies. To illustrate its ideas he told the story of "Sal", a tale about a single mother and how the world evolves around her needs.

"The most profound technologies are those that disappear. They weave themselves into the fabric of everyday life until they are indistinguishable from it" [Weiser, 1991, p. 1].

Weiser believed that the most powerful things are those that are effectively invisible in use. The ideal is to make a computer so embedded, so fitting, so natural, that we use it without even thinking about it. The essence of this vision is making everything easier to do, with fewer mental gymnastics [Weiser, 1991, 1993b].

According to Weiser, the style of computing that has been imposed on users in the first and second modern computing eras (mainframes and PCs, respectively) is too attention consuming, and divorce the users of what is happening around them. In the ubicomp world, as Weiser believed, computation could be integrated with common objects that you might already be using for everyday work practices, rather than considering computation to be a separate activity. If the integration is done well, the envisioned invisibility could be achieved [Weiser, 1993a; Krumm, 2009].

In order to clarify this concept of invisibility, consider the example based on the familiar printed page (inspired in [Krumm, 2009]). To perform a printing it is necessary deposit ink on thin sheets of paper, and a consolidated technology is necessary for that. For a good result it is necessary to ensure that: it must be durable in use; not wick in the paper if wet; among other things. However, we rarely pay attention on the ink technologies when we read printed pages. Instead, we read pages and comprehend ideas, not necessary focusing on the technology, the characteristics of the ink, or the manufacturing process of the paper to be able to use it. In this example, the printing technology got invisible for the user, allowing the higher-level goal of reading a story, or acquiring knowledge. This kind of thinking rarely happens with traditional PCs, which demand the users continuously focus attention on the system, maintaining it and configuring it to complete a task.

Good technology is invisible, staying out of the way of the task, like a good car stays out of the way of driving. Bad technology draws attention to itself, not the task, like a car that needs a tune-up. Computers are mostly not invisible. Ubiquitous computing is about enabling invisibility in computers [Weiser, 1994].

2.1.2 Ubicomp Today

As a promising research area, ubiquitous computing gave us more questions than answers [Weiser et al., 1999], and many of them are still open [Weiser, 1993a]. There are many people around the world working on projects that deal with ubicomp challenges. Those projects range from prestigious computer science Schools, such as MIT (see several projects from Media Lab¹ for some examples), to mainstream computer companies, such as Microsoft (see the website <http://research.microsoft.com/en-us/groups/ubicomp/> with some projects).

In order to have a picture of ubicomp researchers, we collected information about all papers published until 2011 in Ubicomp, Pervasive², and Percom, and performed a data mining process, extracting statistics such as most productive authors and institutions, which include those mentioned above. We also analyzed the collaboration among authors identifying, for instance, the formation of communities. Figures 2.1(a) and 2.1(b) illustrate those results, depicting, respectively, the occurrence of authors and institutions in the analyzed papers. In that analysis authors and institutions are counted just once by paper, and the size of the word reflects its representativeness. As we can see Gregory D. Abowd is the author who published the largest number of papers, and Universities of California and Intel are the most productive institutions. The complete study is presented in the paper [Silva et al., 2012a].



Figure 2.1: Representativeness of authors and institutions.

Since the early days of ubicomp, one of the main concerns was that computer too often remain the focus of attention, rather than being a tool through which we work, disappearing from our awareness [Weiser, 1993a]. We may have not achieved the original Weiser's vision

¹<http://www.media.mit.edu>.

²Now Ubicomp and Pervasive merged in one single conference.

about Ubicomp yet. But we can say that the key ingredients are evolving in a favorable direction for it. Many critical items that were rare in early 1900s are now commercially viable. Each year more possibilities for the mainstream application of ubiquitous computing open up.

The future envisioned by Weiser, ubiquitous computing, considers a computing environment in which each person is continually interacting with many wirelessly interconnected devices [Weiser, 1993a]. Today it is easy to find several microprocessors at home, available, for instance, in alarm clocks, the microwave ovens and in the TV remote controls. They do not qualify as ubicomp devices mainly because they do not communicate with each other, but if we network them together they are an enabling technology for ubicomp [Weiser and Brown, 1996]. It soon may become a reality. For example, Google has announced in the event Google IO'11³ an initiative called Android@Home, which allows Android⁴ applications to discover, connect and communicate with appliances and devices inside the house. After connecting together several information sources with many information delivery systems we will start to have things, such as, clocks that find out the correct time after a power failure, and microwave ovens that download new recipes.

Besides that, some of our computing technology are becoming ubiquitous, for instance smart phones, which are taking center stage as the most widely adopted and ubiquitous computer [Krumm, 2009]. When we get used to the possibility of accessing a GPS-connected map, social networks and the Internet anywhere at anytime, we will realize the value of this and it will become part of our lives.

“Applications are of course the whole point of ubiquitous computing.” [Weiser, 1993a, p. 80]

We have to keep in mind that is not just one service that will make computing a disappearing technology, but the combination of many. Those services have to be available as needed without extraordinary human intervention [Abowd et al., 2002]. The challenge is to create a new kind of relationship between people and computers, where computers do not demand too much attention, letting people live their lives [Abowd and Mynatt, 2000]. Application will go beyond the big problems like corporate finance, to the little annoyances such as: where are the car-keys? Can I get a parking place? What is the best route to take now? Which pub should I go in a certain area of the city? [Weiser and Brown, 1996].

Since ubiquitous computing has intersections with many areas of computing, several research fields can contribute to its development, including distributed computing, mobile computing, sensor networks, and machine learning. In this direction we analyzed all papers

³<http://www.google.com/events/io/2011>.

⁴<http://code.google.com/android>.

published in 2010 and 2011 in Ubicomp, Pervasive, and Percom, creating a taxonomy of recent ubicomp research, more details can be found in [Silva et al., 2012a]. We can see in that study that context-aware computing is a key area of research that can help us to meet the original design goals of ubicomp.

2.1.3 Context-Aware Computing

Several context definitions have been proposed. Among them, those presented by [Schilit et al., 1994], [Dey et al., 1998], and [Pascoe, 1998] are close to the definition considered by most people as the ideal one. The problem is that those definitions lack generality. [Dey and Abowd, 2000] proposed the following definition of context:

“Context is any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves.” [Dey and Abowd, 2000]

This is one of the most accepted and accurate definitions currently used by researchers. It can be observed that the definition is very general when considering what types of data are context, being wide enough to accept the different needs of each application. In addition, it is interesting to note that the definition is precise, not requiring a list of specific types or classes of contexts.

In this work we consider participatory sensing. In this case, humans are responsible for sharing data, acting as sensors in a network (this is also known as participatory sensor network, as explained in Chapter 3). The data shared by the “sensors” (humans plus his/her portable device) can be then transformed in a context used to study city dynamics and urban social behavior. In the next section, Section 2.2, we discuss the model and approaches used to transform raw data shared by users into context information.

2.2 City Dynamics and Urban Social Behavior

In this section we discuss the approaches and models used to extract and generate contextual information from participatory sensor networks data in order to study city dynamics and urban social behavior. This section discusses five different classes of studies to identify contextual information. Section 2.2.1 discusses studies related to the analysis of mobility patterns. Section 2.2.2 considers studies that focus on the better understanding of city dynamics. Section 2.2.3 discusses the study of social patterns. Section 2.2.4 discusses studies

concerned in event detection. Section 2.2.5 presents studies related to human behavior. And, finally, Section 2.3 discusses how this thesis differs from the previous studies.

It is worth mentioning that each class of study is not necessarily mutually exclusive. For example, [Long et al., 2012] used a Foursquare dataset to classify venues based on users' trajectories. This work has intersections with the class "Mobility patterns" (Section 2.2.1), but instead of being classified in that class, it was assigned to the class "Understanding cities" (Section 2.2.2), since it is more concerned in the analysis of city dynamics.

2.2.1 Mobility Patterns

This class of work focuses on studying mobility patterns of users from their logs generated from social media websites. These logs usually include spatio-temporal information, e.g., check-ins and geolocated photos. The study of mobility is useful for many purposes. For example, it is possible to understand how human allocate time to different activities, thus being a fundamental and traditional question in social science [Giannotti et al., 2012]. As another example, one could design new tools to help traffic engineers to understand the flow of people.

The modeling of mobility patterns has been attracting the attention of researchers in different fields, such as physics and ubiquitous computing [Brockmann et al., 2006; Zheng et al., 2009; Gonzalez et al., 2008]. For example, Song et al. [2010] analyzed 50,000 cellphone users and showed that user mobility presents high predictability. It is important to point out that data derived from social media is different from GPS tracking or cellphone usage data, such as phone calls, and present special features and varied contexts. For example, check-ins in location sharing services or photos shared in a photo sharing service bring extra information of a particular place. For instance, a check-in is associated with a type of venue, e.g. pub, and a photo may bring the information about the current situation inside this venue. Again, throughout this work our focus is on studies that analyze data from social media.

Cheng et al. [2011] analyzed 22 million check-ins posted from several location sharing services (Foursquare is responsible for 53.5% of the total). They found that users follow simple and reproducible patterns, and also that social status, in addition to geographic and economic factors, are coupled with mobility. **Approach:** to make their analysis they used three statistical properties to study and model human mobility patterns: *displacement*; *radius of gyration*; and *returning probability*. The *displacement* of check-ins is the distance between consecutive check-ins, measuring how far a user has moved. The *radius of gyration* measures the standard deviation of distances between the users' check-ins and the users' center mass. This measure indicates how far and frequently a user has traveled. *Returning probability* is a measure of periodic behavior in human mobility, since periodic behavior tends to happen

frequently due to human routines. Besides that, the authors also studied factors that could influence mobility, such as social status and geographic and economic constraints.

Cho et al. [2011] investigated patterns of human mobility in three datasets: check-ins, in two location sharing services, and cellphone location data. They were particularly interested in determining how often users move and where they go to, as well as how social ties may impact their movements. They observed that short-ranged travel is periodic both spatially and temporally and is not affected by the social network structure, while long-distance travel is more influenced by social network ties. **Approach:** based on their empirical findings they built a model named Periodic & Social Mobility Model to predict mobility of users. This model is composed by three parts: (1) a model of spatial locations that a user usually visits based in a two-state mixture of Gaussians with a time-dependent state prior; (2) a model of temporal movement between these locations based on a truncated Gaussian distribution parameterized by the time of the day; (3) a model of movement that is influenced by the ties of the social network, e.g. encountering friends. In this specific model, if a user performs a check-in, then it will more likely be close in space and time to one of his/her friend's check-ins. Their model is able to predict the exact user location at any time with 40% accuracy.

Nguyen and Szymanski [2012] used Gowalla, a location-based social network, to create and validate models of human mobility and relationships. In that work, the authors proposed a friendship-based mobility model (FMM) that take into account social links in order to provide a more accurate and complex model of human mobility. With this model the authors were able to study how frequently friends travel together. This model may improve the accuracy of a varied number of applications, such as traffic engineering in communication networks, transportation systems, and urban planning. **Approach:** the proposed mobility model uses a Markov Model where the states represent locations of check-ins and the links represent the probability of going from one place to another. For example, the probability of going from work to pub is defined as the ratio between the number of times a given user performs a check-in in a pub right after a check-in at work, and the number of times that user performs a check-in at work.

Zheng et al. [2012] studied tourist mobility and travel patterns from geotagged photos shared on Flickr. In order to extract the travel patterns, the authors focused the analysis on tourist movement according to regions of attraction and topological characteristics of travel routes by different tourists. The authors demonstrated its potential by testing the approach on four cities. **Approach:** first it is built a database of touristic travel paths based on the concept of mobility entropy (considering Shannon's entropy), used to discriminate the touristic and non-touristic movement. Then, a significance test is applied to ensure that the resulting path is statistically reliable. For that, they devised two methods, one based on a Poisson distribu-

tion and the other on a normal distribution. Next, it is proposed a method to discover regions of attraction in a city, using for this the DBSCAN clustering algorithm. To study the touristic movement the authors considered a Markov chain model created from the visiting sequence of regions of attraction discovered by the proposed method. With that, they can estimate statistics of visitors traveling from one region to another. In order to study the topological characteristics of tour routes, the authors perform sequence clustering on travel routes, applying a modified version of the longest common subsequence as a similarity metric to minimize noise.

2.2.2 Understanding Cities

Information obtained from participatory sensing systems have the power to change our perceived physical boundaries and notions of space, as well as to better understand city dynamics [Bilandzic and Foth, 2012]. This section focuses in presenting studies in these directions. Many potential applications can benefit from these types of studies, such as tools for city planners to provide new manners to see the city, or for end users who are looking for new ways to explore the city.

Cranshaw et al. [2012] presented a model to extract distinct regions of a city that reflect current collective activity patterns. The idea is to expose the dynamic nature of local urban areas considering spatial proximity (derived from geographic coordinates) and social proximity (derived from the distribution of check-ins) of venues. **Approach:** in their study the authors considered data from Foursquare. In order to explore this data, the authors developed a model based on spectral clustering. One of the main contributions is the design of an affinity matrix between venues that effectively blends spatial proximity and social proximity. The similarity of venues is then obtained by comparing pairs of these dimensions. After that, this is used to compute the clusters that may represent different geographical boundaries of neighborhoods. The clustering method is a variation of the spectral clustering proposed by Ng et al. [2002], introducing a post processing step to clean up any degenerated cluster.

Noulas et al. [2011b], proposed an approach to classify areas and users of a city by using venues' categories of Foursquare. This could be used to identify users' communities that visit similar categories of places, useful to recommendation systems, or in the comparison of urban areas within and across cities. **Approach:** their approach is based on spectral clustering algorithm [Luxburg, 2007; Ng et al., 2002]. More specifically, the authors divide the area of a city to be analyzed into a number of equally sized squares, each of them will be a datapoint input for the clustering algorithm. For each area it is represented the activity performed on it based on check-ins in each existing category on that area. Then, it is calculated the similarity between two areas as the cosine similarity between their corresponding

activity representation. Having the similarity information, the authors apply it in the spectral clustering algorithm.

Long et al. [2012] used a Foursquare dataset to classify venues based on users' trajectories. The premise is that the venues that appear together in many users' trajectories will probably be taken as geographic topics, for example representing restaurants people usually go to after shopping at a mall. The approach can be applied, for instance, to understand users' preferences to make recommendation of venues. **Approach:** the authors used the Latent Dirichlet Allocation (LDA) [Blei et al., 2003] model to discover the local geographic topics from the check-ins. With this approach, it is possible to dynamically categorize the venues in Foursquare according to the users' trajectories, what indicates the crowds' preferences of venues. LDA is usually used to cluster documents based on the topics contained in a corpus of documents. For this reason, some terms used to describe the modeling make reference to this context. The authors considered that a single check-in represents a word, which is the basic unit in the LDA. A trajectory of a user consists of all the venues visited by him/her, and this represents a document in the analogy.

Kisilevich et al. [2010] used geo-tagged photos obtained from Flickr to analyze and compare temporal events that happened in a city, and also to rank sightseeing places. More specifically, the authors presented a way to assess the attractiveness of places based on their positions in a ranking, and suggested a set of visual analytic methods that mixes computational techniques with visual interactivity in order to support analysis of the data. **Approach:** to find the attractiveness of places the authors applied the algorithm DBSCAN [Ester et al., 1996]. In order to highlight areas of people's activities within a cluster, the authors applied density maps. From the clusters obtained in the clustering step, the weight of every geo-tagged photo is calculated using a density function based on the relative position of photos of other users in a cluster. The calculated weight is mapped to a color, facilitating the visual inspection.

Frias-Martinez et al. [2012] used a dataset from Twitter and proposed a technique to determine the type of activities that is most common in a city by studying tweeting patterns. They also proposed another technique to automatically identify landmarks in a city. **Approach:** to automatically identify urban land usage, the authors apply two methods. The first one is land segmentation. For that it is applied Self-Organizing Maps [Kohonen, 1990], which is an unsupervised neural network. After training the network, it is obtained a map that segments the urban land into geographical areas with different concentrations of tweets. Each neuron of the network represents a pointer to a region with a high density of tweets. With that, the authors apply Voronoi tessellation considering the location of the neurons to compute the land segments. Next, the authors use the segments found to detect different land usages considering the average tweet usage on them. So, for each land segment is built

a unique tweet-activity vector that represents the average tweeting temporal behavior. To characterize urban land usage, it is applied the k-means algorithm, which shows common tweeting behavior across land segments. To identify the landmarks, the authors used the mean-shift clustering technique [Cheng, 1995]. The authors considered in this algorithm that every tweet is assigned to a local maxima and a cluster represents a potential landmark. After the execution, if the resulting clusters are ranked by the number of tweets on them, then the result represents a list of the most popular landmarks.

Ji et al. [2009] mine blog-based sight photos in order to discover and summarize city landmarks. Their main contribution is a generalized graph modeling framework. This study is useful, for example, for personalized tourist suggestions. **Approach:** first the authors have to extract locations of photos. For that, they collect photos with different descriptors. To identify their locations they use an application called Gazetteer [Wang et al., 2005], which is able to identify location from web resources. Then they create a hierarchical visual-textual clustering scheme to organize sight photos into a “scene-view” structure for each city. For this purpose it is used the concept Bag-of-Visual-Words [Nister and Stewenius, 2006] to generate the content descriptor of photos. Bag-of-Visual-Words are clustered by their similarity measured by the cosine distance, generating then “views”. After that the authors create a “scene-view”, using textual clustering to aggregate “views” into “scenes”. Next, they model two different graphs. The first one represents a scene, where each node is a photo and an edge exist if there is at least one word identical in the photos descriptors. For this graph they present an algorithm, PhotoRank, to discover representative views within a scene. Finally, the authors create another graph to represent the city, that encompasses a scene layer, and present an algorithm to discover city landmarks on it, which explores the PhotoRank algorithm and is inspired in [Kleinberg, 1999].

2.2.3 Social Patterns

This class of studies concentrates in the analysis of data from social media to understand social patterns. Data from social media enables unprecedented opportunities to study human relationships in a global scale, at a relatively low cost. Examples of possibilities include community detection, products recommendation based on the discovery of similar socio-economic behavior, and new definitions of network centrality.

Scellato et al. [2011] presents a study of the spatial properties of the location-based social networks arising among users. Among the results, the authors reported, for instance, that 40% of social links happens below 100 km, and that there is strong heterogeneity across users related to both social and spatial factors. **Approach:** to extract properties and verify their hypothesis, the authors analyzed datasets of three location based services: Foursquare,

Gowalla, and Brightkite. In their study, the authors used two randomized models, a social model and a spatial/geo model, to assess the statistical significance of the empirical spatial properties of the networks analyzed. The social model keeps the social connections as they are, randomizing all user locations. The geo model keeps the user locations unmodified and then assigns every social link between two users at a certain distance according to the relative probability of friendship, observed in their analysis.

Cranshaw et al. [2010] introduced a new set of features of human location trail for analyzing the social context of a geographical region. They demonstrated the applicability of these features by presenting a model for predicting friendship between two users, showing significant gains over previous models for the same purpose. **Approach:** the authors used a dataset from Locaccino⁵, a system that allows users to share his/her current location with other Locaccino users through Facebook⁶. For the co-location analysis, the authors split the space in grids of $0.0002^\circ \times 0.0002^\circ$ latitude/longitude, which means approximately 30 meters x 30 meters. The time was considered in slots of 10 minutes. In this way, a user is co-located with another user if they are located in the same grid within a slot of time. To model the co-location of users, it is applied three diversity measures: frequency, user count, and entropy (Shannon's entropy). The frequency measure captures the raw count of users who visit a location. The user measure considers the total number of unique users in a location. The entropy measure considers the number of users observed at the location, as well as the relative proportions of observations. High entropy means that many users were observed at the location with equal proportion.

Quercia et al. [2012] study how social media communities resemble real-life ones. They tested whether established sociological theories of real-life social networks still hold in Twitter. They found, for example, that social brokers in Twitter are opinion leaders who take the risk of tweeting about different topics. They also discovered that most users have geographically local networks, and that social brokers express not only positive but also negative emotions. **Approach:** the authors applied network metrics about topic, geography, and emotions, regarding to parts of one's social world. These metrics include reciprocity, simmelian ties, and network constraint. Reciprocity is the proportion of edges in a network that are bidirectional. Simmelian ties are a measure that considers triadic relationships. Network constraint measure brokerage opportunities in the network, where high network constraint means less brokerage opportunities. They used Burt's formulation [Burt, 1992] in this specific case.

Java et al. [2008] studied blog communities. For that they present a technique for clustering communities by using both the hyperlink structure of blog articles and tag information

⁵<http://www.locaccino.org>.

⁶<http://www.facebook.com>.

available on them. The technique was tested in a real network of blogs and tag information, as well as in a citation network. **Approach:** the authors define a community as a set of nodes in a graph that link more frequently within this set than outside it, and they also share similar tags. Their technique is based on the Normalized Cut (NCut) algorithm [Shi and Malik, 2000].

Sadilek et al. [2012] studied the interplay between people's location, interactions, and their social ties, presenting a technique for inferring link and location information from a stream of message updates. The authors demonstrated, by analyzing users from New York City and Los Angeles, that their technique significantly outperforms other current comparable approaches. **Approach:** for link prediction their approach infers social ties by considering patterns in friendship formation, the content of people's messages, and user location. For location prediction, their technique implements a probabilistic model of human mobility, where it treats users with known GPS positions as noisy sensors of the location of their friends.

2.2.4 Event Detection

This class of work is focused in the identification of events through data shared in social media. This task is especially favorable due the real-time nature of certain types of social media, such as Twitter. Events might be natural ones, such as earthquakes, or not natural ones, such as the identification/prediction of stock market changes.

Bollen et al. [2011] studied whether collective mood states derived from Twitter feeds are correlated to the value of the Down Jones Industrial Average (DJIA) over time. Their findings indicate that it is possible to obtain an accuracy of 86.7% in predicting the daily up and down changes in the closing values events of the DJIA. This is possible by choosing specific mood dimensions, but not all that were considered. **Approach:** to extract the sentiment expressed by the users in the tweet the authors used two tools. The first one is the OpinionFinder (OF)⁷, which extract negative or positive sentiments from the message. The second tool, Google-Profile Mood State (GPOM), extract six-dimensional daily time series of public mood. The authors use Granger causality analysis in which it is correlates DJIA values to GPOMs and OF values of n past days. The authors also trained a Self-Organizing Fuzzy Neural Network to predict DJIA values on the basis of various combinations of past DJIA values and OF and GPOMS public mood data.

Gomide et al. [2011] analyzed how Dengue epidemic is reflected on Twitter and to what extent that information can be used for the sake of surveillance. Gomide et al. showed that Twitter can be used to predict, spatially and temporally, dengue epidemics by means of

⁷<http://mpqa.cs.pitt.edu/opinionfinderrelease>.

clustering. **Approach:** The authors introduce an active surveillance framework that analyzes how Twitter reflects epidemics based on four dimensions: volume, location, time, and public perception. Specifically, they study how users refer to dengue in Twitter with sentiment analysis and use the result to focus only on tweets that somehow express personal experience about dengue. Then, Gomide et al. constructed a linear regression model for predicting the number of dengue cases using the proportion of tweets expressing personal experience.

Sakaki et al. [2010] studied the real-time interaction of events in Twitter (e.g. earthquakes), and propose an algorithm to monitor tweets to detect a target event. To demonstrate the effectiveness of their method, the authors built an earthquake reporting system in Japan, which was capable to detect 96% of earthquakes reported by the Japan Meteorological Agency (JMA) with seismic intensity scale of 3 or more. Notification to registered users was delivered faster than the announcements that are broadcast by the JMA. **Approach:** the authors devise a classifier of tweets based on features such as the keywords in a tweet, the number of words, and their context. After that, they produced a probabilistic spatio-temporal model for the target event that can find the center and the trajectory of the event location.

Lee and Sumiya [2010] present a geo-social event detection system to identify local events (e.g., local festivals) by monitoring crowd behaviors indirectly via Twitter. The system was created on the hypothesis that users probably write many posts about these local events. **Approach:** first the authors decide what the usual status of crowd behaviors is in a geographical region in terms of tweeting patterns. After that, a sudden increase in tweets in a geographical region can be an important clue. Another hint might be the increasing number of Twitter users in a geographical region in a short period of time. The authors also consider if the movements of the local users become unexpectedly elevated. The detection of unusual events in the study uses the concept of boxplot [McGill et al., 1978], which is applied to create ranges to determine the cases desired to be detected.

Becker et al. [2011] analyze streams of Twitter messages to distinguish between messages about real-world events and non-event messages. They identify each event and its associated Twitter messages. **Approach:** the authors use an online clustering technique that groups together similar tweets. With that, they extract features for each cluster to help determine which clusters correspond to events. Next, the authors use these features to train a classifier to distinguish between event and non-event clusters.

Ginsberg et al. [2009] presented a method for analyzing large numbers of Google search queries to track flu illness in a population. The method can accurately estimate the current level of weekly influenza activity in each region of the United States, with a reporting lag of about one day. **Approach:** By analyzing Google queries usage the authors found a close relationship between how many people search for flu-related topics and how many people actually have flu symptoms. Given that, they developed a simple model that uses

aggregated Google search data to estimate current flu activity.

2.2.5 Human Behavior

This group of studies focus on the study of human behavior through the data shared in social media, which, as we mentioned before, can be seen as signals given by users. This type of study can be applied, for example, to the discovery of individual social profiles, the discovery of collective behaviors, the analysis of sentiment and opinion evolution, and a better understanding of why individuals take certain actions.

Joseph et al. [2012] analyzed a Foursquare dataset to identify groups of people and the places they go. Their model was able to identify groups of people which represent both geo-spatially close groups and people who appear to have similar interests. **Approach:** their model is based on the idea of topic modeling. For that they applied the Latent Dirichlet Allocation [Blei et al., 2003]. In the model instantiation, each check-in for a user can be thought of as a word in a document. Similar to text documents, where a “document” can have multiple words, the authors defined a multinomial distribution for the check-ins for each user by using the number of check-ins in each venue as features.

Naaman et al. [2012] focused their study in the characterization of tweeting patterns in different cities located in the USA, envisioning to provide a framework for reasoning about activities performed in cities. This study might be useful to deal with challenges such as transportation or resource planning faced in urban studies. **Approach:** first the authors selected tweets from some US cities. Then, they selected the top 1000 words from the resulting dataset, and made a cleaning procedure in this dataset using the NLTK toolkit⁸, removing, for example, stopwords. After that, the authors performed a study of keyword-based diurnal patterns in the considered locations. Besides that, the authors applied the concept of Shannon’s entropy and Mean Absolute Percentage Error (MAPE), to measure the variability of the data within days and across days, respectively.

Poblete et al. [2011] analyzed a twitter dataset aiming the discovery of insights of how tweeting behavior varies across countries, as well as the possible explanations for these differences. **Approach:** first the authors selected the top ten most active countries. Then, they extracted differences in the number of twittes per user, languages used per country, sentiment analysis (happiness), using the Affective Norms for English Words (ANEW) [Bradley and Lang, 1999] and also a Spanish version of it [Redondo et al., 2007], and the content of the tweet. Moreover, they studied the social network properties for each country applying metrics, such as, clustering coefficient, diameter, and shortest paths.

⁸<http://www.nltk.org>.

Gao et al. [2012] propose a model to address the “cold start” location prediction problem, by using the social network information. Results in an experiment based on a real-world location-based social network show that the approach is effective for the studied problem. **Approach:** the authors’ strategy encompasses the investigation of the check-ins behavior to understand the correlations in the context of the user’s social network and geographical distance. For this analysis, they considered four social cycles. With that, the authors modeled the geo-social correlations of “new check-in” behavior considering the intrinsic patterns of users’ check-ins and his/her social cycles.

Yu et al. [2012] used the users’ behavioral patterns extracted from Sina Weibo⁹ to investigate how users’ frequent activities reflect their sleeping time and living time zones. The authors showed that may be possible to detect the sleeping time of users. Their results could also be used as an alternative way to estimate time zones. **Approach:** based on the time series of Sina Weibo usage the authors applied a simple statistical method, assuming that users keep a daily routine, going to bed and waking up on time, to detect long periods of inactivity.

2.3 Discussion

A fundamental step to achieve the Ubiquitous Computing vision is to sense the environment. The research in Wireless Sensor Networks has provided several tools, techniques and algorithms to solve the problem of sensing in limited size areas, such as forests or volcanoes. However, sensing large scale areas, such as large metropolises, countries, or even the entire planet, brings many challenges. For instance, consider the high cost associated with building and managing such large scale systems. Thus, sensing those areas becomes more feasible when people participate sharing sensed data using their portable devices (e.g., sensor-enabled cell phones), forming what is called participatory sensor networks (PSNs) (more details in the next chapter, Chapter 3).

Our work differs from previous ones in several ways. Despite the concept of PSNs be relatively old, emerging on 2006, few properties are known of this type of network. Given that, one step that differentiate our study is the identification of fundamental new properties of PSNs (considering different kinds of PSNs), from a sensor network viewpoint, and the discussion of the challenges and implications when dealing with them. As far as we know, we performed the first large scale study of Instagram analyzing photos shared by users, and also the first study of Waze analyzing alerts shared users (more details in Chapter 4). Our work also differentiate from others because we propose new techniques and methodologies,

⁹A popular Chinese micro-blogging service.

relying on PSNs, to the study of city dynamics and urban social behavior (more details in Chapter 5). Besides that, our study is the first to compare two PSNs derived from different systems, particularly Instagram and Foursquare. The aim is to investigate whether data from one PSN could complement the other, or if they are compatible regarding the study of city dynamics and urban social behavior (details in Section 4.4). This comparison, among other results, gave us insights about the potential for joint use of data from these applications, considering each PSN as a sensing layer. With that, another difference of this work is a framework proposition for integrating multiple sensing layers, which was illustrated in the construction of two applications for the study of city dynamics and urban social behavior (more details in Chapter 6).

Chapter 3

Humans in the Sensing Process

The focus of this thesis is on systems that rely on humans' participation in the sensing process, where they are responsible for local data sharing. Such data can be obtained with the aid of sensing devices such as sensors embedded into *smartphones* (e.g., GPS) or by human sensors (e.g., vision), being subjective observations produced by them [Srivastava et al., 2012].

This chapter is organized as follows. Section 3.1 covers particularities of participatory sensor networks and Section 3.2 discuss the challenges that emerge when dealing with PSNs.

3.1 Participatory Sensor Networks

Participatory sensing aims at monitoring large scale phenomena and require the active involvement of people to voluntarily share contextual information and/or make their sensed data available [Burke et al., 2006]. It differs from opportunistic sensing [Lane et al., 2010] mainly by the user participation, which is minimal in the latter case.

Participatory sensing systems (PSSs), such as Instagram and Foursquare, combine the features of online social networks with location-based services, for this reason this type of system have been also called location-based social media. PSSs allow people connected to the Internet to provide useful data about the context in which they are at (near) real time, building new virtual environments that integrate user interactions. Recently, due to the widespread adoption of smartphones and the Internet access through these devices, such systems are becoming increasingly popular, offering unprecedented opportunities of access to planetary scale sensing data.

One important aspect of PSSs is the data the users share, in particular location-related data [Smith et al., 2005]. A data shared in a participatory sensing system is: (i) obtained through physical sensors (e.g., GPS) or human observations (e.g., road congestion report);

(ii) defined in time and space; (iii) obtained automatically or manually; (iv) structured or unstructured; and (v) voluntarily shared or not. To illustrate this type of system, consider an application for traffic monitoring, such as Waze. Users can share reports about accidents or congestion manually. It is still possible to calculate the speed of the car and automatically share the car's route with the aid of the GPS. With speed measurements of different vehicles sampled in a particular area, it is possible to infer, for example, congestion. In this case, users manage the application, which was created for this purpose, and the sensed data are structured. But if users use a microblogging service, such as Twitter¹, the sensed data are unstructured. For example, the user "Bob" sends a message "I am facing slow traffic near the entrance of the campus."

Participatory sensor networks (PSNs) can be derived from participatory sensing systems [Burke et al., 2006]. PSNs have users with their portable devices as the fundamental building block. Individuals carrying these devices are able to sense the environment and to make relevant observations at a personal level. Thus, each node in a PSN consists of the user plus his/her mobile device, sending context data to the systems. For example, in a PSN derived from Instagram, the sensed context data is a picture of a specific place where the user is located.

In traditional wireless sensor networks, the high costs associated with building and managing large scale topologies are prohibitive. In contrast, PSNs allow access to useful data about diverse contexts that users worldwide are inserted in at (near) real time, making them potential sources of sensing at global scale. This opens an unprecedented opportunity to revolutionize the way social science is done. Unlike traditional methods that rely on survey data, new techniques can be designed to exploit participatory data, which is much cheaper, more dynamic as it reflects current situations in real time. Moreover, as we argue here, PSNs have the potential to be a fundamental tool to better understand human urban interaction in the future, leveraging our awareness to different aspects of our lives in urban scenarios. This is useful in many cases, for example, to build smarter context aware applications.

Similar to WSNs, the sensed data in PSN are sent to a server, or "sink node", where data can be accessed (using systems APIs for PSNs, such as Instagram API ²). But unlike WSNs, PSNs have the following characteristics: (i) nodes are autonomous mobile entities, i.e., a person with a mobile device; (ii) the cost of the network is distributed among the nodes, providing a global scale; (iii) sensing depends on the willingness of people to participate in the sensing process; (iv) nodes transmit the sensed data directly to the sink; (v) nodes do not suffer from severe power limitations; and (vi) the sink node does not have direct control over the nodes.

¹<http://www.twitter.com>.

²<http://instagram.com/developer>.

Indeed, PSNs have the potential to complement WSNs in many other aspects besides providing larger scalability. For instance, WSNs are subject to failure, since their operations depend on proper coordination of actions of their sensor nodes, which have severe hardware and software restrictions. On the other hand, as PSNs are formed by independent and autonomous entities, i.e., humans, the task of sensing becomes highly resilient to individual failures. Obviously, PSNs brings also many new challenges, for instance, their success is directly connected to the popularization of the *smartphones* and social media.

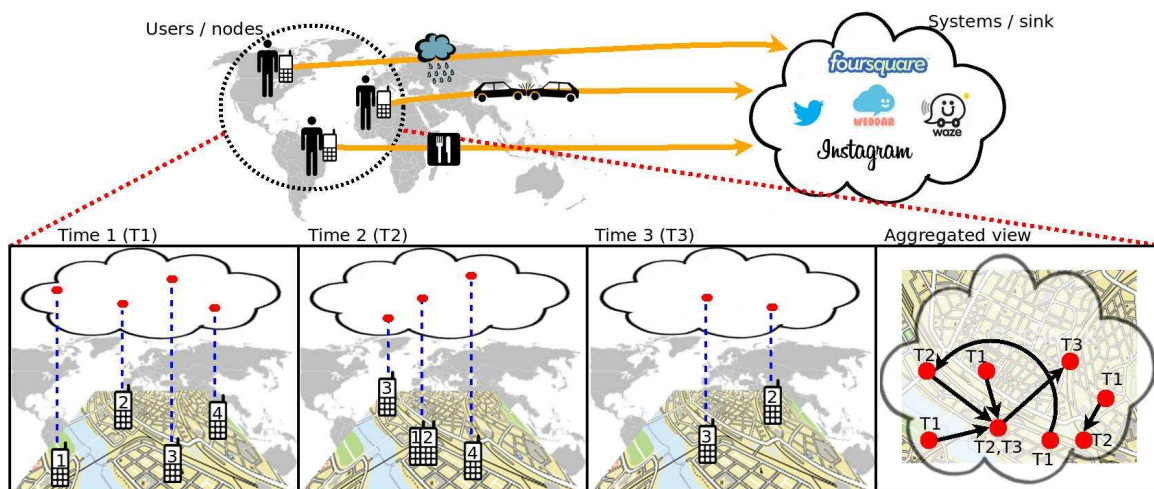


Figure 3.1: Participatory sensor network illustration.

Figure 3.1 illustrates a PSN built from users with their portable devices sending sensed data about their locations to PSSs. The figure shows the sharing activities (represented by red dots) of four users at three different points in time (labeled “Time 1”, “Time 2”, and “Time 3”). Note that a user does not necessarily participate in the system at all times. After a given time, we can analyze this data in different ways. For instance, the bottom rightmost portion of the figure shows, as an aggregated view, a directed graph with nodes representing locations where data was shared and edges connecting locations that were shared by the same user. Using this graph we can extract, for instance, user mobility patterns, information that could be used, for example, to perform load management more efficiently in urban wireless network infrastructure. In fact, knowledge discovery in PSNs walks together with a wide range of studies that use graph theory for social network analysis (SNA) [Scott and Carrington, 2011]. As we show in Chapter 5, well known techniques used for SNA may be directly applied to analyze social oriented graphs derived from PSNs.

PSNs are an example of the interplay between technological networks and social networks, since a key element in a PSN is the human being. The main components of this emerging type of network are illustrated in Figure 3.2. This figure highlights the three most

important components, namely: (i) participatory sensing; (ii) the big raw data; and (iii) the context information.

The component “Participatory sensing” encompasses users sharing data through participatory sensing systems. The component “Big raw data” is responsible for data management. As we can see in Figure 3.2, the collection process may be repeated, for example, to get redundant or complementary data from the same or other systems. After that, the collected data needs to be processed in order to be stored. Since the amount of data coming from PSSs may be very large, all the components need to be carefully designed if the goal is to get (near) real-time information. A more detailed discussion of some of the challenges is presented in Section 3.2.

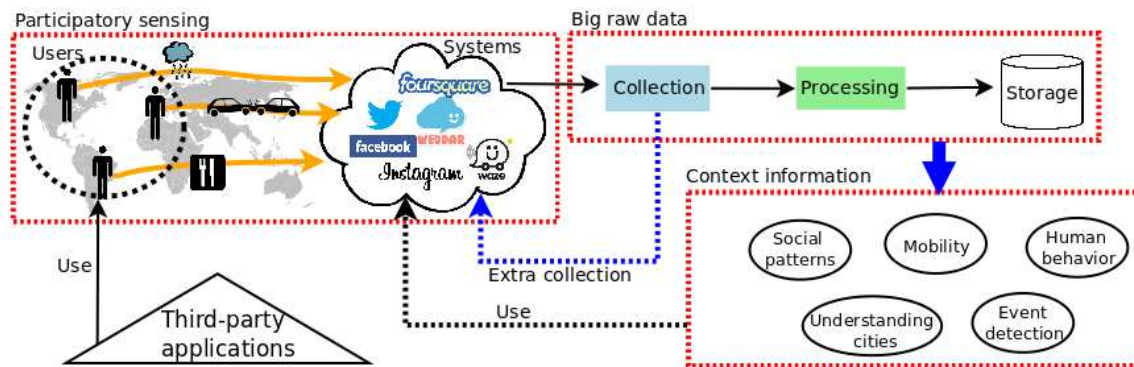


Figure 3.2: Overview of participatory sensor network components.

After the data management stage, the data are ready to be analyzed. The component “Context information” represents five type of analysis that could be performed: Social patterns; mobility; understanding cities; human behavior; and event detection. All these classes of analysis are discussed on Section 2.2.

3.2 Challenges

The construction of a participatory sensor network imposes many challenges. Looking at Figure 3.2 we see that a participatory sensor network could be divided in different blocks. In Section 2.2 we described how researchers have been addressing challenges mainly related in the block named “Context information”, which represents models and approaches to transform big raw data from participatory sensing systems in useful information, to be applied, for example, in applications. In this section we are concentrated in challenges related with the blocks named “Participatory sensing” and “Big raw data”.

Among the challenges present in these blocks we can mention data quality, data collection, data storage, data processing and indexing. The quality of the shared data are a

challenge that has been relatively well tackled in the web domain, however there are unique challenges for controlling the quality of shared data when dealing with ubiquitous user contributions [Mashhadi and Capra, 2011]. For instance, since users can produce sensor readings with relatively little effort, data integrity is not always guaranteed [Saroiu and Wolman, 2010]. Among other initiatives, Saroiu and Wolman [2010] tackled this issue proposing a trusted platform module, which confirms the integrity of sensing devices.

Besides that the shared data through participatory sensing systems in some cases are free text, not presenting structure nor codified semantics, being complex to understand and process. To better interpret such complex data, visualization techniques and tools should be developed. Another issue related to data quality is the liberty given to users in certain PSSs. Sometimes, users can post whatever, even incorrect, information in different formats. This demands mechanisms for data filtering. A reputation system may be very useful in this case.

Data collection is a challenging issue especially from third-party services, such as Foursquare and Waze. By default, data shared in those systems are usually private, unless users decide to make them public, for example sharing it on Twitter. This means that no public data can be available at all. Furthermore, since the data depends on the users will in contribute, there is no guarantee on the delivery of any data. This makes the use of participatory sensing completely out of the control loop of system managers and application developers. Some actions can be taken to ensure that the user participation is sustained over time. An example of action could be an incentive mechanism based on micro-payments, i.e., every time a user perform a given activity, he/she receives an small payment, as proposed by Reddy et al. [2010].

Another important issue is deal with a huge volume of data that PSSs can offer, imposing challenges for *real time* storing, processing, and indexing using traditional database management tools or data processing applications. This makes the offer of real time services using a participatory sensor network a challenge. To tackle this issue we need methods to effectively store, move and process big amounts of data. New algorithmic paradigms, for example map-reduce, should be designed, as well as specific mining techniques should be created according to these new paradigms. Other methods should contemplate data engineering approaches for large networks with up to billions nodes/edges, including effective compression, search, and pattern matching methods [Giannotti et al., 2012]. Fortunately, the research on big data challenges is very active, and has recently made great advances by, for example, relying on parallel platforms (e.g., Hadoop³) for processing large scale datasets.

Furthermore, participatory sensor networks are very dynamic. To illustrate the challenges that emerge with this characteristic we analyze the information flow in PSNs, which

³<http://hadoop.apache.org>.

is depicted in Figure 3.2, particularly the two flows symbolized by arrows labeled with the word “use”, pointing from the Context information to Systems, and from Third-party applications to Users. Users rely on applications, such as Twitter or Waze, to transmit their sensed data. The sensed data are, then, transmitted to the server, or the “sink node”. The Context information component is responsible for processing the shared data and generating useful information, or contexts (Section 2.1.3). Systems, such as Waze, by their turn, may be fed back with the generated contexts and, from this, they may provide useful information to the users. Contexts can also be generated by third-party applications. For example, in Section 5.3, we describe an example of application that enables the identification of regions of interest in a city, which exemplifies a type of context. After using this application, users may choose to change their behavior, e.g., to visit preferably popular areas, which may ultimately impact the number of potential shared data in those places. This gives an idea of how dynamic a participatory sensor network is and the challenges that emerge when dealing with this dynamism.

Besides these problems there is still the problem of user’s privacy. This challenge is very broad, being present in many layers of the system. Data privacy in social media systems has been currently discussed in several studies, such as: [Pontes et al., 2012; Toch et al., 2010; Brush et al., 2010].

A wide range of novel applications opens up after dealing with the challenges of this research field. Some of the opportunities are illustrated in the next chapters.

Chapter 4

Properties of PSNs

Many questions arise from the emerging concept of participatory sensor networks (PSNs). What are the properties of PSNs? What types of applications can we apply PSNs in? What are their limitations? As the data provided by PSNs may be very complex, a fundamental step in any investigation is to characterize the collected data in order to understand its challenges and usefulness.

In this chapter we analyze participatory sensor networks derived from three location sharing services, namely Foursquare, Gowalla and Brightkite (results presented in Section 4.1). We also analyze a PSN derived from a photo sharing service, namely Instagram (results presented in Section 4.2), and a PSN derived from a traffic alert service, namely Waze (results presented in Section 4.3). Section 4.4 compares different PSNs. A discussion about the results is presented in Section 4.5.

4.1 PSN from Location Sharing Services

This section investigates PSNs derived from location sharing services. First, Section 4.1.1 describes the datasets considered. Then, Section 4.1.2 analyzes the coverage of the analyzed PSN at different spatial granularities, starting from the entire planet, going to continents, cities until individual venues. Next, Section 4.1.3 looks at the frequency which nodes share data in individual locations of our dataset. Finally, Section 4.1.4 discusses the sensing seasonality.

4.1.1 Data Description

The analyzed PSNs are derived from four datasets collected from 3 location sharing services, namely Foursquare, Gowalla and Brightkite. Three of these datasets, one for each

| System | # of check-ins | Interval | # of Venues | Categories |
|--------------------|-------------------|---------------------|-------------|------------|
| Foursquare-Year | 11,743,781 | Feb2010 - Jan2011 | 490,079 | no |
| Foursquare-Crawled | 4,672,841 | April 2012 (1 week) | 1,929,237 | yes |
| Gowalla | 6,442,890 | Feb2009 - Oct2010 | 1,280,969 | no |
| Brightkite | 4,491,143 | Apr2008 - Oct2010 | 772,966 | no |
| Total | 27,350,655 | | | |

Table 4.1: Dataset information.

system, are publicly available [Cho et al., 2011; Cheng et al., 2011]. Moreover, since we are also interested in the information about the categories of the venues, we collected a fourth dataset from Foursquare. We collected this data directly from Twitter, since Foursquare check-ins are not publicly available, by default. Approximately 4.7 million tweets containing check-ins were extracted from Twitter, each one providing a URL to the Foursquare website, where information about the geographic location of the venue was acquired. To differentiate the two datasets from Foursquare we refer to the one obtained from [Cheng et al., 2011] as **Foursquare-Year**, and to the one we crawled as **Foursquare-Crawled**.

In location sharing services the basic activity users can perform is called check-in, which is an action to announce in the system where you are at a certain moment. Other actions could also be allowed. For instance, in Foursquare users can post tips in specific places aiming at sharing information on any aspect related to the venue with others [Vasconcelos et al., 2012]. In this chapter we focus on users' check-ins. In all four datasets, each check-in consists of the latitude, longitude, venue's id, and time. As we mentioned, our collected Foursquare-Crawled dataset also includes the venue category. Table 4.1 summarizes the four datasets. Note that the Foursquare-Crawled dataset has approximately 40% of data of Foursquare-Year dataset, despite the interval of collection being much shorter. This is explained by the way we performed our collection. The Foursquare-Year dataset also extracted information about check-ins from Twitter, but instead of using the URL available in the tweet to acquire the geographic information in the Foursquare website, the authors considered only tweets with geo-tagged updates, which are less frequent.

4.1.2 Network Coverage

Figure 4.1 depicts the coverage in the PSN formed from our datasets, which can be very comprehensive in a planetary scale. However, despite the global magnitude of the coverage, observe in Figure 4.2 the total number of check-ins per continent and per interval of time. Note that the sensing activity in some continents, such as North America and Europe, are significantly higher than in others, such as Oceania and Africa. However, observe that Africans

are increasing their participation, probably because of the recent investments in mobile infrastructure in Africa [TheEconomist, 2012]. Observe also that for Foursquare-Crawled, the most recent dataset, the participation of Asians and Latin Americans is at least equivalent, if not larger in the case of Asians, than the participation of North Americans.

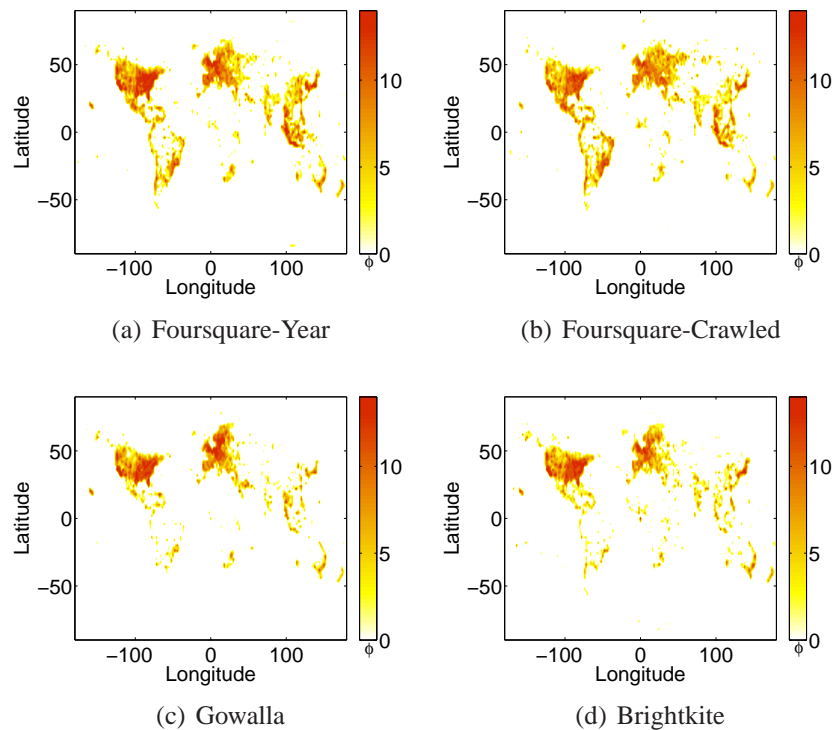


Figure 4.1: All sensed locations. The number of locations n per pixel is given by the value of ϕ displayed in the colormap, where $n = 2^\phi - 1$.

We now turn our attention to six large and populous cities located in five continents: New York City (U.S.A.), Rio the Janeiro city (Brazil), Paris (France), Sydney (Australia), Tokyo (Japan) and Cairo (Egypt). Figure 4.3 shows, for each city, the heatmap of the sensing activity in these cities. In the heatmap, the darker the color, the higher is the number of check-ins in that area. Figures 4.3a (New York), 4.3c (Paris) and 4.3e (Tokyo) show that the coverage of the PSN in these cities is high. Now, looking at Figure 4.3f, we see that the coverage in Cairo is very low (approximately only 10%), despite the its equivalent population.

Economic factors might impact the usage of mobile devices by the local population, ultimately impacting sensing coverage. If most people living a given area cannot afford to buy a smartphone (or any other mobile device), the local coverage may be low. Besides the economical aspect, cultural differences are also an important aspect that must be considered. The cultural differences in Cairo, compared to the previous cities mentioned, might have an impact on the adoption and use of location sharing systems. People from certain cultures

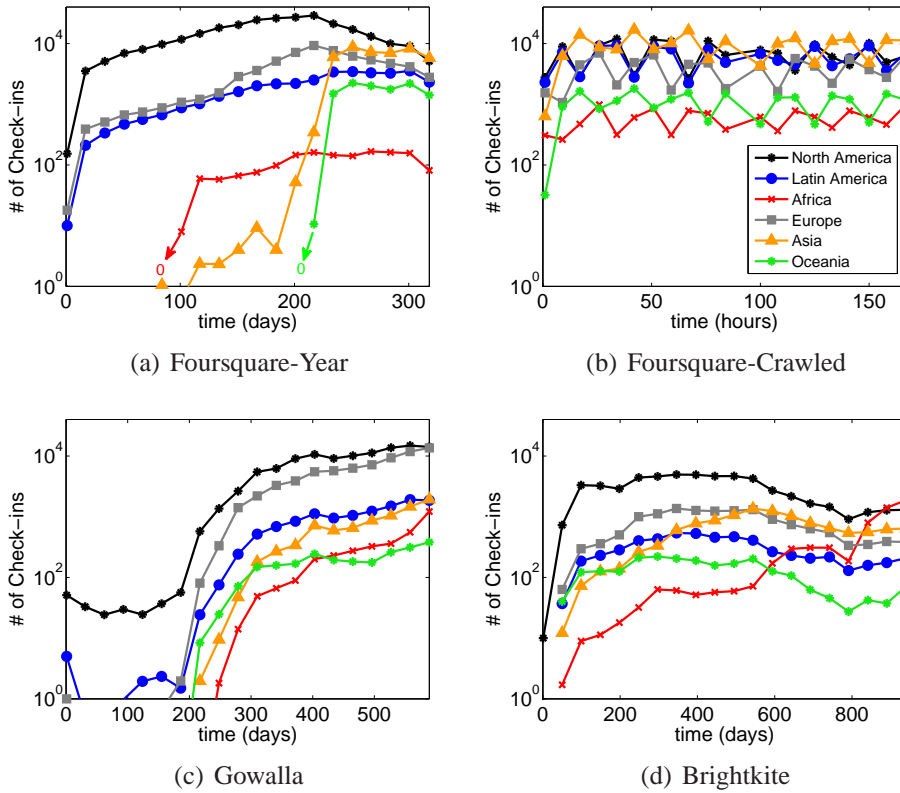


Figure 4.2: Temporal variations in the number of check-ins per continent.

might be more aware of (and worried about) privacy issues than others, and this might impact their contributions to the PSN in terms of data sharing.

Moreover, analyzing Figures 4.3b (Rio) and 4.3d (Sydney), we can see that the coverage is not as homogeneous as in Paris, Tokyo or New York. Rio and Sydney share some geographic aspects in common. Rio has the biggest urban forest in the world, located in the middle of the city, and many hills of difficult human access. Since a central element of a PSN sensor is a human being, areas with low population density, such as rural areas, or areas with difficult access are expected to have fewer data sharing (and thus lower coverage). Residential areas with few commercial venues also contribute for a low sensing rate.

Now we analyze the number of check-ins in particular venues. Figure 4.4 presents the complementary cumulative distribution function (CCDF) of the number of check-ins per venue. First, observe that a power law fitting is appropriate to explain this distribution. Second, note that for all datasets the majority of locations have only a handful of check-ins, while there are few locations with hundreds of them. These findings are consistent with previously reported results [Noulas et al., 2011a]. As we are analyzing location sharing systems it is natural that some locations are shared more than others. For example, locations representing a restaurant or a coffee shop are more likely to be shared than a post office,



Figure 4.3: [Best viewed in color]. All sensed locations in six international cities (Foursquare datasets). The number of check-ins in each area is represented by a heatmap. The color range from yellow to red (high intensity).

despite the fact that post offices are usually very popular as well. If our application needs a more comprehensive contribution per area, we have to incentive users to participate in places that usually they would not. A punctuation system is one of many types of incentive that might work in this case.

We have seen that PSN can cover a planetary scale area. Now we verify, in Figure 4.5, the number of places that are active in a given time interval. The Foursquare-Year, Foursquare-Crawled, Gowalla, and Brightkite datasets have, respectively, approximately 490 thousands, 1,9 million, 1,3 million, and 773 thousands distinct venues. Considering the total number of distinct venues in each dataset, we find that the maximum number of active venues per day, or per hour, for Foursquare-Crawled, corresponds to only 6%, 2%, 3.3% and 0.7% of this maximum for Foursquare-Year, Foursquare-Crawled, Gowalla and Brightkite, respectively. This indicates that the instant coverage of PSN is very limited considering all locations they can reach, i.e., the probability of a random location be active in a given day is very small.

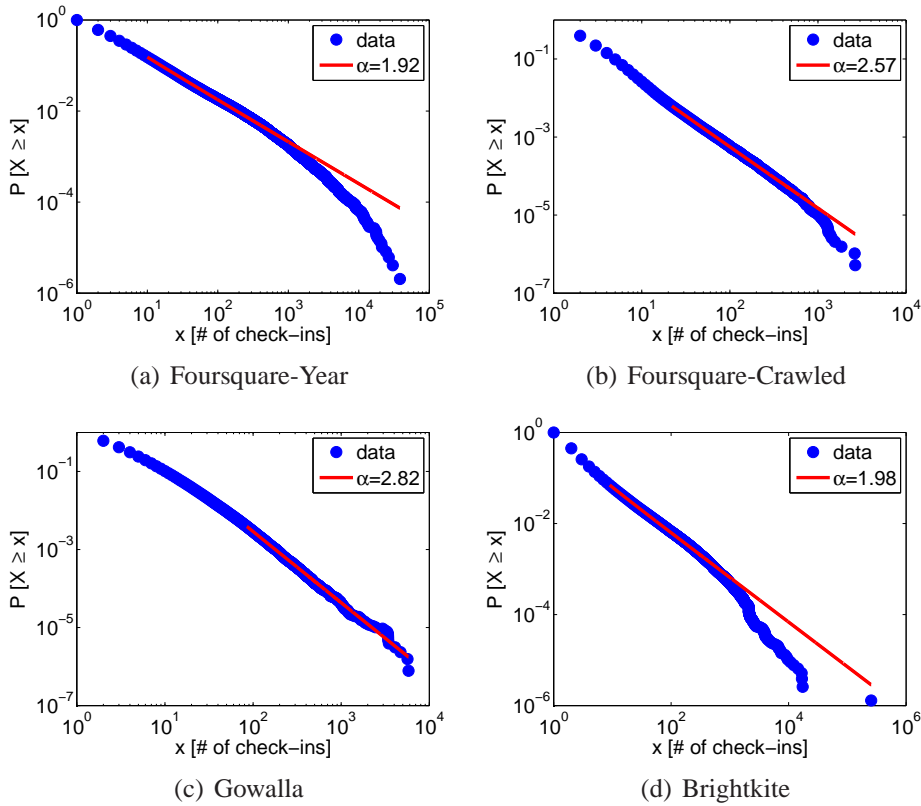


Figure 4.4: The complementary cumulative distribution function of the number of check-ins per venue.

4.1.3 Sensing Interval

Participatory sensor networks are very scalable because their nodes are autonomous, i.e., users are fully responsible for their own functioning. Since the cost of the network infrastructure is distributed among the participants, this enormous scalability and coverage are achieved without significant costs. The key challenge to the success of this type of network is to have sustained and high quality participation. In other words, the sensing is efficient as long as users are kept motivated to share their resources and sensed data frequently.

Thus, now we investigate the frequency at which users perform data sharing. Figures 4.6a, 4.6b, 4.6c, and 4.6d show the histograms of the inter-event times Δ_t between consecutive check-ins of one popular venue for the four analyzed datasets. Note that a log-logistic distribution¹ [Fisk, 1961] fits well the data. Observe the bursts of activity and the long periods of inactivity in all datasets, i.e., a large number of check-ins separated by a few minutes and also consecutive check-ins separated by several days. This may suggest that most of the data sharing, in these particular places, happen in specific intervals of time,

¹Probability Density Function: $f(x|\mu, \sigma) = \frac{1}{\sigma} \frac{1}{x} \frac{e^{-z}}{(1+e^{-z})^2}$; $x \geq 0$, where $z = \frac{\log(x)-\mu}{\sigma}$.

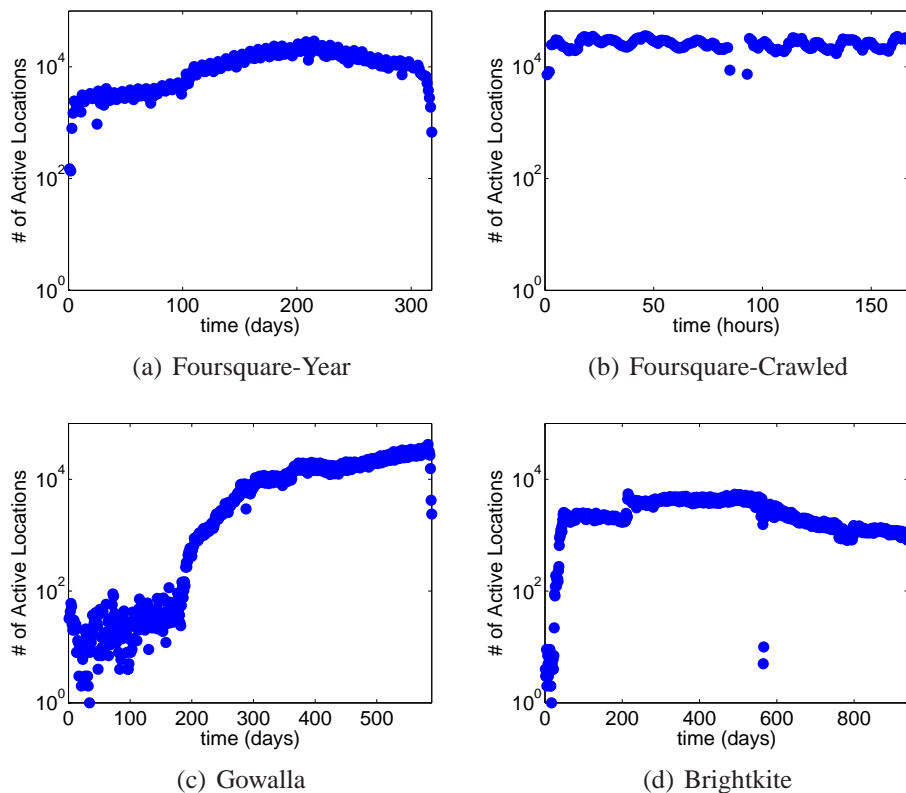


Figure 4.5: The number of locations that were active in a given day.

probably related to the time that people usually visit them (e.g., in restaurants people check-in for lunch and dinner mostly). If, for instance, an application depends on sensed data from a beach area (e.g., real-time weather), it has to be aware that very few people go to the beach at night, so the sensing data will be rare.

Another interesting observation related to the inter-event times Δ_t can be drawn from Figures 4.6e, 4.6f, 4.6g, and 4.6h. In these figures, we show the Odds Ratio (OR) function of the inter-event times Δ_t . The OR is a cumulative function where we can clearly see the distribution behavior either in the head or in the tail, and its formula is given by $OR(x) = \frac{CDF(x)}{1-CDF(x)}$, where $CDF(x)$ is the cumulative density function. As in [Vaz de Melo et al., 2011], which analyzed phone SMS usage, the OR of the inter-event times between check-ins is well fitted by a straight line with slope $\rho \approx 1$ in a plot with logarithmic scales, suggesting that the data is well modeled by a Log-logistic distribution. This is fascinating, since it suggests that the mechanisms behind human activity dynamics may be more simple and general than we know [Barabási, 2005; Malmgren et al., 2008].

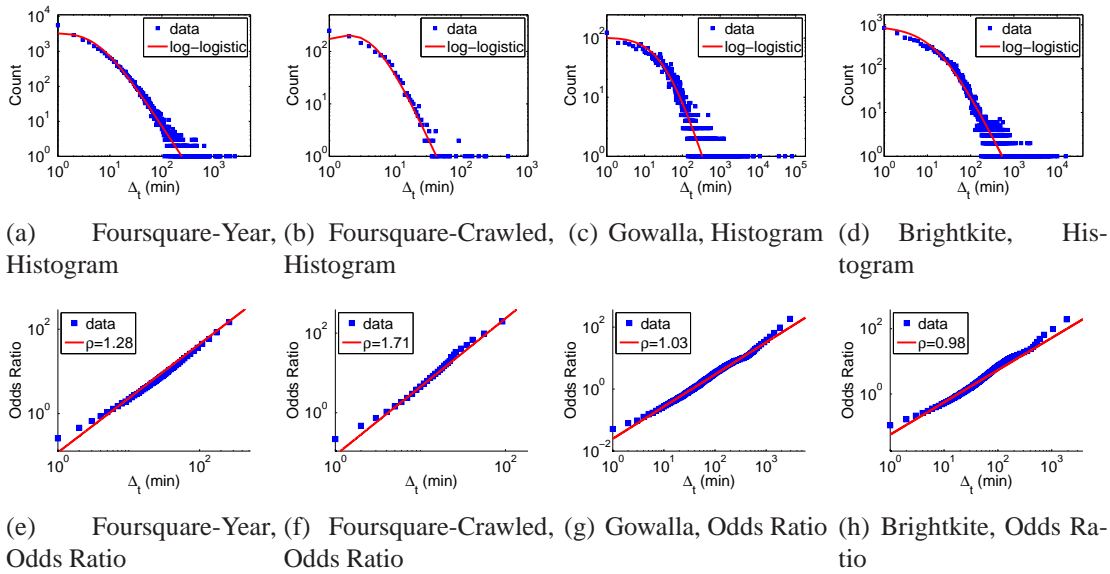


Figure 4.6: The distribution of the inter-event times between consecutive check-ins of one popular venue of each dataset.

4.1.4 Seasonality

We now analyze how the seasonal behavior of humans affects the data sharing. Figure 4.7 shows the weekly location sharing pattern for all analyzed datasets². As expected, the network actuation presents a diurnal pattern, meaning that during the dawn the sensing activity is very low. We can also observe that there are two classes of behavior: weekdays and weekends. Considering weekdays, we can note, in all datasets, an increase in the activity from Monday to Friday, as verified also by Cheng et al. [2011]. It is also possible to observe three peaks during the day, around breakfast, lunch, and dinner times. These peaks occur on every weekday, except on Friday morning. On that specific day there is no significant peak around the breakfast time. This might be due to specific behavior patterns, e.g., going out on Thursday night and waking up late on Friday morning.

We further analyze the different behavioral patterns on weekdays and weekends, focusing now on the two Foursquare datasets, as similar characteristics were observed for all analyzed systems, in this thesis and also in [Scellato et al., 2011]. Figure 4.8a shows the average number of check-ins of each hour from Monday to Friday. Figure 4.8b shows the same information for Saturday and Sunday. As we observe, the peaks during weekdays happens on 8:00 a.m. (breakfast), 12:00 p.m. (lunch), and 6:00 p.m. (dinner). On weekends, there is no peak activity in the morning, the lunch peak happens around 1:00 p.m., and the dinner peak is flatter (comprising 6:00 p.m. to 7:00 p.m.). We can also observe that the activity is

²The timestamps were normalized to the local time of the check-in.

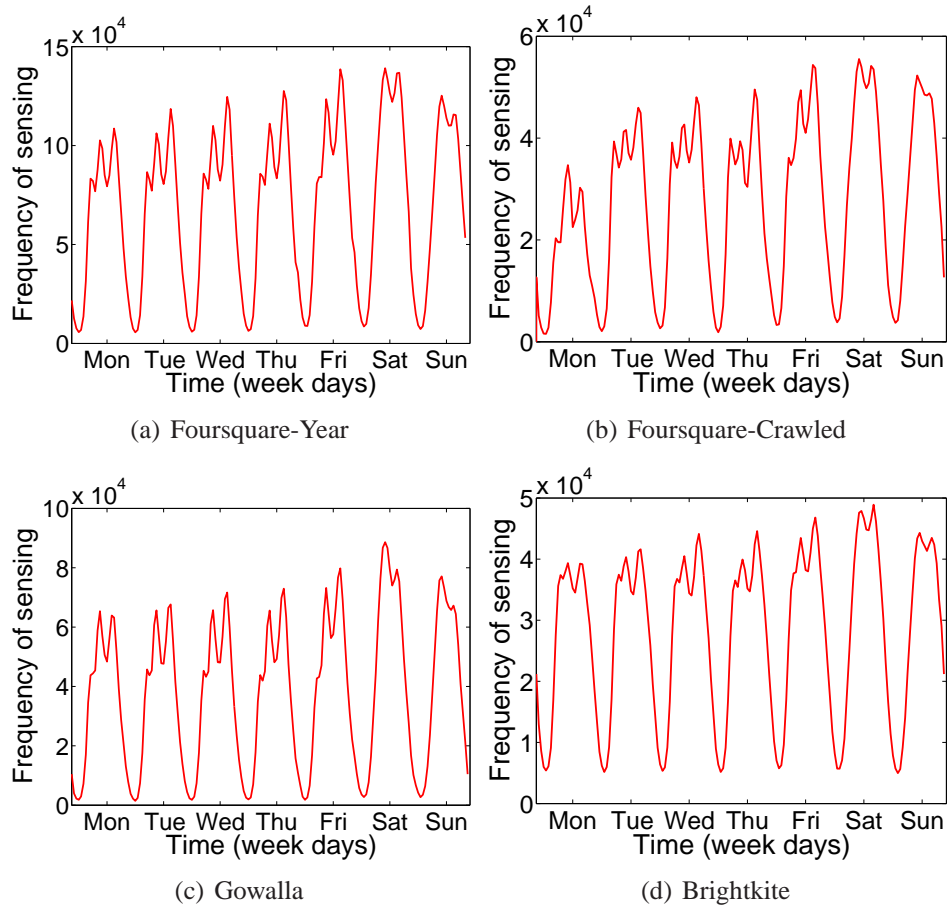


Figure 4.7: Weekly location sharing patterns.

more intense on weekends. It is worth noting that routines, usually performed on weekdays, affects considerably the data sharing.

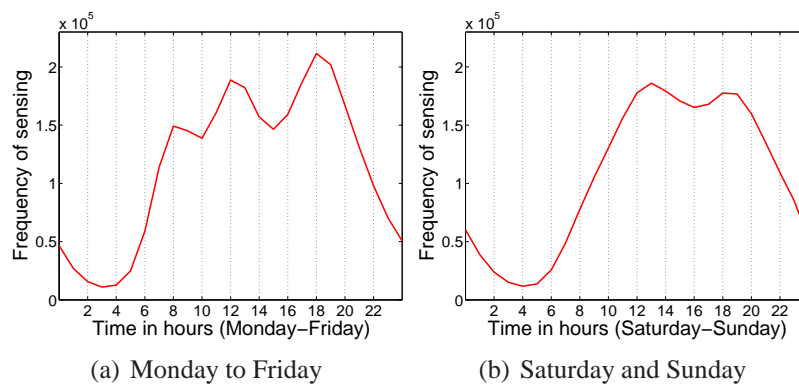


Figure 4.8: Weekdays and weekend location sharing patterns.

4.2 PSN from Photo Sharing Services

This section investigates the participatory sensor network derived from Instagram. Section 4.2.1 describes the datasets considered. Section 4.2.2 analyzes the coverage of the analyzed PSN at different spatial granularities. Section 4.2.3 looks at the frequency which nodes share data in individual locations of our dataset. Section 4.2.4 discusses the sensing seasonality, and finally Section 4.2.5 analyzes the sensing activity of each individual node (i.e., user plus *smartphone*) in the PSN.

4.2.1 Data Description

Instagram, created in 2010, is a photo sharing service that allows users to take pictures, and share them on a several social networking services, such as Twitter. Currently, Instagram users can create Web profiles featuring recently shared pictures, biographical information, and other personal details. Instagram is a very popular photo-sharing service. In February 2013 Instagram announced that they had 150 million users, and in 2014 this number reached 200 million users Instagram [2014].

The data was collected via Twitter, which enables users to announce photos available at Instagram. In this case, photos of Instagram announced on Twitter become available publicly, which by default does not happen when the picture is published solely on the Instagram system.

Between June 30 and July 31 of 2012, we collected 2,272,556 tweets containing geo-tagged photos, posted by 482,629 users. Each tweet consists of GPS coordinates (latitude and longitude) and the time when the photo was shared.

4.2.2 Network Coverage

In this section, we analyze the coverage of the PSN of Instagram at different spatial granularities, starting around the planet, then by continents and cities and ending up at neighborhoods. Figure 4.9a shows the coverage on the planet by the PSN of Instagram as a heat map of user participation: darker colors³ represent larger numbers of photos shared in the particular area. The results are similar to those observed in the Section 4.1.2. We also observe that, despite being a fairly comprehensive coverage on a planetary scale, it is not homogeneous. Figure 4.10b shows the number of photos shared by continent along the time. Note that the sensing activity in the Americas (North and South), Europe and Asia is at least an order of magnitude greater than in Africa and Oceania. Moreover, it can be observed that the par-

³Colors of the heat map for all subfigures are in the same scale.

ticipation of users in North America is slightly higher than in Latin America, Europe and Asia.

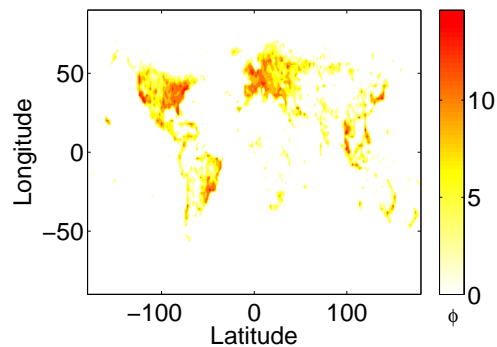


Figure 4.9: All photos shared. Number of photos n per pixel obtained from the value of ϕ shown in the figure, where $n = 2^\phi - 1$.

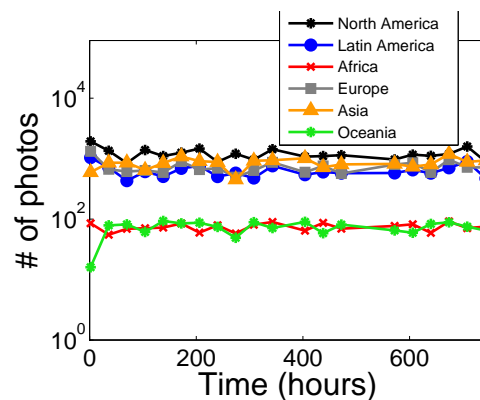


Figure 4.10: Temporal variation of the number of photos shared by continent.

Now we evaluate the participation of users in Instagram PSN in the same eight large and populous cities in five continents analyzed in Section 4.1.2. Figure 4.11 shows the heat map of the sensing activity (photo sharing) in each one of these cities. Again, darker colors represent a greater number of pictures in a given area. As in Section 4.1.2, we here can also observe a high coverage for some cities, as shown in Figures 4.11a (New York), 4.11e (Paris) and 4.11g (Tokyo). However, we can see in Figure 4.11f that the sensing in Cairo, which also has a large number of inhabitants, is significantly lower. Such difference in coverage may be explained by the same factors mentioned earlier. Besides the economic aspects, differences in the culture of the inhabitants of this city when compared with cultures present in the other cities analyzed may have a significant impact on the adoption and use of Instagram [Barth, 1969].

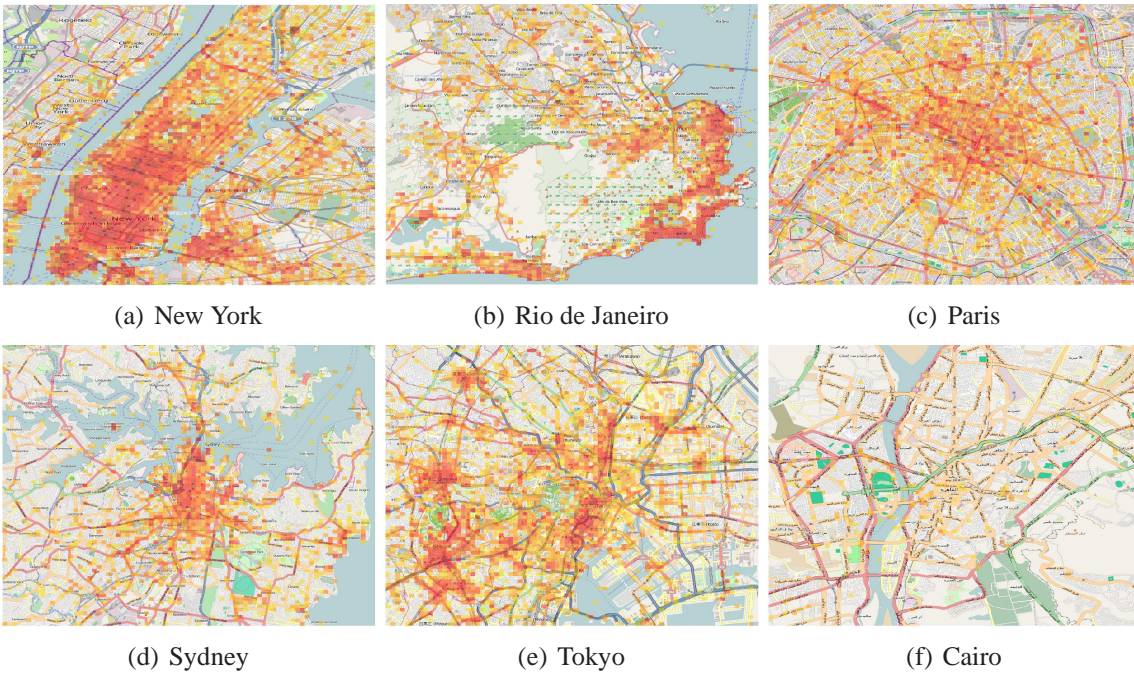


Figure 4.11: Spatial coverage of Instagram in eight cities for all shared photos. The number of pictures in each area is represented by a heat map, where the scale varies from yellow to red (more intense activity).

Furthermore, we can see that the coverage in Rio de Janeiro and Sydney is more heterogeneous compared with the coverage in Paris, Tokyo and New York. This is probably because of the geographical aspects that these cities have in common, i.e., large green areas and large portions of water, as we pointed out in Section 4.1.2. Moreover, in both cities the points of public interest such as tourist spots and shopping centers are unevenly distributed throughout the city. There are large residential areas with few points of this type, while other areas have large concentrations of these points. This observation demonstrates the potential of Instagram as a tool for participatory sensing in large urban regions.

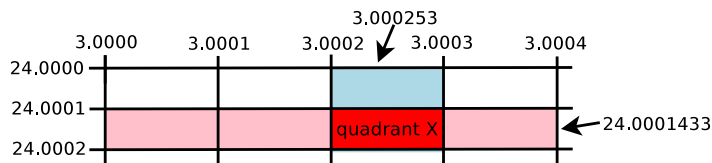


Figure 4.12: Example of identification of a quadrant.

As the users' participation can be quite heterogeneous within a city, we propose to divide the area of the cities into smaller rectangular spaces, as in a grid. We call each rectangular area of a *quadrant* within a city and, from this, we analyze the number of photos shared in these quadrants. In this thesis, we consider that a quadrant has the following delimitation:

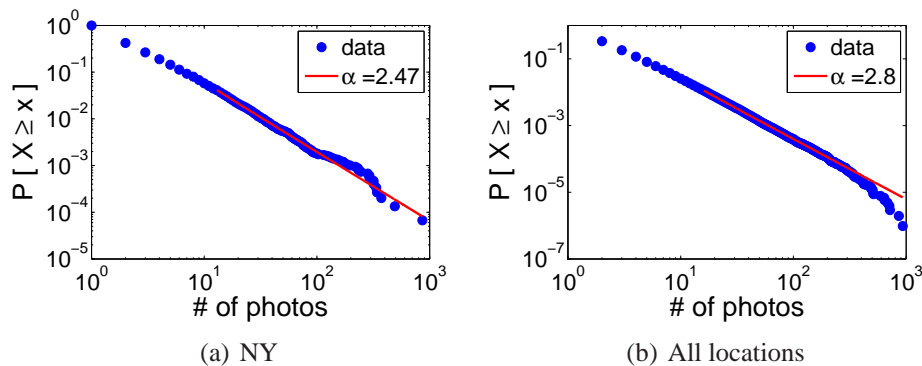


Figure 4.13: Distribution of the number of photos in quadrants.

10^{-4° (latitude) $\times 10^{-4^\circ}$ (longitude). This represents an area of approximately 8×11 meters in New York City and 10×11 meters in Rio de Janeiro. For other cities, the areas can also vary slightly, but this does not affect the analysis. We believe that this is a reasonable size to represent an area of a venue, enabling then analysis of users' activity at venue level in a city. Figure 4.12 illustrates the process of dividing the area of a city in quadrants and how it is the association of geographic coordinate (24.0001433; 3.000253) to a quadrant X.

Figure 4.13 presents the complementary cumulative distribution function (CCDF) of the number of photos shared in a quadrant of the city of New York (Figure 4.13a) and all locations in our database (Figure 4.13b). First, note that in both cases, a power law describes well this distribution. This implies that most of the quadrants have few shared photos, while there are few areas with hundreds. These results are consistent with the results for the participation of users in location sharing services ([Noulas et al., 2011a] and in Section 4.1.2). In systems for photo sharing, as well as systems for location sharing, it is natural that some areas have more activity than others. For example, in tourist areas the number of shared pictures tends to be higher than in a supermarket, although a supermarket is usually a location quite popular. If a particular application requires a more comprehensive coverage, it is necessary to encourage users to participate in places they normally would not. Micro-payments or scoring systems are examples of alternatives that might work in this case.

As previously shown, a PSN can have planetary scale coverage. However, it was also shown that such coverage can be quite heterogeneous, in which large areas are practically uncovered. Figure 4.14 shows the total network coverage considering the temporal dimension, i.e., the number of localities that are active (i.e., sensed) in a given time interval considering all available data. The maximum number of quadrants sensed per hour corresponds to only approximately 0.2% of the total number of areas in our dataset (1,030,558). In other words, the instant coverage of the PSN of Instagram is very limited when we consider all locations that could be sensed on the planet. This means that the probability of a quadrant to be sensed

on a random time is very low.

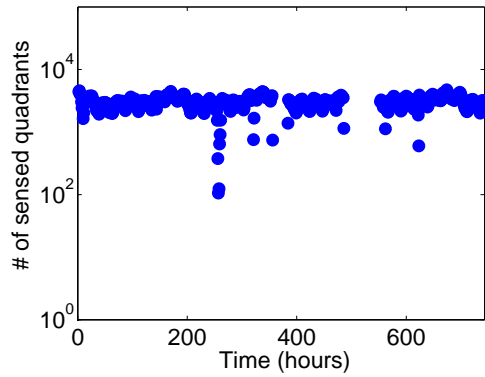


Figure 4.14: Temporal variation in the number of sensed quadrants.

4.2.3 Sensing Interval

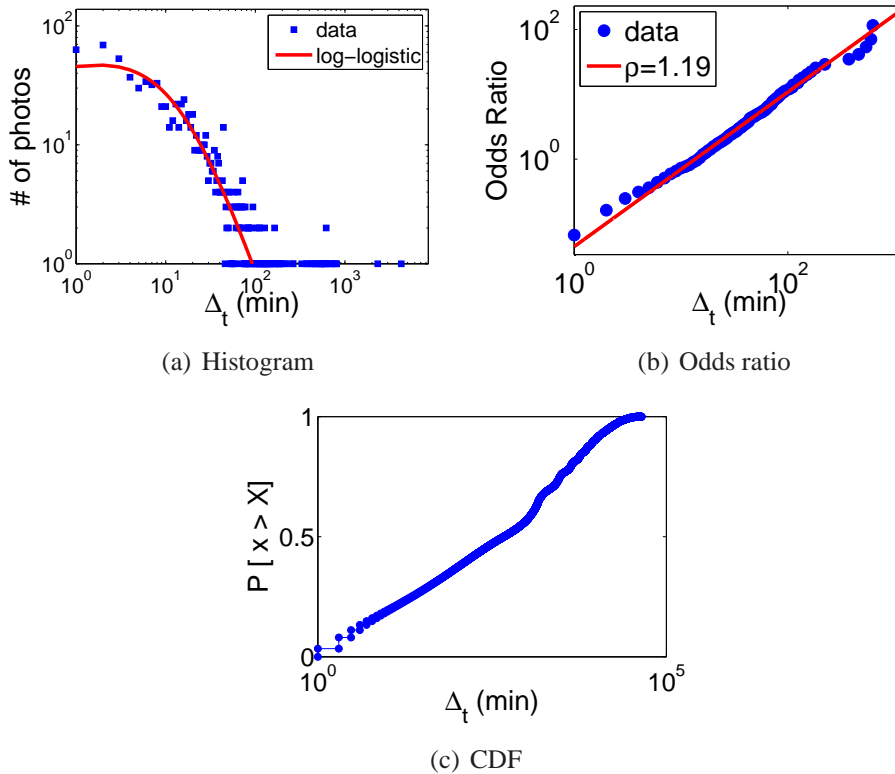


Figure 4.15: Distribution of the time interval between shared photos in a popular quadrant.

We now investigate the frequency in which users share photos in Instagram. Figure 4.15a shows the histogram of the inter-sharing time Δ_t between consecutive photos in a

typical popular quadrant. Note that the histogram is well modeled by a log-logistic distribution ($\mu = 2.605$, $\sigma = 0.839$) that has bursts of activity and long periods of inactivity: there are times when many photos are shared within a few minutes and there are times when there is no sharing for hours. Information observed also in Section 4.1.3. This may indicate that the majority of photo sharing, in this popular area (as in others), occurs at specific intervals, probably related to the time when people usually visit them. For example, sharing photos in restaurants is likely to happen during lunch and dinner times. Applications based on this type of sensing, as for location sharing services, should consider that the user participation can vary significantly along the time. Figure `refig:indiv-areasInstagramPHOTO` shows the odds ratio function (OR) of these intervals (inter-sharing time Δ_t). As found in analyses of phone SMS usage [Vaz de Melo et al., 2011] and location sharing (Section 4.1.3), the OR of the inter-sharing time between photos suggests that the data is well modeled by a Log-logistic distribution.

Based on these facts and also on Figure 4.15c, we can observe that a significant portion of users performs consecutive photo sharing in a short time interval. About 20% of all observed sharing occurs within 10 minutes. As discussed in Section 4.2.5, this suggests that nodes tend to share more than one photo in the same area. Noulas et al. [2011a] also observed a significant portion of check-ins performed in Foursquare within a short time interval. For instance, more than 10% of checkins occur within 10 minutes.

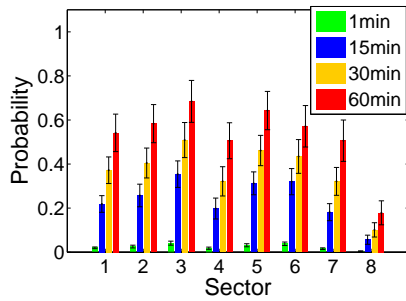
Related to this analysis, it is interesting to verify the feasibility of an application for near real-time visualization of a certain area of a city. For that, a central question is: what is the probability to obtain one picture of an area in a given time? To address this question, we select a popular area of our dataset (south of Manhattan), shown in Figure 4.16a, and divide it in eight sectors of equal size.

Figures 4.16b-e show the mean probability, along with its confidence interval of 95%, of seeing a picture in each of these sectors in the next 1-minute, 15-minutes, 30-minutes, and 60-minutes. All these probabilities are calculated for four different times of the day: dawn (Figure 4.16b), morning (Figure 4.16c), afternoon (Figure 4.16d), and night (Figure 4.16e). We observe that during the afternoon and night the difference between the probability of seeing a picture in the next 15 minutes and 60 minutes are not very high in most sectors. On the other hand, during the dawn and morning this difference is more expressive. This is explained by the low sharing frequency during the dawn and morning periods, as observed in Figure 4.17. Note also that even for a very popular area the probability to obtain a picture in the next minute is very low, for all four periods of the day. This means that applications that need a considerable amount of photos within a small interval have to be aware that this may not be feasible.

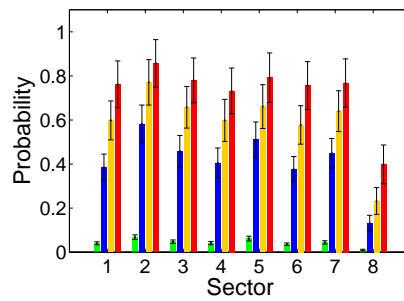
The results in Figure 4.16 can also be used to better understand those sectors. For



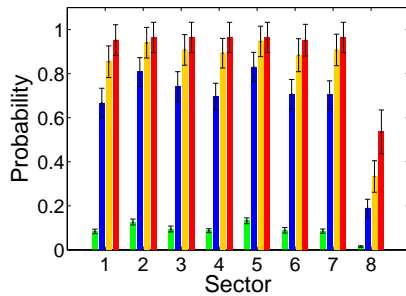
(a) Sectors of NY



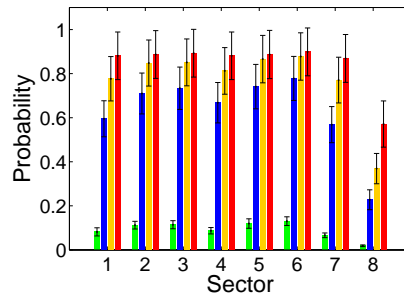
(b) Dawn



(c) Morning



(d) Afternoon



(e) Night

Figure 4.16: Mean probability of obtain a picture in the next 1-minute, 15-minutes, 30-minutes, and 60-minutes, for eight popular areas during the dawn, morning, afternoon, and night.

instance, Sector 8 seems to be the least popular among the others, despite the biggest part of water in that sector. If we analyze the probability of a photo in the next 15-minutes, we can also see that during the dawn, Sectors 3, 5, and 6 are the most popular ones, which might indicate that those sectors have a more intense nightlife. This information could be useful, for example, in a tourist guide, being one feature in an algorithm to recommend areas in a city.

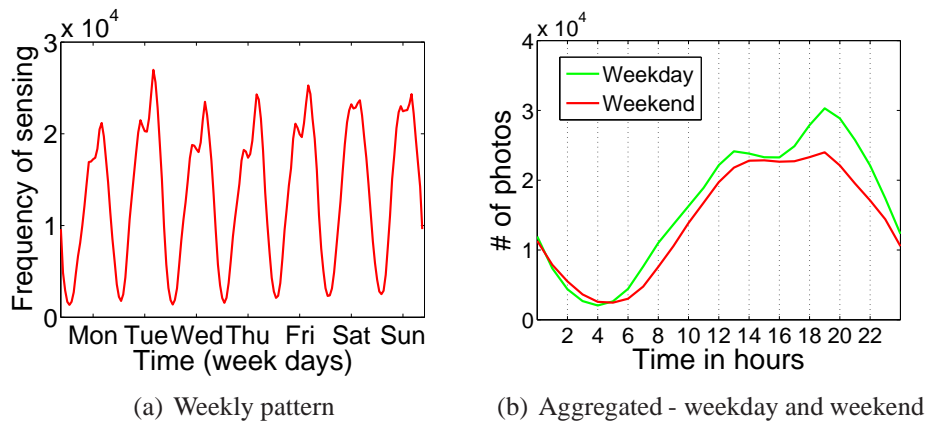


Figure 4.17: Temporal photo sharing pattern.

4.2.4 Seasonality

We now analyze how humans' routines affect the data sharing. First, we study all localities present in our dataset (Section 4.2.4.1), and then we study the sharing pattern for some cities from different continents (Section 4.2.4.2).

4.2.4.1 All Localities

Figure 4.17a shows the weekly pattern of photo sharing in Instagram⁴. As expected, the network participation presents a diurnal pattern, implying that the overnight sensing activity is quite low.

Considering weekdays, we can see a slight increase in activity throughout the week, except for Tuesday, when there is a peak of activity. We, in Section 4.1.4, and Cheng et al. [2011] analyzed location sharing systems and observed the same behavior. This suggests that during the period of data collection, an unusual event may have happened on Tuesday that resulted in an abnormal number of shared photos. Finally, observe two peaks of activity throughout the day, one around lunch and the other at dinner time. Unlike the behavior observed for location sharing (Section 4.1.4 and Cheng et al. [2011]), for photo sharing there is no peak of activity at breakfast time.

We also analyzed the behavioral patterns during weekdays and weekends. Figure 4.17b shows the average number of photos shared per hour during weekdays (Monday to Friday), and also during the weekend (Saturday and Sunday). As we can see, the peaks during weekdays happen around 13:00 (lunch) and 19:00 (dinner), but on weekends there is no peak of activity at lunchtime. Rather, the activity remains intense throughout the afternoon until early evening, with a slight increase at 19:00.

⁴The time of sharing was normalized according to the location where the photo was taken.

4.2.4.2 Selected Areas

We now turn our attention to the photo sharing pattern throughout the day in Rio de Janeiro, Sao Paulo, Osaka, Tokyo, Barcelona, Madrid, Chicago and New York City during weekdays and weekends. These results are shown in Figure 4.18⁵. It is interesting to note that, even when we analyze separate cities, we still do not observe, for most of the cities, a clear peak of photo sharing around the breakfast time, as observed for location sharing.

Studying weekdays first, we can see that cities from Japan (Figure 4.18c), Spain (Figure 4.18e) and USA (Figure 4.18g) present peaks of photo sharing that reflect typical lunch and dinner times. On the other hand, not all peaks in the Brazilian curves (Figure 4.18a) represent typical meal times. This might indicate that Brazilians share photos in uncommon moments. We conjecture that the peak of 6:00 p.m. is due a “happy hour” and the peak of 9:00 p.m. is due to a leisure activity that happens in a pub, theater, concert, etc. Another difference is that, in general, the Brazilian activity is more intense late at night. During weekdays it is possible to observe a certain similarity of sharing patterns between Japanese, Spanish, and American cities.

However, during the weekends these patterns are very distinct. The Brazilian curve still presents an unusual peak at 5:00 p.m. and the Spanish and American curves now present more intense activity around the “brunch”/lunch time. These observed patterns might express cultural behaviors of inhabitants of those countries, presenting somehow the signature of a certain culture. This hypothesis is reinforced because we surprisingly see that the pattern for each city in the same country is fairly similar on weekdays, and also on weekends, at the same time, being distinct from patterns observed for other countries.

4.2.5 Node Behavior

In this section we analyze the sensing activity of each individual node (i.e., user plus *smartphone*) in the PSN. Figure 4.19 shows that the distribution of the number of photos shared by each user of our database has a heavy tail, meaning that user participation may vary widely. For example, about 40% of users contribute with only a photo during the considered period, while only 17% and 0.1% of users contribute more than 10 and 100 photos, respectively. A heavy tail in the distribution of the number of performed check-ins was also observed in [Noulas et al., 2011a]. About 20% of users have just one check-in, with 40% above 10, whereas there is a set of approximately 10% that has more than 100 check-ins.

We also analyze the geographical distance between two consecutive photos shared by the same user, according to the geographic coordinates associated with each photo. Fig-

⁵Each curve is normalized by the maximum number of photos shared in a specific region representing the city.

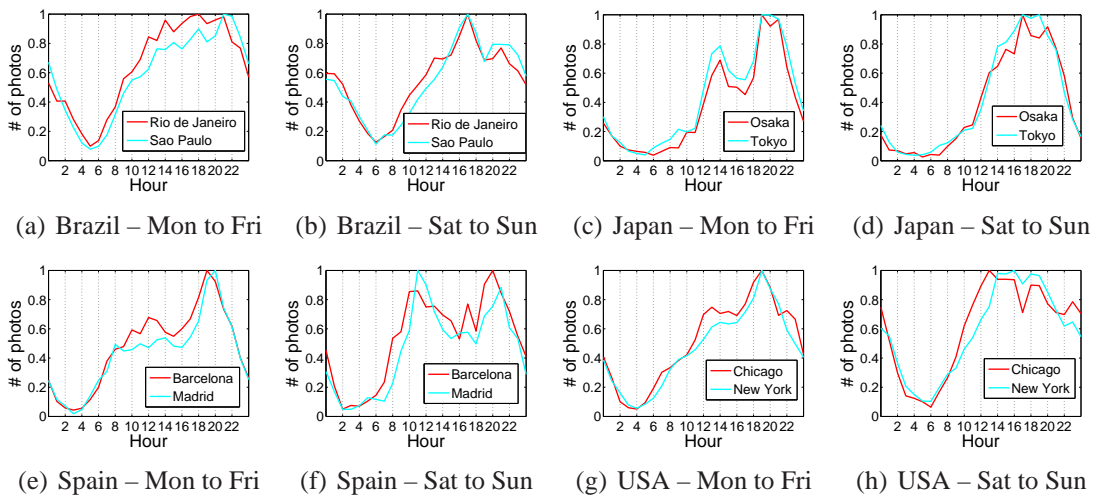


Figure 4.18: Photo sharing throughout the day in Rio de Janeiro, Sao Paulo, Osaka, Tokyo, Barcelona, Madrid, Chicago and New York City.

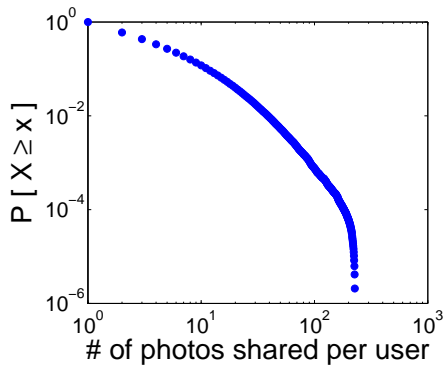


Figure 4.19: Distribution of the number of photos shared by people.

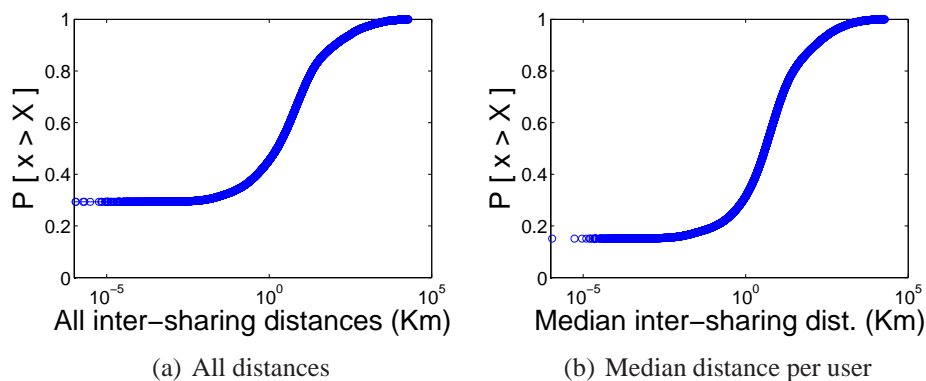


Figure 4.20: Distribution of the geographical distance between consecutive pictures of the same person.

Figure 4.20a shows the cumulative density function of the geographic distance between each pair of consecutive photos shared by each user in our dataset. It can be observed that a significant portion (about 30%) of the distances between consecutive photos are very short (less than 1 meter). This indicates that users tend to share multiple photos in the same location. This hypothesis is reinforced by the significant portion of time intervals between consecutive pictures of short duration shown in Figure 4.15c: 20% of these intervals (Δ_t) do not exceed 10 minutes. This was not observed in the same proportion for location sharing. Noulas et al. [2011a] observe that 20% of the shared locations happen up to 1 km away. For shared photos, this value is approximately 45%. This result can be explained by the simple fact that a photo can contain much more information than one location. For example, in a restaurant users could share photos of his/her friends at the place, food, or a particular situation, but tend to share their location only once.

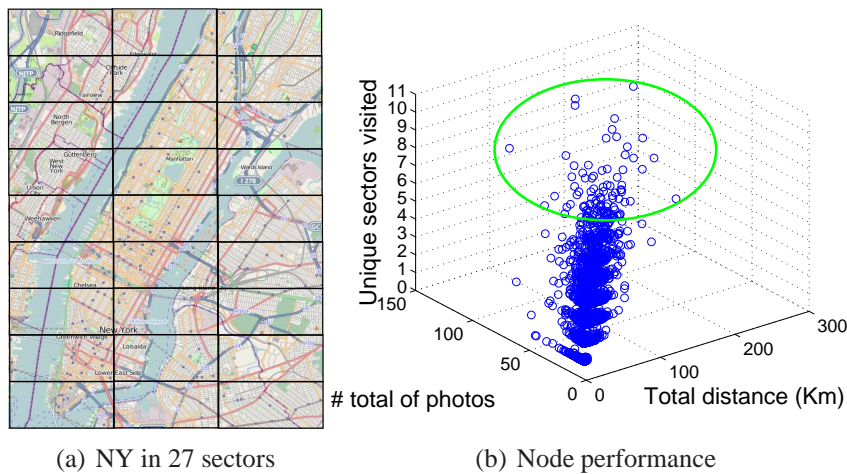


Figure 4.21: Contribution of nodes, distance traveled, and coverage.

We now analyze each user separately. Figure 4.20b shows the distribution of the median distance between consecutive sharing computed for each user. Note that at least 50% of consecutive photos of a significant portion of users (about 20%) are taken at a very short distance (around 1 meter).

Finally, we study the performance of nodes considering the total traveled distance, the coverage in the city of New York (NY), and total number of contributed photos. To analyze the coverage, we consider the area of NY (Figure 4.21a), which was divided into 27 sectors of equal size. Figure 4.21b shows a 3-D plot for the three dimensions considered. We are able to observe the existence of “super nodes” in the system, indicated by a green circle. This nodes share a lot of photos, travel long distances, and visit many different areas in the city (observed by the number of unique visited sectors). The identification of this type of users is

important for several reasons. As the success of a PSN relies on a continuous contribution, it is interesting to award this type of user to keep them active in the network. Besides that, nodes of this type might be good candidates to be selected, for example, in a network for information dissemination a city.

4.3 PSN from Traffic Alert Services

This section investigates the participatory sensor network derived from Waze. Section 4.3.1 describes the used dataset. Section 4.3.2 studies the coverage of the considered PSN. Section 4.3.3 analyses the frequency that users share alerts. Section 4.3.4 studies how user routines affect the temporal frequency of alert sharing. Finally, Section 4.3.5 analyses the contribution of individual users in the PSN derived from Waze.

4.3.1 Data Description

Waze is a popular navigation system that uses crowdsensing to offer near real-time traffic information and routing. The system was created in 2008 and registered approximately 50 million users in 2013. Waze periodically collects data from the built-in GPS typically found in smart phones, and uses it to compute the speed of the device. With that, Waze can provide useful information about traffic conditions in different areas. The system also offers to its users predefined alerts stating incidents such as traffic jams and police traps, which extends the information about traffic conditions. It is also possible to use subcategories of incidents to better specify them, for example, “heavy traffic jam” instead of just “traffic jam”.

Here, we are interested in characterizing user participation in the dissemination of alerts about traffic. To that end, we collected a dataset of Waze alerts directly from Twitter, since Waze traffic information is not publicly accessible by an API. Our dataset covers the period from December 21st, 2012 to June 28th, 2013, and consists of 212,814 tweets containing alerts about traffic shared by Waze users, each one providing the user id, type of incident (e.g., traffic jam), and the address of the incident. In order to obtain the latitude and longitude of the provided address, we performed a geocoding process using the Bing Maps API⁶, which provides the confidence of the result’s quality: low, medium, and high. We excluded all results classified as low. After this filtering process, we extracted 162,212 tweets containing alerts, shared by 21,852 users.

In Figure 4.22, we provide an overview of types of alerts reported by users of our dataset, using word clouds to represent the relative frequency⁷. Alerts were translated into

⁶<http://www.microsoft.com/maps/developers/web.aspx>.

⁷The size of the word indicates its popularity.

hazard
 accident traffic-jam police

Figure 4.22: Overview of reported alerts.

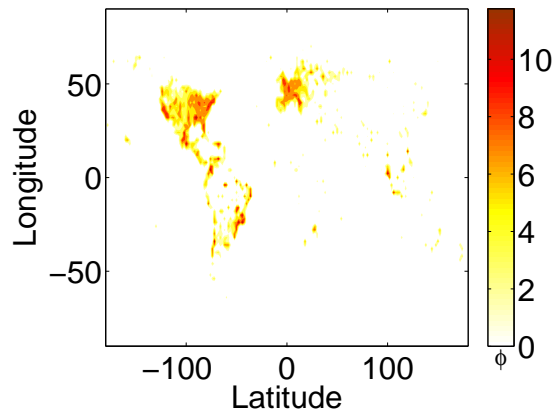


Figure 4.23: All sensed locations. The number of locations n per pixel is given by the value of ϕ displayed in the colormap, where $n = 2^\phi - 1$.

English using a manually created dictionary of translation. As we can see, the most common type of reported alert is traffic jam⁸, though police and hazard are also very popular.

4.3.2 Network Coverage

In this section, we discuss the spatial coverage of the PSN derived from Waze. In this direction, we first built a heatmap with all alerts shared by users in our dataset, shown in Figure 4.23. As in location and photo sharing services (Sections 4.1.2 and 4.2.2, respectively), we note that user participation in Waze is global, however, it is low in certain regions, particularly Asia. Then we selected the most popular cities for further analysis.

A popular city from our dataset is shown in Figure 4.24. In this figure we show the number of alerts in different regions of Rio de Janeiro by a heat map, where the scale varies from yellow to red (more intense activity)⁹. The spatial coverage is not as proliferated as the one observed in location and photo sharing systems (Sections 4.1.2 and 4.2.2). A factor

⁸Alerts containing a subcategory of an incident were unified to its main category, for example, “heavy traffic jam” was associated to the word “traffic jam”.

⁹The darkest red represents a region with 508 alerts.



Figure 4.24: Spatial coverage of Waze in Rio de Janeiro.

that might help to explain it is the user population of our dataset, which is smaller than those reported in the mentioned studies. Another factor is that users might have fewer opportunities to share traffic alerts, compared to opportunities to share photos or check-ins.

In order to evaluate user participation across different regions at a finer granularity, we propose to divide the geographical area of each city into smaller rectangular spaces (or quadrants), as performed in the analyzes for the PSN derived of photo sharing services (Section 4.2.2).

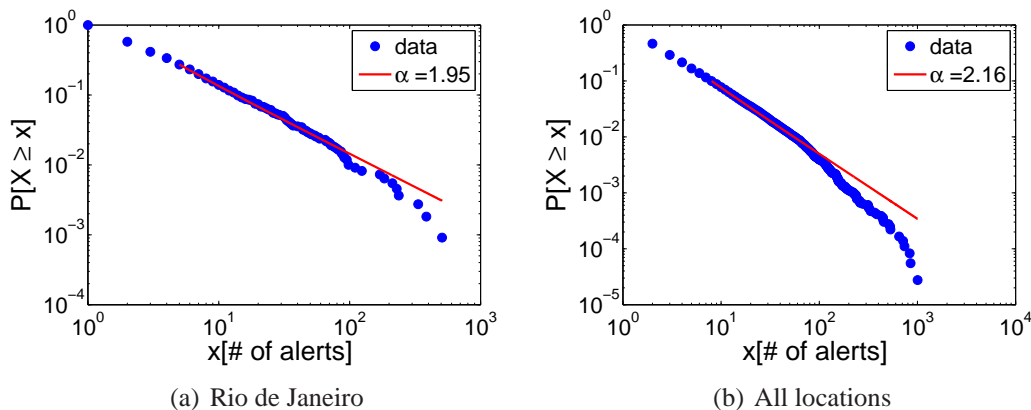


Figure 4.25: Distribution of the number of alerts.

The complementary cumulative distribution functions of the number of alerts shared in a quadrant of the city of Rio de Janeiro, as well as across all locations in our dataset are presented in Figures 4.25a and 4.25b, respectively. Note that a power law describes well this distribution in both cases. This means that few areas have hundreds of shared alerts, while most of the quadrants have just a small number. This finding is consistent with previous results about user participation in location sharing services, as shown in Section 4.1.2 and

by Noulas et al. [2011a], and photo sharing services, shown in Section 4.3.2. As in those other services, it is likely that some areas, such as large avenues in downtown, have more activity of traffic alerts. Note that the number of vehicles circulating on each region greatly impacts the local coverage of a traffic alert sharing system such as Waze, as shared alerts often refer to traffic jams and hazards, or even police traps (see Figure 4.22), which tend to occur more often in locations with heavier car flow. This is in contrast to location and photo sharing services, where places often visited by a large number of people are not necessarily covered by a large amount of shared data, because the motivation of users to share data in such systems is different from Waze. For example, a large supermarket may be visited by a large number of people on a daily basis, but it is not likely that those people will share many check-ins or photos at it.

4.3.3 Sensing Interval

We now analyze the frequency in which users share alerts in Waze. The histogram of the inter-sharing time Δ_t between consecutive alerts (performed not necessarily by the same user), in a popular quadrant is shown in Figure 4.26a. Note that a log-logistic distribution ($\mu = 2.931$, $\sigma = 1.065$) fits well the data, reflecting the fact that there are times when many alerts are shared within a few minutes and there are times when there is no sharing for hours. As also observed for location (Section 4.1.3) and photo (Section 4.2.3) sharing services, this result may indicate that the majority of alert sharing occurs at specific intervals. For instance, alerts are more likely to be common in urban areas during rush hours.

In Figure 4.26b, we show the odds ratio function (OR) of inter-sharing time Δ_t . As also observed in previous analyses of phone SMS usage [Vaz de Melo et al., 2011], location sharing (Section 4.1.3), and photo sharing (Section 4.2.3), the OR function also suggests that the inter-sharing time between alerts also is well modeled by a Log-logistic.

The CDF of all observed inter-sharing times performed by any user in the same quadrant is shown in Figure 4.26c. As we can observe, a considerable portion of users perform consecutive alert sharing in a short time interval. This was also observed for photo sharing (Section 4.2.3) and location sharing [Noulas et al., 2011a]. This is expected to happen for traffic alert services because, for example, when an accident happens many users tend to share it in a short interval.

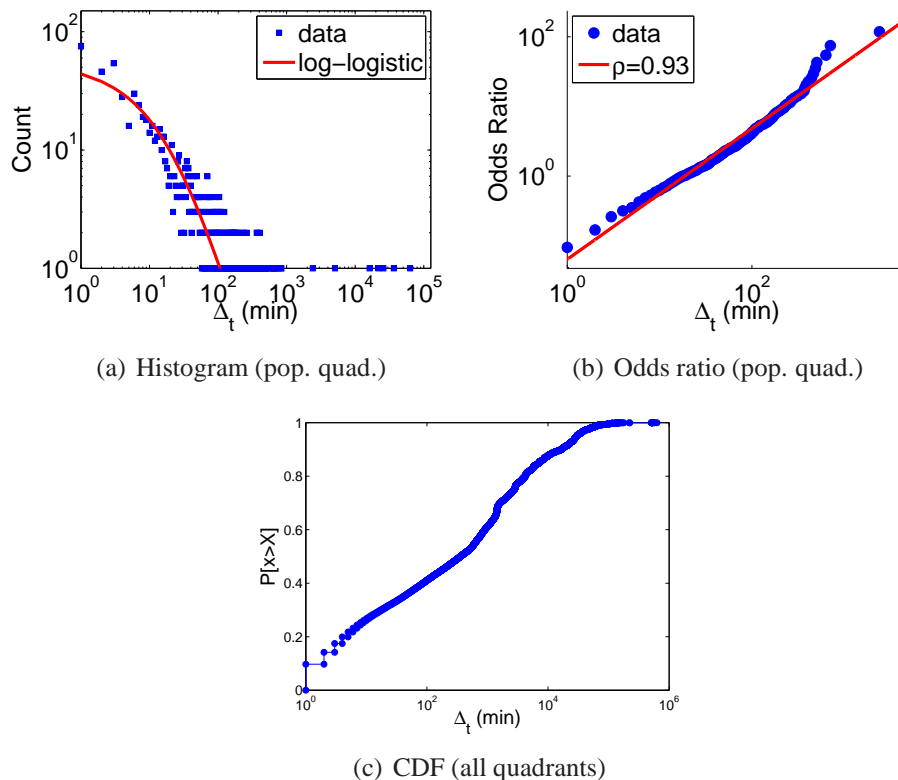


Figure 4.26: Time intervals between consecutive alerts, not necessarily done by the same user.

4.3.4 Seasonality

In this section, we study how user routines affect the temporal frequency of alert sharing. In Figure 4.27a, we show the temporal variations¹⁰ of the number of alerts shared throughout the week (Monday to Sunday), for all locations of our dataset. As expected, user participation presents a diurnal pattern, and the activity during late night hours and dawn is much lower than previously observed in location and photo sharing patterns (Sections 4.1.4 and 4.2.4, respectively). During that period, traffic problems are typically rare, whereas users have more opportunities to share data in location and photo sharing systems (e.g., in a night club or in a concert).

Intense user activity during the weekends, as observed in location sharing services (Section 4.1.4 and [Cheng et al., 2011]) and photo sharing services (Section 4.2.4), is not observed for traffic alerts. This might indicate that the reasons motivating users to contribute alerts are distinct from the ones to perform check-ins. In Figure 4.27b, we show the average number of data sharing throughout the day, separately for weekdays (Monday to Friday) and

¹⁰The time of sharing was normalized according to the timezone where the alert was shared.

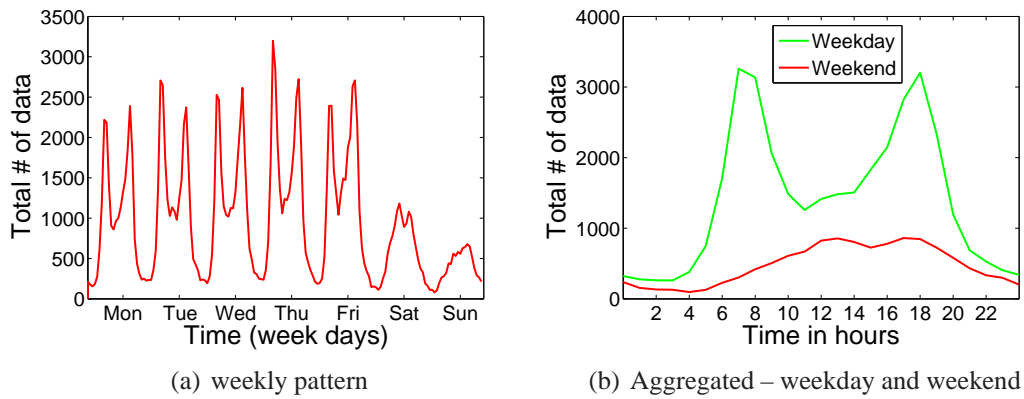


Figure 4.27: General temporal sharing pattern (all locations).

weekends (Saturday and Sunday). Note the two clear peaks of activity, one around 7 to 8 a.m. and the other around 6 p.m., coinciding with typical rush hours in urban areas. This result is different from the three clear peaks previously observed in location sharing services (Section 4.1.4 and [Cheng et al., 2011]), around breakfast, lunch and dinner times, as well as from the two peaks during lunch and dinner times in photo sharing (Section 4.2.4).

We now analyze the hourly variations of alert sharing in six large cities: Chicago and New York (Figure 4.28a¹¹) in USA; Belo Horizonte and Sao Paulo in Brazil (Figure 4.28b); and London, and Paris (Figure 4.28c) in Europe. Note that the curve of each city follows the general trend observed for all locations (Figure 4.27b).

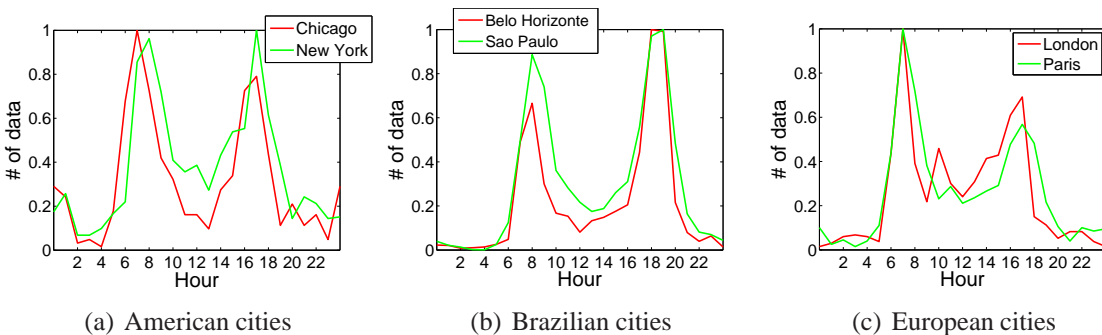


Figure 4.28: Alerts sharing throughout the day in different cities around the world.

We can also observe that the peaks reflect distinct rush times that are related to the common working hours of different cities. In Chicago (Figure 4.28a) the morning peak is around 7 a.m., as in the two European cities (Figure 4.28c). In contrast, in New York and in the Brazilian cities (Figure 4.28b), the morning peak is usually one hour later, suggesting that people tend to leave later to work in those cities. The second most expressive peak in

¹¹Each curve is normalized by the maximum number of alerts shared in the city in question.

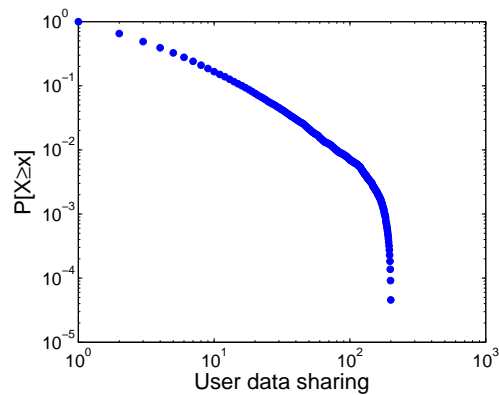


Figure 4.29: CCDF of the number of shared alerts (same user).

both American cities is around 5 p.m., which is similar to the European cities. However, this is distinct from the Brazilian cities, which have a peak of activity around 6 p.m..

To complement this analysis, we performed, from July 16th to July 18th, an hourly collection of traffic conditions of Paris, using Google Maps. We note that the time of the observed peaks reflects relatively well intense traffic conditions reported by Google Maps, whereas the reduced activity prior and after the peaks also reflects better traffic conditions. This suggests that this information could be used to assure the quality and improve traffic condition information services, such as those offered by Google Maps.

4.3.5 Node Behavior

We now analyze the contribution of individual nodes in the PSN derived from Waze. In Figure 4.29, we show that the distribution of the number of alerts shared by each user of our dataset has a heavy tail, as observed for photo sharing (Section 4.2.5) and location sharing [Noulas et al., 2011a]. This implies in a great variability of user participation. For instance, 35% of the users contributed with only one alert during approximately the six-month period covered by our dataset, while 16% and 0.006% of users contributed with more than 10 and 100 alerts, respectively, in the same period. These proportions are similar to those observed in photo sharing.

We now analyze the spatial distance between consecutive alerts by the same user, by taking the distance [Sinnott, 1984] between the geographic coordinates associated with both alerts. In Figure 4.30a, we show the CDF of the distances between consecutive alerts shared by each user, for all users. Note that a large portion of the distances are very short: for instance, around 30% are below 1 meter. Such large fraction of small distances between consecutive sharing were also observed in photo sharing (Section 4.2.5) and, to a lesser extent, location sharing services [Noulas et al., 2011a]. For location sharing 20% of the

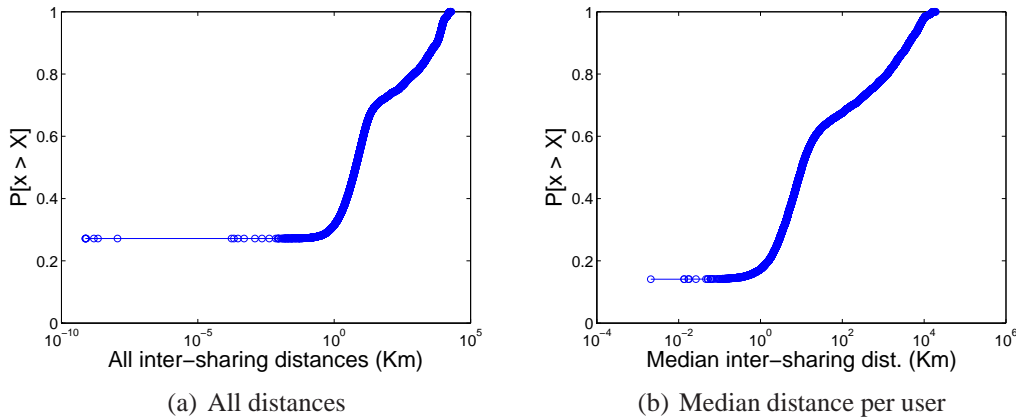


Figure 4.30: Distribution of the geographical distance between consecutive data of the same person.

consecutive sharing by the same user were in locations that were apart from each other by up to 1 km. For photos and alerts, this fraction raises to approximately 45% and 80%, respectively. This suggests that users tend to share multiple alerts in the same location.

In Figure 4.30b, we show similar results for the distribution of the median distance between consecutive sharing for each user. That is, even aggregating results for each user, we still observe that alerts are shared at very short distances: around 15% of users share alerts 1 meter apart from each other.

4.4 Comparing PSNs from different systems

Social networks and social software have been driven by two aspects: connections between people who use them and the information they share, in particular location-related information [Smith et al., 2005]. In this section we are interested in comparing different PSNs, two datasets of Foursquare (Foursquare-Crawled and Foursquare-New), and two datasets of Instagram (Instagram-OLD and Instagram-New¹²). Table 4.2 summarizes all the used datasets in this chapter. We analyze those datasets to investigate whether we can observe the same users' movement pattern, the popularity of regions in cities, the activities of users who use those social networks, and how users share their content along the time. In answering those questions, we want to better understand location-related information, which is an important aspect of the urban phenomena.

This section compares the four datasets of the two social networks using location-related information as the main aspect of the analysis, and is organized as follows. Sec-

¹²Note that Instagram-New and Foursquare-New have the time of collection in common, which is not the case for Instagram-OLD and Foursquare-Crawled datasets.

| System | # of data | Interval |
|--------------------|---------------------|-----------------------|
| Foursquare-Crawled | 4,672,841 check-ins | Apr/2012 (1 week) |
| Foursquare-New | 4,548,941 check-ins | 11 May 13 – 25 May 13 |
| Instagram-OLD | 2,272,556 photos | 30 Jun 12 – 31 Jul 12 |
| Instagram-New | 1,855,235 photos | 11 May 13 – 25 May 13 |

Table 4.2: Dataset information. Note that Foursquare-Crawled was already analyzed in Section 4.1, and Instagram-OLD in Section 4.2.

tion 4.4.1 study the user behavior on the considered systems. Section 4.4.2 studies the popularity of different areas. Section 4.4.3 studies how the routines affects the data sharing. Section 4.4.4 analyses the transitions performed by people.

Throughout this section we consider three large and populous cities (New York, Sao Paulo, and Tokyo) in several analyses. Figure 4.31 shows the heat map of the coverage of the datasets for each city, containing all data from Instagram-New and Foursquare-New. The darker the color¹³ in the figure, the higher the number of content shared in that area. The coverage for the datasets Foursquare-Crawled and Instagram-OLD is presented in Sections 4.1 and 4.2, respectively.

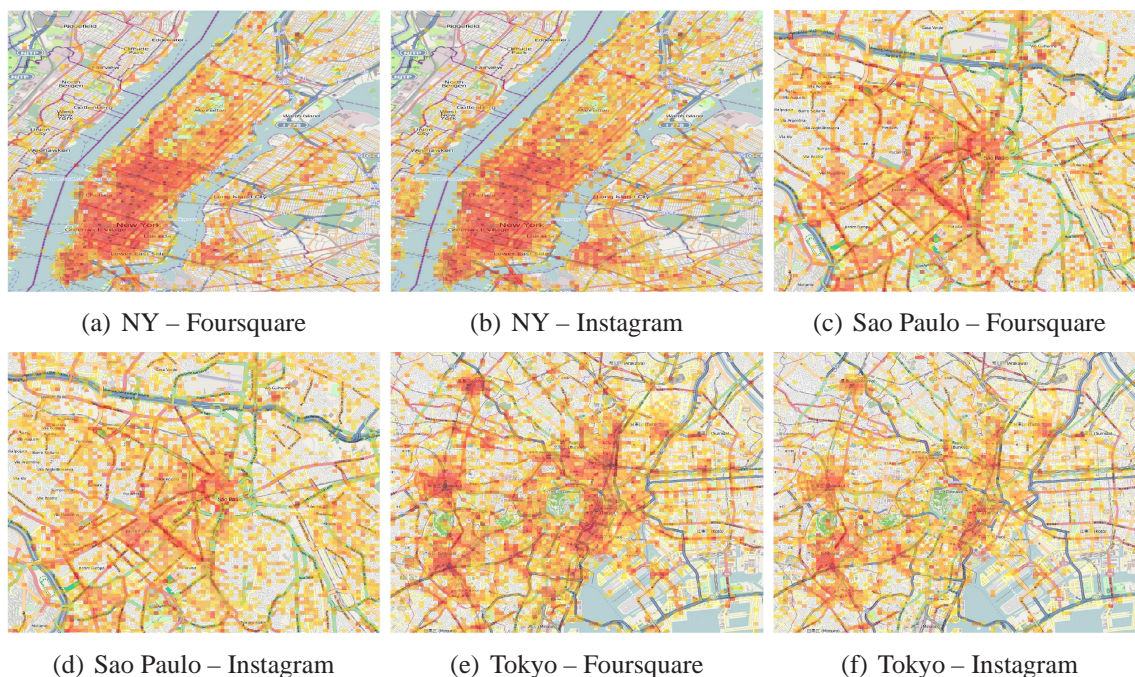


Figure 4.31: All sensed locations in three populous cities. The number of check-ins in each area is represented by a heat map. The color range from yellow to red (high intensity).

¹³Colors of the heat map for all subfigures are in the same scale.

4.4.1 User Behavior

Considering the Instagram-New and Foursquare-New (datasets with a common collection time), we group users in three classes: (1) users that only participated in Instagram; (2) users that only participated in Foursquare; and (3) users that participated in both systems. Figure 4.32a shows the cumulative density function (CDF) of the frequency of sharing content per class, showing the inter-sharing time Δ_t in minutes between consecutive content sharing. We can observe that Class 1 (Instagram only), and Class 3 (both systems) contribute more content in shorter intervals than Class 2. For instance, approximately 20% of users in Class 1 and 3 share a consecutive content in an interval up to 10 minutes. In Class 2, the portion of users that share content up to 10 minutes is approximately 12%. This suggests that users tend to share more content in the same place when using Instagram. This was also observed in Section 4.2. The sharing pattern of Class 3 might be dominated by the use of Instagram, explaining the closer similarity among the curves. It is natural to expect a higher volume of content to be shared in the same place through Instagram than in Foursquare. For instance, in a night club users can share a photo of the place, of a drink, and friends.

Figure 4.32b shows the CDF of the median distance between consecutive uploads for each user. We observe that a significant portion of users from Class 1, around 20%, shared consecutive content at a very short distance, around 1 meter (this was also observed in Section 4.2). This is not observed in the same proportion for the other classes of users. The results for Classes 2 and 3 are 3% and 15%, respectively. This reinforces what was previously observed, i.e., users tend to share content in a shorter distance in Instagram than in Foursquare. For instance, Noulas et al. [2011a] observed that 20% of the shared locations happen up to 1 km away. Again, the behavior of users that participate in both systems (Class 3), is more similar to Class 1. This closer similarity might be explained by a more intense content contribution in Instagram.

The understanding of user behavior is the first step to model it. With models that explain the user behavior we can make predictions of actions and develop better capacity planning of the system that supports the service.

4.4.2 Popularity of Areas

How is the popularity of regions across PSNs derived from Instagram and Foursquare? This is probably one of the main issues in an urban scenario. To answer this question we divided the areas of New York, Sao Paulo, and Tokyo in a 10×10 grid, as shown in Figure 4.33. After that, we verified the number of content (photo or check-in) shared in each cell of the grid for all four considered datasets. Then, we correlated the number of content in each cell using the Pearson correlation. This result is shown in Figure 4.34. As we can see the

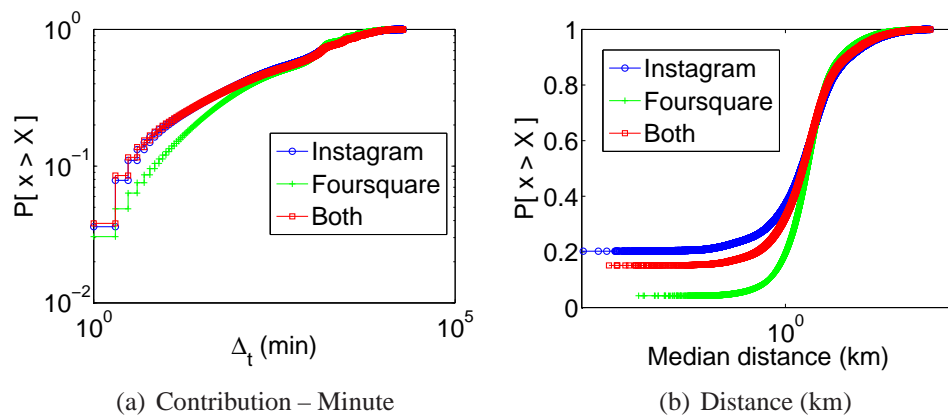


Figure 4.32: Analysis of classes of users.

correlation is very high among all datasets. The lowest correlation, although still high, was observed in Tokyo with respect to the correlation of Instagram and Foursquare (both old and new). This might indicate that the popularity of regions inside cities is consistent regardless of the system, and over the time. Recall that we use two datasets with the same collection time (Instagram-New and Foursquare-New) and two datasets with different collection time (Instagram-OLD and Foursquare-Crawled). Besides that, the difference of time between the “new” datasets and the “old” ones are of approximately one year. Maybe what is popular in the city tend to remain popular for a long time and is captured by both systems, since they allow users to express their routines freely.

Next, we verified if the popularity of a city is consistent across the systems. Popularity in this case is measured by the number of content shared in the city. For that we considered 29 cities around the world¹⁴: Latin American cities (Belo Horizonte, Buenos Aires, Mexico City, Rio, Santiago, and Sao Paulo); American cities (Chicago, Los Angeles, New York, San Francisco); European cities (Barcelona, Istanbul, London, Madrid, Moscow, and Paris); Asian cities (Bandung, Bangkok, Jakarta, Kuala Lumpur, Kuwait, Manila, Osaka, Semarang, Seoul, Singapore, Surabaya, Tokyo); and Australian cities (Melbourne, Sydney). We ranked all the cities by the number of content shared on it, then we correlated these ranking using Spearman correlation. Figure 4.35 displays the correlation results. As we can see the popularity of cities tend to be very correlated over time for the same system, but this is not the case for different systems. This means that users may use Instagram and Foursquare in particular ways on different cities. For instance, Foursquare might be very popular in Tokyo, but Instagram might not be as popular. Cultural differences might help to explain these results.

¹⁴Chosen by their popularity and representativeness of different regions.

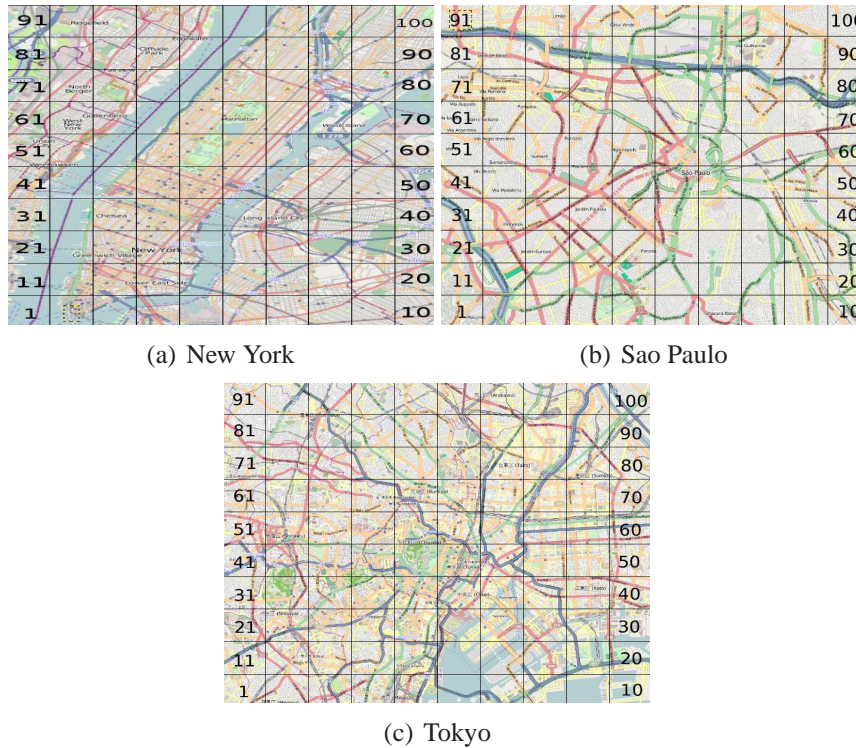


Figure 4.33: Grids for the areas of New York, Sao Paulo, and Tokyo.

4.4.3 Routines and the Data Sharing

Figure 4.36 shows the temporal sharing pattern for Instagram and Foursquare considering the old and new datasets. This figure shows the average number of photos shared per hour during weekdays (Monday to Friday), and also during the weekend (Saturday and Sunday). As previously observed in Section 4.2 for the Instagram-OLD dataset, we can also see two peaks of activity throughout the day, one around lunch and the other at dinner time. But, we cannot see a clear peak at breakfast time, as the one observed in Figure 4.36c and also in [Cheng et al., 2011]. During the weekends we cannot observe clear peaks of activities inherent of routines. Rather, the activity remains intense throughout the afternoon until early evening.

Surprisingly, we see that the sharing pattern for each curve regarding to the old and new datasets, both on Instagram and Foursquare, are very similar, despite the huge gap between collections (approximately one year). This is the case for weekdays and weekends, suggesting that the user behavior in both systems tend to keep consistent over time, reinforcing what was observed in Section 4.4.2. This is an interesting and important result because it shows how we can use different datasets.

In Figure 4.37 we show the correlogram for the temporal sharing pattern of Instagram-New and Foursquare-New datasets, during the weekday (Figure 4.37a) and weekend (Fig-

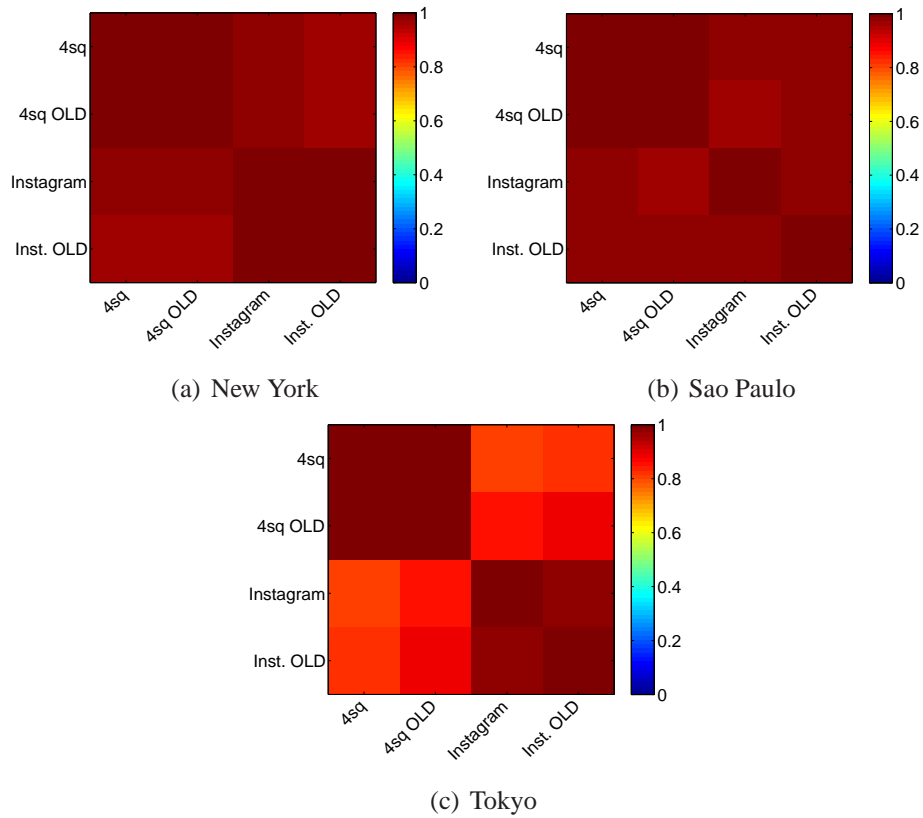


Figure 4.34: Correlation of popularity of sectors inside cities.

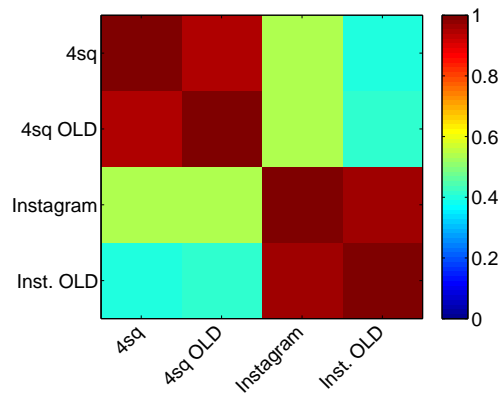


Figure 4.35: Spearman correlation of popularity between cities.

ure 4.37b). The correlogram plots correlation coefficients on the vertical axis, and lag values (in hours) on the horizontal axis, and it is an important tool for analyzing time series in the time domain. As we can see, the lag of one hour in the time series of Instagram-New dataset provides the highest correlation, however it is not 1 (maximum). Analyzing the cross-correlation for weekend, we observe that a lag of 0 provides a correlation of 1, indicating that the time series is already very correlated. This suggests that users have particular shar-

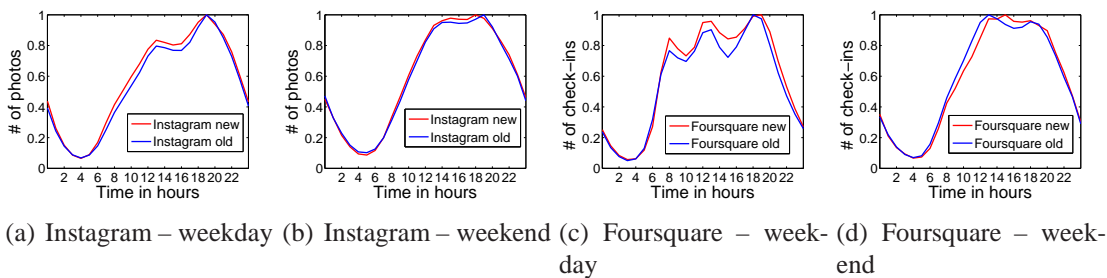


Figure 4.36: Temporal sharing pattern for Instagram and Foursquare – new and old datasets.

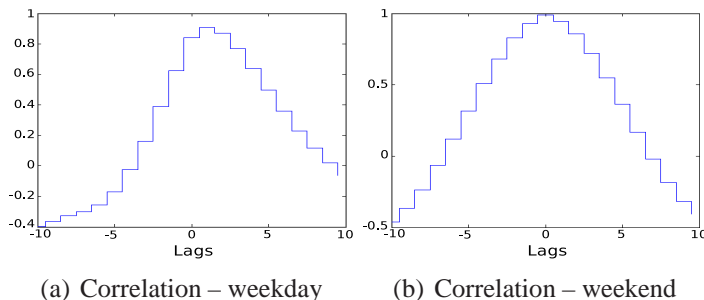


Figure 4.37: Cross-correlation between Instagram-New and Foursquare-New datasets, during weekday and weekend.

ing pattern in each system during weekdays, but it is not the case on weekends. The users' routines performed on weekdays may be the explanation for these results. The act of sharing a photo might be more likely to happen in special occasions that are usually out of the routines of people. For example, during breakfast time it is probably uncommon to happen something interesting to share a photo, but, for example, when you go out at night to have a dinner you have more incentives to share photos.

We now study how routines impact the sharing behavior analyzing the sharing pattern during weekdays, considering the datasets Instagram-New and Foursquare-New for New York, Sao Paulo, and Tokyo. The results are shown in Figure 4.38¹⁵. In all figures we display two cities from the same country for the new collected datasets, and one city for the old dataset as a reference of comparison.

First, observe the distinction between curves of each city in the same system (e.g., Instagram, Figures 4.38a, c, e) and also across different systems (e.g., Figures 4.38a and 4.38b for New York). Next, observe that the sharing pattern for each city in the same country is fairly similar, which might indicate cultural behaviors of inhabitants of those countries, presenting somehow the signature of a certain culture.

¹⁵Each curve is normalized by the maximum number of content shared in a specific region representing the city.

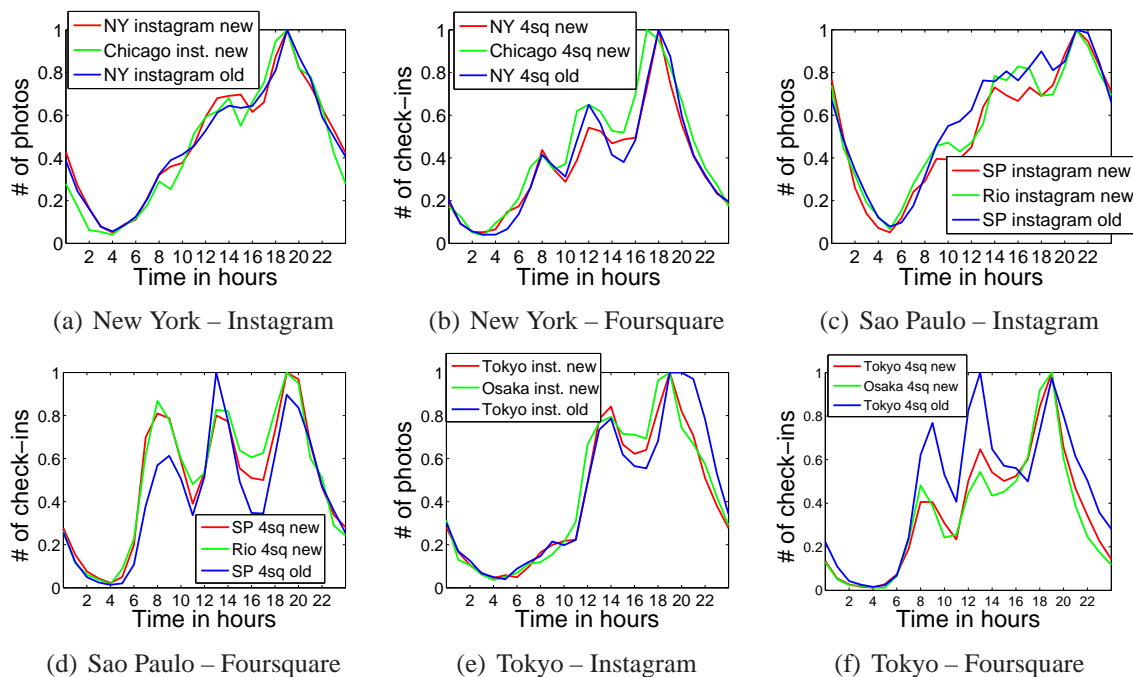


Figure 4.38: Temporal sharing pattern of Instagram and Foursquare for New York, Sao Paulo, and Tokyo during weekdays.

Note that the sharing pattern in Instagram for American cities (Figure 4.38a) and Japanese cities (Figure 4.38e) present peaks that reflect typical lunch and dinner times. This is not the case for the curves that represent the Brazilian sharing pattern in the cities of Sao Paulo and Rio (Figure 4.38c), where not all peaks represent typical meal times, suggesting that Brazilians share photos in atypical moments. Besides that, in general, the Brazilian activity is more intense late at night. This information was also observed considering only the Instagram-OLD dataset in Section 4.2.

The sharing pattern of the new dataset of Foursquare varied more when compared to the old one (Figures 4.38b, d, f), than the variation observed in the Instagram datasets (Figures 4.38a, c, e). Observe also that the sharing pattern in Instagram for each analyzed city is more distinct to each other than the one observed for Foursquare. This suggests that using the sharing pattern from Instagram we might have a more distinguishable “cultural signature” for a certain region, and less susceptible to changes over time.

4.4.4 Mapping Transitions

In a PSN, mobile nodes (users and portable devices) move accordingly to their routines or local preferences sharing data along the way. Looking at data people share it is possible to have a sort of rudimentary location tracking. If we aggregate all transitions performed by all

users we can obtain common paths users tend to take in the city.

Given that observation, a question emerges: can we observe a similar movement of people using a PSN derived from Instagram and Foursquare? In order to address this question we create a directed graph $G(V, E)$, where nodes $v_i \in V$ are a cell in the grid a particular city shown in Figure 4.33. A direct edge (i, j) , representing a transition, exists from node v_i to node v_j if at some point in time a user shared a content in cell v_j just after sharing a content in cell v_i . The weight $w(i, j)$ of an edge is the total number of transitions that occurred from cell v_i to cell v_j . Some features of transitions: (i) the content must be shared consecutively and by the same individual; (ii) continuous content sharing at the same considered venue cell represents a self-loop; and (iii) a transition must have occurred at the same day (we only consider transitions occurred from 6:00 a.m. to 6:00 p.m.).

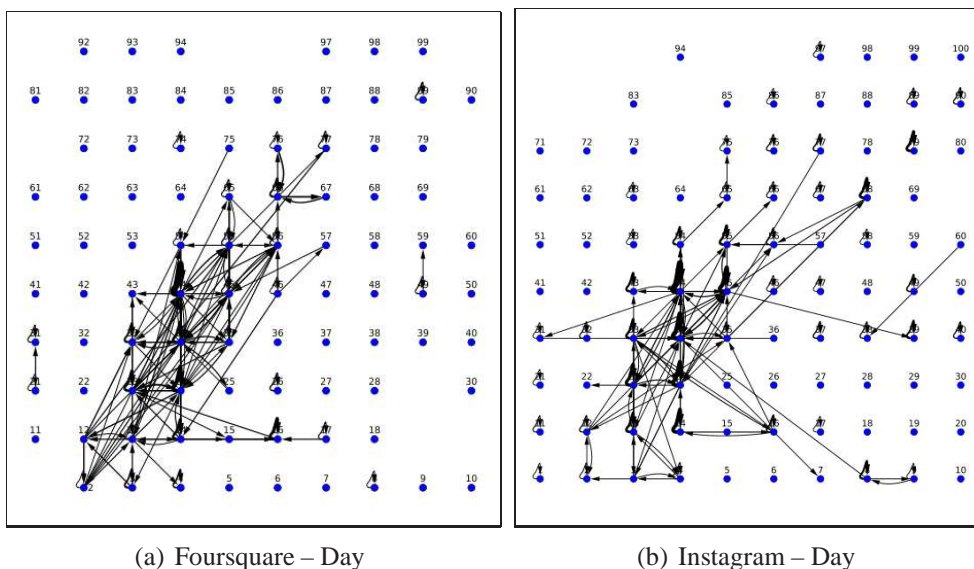


Figure 4.39: Transition graphs – New York.

Figures 4.39, 4.40, and 4.41 show the transition graphs for New York, Sao Paulo, and Tokyo, respectively. In those figures, for better visualization, we excluded all edges with weight $w = 1$. Nodes' positions in the figure are depicted according to the cell position they represent in the city area. Nodes not displayed mean that no one shared content in that particular area of the city.

Note that there are few transitions in the city. In other words, typical movements in the city might not be very diverse. It is also interesting to observe that we could capture more transitions with the Foursquare dataset. This means that check-ins might be more effective to track typical routes of users. However this hypothesis needs further investigation, because this result might be due to the large amount of data obtained in the Foursquare dataset. An

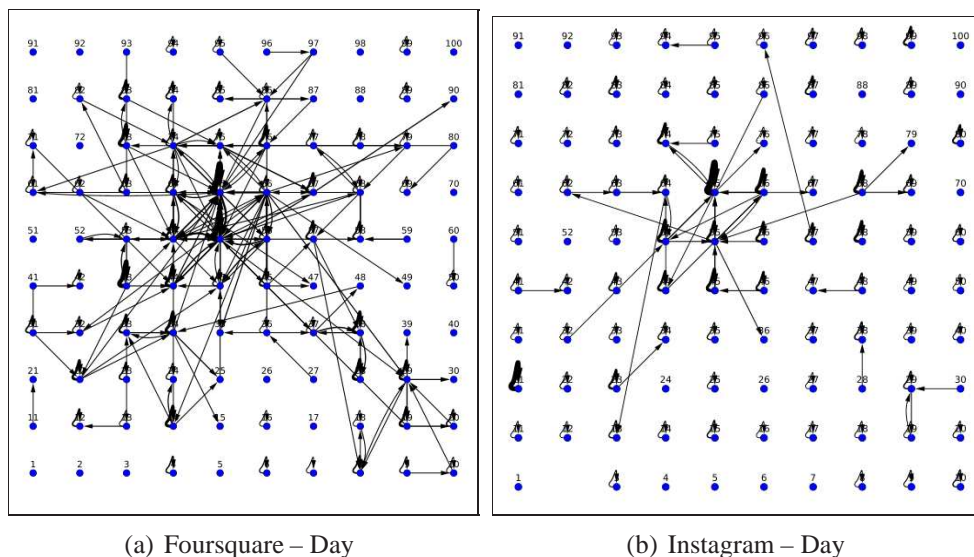


Figure 4.40: Transition graphs – Sao Paulo.

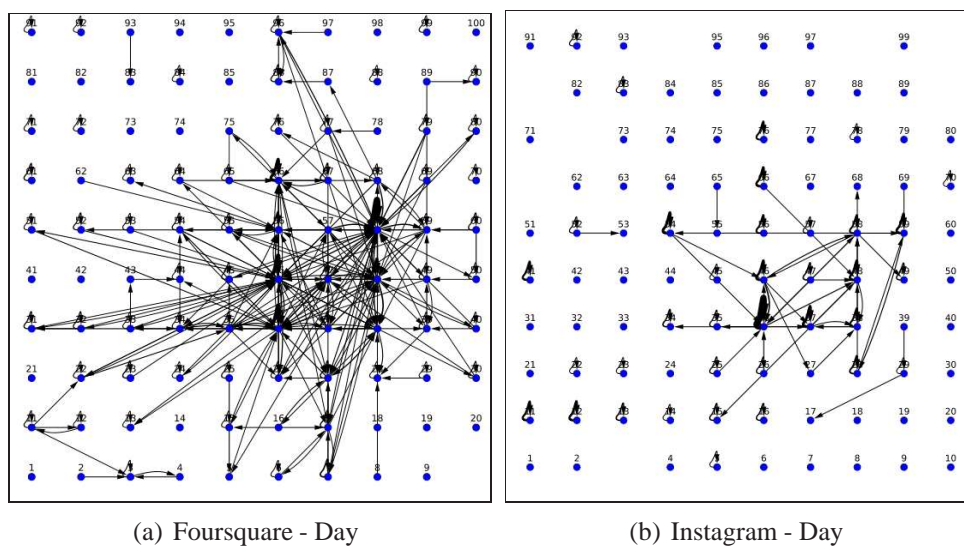


Figure 4.41: Transition graphs – Tokyo.

interesting possibility in this direction is use data mining algorithms, such as [Zheng et al., 2009], on transition graphs to discover movement patterns.

In order to compare the similarity graph, we first discard all self-loops, since those transitions tend to be more likely to happen, then we rank the resulting transitions and select the top ten from each graph. We compare the groups of top transitions between Instagram and Foursquare graphs of each city, analyzing the number of transitions in common. The results show that approximately 70%, 50%, and 70% of the top transitions are similar for New York, Sao Paulo, and Tokyo, respectively. However, if we take now the top twenty

transitions the results for transitions in common are approximately: 59%, 53%, and 50%, for New York, Sao Paulo, and Tokyo, respectively. This indicates that the graphs are not very similar, but popular transitions are more likely to be expressed by both systems.

The similarity graphs are also evaluated in a different way. For this comparison we preserve the graphs without discarding self-loops, then we compute the difference between sets of edges of Instagram and Foursquare for each city. We discovered that graphs for NY have 156 different edges, and these numbers are 237, and 376 for Sao Paulo and Tokyo, respectively. This is a significant difference, and these results are partially explained by the fact that Foursquare graphs captured more transitions.

4.5 Final Considerations

4.5.1 Data Limitations and Bias

It is important to point out some possible limitations of our datasets. First, it reflects the behavior of a fraction of the city citizens. Our collection is based on data shared by users of Foursquare, Instagram, and Waze on Twitter. Therefore biased towards the citizens who use those systems, who are likely to be under 50 year-old, and especially those between 18-29 year-old, owners of smartphones, and urban dwellers [Brenner and Smith, 2013; Duggan and Smith, 2014]. Consequently, urban areas with older and poorer populations tend to have fewer data. Besides that, users might not share data at all of their destinations, for example hospitals, love hotels, and strip clubs. Thus, our dataset might offer a partial view of citizens habits.

Second, our dataset are based on a limited sample of data. This means that we only have a sample of the activities performed. External factors, such as bad weather conditions, might have affected the total number of data we collected for some places, especially outdoor locations.

Third, as mentioned, we collected our datasets from Twitter, which has recently emerged as a popular tool to spread information. This powerful tool opens opportunities for new forms of spam [Benevenuto et al., 2010; Yardi et al., 2009]. Data quality, one of the challenges discussed in Section 3.2, under this circumstances becomes even more serious, because the production of false data might be possible. So far we are not aware of any significant production of false data in the systems we analyze, however this could potentially compromise the results.

4.5.2 Results Discussion

In this chapter we studied properties of PSNs derived from location sharing services, photo sharing services, and traffic alert services. These PSNs share several properties in common, for instance:

- very large scale;
- user participation regarding to the number of shared data, and the number of data shared per location may vary widely;
- and highly unequal frequency of data sharing, both spatially and temporally, which is highly correlated with the typical routine of people.

We also know that data from different PSNs are associated with spatio-temporal contexts that can be correlated or not. With that, a fundamental step to deal with data from different PSNs is to perform a characterization. In this chapter, we also characterized PSNs from Instagram and Foursquare considering the time and location where the content (photo or check-in) was shared. We aimed to understand whether the pieces of data from one system are correlated to be used for the study of city dynamics and urban social behavior. The results show the existence of correlation. This gave us insight about using different PSNs as “sensing layers” of a predefined geographical region. The concept of sensing layers is defined in more details in Chapter 6.

Our findings regarding the comparison of PSN can be summarized in:

- both Instagram and Foursquare datasets might be compatible in finding popular regions of cities;
- the temporal sharing pattern did not vary considerably over time for the same system. However, the sharing pattern for each system during weekdays are distinct;
- both Instagram and Foursquare might be used to capture particular signatures of cultural behaviors, but apparently Instagram offers a more distinguishable “cultural signature”, and is less susceptible to changes over time;
- and Foursquare is apparently better to express typical routes of people inside cities.

These results illustrate the potential of PSN analysis to foster the large scale study of urban social behavior. More broadly, our characterization provides a deep understanding of the properties of those particular PSNs, revealing their potential to drive various studies on city dynamics and urban social behavior, as discussed in the next chapter, Chapter 5. In

the propositions made in the next chapter we take into account the possible data limitations mentioned above.

Chapter 5

Understanding city dynamics and urban social behavior

What is the current best way to study the dynamics of a city? How can we learn about the routines of its citizens, their movement patterns, its points of interest, and its cultural and economic aspects? One might choose to rely on official census data, while others may opt to simply get one or more guide books at their favorite bookstore. Although we are very fond to books and census efforts, do they always offer accurate and comprehensive knowledge about the current patterns and dynamics of a city? Societies are inherently very dynamic, i.e., they change constantly over time, and, as the world gets more and more connected, we believe that these changes tend to be even more frequent. Take, for instance, guide books about large cities involved with the Arab Spring, such as Tunis and Cairo. If they do not capture the changes that came from this period, they are already outdated! Similarly, official census data may quickly become obsolete as such efforts are usually undergone at low frequency (e.g., once every 10 years) due to their high costs.

In contrast, PSNs offer up-to-date views about the locations, opinions, likes and dislikes of their users, and thus have the potential to address the aforementioned questions in near real time and, given their coverage (e.g., Figure 4.1), reaching almost every part of the globe. In this chapter, we elaborate on this potential by presenting various new techniques and methods that exploit PSN data to support studies on city dynamics and urban social behavior.

This chapter is organized as follows. Section 5.1 presents a technique called City Image, which provides a visual summary of the city dynamics based on people movements. Section 5.2 discusses other possibilities to understand better city dynamics through people movements. Section 5.3 presents a technique to extract points of interest in the city. Section 5.4 discusses possibilities to use PSNs to the analysis of social, economic, and cultural

aspects of its inhabitants. Section 5.5 motivates the use of participatory sensor networks to the study of cultural differences. Finally, Section 5.6 discusses the key messages of the chapter.

5.1 Visualizing the Invisible Image of Cities

Similarly to Kostakos et al. [2009], we believe that cities present distinct characteristics and evolve over time. Thus, we propose the City Image visualization technique, which exploits the movements of the city inhabitants. In summary, the City Image is a square matrix that displays a visualization of the dynamics of a city. We start by describing, in Section 5.1.1, a transition graph used to build the City Image. We then describe, in Section 5.1.2, a technique to identify and quantify the most preferred and rejected transitions (i.e., movement patterns) in a city. Finally, in Section 5.1.3, we show, analyze and compare the City Image for several cities.

5.1.1 Transition Graph

As we mentioned before, the sensing activity in a PSN is performed by mobile individuals who choose to share their information. Unlike traditional mobile wireless sensor networks, the nodes in a PSN move according to their routines or local preferences, which are dictated by the city dynamics. Thus, we propose a transition graph to map the movements of individuals in a PSN, and thus represent the city dynamics.

The proposed transition graph is a directed weighted graph $G(V, E)$, where the nodes $v_i \in V$ are the **main categories** of locations, and a direct edge (i, j) exists from node v_i to node v_j if at some point in time an individual performed a check-in at a location categorized by v_j just after performing a check-in at a location categorized by v_i . Thus, an edge represents a transition between two location categories. The weight $w(i, j)$ of an edge is the total number of transitions that occurred from node v_i to node v_j .

To demonstrate this technique we use the PSN derived from Foursquare-Crawled dataset, described in Section 4.1. A transition between location categories is configured according to three requirements. First, the check-ins must be performed consecutively and by the same individual. Second, the check-ins should be performed at different venues¹. Third, the check-ins must occur in the same “social day”, which we define as the 24-hour interval starting at 5:00 a.m. (instead of 12:00 a.m., since we are interested in capturing the nightlife transitions as well). Transitions that cross two different “social days” are considered

¹ The number pairs of consecutive check-ins performed at the same venue is very small, representing at most 1.8% of the total transitions in any analyzed city.

only if the time interval between them is under four hours. We experimented with various policies for characterizing transitions, finding very similar results as only a small percentage of transitions are discarded as we vary the policy.

5.1.2 Preferred and Rejected City Transitions

We here introduce the City Image technique, which is based on the transition graph $G(V, E)$ defined in the previous section. In summary, the City Image is a square matrix that displays a visualization of a city dynamics based on the frequency of transitions that are performed by its inhabitants.

After building the transition graph $G(V, E)$, we create ten random graphs $G_{Ri}(V, E_{Ri})$, where $i = 1, \dots, 10$. Each such graph is built using the same number of *individual* transitions in $G(V, E)$. However, instead of considering the actual transition $v_i \rightarrow v_j$ performed by an individual (as reported in our dataset), we randomly pick a location category to replace v_j , simulating a random walk for this individual.

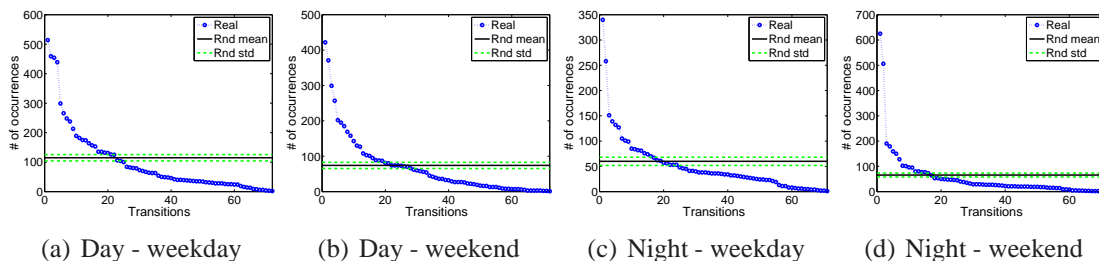


Figure 5.1: Observed transitions occurrences sorted in a descending order for NY city. Periods: weekday and weekend during the day and night.

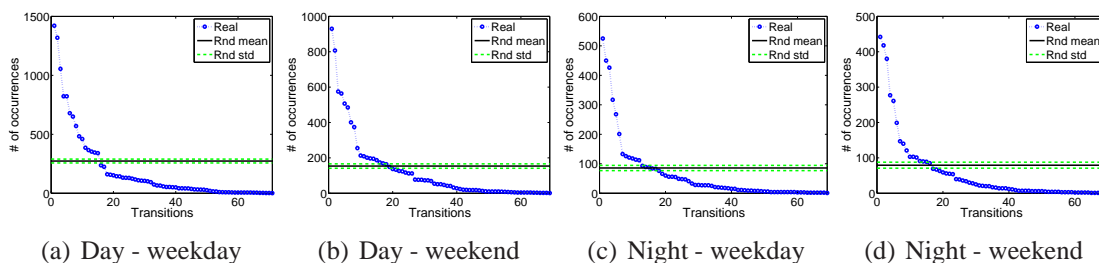


Figure 5.2: Observed transitions occurrences sorted in a descending order for Tokyo. Periods: weekday and weekend during the day and night.

In Figures 5.1 and 5.2 we compare, for each pair of location categories, the number of transitions that were simulated against the number of transitions that were actually made by individuals of New York and Tokyo². In these figures we consider four time periods: week-

²These results are representative of other cities.

day/weekend during the day (from 5:00 a.m. to 6:00 p.m.), and weekday/weekend during the night (from 6:01 p.m. to 4:59 a.m.). The x -axis represents particular transitions, e.g., *Food* \rightarrow *Work*, and the y -axis indicates the frequency of this particular transition. The blue curve (dotted line with a circle marker) represents the real transitions (i.e., represented in G), sorted in descending order of number of occurrences. The black curve (solid line) is the average number of transitions in the random graphs $G_{R1..10}$, and the two green curves (dashed lines) delimit the standard deviation. The results are shown separately for each time period. Note that, for many transitions, the number of real occurrences is significantly larger (i.e., by several standard deviations), than the expected average value in the random graphs. This implies that some transitions reflect more the preferences and habits of users from a certain city than others. There are also transitions that do not occur very often, with the number of real occurrences being much smaller than the average number in the random graphs, indicating that the inhabitants of this city strongly reject these transitions.

Based on these observations, we next identify the most and least favorable transitions to occur in a given city. To that end, we adopt one of two strategies, depending on whether the edge weights of the randomly generated graphs $G_{R1..10}$ follow a Normal distribution $N(\bar{w}, \sigma_w)$. If they are normally distributed, we compute the mean \bar{w} and the standard deviation σ_w of the edge weights. We then define the *indifference range* as the interval $(\bar{w} - 3\sigma_w, \bar{w} + 3\sigma_w)$, which is expected to contain 99.73% of the randomly generated edge weight values, since the edge weights follow a Normal distribution $N(\bar{w}, \sigma_w)$. Analogously, we define the *rejection range* as the interval $[-\infty, \bar{w} - 3\sigma_w]$, and the *favouring range* as the interval $[\bar{w} + 3\sigma_w, \infty]$.

In case the edge weight distribution is not Normal, we calculate the maximum (*max*) and minimum (*min*) values of the randomly generated edge weights. We then define the *indifference range* as the interval (\min, \max) , the *rejection range* as the interval $[-\infty, \min]$, and the *favouring range* as the interval $[\max, \infty]$.

For all the cities analyzed in the next section, the edge weights of the randomly generated graphs do follow a Normal distribution, as illustrated in Figures 5.3 and 5.4 for New York and Tokyo, respectively. These figures show both the histogram of the edge weights and the fitting of the Normal distribution (red curve). Note that, for New York city, the fitted Normal distribution has parameters $\bar{w} = 114.85$ and $\sigma_w = 10.712$, which are the values used to delimit the *rejection range*, *indifference range* and *favouring range* for the transitions for that city in that particular time period.

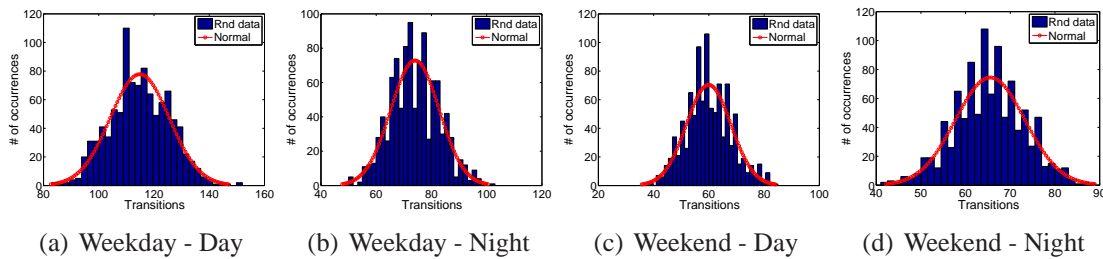


Figure 5.3: Histogram of random generated transitions for NY with a Normal fitting.

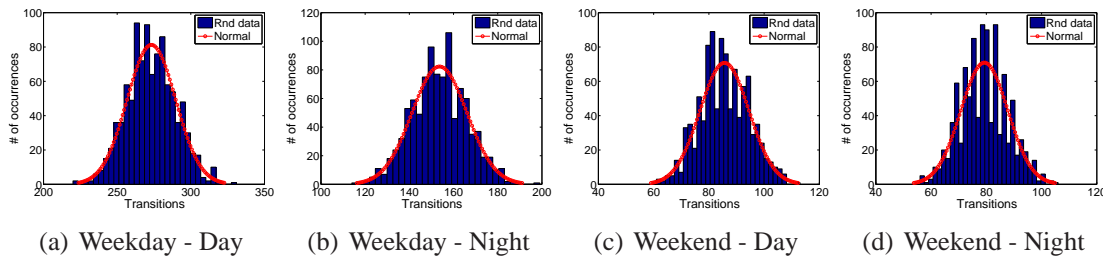


Figure 5.4: Histogram of random generated transitions for Tokyo with a Normal fitting.

5.1.3 Building the City Images

Having defined the ranges for preferred, rejected and indifferent transitions in a given city, we construct a square matrix that represents the movement patterns of the city, which is here called the City Image. In this matrix, each cell (i, j) represents the willingness of a transition from category i (line i of the matrix) to another category j (column j of the matrix). To better visualize this, we color cells that represent transitions that are *not* likely to occur in a city, i.e., transitions whose edge weight fall in the *rejection range*, in *red*. We also color transitions that are more likely to occur, i.e., transitions that fall in the *favouring range*, in *blue*. Finally, white color are used in cells that represent transitions that fall in the *indifference range*.

We built the City Image for 30 cities around the world. The cities and the number of check-ins available in our dataset in each of them are presented in Table 5.1. Appendix A shows the City Image, for all analyzed cities, built using aggregated data across all time periods. These images provide a general picture of each city, and serve to illustrate broad differences across cities.

Delving further into each city, we also analyze the City Image for each time period separately. Figure 5.5–5.12 present the City Image for London, Kuwait, Belo Horizonte, Chicago, Surabaya, New York, Sydney, and Tokyo. Each figure shows the City Image for one of the four time periods: day, from 5:00 a.m. to 6:00 p.m., on a weekday; day on a weekend; night, from 6:01 p.m. to 4:59 a.m., on a weekday; and night on a weekend.

The City Image captures the city dynamics in a very summarized way. Nevertheless, it

| City | # of check-ins | City | # of check-ins |
|------------------------|----------------|---------------------|----------------|
| Bandung/Indonesia | 59,332 | Mexico City/Mexico | 85,721 |
| Bangkok/Thailand | 67,075 | Moscow/Russia | 59,654 |
| Barcelona/Spain | 9,083 | New York/USA | 86,867 |
| Belo Horizonte/Brazil | 18,280 | Osaka/Japan | 27,396 |
| Buenos Aires/Argentina | 17,762 | Paris/France | 11,746 |
| Chicago/USA | 27,446 | Rio/Brazil | 27,222 |
| Istanbul/Turkey | 103,456 | San Francisco/USA | 17,840 |
| Jakarta/Indonesia | 158,732 | Santiago/Chile | 79,733 |
| Kuala Lumpur/Malaysia | 109,048 | Sao Paulo/Brazil | 85,640 |
| Kuwait City/Kuwait | 34,195 | Semarang/Indonesia | 10,518 |
| London/UK | 15,671 | Seoul/Korea | 26,073 |
| Los Angeles/USA | 21,961 | Singapore/Singapore | 65,534 |
| Madrid/Spain | 13,004 | Surabaya/Indonesia | 38,021 |
| Manila/Philippines | 47,343 | Sydney/Australia | 6,390 |
| Melbourne/Australia | 6,182 | Tokyo/Japan | 118,788 |

Table 5.1: Distribution of check-ins across the selected cities.

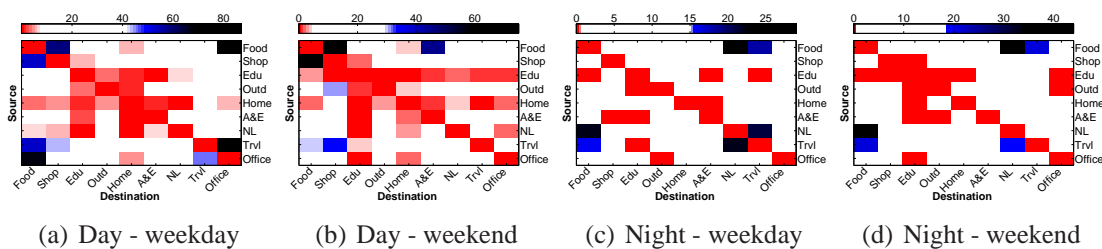


Figure 5.5: The City Image of London for different periods.

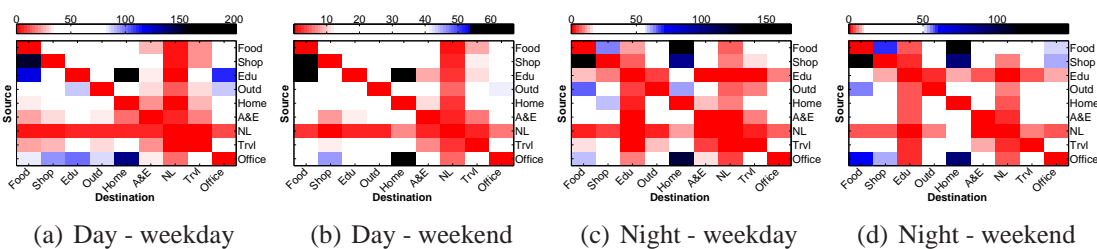


Figure 5.6: The City Image of Kuwait for different periods.

can reveal striking differences in the dynamics of the same city across different time periods (weekdays and weekends, day and night), as well as across different cities. Moreover, note that the main diagonal of each matrix indicates a tendency of not having consecutive check-ins at the same category. The City Image also provides an easy way to learn the most and least favored places and transitions of each city in a given time period.

In general, using the City Image it is possible to distinguish the routines of the inhabi-

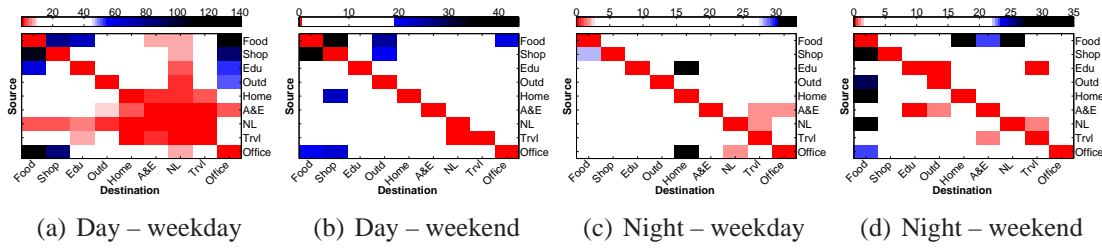


Figure 5.7: The City Image of Belo Horizonte for different periods.

tants of two particular cities. For instance, in Kuwait (Figure 5.6) and Surabaya (Figure 5.9) we observe the lack of favorable transitions considering the category *nightlife* for all analyzed periods. On the other hand, *nightlife* transitions are strongly favorable to happen in Chicago (Figure 5.8) and New York (Figure 5.10), not only on weekend nights but also on weekday nights. Moreover, on weekends at night inhabitants from Kuwait and Surabaya are very favorable to perform the transitions $shop \rightarrow food$ and $food \rightarrow home$. This might be explained by cultural differences that exist among these cities.

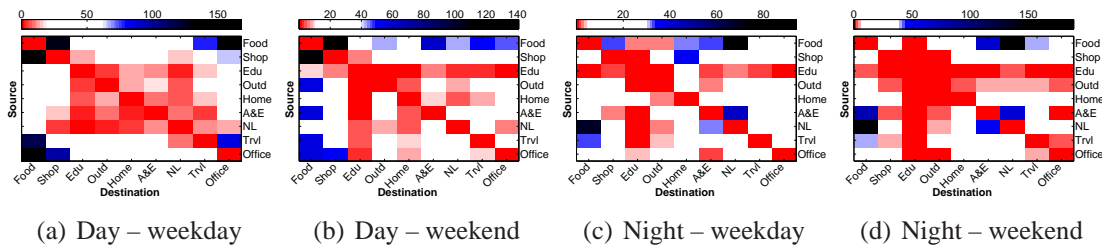


Figure 5.8: The City Image of Chicago for different periods.

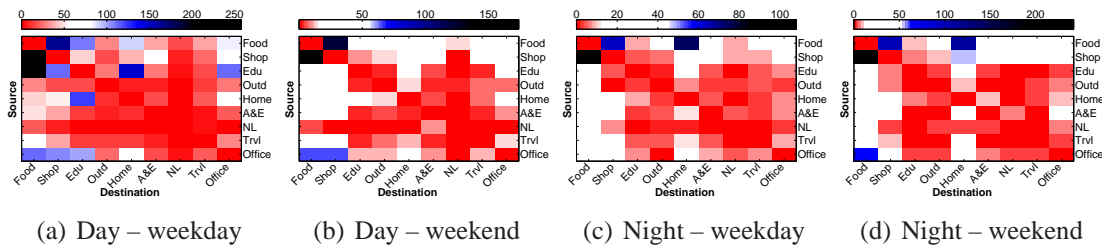


Figure 5.9: The City Image of Surabaya for different periods.

As another example, note that inhabitants of Belo Horizonte (Figure 5.7) are highly favorable to perform transitions containing the category *education*. This comes with no surprise since this city is an important hub of education in Brazil. In this particular City Image it is also worth noting that the transition $education \rightarrow office$ is favorable. This is because, many students in Belo Horizonte do keep a (part-time or full-time) job. This also explains the favorable transition $education \rightarrow home$ on weekdays at night, as many students

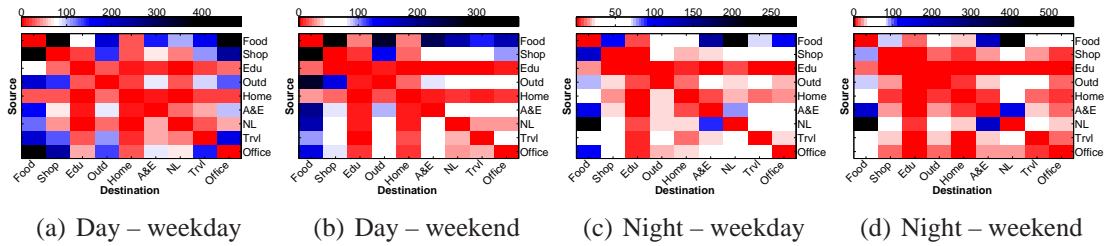


Figure 5.10: The City Image of New York for different periods.

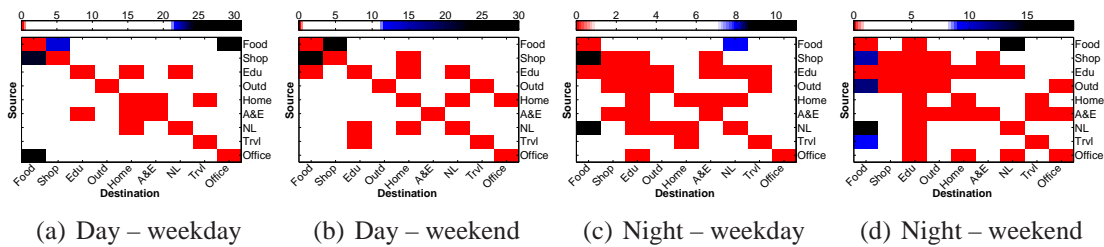


Figure 5.11: The City Image of Sydney for different periods.

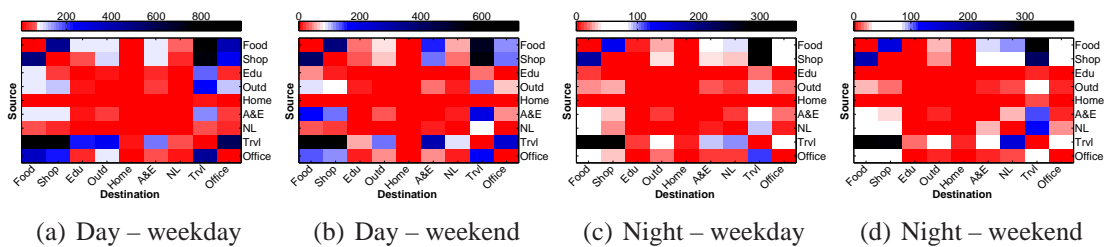


Figure 5.12: The City Image of Tokyo for different periods.

who have a full-time job go to school at night. In contrast, we find that Chicago residents tend to reject any transition involving the category *education* for all analyzed periods. This is surprising, since Chicago has been a world center of higher education and research, with several universities located in the city.

We also note that one of the most favored transitions in London (Figure 5.5) on weekdays during the day is *travel* \rightarrow *office*. A similar trend also happens in other cities, such as New York and Tokyo (Figure 5.12). On the other hand, some cities, such as Belo Horizonte, Sydney (Figure 5.11), Kuwait and Surabaya, do not present favorable transitions containing the category *travel* on weekdays during the day. This could be associated with a larger number of people who choose to drive to get to their destinations, instead of taking public transportation.

The City Image technique, as illustrated above, is an interesting way to better understand the invisible image of a city. It provides a useful tool in various contexts, ranging from helping city planners to better understand the actual dynamics of a city, to providing tourists

another source of information that might help them make their travel choices. The transition tendencies further serve as a source of fundamental information for social behavior study.

One possible limitation of our dataset is the covered time interval, one week, which might be considered short. In order to assess to which extent this might impact the conclusions drawn from the City Images, we collected the check-ins performed on the cities of Belo Horizonte, Chicago, London, and Surabaya in the week following the period covered by our original dataset. We then recalculated the City Images for each of these cities using all the data available, thus covering a time interval of two weeks. We show the results for weekdays during the day, which is the period where most of the routines are performed, in Figure 5.13. We can observe that the new City Images are very similar to the corresponding ones produced using our original one-week dataset (Figures 5.7a, 5.8a, 5.5a, and 5.9a for Belo Horizonte, Chicago, London, and Surabaya, respectively). The strong favorable or rejection transitions remain basically the same, whereas the changes, if observed, occur in some transitions classified in the indifference range. These particular changes are expected because the larger dataset enables a clearer image of the analyzed city. The same strong similarities were observed for the City Images produced for the other periods of time (e.g., weekend night). Thus, even with a single week of data, the City Image technique is able to reveal remarkable and consistent patterns of each analyzed city.

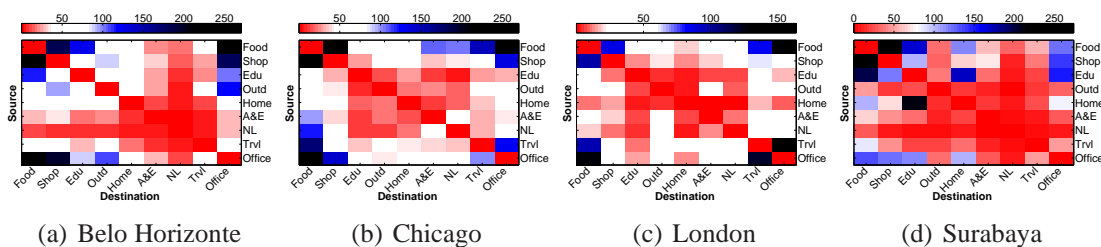


Figure 5.13: The City Image of cities in different regions of the world during the day on weekdays.

An application that naturally emerges from the City Image technique is the numerical comparison and clustering of different cities, by exploiting the values in each square matrix. This application helps to validate the City Image results, besides other applicability. We present this application on Appendix B. As we can see in that appendix, the results agrees with common knowledge.

5.2 Insights into People Movement Patterns

Another possible visualization of city dynamics based on data collected by PSNs is illustrated in Figure 5.14. It shows a heatmap of the sensing activity for the city of Belo Horizonte,

Brazil (Figure 5.14a) and New York, USA (Figure 5.14b) for the PSN derived from the Foursquare-Crawled dataset. The darker the color, the higher the number of check-ins in the area. These heatmaps by themselves convey information related to the popularity of specific areas, being thus only partially informative about the city dynamics. Richer information can be obtained by making a small change in the City Image transition graphs presented above in Section 5.1. Thus, in this section we explore the concept of transition graphs to draw valuable insights about crowd mobility in cities.

5.2.1 Using Centrality Metrics

Many metrics of node centrality can be used to estimate the relative importance of a node within the graph. Although most of these metrics were first developed in social network analysis [Newman, 2010], they can also be applied to a transition graph, similar to the one proposed in Section 5.1.1, enabling the study of city dynamics. Thus, in this section we build a transition graph where each node represents a specific location (and not location category, as in Section 5.1.1), and a direct edge (i, j) exists if someone performed a check-in at location j after a check-in at location i . The weight of the edge reflects the number of transitions between the two specific locations. These transitions are configured according to the same requirements defined in Section 5.1.1.

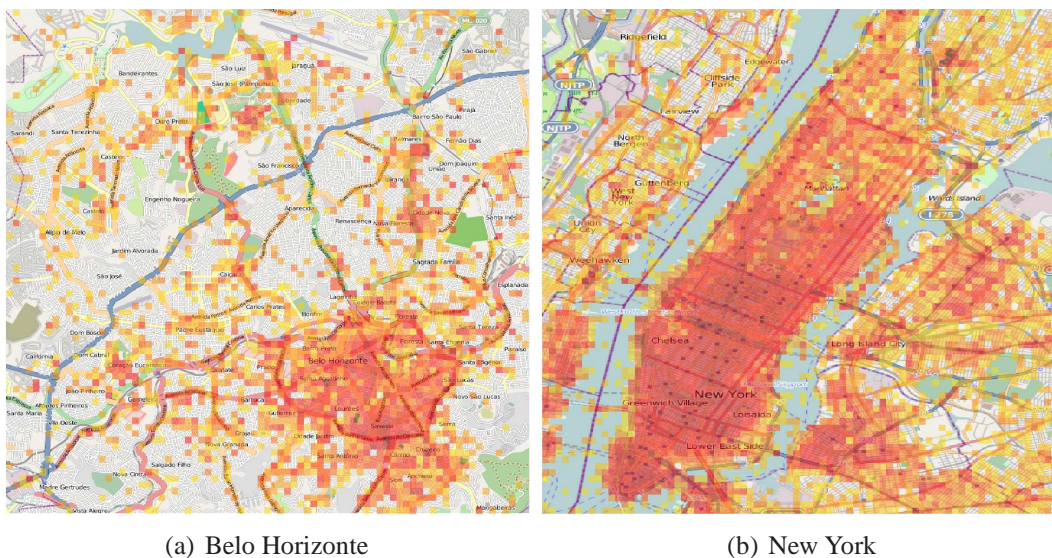


Figure 5.14: Heatmap of the number of check-ins, where the color range from yellow to red (high intensity).

Traditionally used centrality metrics are degree, closeness and betweenness centrality [Newman, 2010]. These metrics aim to identify nodes that have central locations within the network structure. Since nodes in our networks represent locations, a central node may

| Day | | | | | |
|--------|---------------|-----------|----------------|------------------|---------------|
| Degree | | Closeness | | Node Betweenness | |
| Value | Venue | Value | Venue | Value | Venue |
| 0.04 | Yankee S. | 0.18 | Yankee S. | 0.1 | Yankee S. |
| 0.02 | Penn S. | 0.18 | Penn S. | 0.05 | Penn S. |
| 0.02 | Grand C. | 0.18 | Times S. | 0.04 | Grand C. |
| 0.02 | Mad. S. G. | 0.17 | Grand C. | 0.03 | Times S. |
| 0.02 | Times S. | 0.17 | Mad. S. G. | 0.03 | Mad. S. G. |
| 0.01 | Bryant | 0.17 | Bryant | 0.03 | Union S. |
| 0.01 | Union S. | 0.17 | Union S. | 0.03 | Bryant |
| 0.01 | Wash. S. | 0.17 | Int. Auto Show | 0.02 | Wash. S. |
| 0.009 | MoMa | 0.17 | Rockef. C. | 0.02 | Mad. Sq. P. |
| 0.008 | Port A. | 0.16 | Port A. | 0.01 | Port A. |
| Night | | | | | |
| Degree | | Closeness | | Betweenness | |
| Value | Venue | Value | Venue | Value | Venue |
| 0.01 | Yankee S. | 0.06 | Yankee S. | 0.02 | Yankee S. |
| 0.007 | Penn S. | 0.06 | Penn S. | 0.01 | Penn S. |
| 0.007 | Times S. | 0.06 | Mad. S. G. | 0.007 | Tribeca F. F. |
| 0.006 | Mad. S. G. | 0.06 | Times S. | 0.007 | Mad. S. G. |
| 0.005 | Tribeca F. F. | 0.06 | Tribeca F. F. | 0.007 | Times S. |
| 0.005 | Grand C. | 0.06 | Grand C. | 0.006 | Grand C. |
| 0.004 | Webster H. | 0.06 | Bowery B. | 0.004 | Webster H. |
| 0.003 | Union S. | 0.06 | Term. 5 | 0.003 | Bryant |
| 0.003 | Bowery B. | 0.06 | Brook. Bowl | 0.003 | Pacha |
| 0.003 | Port A. | 0.06 | Pacha | 0.003 | Radio City |

Table 5.2: Centrality metrics for NY during the day and night.

indicate a strategic point in the city, according to a specific metric. For example, the main idea behind the degree centrality is to identify the total number of links incident to a node, i.e., the number of incoming and outgoing edges that a node has. In our transition networks, a node with high degree indicates a location where people may arrive and depart with a high probability. Thus, degree centrality is a good measure to identify popular places in the city. These locations can be seen as city hubs.

The closeness centrality metric is related to how close a node is to all other nodes in the network, i.e., the number of edges separating a node from the others. In the context of information dissemination, the higher the closeness of a place, the higher the probability that a piece of information being disseminated from that place reaches the whole network in the least amount of time. In the perspective of a transition graph, the closeness centrality may indicate favorable locations in the network structure to start the dissemination of information

to the whole network. These locations may be strategic places to install public information centers to disseminate, for example, alerts using users' portable devices in an ad hoc manner.

Finally, the main idea behind the betweenness centrality is to show how often a node is in the shortest path between any two other nodes. In our transition networks, it may indicate the most interesting locations to act as bridges to carry information among different places or regions of places (set of places). That is, the higher the betweenness of a location, the higher the chance that a user passes through that particular location. One could explore these central nodes to sign a commercial agreement to increase their revenues by, for instance, making an advertising in order to direct flow of users to other independent business venues in the city.

We illustrate the use of these centrality metrics by showing in Table 5.2 the top-10 locations with the largest degree, betweenness, and closeness centrality values in New York. The table presents results for two time periods, day (5:00 a.m. to 7:00 p.m.) and night (6:00 p.m. to 6:00 a.m.)³, aggregating results for weekdays and weekends for the sake of avoiding hurting the presentation with excessive data. Note that most top-10 locations, according to all metrics, are widely known. Some of these locations, such as Yankee Stadium (Yankee S.), are in the top-10 according to all metrics and in both analyzed periods, whereas others appear in the top-10 list of only one metric, such as MoMa which is listed only in the degree centrality column. This demonstrates that different centrality metrics may identify different central places.

We note that the Tribeca Film Festival (Tribeca F. F.) was identified as a central place in all metrics during the night. Foursquare encouraged users to check-in in this event offering a special badge for it. This justifies the large number of check-ins and, thus, the increase of centrality. Since in the studied network nodes are venues and venues tend to be dynamic, a temporal analysis when studying centrality is desirable. In this case, it would be possible to identify that Tribeca F. F. was a temporary venue, and thus avoid considering it a central location after its expiration date.

We also note the greater diversity of central locations across metrics for the night period. In other words, there is a larger number of locations that appear among the top ten according to only one or two metrics during the night. The type of these locations might help explain the results. Observe that nightlife places, such as Pacha and Brooklyn Bowl, are not listed in the top-10 locations with highest degrees. Yet, they are amongst the locations with highest betweenness and closeness values. This could be explained by the routine of people, who usually go to a pub or a restaurant before going to a nightlife spot. This first visited location might not be very popular, e.g. a random place close to the user's house that

³If one transition happened in the overlapped hours (5:00 a.m. to 6:00 a.m., or 6:00 p.m. to 7:00 p.m.), it is considered a transition of day and night periods, respectively. NY has 49,849 check-ins during the day and 19,491 check-ins during the night.

might be far away from the target place (nightlife spot). This could connect different regions from the network, helping to increase the betweenness of the first location. Alternatively, the first visited location could be a popular place, helping to increase the closeness.

5.2.2 Network Visualization

The visualization of transition graphs, specially highlighting central places, is interesting because it gives fascinating insights into how people move and interact with the city. The edges in the transition graphs represent somehow a rudimentary GPS tracking. With that, the final network, after aggregating the transitions performed by all users, enables the reconstruction of typical paths that users take to move in the city. When representing the information of centrality of a place in this network we are also able to visualize and understand better how users interact with the city. Figures 5.15, 5.16, and 5.17⁴ show such networks for Belo Horizonte and New York, during the day and night, for the degree, betweenness, and closeness centrality, respectively. Each color means a category of place, as defined in the caption of Figure 5.15.

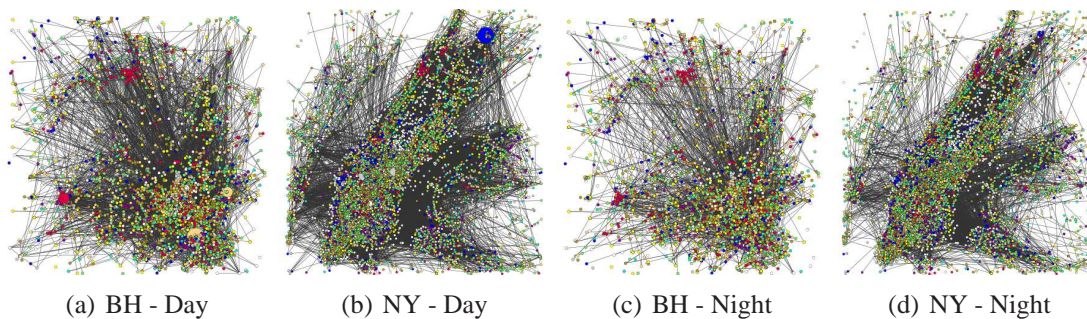


Figure 5.15: Node degree - For two cities in different countries. Each node color represents an specific category of places. Blue=Arts& Entertainment; Red = College & Education; Light Green = Food; Yellow = Home; Green Moss = Office; Purple = Nightlife Spot; White = Great Outdoors; Beige = Shop & Service; Grey = Travel spot; Cyan = no category.

Studying the results for New York, for example, it is possible to observe that during the day there is an intense movement of people between Manhattan, New Jersey, Brooklyn, and Queens, where Manhattan is the central destination. However, during the night the movement of people between Manhattan and New Jersey is much lower, but the movement between Manhattan, Brooklyn and Queens is still quite intense. This might indicate that people from New Jersey tend to go to Manhattan more often to work during the day than for leisure time at night.

⁴The area represented by those networks is the same as the one shown in Figure 5.14. Nodes disposition respects their geo-location in the city.

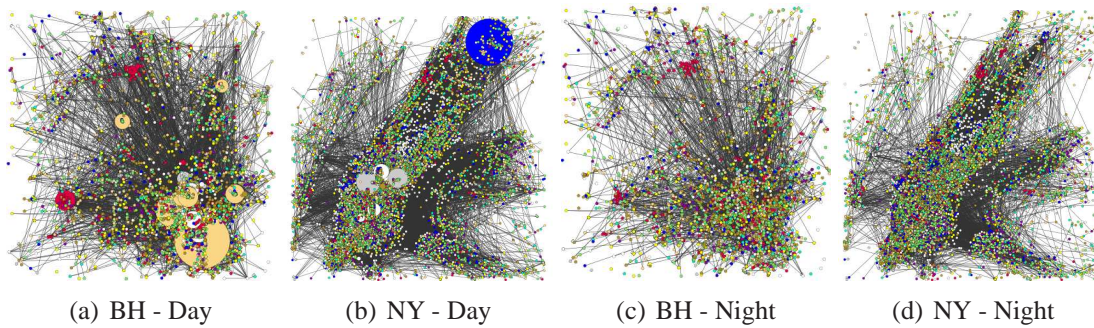


Figure 5.16: Node Betweenness - For two cities in different countries. Colors legend: see caption of Figure 5.15.

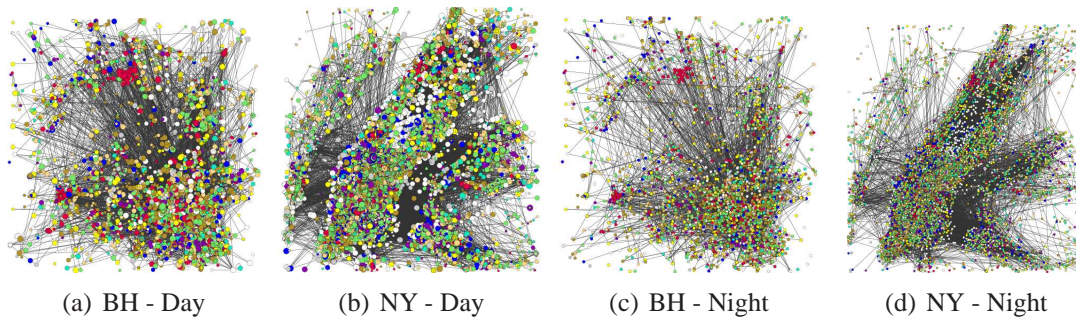


Figure 5.17: Node Closeness - For two cities in different countries. Colors legend: see caption of Figure 5.15.

As another example, Figures 5.16a and 5.16b show that, during the day, both New York and Belo Horizonte have a few places that stand out with higher betweenness values than the others in the same city. This does not happen in the same proportion for the degree centrality, as shown in Figures 5.15a and Figure 5.15b. Moreover, the same discrepancies cannot be observed for neither centrality metric during the night (Figures 5.15c, 5.15d, 5.16c, and 5.16d), which might be explained by the lack of peoples' routines.

Regarding the closeness metric, we can see a large number of places with high closeness during the day in both cities (Figures 5.17a and 5.17b), implying that there are many options of places to select in case one wishes to install alert dissemination schemes in the city, for example. Note also that places with high closeness are relatively well spread in both cities during the day. However, this is not the case during the night (Figures 5.17c and 5.17d). The results in this period follow the same tendency observed for the other metrics and the explanation might be the same, i.e., lack of well-defined routines.

The network visualization can be also done in other ways, which could potentially ease the understanding of certain aspects of the city. For example Figures 5.18a and 5.18b show the top 50 heavy weighted edges and the top 50 hub nodes (largest degrees) for the

Belo Horizonte and New York, respectively. Stars represent the hubs, black arrows represent edges, and black circles represent self-loops. The larger the symbol, the larger the associated value (edge weight or node degree) associated. Note that the flow of people tend to be very concentrated and skewed, as expected, with a small fraction of the city areas having most of the heavy weighted edges and hubs. Note also that for Belo Horizonte (Figures 5.18a), most of the heavy weighted edges are self-loops and short distance edges, suggesting that people tend to perform activities in their neighborhoods. In contrast, cities that are known for their fast public transportation systems, such as New York, favor the existence of some long distance heavy weighted edges along the public transport links, as shown in Figure 5.18b.

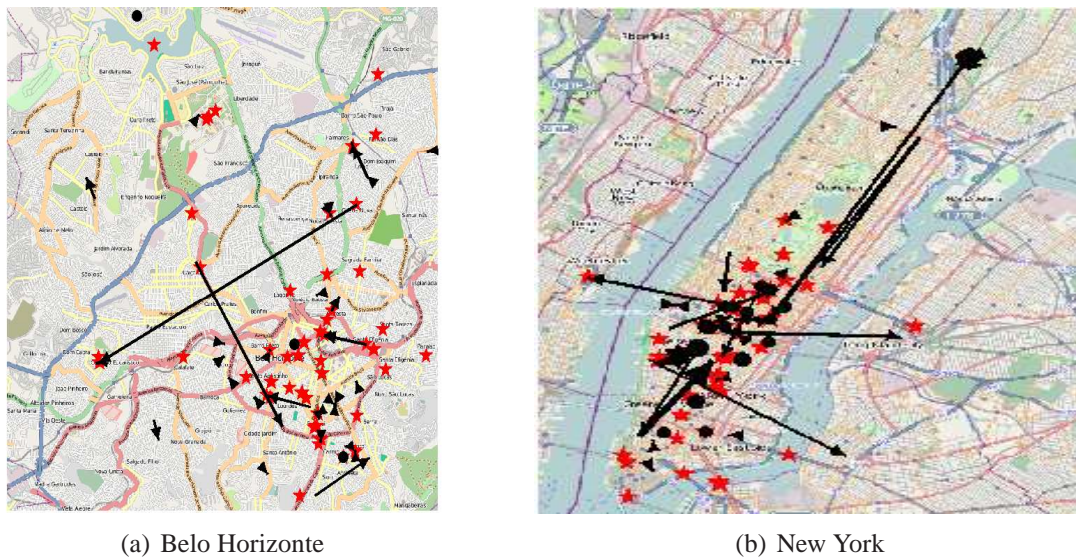


Figure 5.18: Top 50 edge weights and node degrees (hubs); stars represent hubs, black arrows edges, and black circles self-loops. Featured places (nodes) and transitions (edges).

5.2.3 Information Summarization

Tables 5.3, 5.4, and 5.5 show the summarization of values of each centrality metric (degree (D), betweenness (B), and closeness (C)), calculated for all places during the day and night in Belo Horizonte, New York, and Tokyo, respectively. The summarization is expressed by the percentage relative to the total of values by category of places. For example, in Table 5.3 we can see that, during the day, all places of the category Food represent 17.7% of all degree centrality observed. These tables help us to visualize the cities by their most important classes of places. Analyzing the top degree centrality during the day we can observe that inhabitants of Belo Horizonte concentrate a lot of activities in education, shopping and work-

| Categ. | D. (%) | B. (%) | C. (%) |
|--------|--------|--------|--------|
| Day | | | |
| Food | 17.7 | 14.9 | 21.6 |
| Shop | 13.8 | 22.5 | 12.9 |
| Edu | 14.5 | 15.8 | 10.6 |
| Outd | 9.1 | 15.3 | 6.3 |
| Home | 9.1 | 4.7 | 11.04 |
| A&E | 3.4 | 3.6 | 4.3 |
| NL | 4 | 3.2 | 5.7 |
| Trvl | 5.3 | 6.5 | 4.7 |
| Offi | 20.4 | 12.7 | 20 |
| none | 2 | 1 | 2.9 |

| Night | | | |
|-------|------|------|------|
| Food | 18.7 | 19.5 | 23.3 |
| Shop | 9.5 | 16.1 | 7.9 |
| Edu | 11.3 | 11.9 | 9.1 |
| Outd | 9.26 | 16.5 | 8.6 |
| Home | 15.3 | 6.2 | 14 |
| A&E | 5.5 | 5.6 | 5.2 |
| NL | 10.3 | 14.1 | 13.6 |
| Trvl | 3.9 | 3.4 | 4 |
| Offi | 14.6 | 6.1 | 13 |
| none | 1.5 | 0.3 | 1.4 |

Table 5.3: Percentage of centrality metrics for all categories of places for BH (day and night). D=degree, B=betweenness, C=closeness.

| Categ. | D. (%) | B. (%) | C. (%) |
|--------|--------|--------|--------|
| Day | | | |
| Food | 29.5 | 21.8 | 33.3 |
| Shop | 13.5 | 13.2 | 15.1 |
| Edu | 2.5 | 1.9 | 2.5 |
| Outd | 8.8 | 15.2 | 6.1 |
| Home | 2 | 1 | 3.1 |
| A&E | 9.5 | 15.6 | 6.2 |
| NL | 10.2 | 8.4 | 10.4 |
| Trvl | 7 | 10.9 | 5.7 |
| Offi | 14.7 | 11 | 14.2 |
| none | 2 | 0.9 | 3.3 |

| Night | | | |
|-------|------|------|------|
| Food | 31.1 | 22.7 | 36.4 |
| Shop | 7.4 | 6.4 | 7.8 |
| Edu | 1.5 | 0.6 | 1.5 |
| Outd | 5.8 | 8.3 | 5 |
| Home | 3.2 | 1.1 | 3.6 |
| A&E | 10 | 17.3 | 7.2 |
| NL | 23.4 | 27.1 | 23.7 |
| Trvl | 6.6 | 9.9 | 6.1 |
| Offi | 9.4 | 6.3 | 6.8 |
| none | 1.6 | 0.5 | 2 |

Table 5.4: Percentage of centrality metrics for all categories of places for NY (day and night). D=degree, B=betweenness, C=closeness.

| Categ. | D. (%) | B. (%) | C. (%) |
|--------|--------|--------|--------|
| Day | | | |
| Food | 25.4 | 15.7 | 39.2 |
| Shop | 16.3 | 13.9 | 18 |
| Edu | 3.1 | 1.8 | 3 |
| Outd | 4 | 4.2 | 4.8 |
| Home | 0.2 | 0.1 | 0.4 |
| A&E | 5.1 | 4.2 | 4.8 |
| NL | 2.9 | 1.3 | 5.7 |
| Trvl | 32.8 | 50.8 | 11.8 |
| Offi | 8.8 | 7.4 | 10 |
| none | 1.3 | 0.6 | 2.4 |

| Night | | | |
|-------|------|------|------|
| Food | 26.9 | 10.8 | 35.4 |
| Shop | 13.3 | 11.1 | 15.7 |
| Edu | 1.1 | 0.6 | 1.1 |
| Outd | 3 | 3.2 | 3.7 |
| Home | 0.4 | 0.1 | 0.7 |
| A&E | 5 | 3.2 | 5.4 |
| NL | 7.5 | 3.6 | 10.8 |
| Trvl | 35.8 | 63.5 | 20 |
| Offi | 5.4 | 2.8 | 5.4 |
| none | 1.4 | 0.6 | 1.8 |

Table 5.5: Percentage of centrality metrics for all categories of places for Tokyo (day and night). D=degree, B=betweenness, C=closeness.

ing (represented by the category Office), having the categories of places Food⁵, Office, Shop and Education as the most popular. Following the same analysis, places related to working, shopping, and nightlife are quite central in New York. Studying now the centrality in Tokyo it is interesting to observe the high amount of activity in Travel places, probably related to public transportation spots. Note the high value for betweenness and the considerable lower value for closeness. This means that inhabitants of Tokyo might use public transportation to move to areas with not many central places, such as suburbs, justifying the values observed for betweenness and closeness.

Regarding to privacy issues, observe the centrality in the category of places Home. In Belo Horizonte the number of check-ins is expressive in this category. However, in NY

⁵We consider that food activities are complementary to a main activity, such as work or study, for this reason we are not mentioning it as a main activity.

and mainly in Tokyo people do not appear to have the same behavior. This fact might be explained to cultural differences. It is known that Japanese people are concerned with privacy issues, and apparently Brazilians are not as concerned.

Differences in the habits of inhabitants of the cities can also be captured by those tables. During the night, places related to education are still quite central in Belo Horizonte, but not in NY or Tokyo. This is explained because night courses in schools and universities are common in Belo Horizonte, since many people have to work during the day to pay their studies. In New York, as expected, the centrality of places related to nightlife and arts & entertainment is high. On the other hand, shopping places have high centrality in Tokyo for this considered period. This analysis illustrates how we can visualize characteristics of cities, and the potential of using it to differentiate them.

5.3 Points of Interest

It is quite common to find particular areas in a city that attract more attention of residents and visitors, here called *points of interest* (POI). Among the most visited POIs, we can mention the sights of the city. However, not all POIs are sights of a city. For example, an area of bars can be quite popular among city residents, but not among tourists. Furthermore, POIs are dynamic, in other words, areas that are popular today may not be tomorrow. In Section 5.3.1 we present an algorithm to identify POIs and in Section 5.3.2 we perform temporal analyses of data shared on POIs.

5.3.1 POI Identification Algorithm

Another example of application that naturally emerges from analyzing PSN data are related to the identification of POIs in a city. Let's consider a PSN derived from Instagram. The identification of points of interest is possible because each picture represents, implicitly, an interest of an individual at a given moment. So, when many users share photos in a particular location at a given moment, it can be inferred that this place is a POI (see Figure 4.13).

More specifically, using as an example data of a PSN derived from Instagram, the Algorithm 1 formalize the process of identifying POIs by the following steps:

1. Each pair i of coordinates (longitude, latitude) $(x, y)_i$ is associated with a point p_i ;
2. calculate the distance [Sinnott, 1984] between each pair of points (p_i, p_j) ;
3. group all the points p_i that have a distance smaller than 250 m into a cluster C_k . This distance threshold was obtained by the method Complete-Linkage [Sørensen, 1948].

Algorithm 1: Identification of points of interests (POIs).

```

input : a data dictionary  $M$  with shared data. The keys are the locations and the values are the data attributes
output: a data dictionary  $M_{POIs}$  containing POIs
1  $D \leftarrow \emptyset$ ; // Distance between all points
2  $C \leftarrow \emptyset$ ; // All identified clusters
3  $C_{alt} \leftarrow \emptyset$ ; // Alternative clusters
4  $threshold \leftarrow \emptyset$ ;
5  $C_{POIs} \leftarrow \emptyset$ ; // Clusters representing POIs
6  $P \leftarrow$  contains all geographic coordinates of  $M$ ;
7 foreach  $p \in P$  do
8   |  $D.insert(\text{distance of } p \text{ between all coordinates of } P)$ ;
9 end
10  $C \leftarrow identifyClusters(D)$ ;
11 foreach  $c_k \in C$  do
12   | consider only one photo per user that shared data in  $c_k$ ;
13   | // creates alternative empty clusters for each  $c_k$ 
14   |  $C_{alt}.insert(\emptyset)$ ;
15 end
16 foreach  $f_i \in [all\ photos\ of\ C]$  do
17   | select a random cluster  $c_r \in C_{alt}$ ;
18   |  $c_r.insert(f_i)$ ;
19 end
20 calculate the normal distribution of # of photos of each  $c_i \in C_{alt}$ ;
21 //  $\mu$  and  $\sigma$  refer to the normal distribution
22  $threshold \leftarrow [-\infty; \mu + 2\sigma]$ ;
23 foreach  $c_k \in C$  do
24   | if # of photos in  $c_k \notin threshold$  then
25   | |  $C_{POIs}.insert(c_k)$ ;
26   | end
27 end
28  $M_{POIs} \leftarrow \{area\ of\ each\ c_i \in C_{POIs} : [all\ users\ that\ shared\ data\ in\ c_i,\ time\ of\ the\ data\ shared\ considered]\}$ ;

```

This step is represented by the function *identifyClusters* in the Algorithm 1. The result of this procedure is shown in Figure 5.19a, in which different colors represent different clusters k for the city of Belo Horizonte;

4. for each cluster C_k , we consider only one point (photo) per user. With that, the popularity of a cluster is now based on the number of different users that shared a photo in the cluster area. This procedure avoids considering areas visited by very few users, e.g., homes, as popular ones;
5. finally, for each cluster C_k , we create an alternative cluster C_r . Then, for each photo f_i , we randomly choose an alternate cluster C_r and we assign f_i to C_r . The number of photos assigned to each cluster from that process follows a normal distribution with mean μ and standard deviation σ . Thus, from the original clusters C_k found in the previous step, we exclude those in which the number of photos is within the distance 2σ above from the average μ , or is in the range $[-\infty; \mu + 2\sigma]$. The idea of this step is to exclude those clusters that may have been generated by random situations, i.e., those that do not reflect the dynamics of the city.

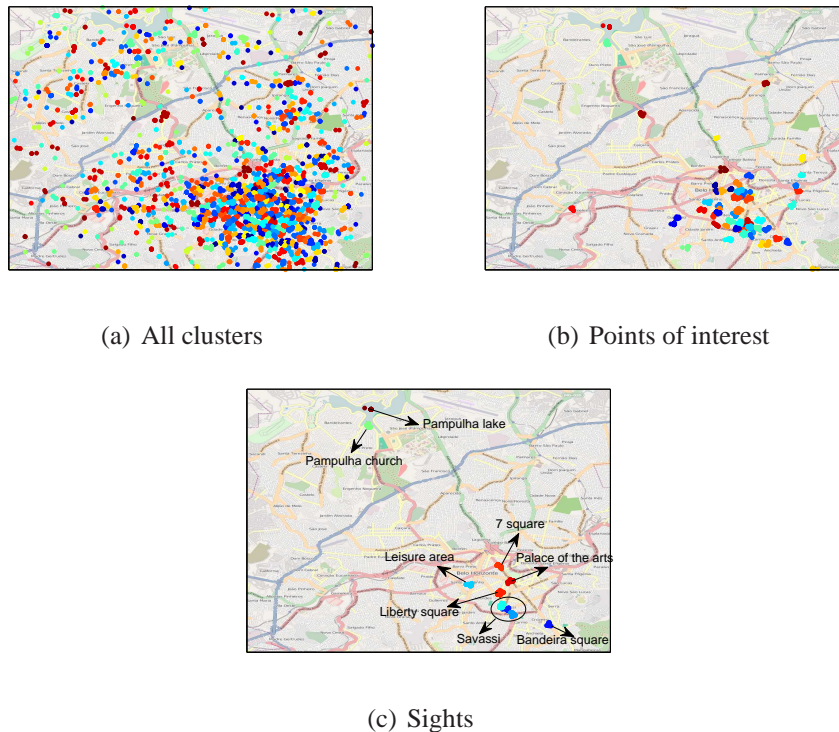


Figure 5.19: Points of interest of Belo Horizonte.

Figure 5.19b shows POIs obtained through this process. Observe the significant smaller number of points compared with the ones shown in Figure 5.19a. Besides identifying POIs in a city, we can also separate the sights from POIs. For this, first we generate a graph $G(V, E)$, where the vertices $v_i \in V$ are all POIs and there is an edge (i, j) from the vertex v_i to the vertex v_j if in a given time a user shared a photo on a POI v_j , after having shared a photo on POI v_i .

The weight $w(i, j)$ of an edge represents the total number of transitions performed from POI v_i to POI v_j considering transitions of all users. To identify sights, we consider that most tourists follow a well-known path within the city, being guided by the main sights of it. Moreover, at each point of interest he/she takes one or more photos and goes to the next tourist spot. Thus, we consider that edges (i, j) with high weights $w(i, j)$ denote these frequent transitions from one sight to another in a city.

After this, we exclude from G all edges (i, j) with weights $w(i, j)$ smaller than a threshold t , which is given by the probability of generating $w(i, j)$ randomly in a random graph $G_R(V, E_R)$. The identification of the value that separates edges with high weights from low weights is made as follows. First, we create a random graph $G_R(V, E_R)$ containing the same nodes of G . Then, for each sequence of n_u photos $f_u^1, f_u^2, \dots, f_u^{n_u}$ of each user u , we randomly assign a POI to each photo, what generates the random edges E_R of G_R . Thus,

the sequence of locations where the photos were taken is random, but the total number of photos that were taken is preserved. The idea is to simulate random walks in a city. In this random fashion, the distribution of edge weights follows a normal distribution $N_w(\mu_w, \sigma_w)$ with mean μ_w and standard deviation σ_w .

When the probability p_w of generating an edge weight $\geq w_t$ in $G_R(V, E_R)$ is, according to $N_w(\mu_w, \sigma_w)$, close to zero, then all transitions $v_i \rightarrow v_j$ with $w(i, j) \geq w_t$ are popular, in which, according to our conjecture, are transitions between sights. For our dataset, the value of w_t which provides a probability p_w close to 0 is $w_t = 10$. As we can see in Figure 5.19c, the vertices (POIs) of the resulting graph represent practically all the sights of Belo Horizonte. The areas of the resulting POIs cover seven out of all the eight Landmarks recommended by TripAdvisor⁶ as the most important cultural and leisure areas of Belo Horizonte.

Notice the difference between Figures 5.19b and 5.19c, the first containing all POIs and the second only the sights of the city of Belo Horizonte. This means that inhabitants could also use this application to explore the city. Again, this application is interesting because it is able to identify POIs in a spatio-temporal context, which is fundamental, since POIs are dynamic and change over time.

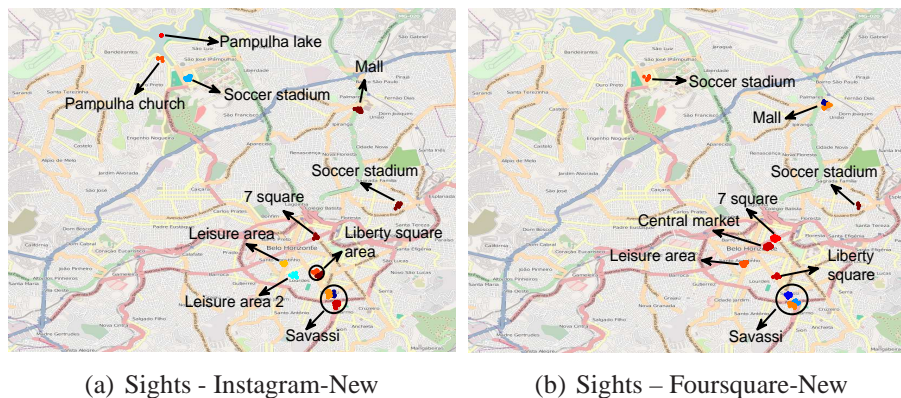


Figure 5.20: Sights identified in different datasets.

Inspired by the analysis performed in Section 4.4, we now have two goals: (i) verify whether with the Instagram-New dataset we can identify the same sights showed in 5.19c, which used the Instagram-OLD dataset; and (ii) verify whether the Foursquare dataset can also be used for this purpose, by using the Foursquare-New dataset. Following the steps described earlier.

Figure 5.20 shows sights identified for different PSNs. As a baseline of comparison, consider 5.19c. Figures 5.20a and 5.20c show the sights identified for Instagram-New and

⁶www.tripadvisor.com.

Foursquare-New datasets, respectively. During the collection of Instagram-OLD Belo Horizonte was not receiving soccer games. This explains why no soccer stadium was identified. Apart of that, we can see that many of the sights identified are in common in all three datasets, for example, Liberty Square, one of the most important sights of Belo Horizonte. The sights that were only previously identified, Palace of the Arts, and Bandeira Square, might not have been identified in the new datasets because no special event happened in those places. Palace of the Arts is a gallery with itinerant expositions, and Bandeira Square is not a spot that attracts naturally many people, especially tourists. It is interesting to note that, all social networks identified relevant sights of the city of Belo Horizonte, and they might be able to complement each other, since no one found all sights.

5.3.2 The Vibe of POIs

Figures 5.19b and 5.19c, for example, show that a particular area (southeast) of the city has a high concentration of POIs. This can be useful to guide tourists in the city, for example, when choosing a hotel location. Another interesting information for city explorers is the time when certain POI is more popular. Intuitively, we know that certain types of places are frequented by people only at specific times of particular days. Figure 5.21 shows the number of shared photos per hour for all days of our dataset in different types of places. Figure 5.21a shows a soccer stadium. In that figure, the word “WK” indicates that the delimitation for dashed lines represents a weekend, five in total. Most of the activities shown represent games that happened during the analyzed interval. Observe also the lack of activity between games, indicating that this is an *event-oriented* POI. Other types of POIs are also event-oriented: night clubs (Figures 5.21b and 5.21c), and a convention center (Figure 5.21d). Note that the activities in night clubs concentrate more during weekends, on the other hand in a convention center most of the activity happens during weekdays.

Concerning other types of POIs, we can see in Figures 5.21e and 5.21f that people share photos in a mall in many different times of the day, during weekdays and weekends. This is expected due to the high number of different attractions that a mall usually offers every day of the week. We also show the frequency of two of the most famous touristic attractions of Belo Horizonte in Figures 5.21g and 5.21h. The sharing pattern in touristic spots are not as intense as POIs with a high concentration of people and attractions such as malls, or as periodic as an event oriented POI, such as night clubs. These are powerful features for classifying POIs by their type and suggesting users about the best time and day to make a visit to it.

Finally, as we can see, the temporal photo sharing pattern presents somehow a signature of POIs, meaning that may be possible to automatically identify anomalous events. This

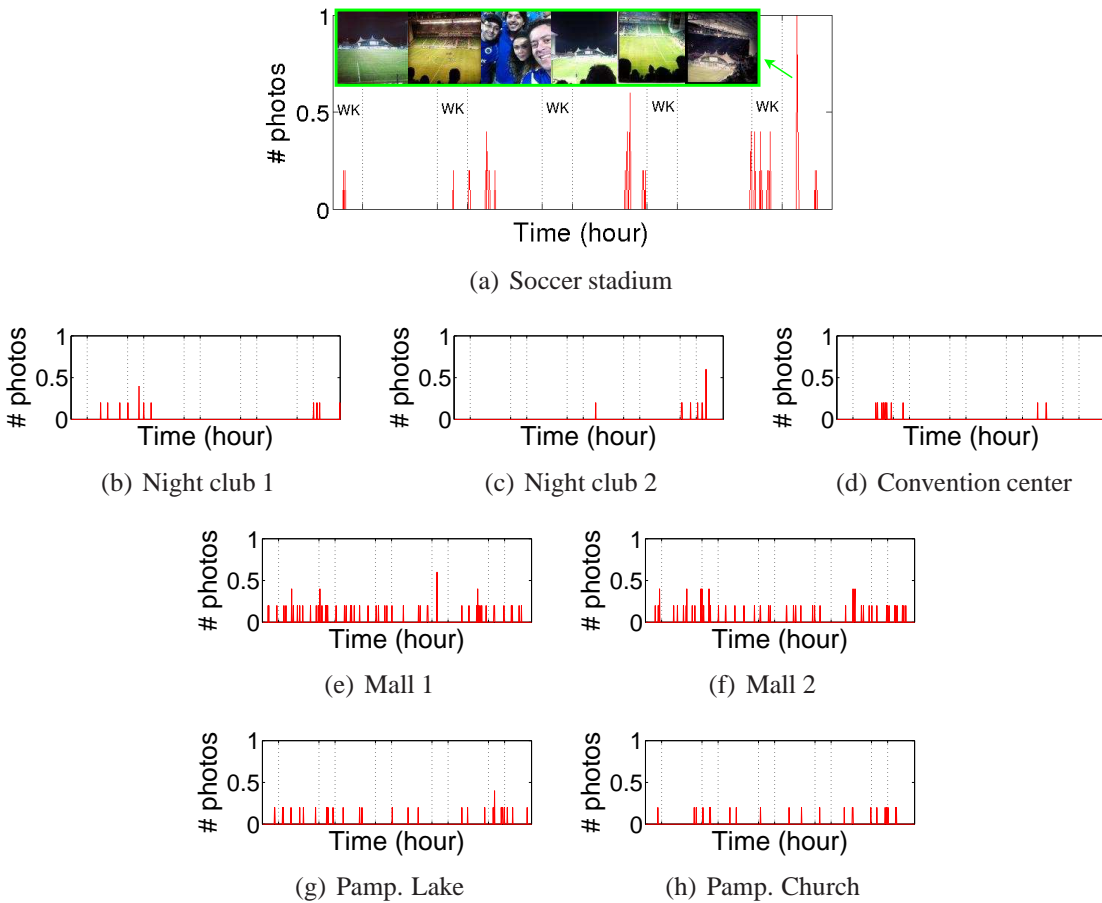


Figure 5.21: The temporal photo sharing pattern for different types of POIs.

can be used to capture in near real time unexpected events, such as an accident, or an event happening in an unusual place, for instance a street party or a concert on a park. After identifying those events, we could use the shared pictures to check, in near real time, snapshots of those events. Figure 5.21a illustrates the potential of this application, showing some pictures for the greatest peak of activity in that POI. In this case, a user could be aware that this event is a game of the Cruzeiro soccer team.

5.4 Socio-Economic Aspects

Data collected from social media applications can be used to infer the social network topology and dynamics of entire cities, ultimately enabling the analysis of social, economic, and cultural aspects of its inhabitants.

Semantic location services will be critical for the next wave of killer applications [Kim et al., 2011], and there are many possibilities to design them. The possibilities

listed here exploit the information about category of the venues present in the Foursquare-Crawled dataset.

Together with geographic neighborhoods, cities can be divided into semantic neighborhoods. To illustrate this idea, consider Figure 5.22a. This figure shows a heat map for two categories of venues: Arts & Entertainment, ranging from yellow to red, and Great Outdoors, ranging from light to dark blue. Again, darker colors represent larger numbers of check-ins. Note that it is possible to distinguish popular areas of venues related to the Arts & Entertainment and Great Outdoors categories. Using simple clustering algorithms to classify these regions, such as the one in [Cranshaw et al., 2012], it may be possible to offer to a tourist, for instance, an intuitive and automatic visualization of the points of interest in a given city.

Moreover, one might argue that a small coverage of a certain area by a PSN (i.e., only a small amount of data shared in that area) might indicate a lack of technology access by the local population, since the frequent use of location sharing services often relies on smartphones and 3G or 4G data plans, which, usually, are expensive. The preliminary results in the use of PSN in these scenarios demonstrate good opportunities to enable the visualization of interesting facts, some of them discussed in Section 4.1.2. For instance, analyzing carefully the data for the particular case of Rio de Janeiro, illustrated in Figure 5.22b, we observe that it is common to find very poor areas next to wealthy ones. Note the small sensing activity in the circle areas indicated as poor. This information may be useful to guide better public politics in those areas. The same information can be obtained using traditional methods, such as surveys, but in this new way we may be able to obtain the same results more quickly and cheaply.

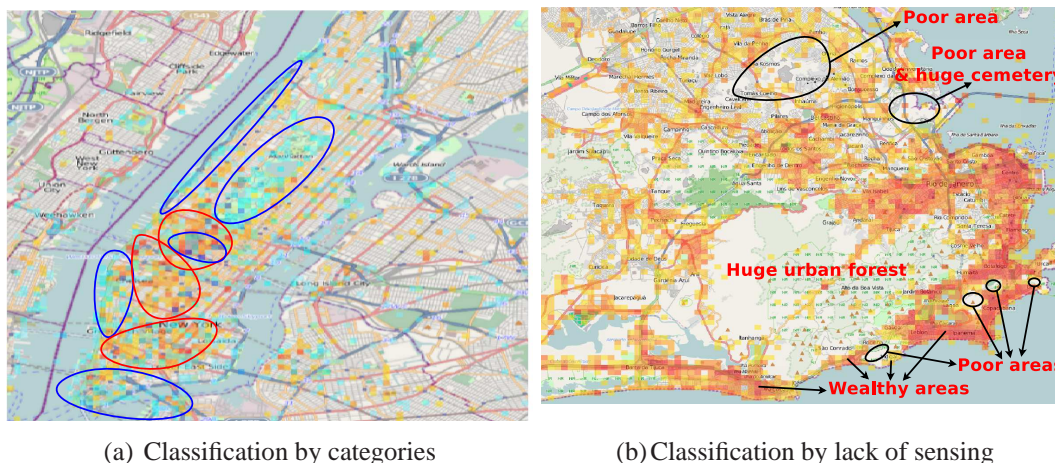
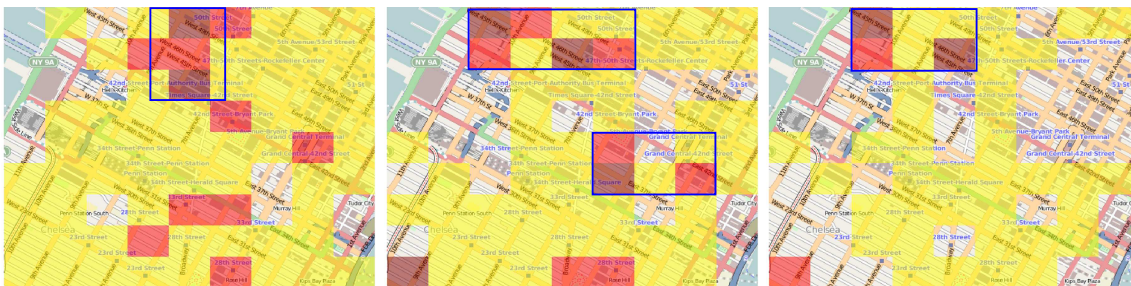


Figure 5.22: Examples of possible area classifications.

Other possibilities to classify areas emerge when jointly considering the time and

venue where the check-ins are performed. It may be possible to visualize crowds in a city in near real-time. Besides that, we observed in Section 4.1.4 that the seasonal patterns may be due to the circadian rhythm present in human routines. This seasonality has a great potential for prediction applications, since it is very likely that people repeat their activities in a periodic manner. We do believe that there are fruitful opportunities for prediction given by the circadian rhythm of people, enabling the prediction, for instance, of how crowded a place will be. This type of information is valuable in many scenarios, such as services for smart cities to avoid traffic in certain areas and offer alternative routes for users.

The following scenario illustrates another possibility that exploits the same data. For that, we created a simple method to estimate the number of check-ins in certain time and space. This method average the number of check-ins for the area of interest at a given time, taking into account every category separated. Figures 5.23a, 5.23b, and 5.23c show the check-ins estimation for “Food” places at 7:00 p.m., “Nightlife” category at 11:00 p.m., and “Nightlife” category at 1 a.m. for the same area respectively. Consider that Bob and Alice have tickets to watch their favorite rock band at Madison Square Garden, located in the area depicted, on Saturday at 8:30 p.m. to 10:30 p.m.. They want to have dinner in a popular place before the concert, and after that go clubbing nearby the arena. Since they do not know New York, they decide to use the information provided by an imaginary application represented on Figures 5.23a, 5.23b, and 5.23c. A candidate area to have dinner is marked by a blue rectangle in the Figure 5.23a. Regarding to where to go clubbing after the concert, the result shown in Figure 5.23b indicates at least two potentially good areas (blue rectangles). Since the couple plan to club until late at night, a tiebreaker criterion could be the estimation of the number of check-ins late at night for the same category, as shown in Figure 5.23c. The result indicates one of the two areas as the best choice.



(a) Food, Sat. 7 p.m.

(b) Nightlife, Sat. 11 p.m.

(c) Nightlife, Sun. 1 a.m.

Figure 5.23: Check-ins estimation for different times and type of places.

5.5 Cultural Differences

When studying the social behavior of particular areas, one of the first questions that emerges is: how different is one's culture from another? To address this question, it is necessary to define culture first. However, culture is such a complex concept that no simple definition or measurement can capture it. Among the various aspects that define the culture of a society (or person), one may cite its arts, religious beliefs, letters, and manners. Moreover, eating and drinking habits are also fundamental elements in a culture and may significantly mark social differences, boundaries, bonds, and contradictions [Carole, 1997; Cochrane and Bal, 1990]. Thus, we use this aspect to study the idiosyncrasies of different societies.

In this section, we propose a new methodology for identifying cultural boundaries and similarities across populations using self-reported cultural preferences recorded in PSNs. Our methodology, which is here demonstrated using data collected from Foursquare, consists of the following steps. First, we map food and drink check-ins extracted from Foursquare into users' cultural preferences. By exploring this mapping, we are able to identify particular individual preferences, such as the taste for barbecue or sake. Food and drink individual preferences, as shown in this thesis, are good indicators of cultural similarities between users. We then show how to extract features from Foursquare data that are able to delineate and describe regions that have common cultural elements, defining signatures that represent cultural differences between distinct areas around the planet. To that end, we investigate two properties of food and drink preferences: geographical and temporal characteristics. Next, we apply a simple clustering technique, namely k -means, to show the "cultural distance" between two countries, cities or even regions of a city, allowing us to draw cultural boundaries across them.

Unlike previous efforts, which used survey data, our work is based on a dynamic and publicly available Web dataset representing habits of a much larger and diverse population. Besides being globally scalable, our methodology also allows the identification of cultural dynamics more quickly than traditional methods (e.g., surveys), since one may observe how countries or cities are becoming more culturally similar or distinct over time.

The correct identification of cultural boundaries is useful in many fields and applications. Rather than using traditional methods to identify cultural differences, the proposed method is an easier and cheaper way to perform this task across many regions of the world, because it is based on data voluntarily shared by users on Web services. Moreover, since culture is an important aspect for economic reasons [Arrow, 1972; Garcia-Gavilanes et al., 2013], our methodology is valuable for companies that have businesses in one country and want to verify the compatibility of preferences across different markets. Another application that could rely on our methodology is a place recommendation system, which is useful

for visitors and residents of a city. Foursquare estimates that only 10% to 15% of searches on Foursquare are for specific places [Chaey, 2012]. Much more often users are searching within broader categories, such as “sushi” [Chaey, 2012]. Based on this information, systems like Foursquare and other location-based search engines, as the one proposed in [Shankar et al., 2012], could benefit from the introduction of new criteria and mechanisms in their recommendation systems that consider cultural differences between areas. For instance, a person who enjoyed a specific area of Manhattan could receive a recommendation of a similar area when visiting London.

It is important to emphasize that cross-cultural studies (i.e., the study of cultural differences) do not constitute a new research area. Indeed, they have been carried out by researchers working in the social sciences, particularly in cultural anthropology and psychology [Murdock, 1949]. Despite globalization and many other technological revolutions [Blossfeld et al., 2005], group formation might lead to the emergence of cultural boundaries that exist for millennia across populations [Barth, 1969]. Axelrod [1997] proposed a model to explain the formation and persistence of these cultural boundaries, which are basically a consequence of two key phenomena: social influence [Festinger, 1967] and homophily [McPherson et al., 2001]. While homophily dictates that only culturally similar individuals are likely to interact, social influence makes individuals more similar as they interact. In a long term, these two phenomena lead to very culturally distinct groups of individuals, delimited by the so-called *cultural boundaries*.

The rest of this section is organized as follows. Section 5.5.1 describes our dataset and the core of our methodology for extracting cultural preferences from participatory sensor networks. Section 5.5.2 shows how to extract cultural signatures for different areas of the globe and explore the similarities among them, while Section 5.5.3 applies this knowledge to analyze the implicit cultural boundaries that exist for different cultural aspects of the society.

5.5.1 Extracting Cultural Preferences

5.5.1.1 Mapping User Preferences

One of the biggest challenges in the analysis of cultural differences among people and regions is finding the appropriate empirical data to use. The common approach to overcome this challenge is the use of surveys based on questionnaires filled during face-to-face interviews [Valori et al., 2012], such as the Eurobarometer dataset [Schmitt et al., 2005]. Through these questionnaires, individual preferences, such as the taste for coffee and fast food, can be mapped into multidimensional vectors representing (and characterizing) each interviewee. From these vectors, it is possible, for instance, to quantify how similar or different two individuals are.

Although survey data are broadly used in the analysis of cultures, there are some severe constraints in its use, which are well known to researchers. First, surveys are costly and do not scale up. That is, it is hard to obtain data of millions, or even thousands of people. Second, they provide static information, i.e., they reflect the preferences of users at a specific point in time. If some of the preferences change for a significant amount of the interviewed people, such as the taste for online gaming instead of street ball playing, the data is compromised.

In order to overcome the aforementioned constraints, we propose the use of publicly available data from PSNs to map individual preferences. PSNs can be accessed everywhere by anyone who has an Internet connection, solving the scalability problem and allowing data from (potentially) the entire world to be collected. Moreover, these systems are dynamic, being able to capture the behavioral changes of their users when they occur, which solve the second mentioned constraint. However, data from such systems can be used if and only if they meet the requirements:

- **[R1]** It is possible to associate a user to its location;
- **[R2]** It is possible to extract a finite set of preferences from the data that is generated by the system;
- **[R3]** It is possible to map users' actions in the system into the preferences defined in **[R2]**.

Considering that these requirements are met, a dataset containing individual activities of N users of a PSN can be used to map preferences as follows. First, associate each user n_i with a location l_i , which may be a country, a city or even a region within a city. Then, define a set of m individual preferences (or features) f_1, f_2, \dots, f_m that can be extracted from the dataset, which may represent the taste for the most varied things, such as Japanese food or a certain football team. Finally, map the activities of each individual n_i into an m -dimensional vector of preferences $F_i = f_1^i, f_2^i, \dots, f_m^i$ that characterizes the person's tastes, the same type of vector that is usually created from survey data [Valori et al., 2012].

Since the preference vector F_i is generated from self-reported temporal data of an individual n_i , we may populate and modify it in various ways. For instance, we can use a binary representation, where $f_k^i = 0|1$ represents whether user n_i has or not preference f_k (e.g., whether a person likes/dislikes a certain type of food), respectively. Alternatively, we may consider the intensity at which a user likes a feature, inferred from the number of times the corresponding preference is reported in the person's data, i.e., $f_k^i = [0; \infty)$. In Appendix C, we adopt a binary representation. Finally, one can group individuals by

their geographical regions and sum up their preference vectors to characterize their regions. We adopt this approach in Section 5.5.2 to build preference vectors for regions (instead of individuals).

5.5.1.2 Data Description

Here we consider the dataset Foursquare-Crawled, described in Section 4.1. Since we are primarily interested in food and drink habits, we manually grouped relevant subcategories of the Food and Nightlife Spots categories into three classes: Drink, Fast Food, and Slow Food places. We did this by excluding some subcategories that are not related to these three classes (e.g. Rock Club and Concert Hall) and moving some subcategories (e.g. Coffee Shop and Tea Room) from the Food category to the Drink class. Besides that we also disregard the category Restaurant, because it is a sort of meta category that could fit in any of the two classes of food. After this manual classification process, the Drink class ended up with 279,650 check-ins, 106,152 unique venues and 162,891 unique users; the Fast Food class with 410,592 check-ins, 193,541 unique venues, and 230,846 unique users; and the Slow Food class with 394,042 check-ins, 198,565 unique venues, and 231,651 unique users. Moreover, the Drink class has 21 subcategories (e.g., brewery, karaoke bar, and pub), whereas the Fast Food class has 27 subcategories (e.g., bakery, burger joint, and wings joint) and the Slow Food class has 53 subcategories, including Chinese restaurant, Steakhouse, and Greek restaurant.

To provide an idea about the size of the user population PSNs can reach, consider the World Values Survey⁷ project. That study is maybe the most comprehensive investigation of political and sociocultural change worldwide, which was conducted from 1981 to 2008 in 87 societies, with about 256,000 interviews. Observe that our one-week dataset has a population of users of the same order of magnitude of the number of interviews performed in that project in almost three decades.

5.5.1.3 Mapping Foursquare Data into User Preferences

Several characteristics of human beings are not directly observable, such as personality traits. Thus, we rely on face-to-face interactions or online signals to discover the presence of those hidden qualities [Pentland, 2010]. In this direction, a check-in given in a PSN can be considered as a signal because it is a perceivable feature/action that expresses the preference of a user for a certain type of place. With that in mind, we use Foursquare check-ins to represent user preferences regarding food and drink places. Specifically, we use the three main classes defined in Section 5.5.1.2, namely, *Drink*, *Fast Food*, and *Slow Food*.

⁷<http://www.worldvaluessurvey.org>.

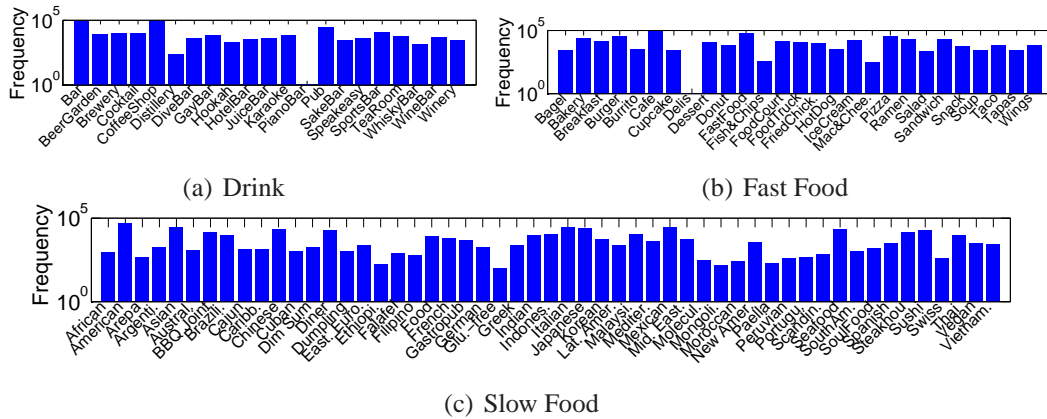


Figure 5.24: Frequency of check-ins at all subcategories of the three analyzed classes. The names of some places are abbreviated but the semantics of the names is preserved.

Figures 5.24a, 5.24b, and 5.24c show the frequency of check-ins at each subcategory of the Drink, Fast Food, and Slow Food classes, respectively, so we can have a general idea about the popularity of user preferences for different food and drink related places. These figures show the popularity of different places according to people’s preferences worldwide. Note that Coffee Shop and Bar are the two most popular subcategories of Drink places, with 86,310 and 81,124 check-ins, respectively. The two most popular Fast Food subcategories are Café⁸ and Fast Food Restaurant, with 91,303 and 56,648 check-ins, respectively. Finally, American Restaurant (47,373 check-ins), and Mexican Restaurant (28,712 check-ins) are the two most visited subcategories of Slow Food places.

In this dataset, a user is represented by a vector of $m = 101$ features corresponding to the 101 subcategories that comprise the three classes we have defined. A feature $f_i \in F = \{f_1, f_2, \dots, f_{101}\}$ is equal to 1 if a user made at least one check-in at f_i , and 0 otherwise. In this way, a feature vector represents the positive and negative preferences of a user for fast food, slow food and drink subcategories. With that, a finite set of preferences is extracted (requirement **[R2]**, see definition in Section 5.5.1.1) and users’ actions are mapped into this set (requirement **[R3]**). To associate a user with a location (requirement **[R1]**), we analyzed the GPS coordinates of all check-ins performed by the user. If all check-ins performed are from the same country, according to the free reverse geocoding API offered by Yahoo⁹, we assume that the user taken into consideration is from that country. Otherwise, we do not consider the user in our analysis. In this way, we minimize the wrong association of a user with a country. Following this procedure, approximately 1% of the users were disregarded from our analysis.

⁸Like in many European countries, this term is referred as a restaurant primarily serving coffee as well as pastries.

⁹<http://developer.yahoo.com>.

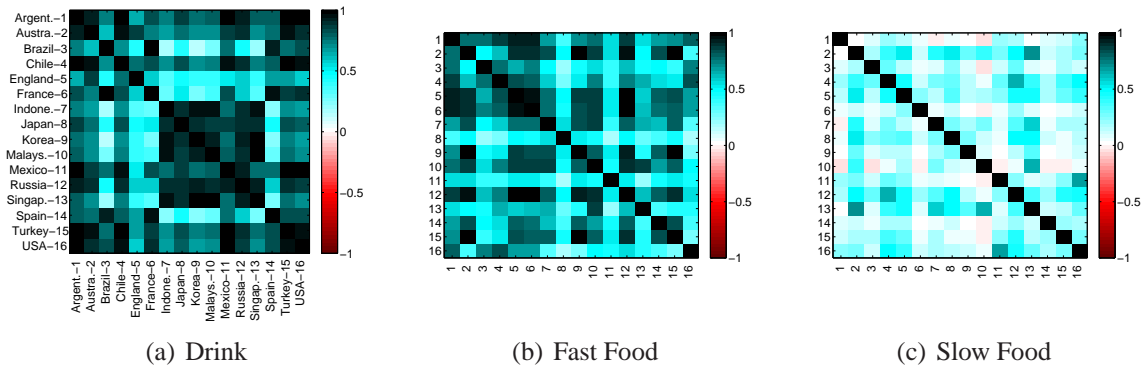


Figure 5.25: Correlation of preferences between countries.

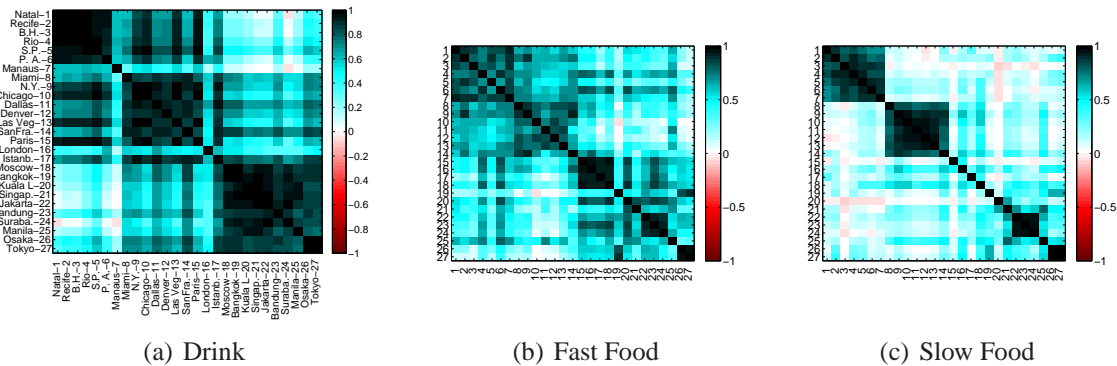
5.5.2 Extraction of Cultural Signatures

Appendix C analyzes the individual preferences of users, showing, among other results, that food and drink preferences are good indicators of cultural similarities. Given that, we hypothesize that it is possible to define cultural signatures of different areas around the planet. In this section, we show how to extract features from Foursquare data that are able to describe regions from their cultural elements. In particular, we investigate two properties of food and drink preferences: their geographical (Section 5.5.2.1) and temporal (Section 5.5.2.2) aspects.

5.5.2.1 Spatial Correlations

Here our goal is to define a set of features that are able to characterize the cultural preferences of a given geographical area in the planet, such as a country, a city or a neighborhood. Thus, for a given delimited area a (e.g., the city of Chicago), we sum up the values of the features in the preference vectors of the users who checked in at venues of that area. In other words, we count the number of check-ins $C^a = c_1^a, c_2^a, \dots, c_{101}^a$ performed in venues of each of the 101 subcategories s_1, s_2, \dots, s_{101} of the Fast Food, Slow Food and Drink classes (Section 5.5.1.2) that are located within the perimeter of area a . Next, we represent each area a by a vector of 101 features $F^a = f_1^a, f_2^a, \dots, f_{101}^a$, where each feature f_i^a is equal to $c_i^a / \max(C^a)$. That is, we normalize the number of check-ins at each subcategory by the maximum number of check-ins performed in a single subcategory in area a ($\max(C^a)$). Thus, each area a is represented by a feature vector F^a containing values from 0 to 1, indicating the preferences of people who visited that area, i.e., the profile of preferences for that area. From now on, we use F_{drink}^a , $F_{slow\ food}^a$ and $F_{fast\ food}^a$ to refer, respectively, to the subset of features that correspond to subcategories belonging to the Drink, Slow Food and Fast Food classes in area a .

In order to verify if two areas a and b are culturally similar, we compute the Pearson's correlation coefficient between the two feature vectors F^a and F^b of those areas. We



(a) Drink (b) Fast Food (c) Slow Food

Figure 5.26: Correlation of preferences between cities.

compute the correlation considering all features (F^a and F^b) as well as a subset of them (e.g., F_{drink}^a and F_{drink}^b). In particular, Figure 5.25 shows the correlations between areas corresponding to 27 different popular countries for the Drink (5.25a), Fast Food (5.25b), and Slow Food (5.25c) classes; the darker the color, the stronger the correlation (blue for positive correlations, red for negative correlations). The same correlations computed for city level areas (16 cities around the world) are shown in Figure 5.26.

Analyzing the results for the Drink class (Figure 5.25a), we find countries with very strong correlations, such as Argentina and Chile, as well as countries with low correlation, such as Brazil and Indonesia. Moreover, although regions close geographically tend to have stronger correlations, this is not always the case. For example, the correlation between Brazil and France is stronger than the correlation between England and France, which are geographically closer. Similarly, Figure 5.26a¹⁰ shows that cities in the same country tend to have very correlated drinking habits in most cases, but there are exceptions: Manaus (Brazil), for instance, has weak correlation with other cities in Brazil. This might be due to this city being located in the North region of Brazil, which is known for having a strong cultural diversity compared to other parts of the country.

Turning our attention to food practices, we observe in Figures 5.25b and 5.26b the global penetration of fast food venues, at both country and city levels, explained by the diffusion of fast food places worldwide [Watson, 2006]. This is not observed in the same intensity for the Slow Food class (Figures 5.25c and 5.26c). The Slow Food class presents the highest distinction, or smaller correlation, across most of the countries and cities. This is expected, since Slow Food venues usually are representative of the local cuisine. Note, for instance, that cities from Brazil and USA have highly correlated drinking and fast food

¹⁰The ratio of check-ins per inhabitant is similar among all the cities taken into consideration. For example, comparing Manaus (one of the cities with fewer check-ins) with Sao Paulo (largest number of check-ins in Brazil) we find the following ratios: 0.35×10^{-3} and 0.37×10^{-3} (Drink class); 0.73×10^{-3} and 0.75×10^{-3} (Fast Food class); and 0.54×10^{-3} and 0.71×10^{-3} (Slow Food class).

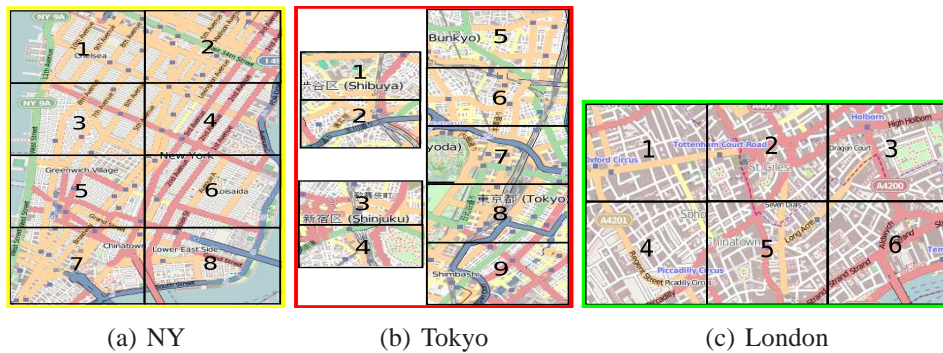


Figure 5.27: Areas of cities taken into consideration: London/England; New York/USA; and Tokyo/Japan.

habits, but almost no correlation in slow food habits.

Finally, we turn our attention to the cultural habits within city boundaries. It is known that, in many cities, there is a strong cultural diversity across different neighborhoods [Cranshaw et al., 2012], reflecting distinct activities typically performed in these areas. To analyze these local cultures, we focus on three populous cities, namely London, New York, and Tokyo. We divide each city’s geographical area using a grid structure. Next, we select the most popular cells in the grid of each city and label them with a number, as shown in Figure 5.27. We then compute the correlation between the selected cells. Note that we here assume a grid with regular (rectangular) cells to show the potential of the proposed analysis. However, our approach can be applied to any other segmentation of the city areas (e.g., by city districts).

Figure 5.28 shows the correlations for pairs of cells within the same city and from different cities. Note that, for the Drink class, different areas within the same city tend to have very strong correlations. There are also areas from different cities with strong correlations (e.g., areas NY-5 and TKO-1). For Fast Food places, the correlations between areas within the same city are much stronger for Tokyo, although the correlations between New York and London areas are fairly moderate. In contrast, there are areas with negative correlation, e.g., NY-3 with most of Tokyo areas.

Finally, for the Slow Food class, once again Tokyo areas are very strongly correlated among themselves. In comparison with the Fast Food class, there is a more clear distinction (weaker correlation) between London and New York areas as well as among distinct areas in London. This last observation is probably due to a specific characteristic of London, which has neighborhoods with a strong presence of a cuisine of a particular region of the globe. Observe also that two specific areas of New York, namely NY-7 and NY-8, are particularly not correlated with the others from this city. This is probably related to the location of Chinatown in those areas (mainly NY-7). Indeed, this particular area (NY-7) has a strong

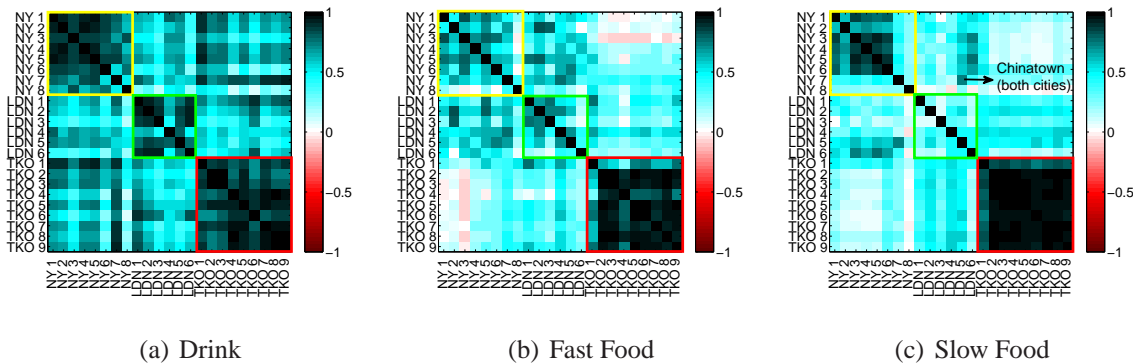


Figure 5.28: Correlation of preferences in regions of London, NYC and Tokyo.

correlation with a particular area of London, LND-5, where Chinatown/London is located.

5.5.2.2 Temporal Analysis

We now turn our attention to the temporal and circadian aspects of cultural habits. The time instants when check-ins are performed in food and drink places may also provide valuable insights into the cultural aspects of a particular region. For example, in a particular area, one may like to drink beer during the weekends but not during the weekdays.

To that end, we first count the number of check-ins per hour during the whole week covered by our dataset in venues of each class (Drink, Fast Food and Slow Food) for different regions. Next, we group days into weekdays and weekends, summing up the check-ins performed on the same hour of the day in each group and for each region. We then normalize this number by the maximum value found in any hour for the specific region, so that we can compare the patterns obtained in different regions. For illustration purposes, we show the results for three countries (Brazil, USA, and England) and for three American cities (Chicago, Las Vegas, and New York) in Figures 5.29 and 5.30, respectively. Results for each class are shown separately for weekdays and weekends.

Focusing first on weekday patterns, Figure 5.29 shows that American and English people have similar peaks of activities, despite differences in their preferences for different categories of places, as previously shown (Figure 5.25). In contrast, Brazilians tend to have significantly different temporal patterns, particularly in terms of activities in Slow Food places (Figure 5.29c): whereas Americans and English people tend to have their main meal at dinner time, Brazilians have it at lunch time. Observe also that Brazilians have their meals later, compared to Americans and English people.

Concerning the times when people go to drink venues, it is possible to note similarities among most of the cities from the same country, but also some different patterns. For example, most of the analyzed cities from USA exhibit a weekday pattern similar to New York

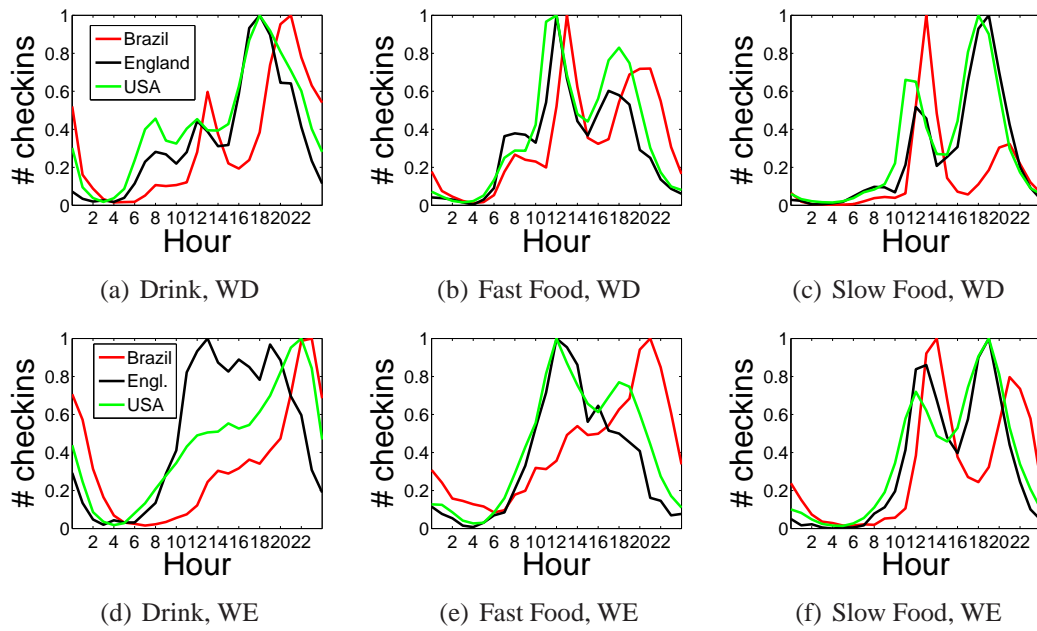


Figure 5.29: Number of check-ins throughout the hours of the day in different countries (WD = weekday; WE = weekend).

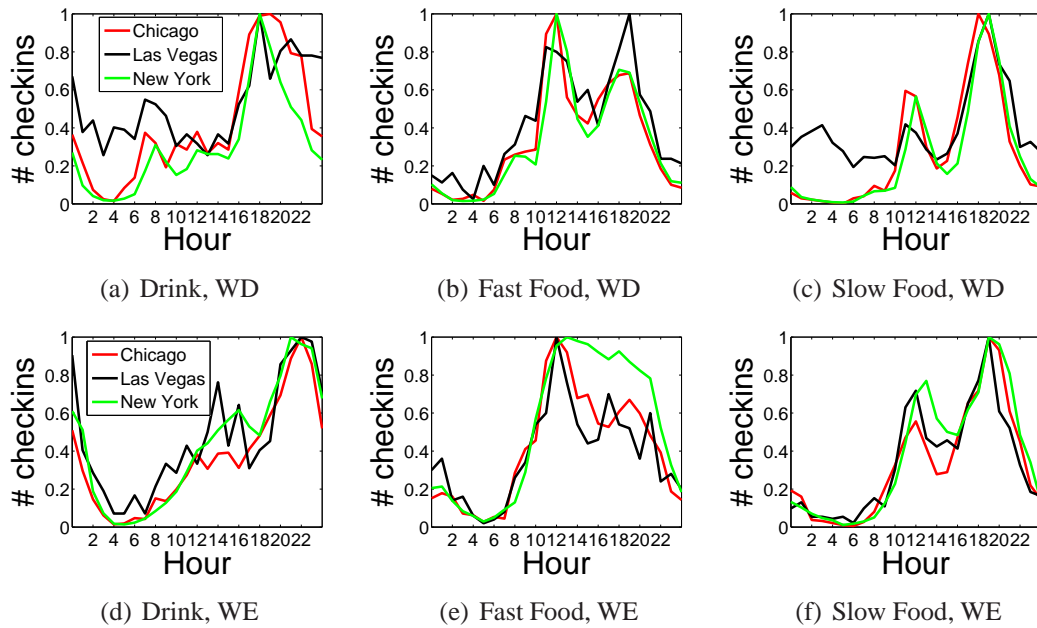


Figure 5.30: Number of check-ins throughout the hours of the day in different American cities (WD = weekday; WE = weekend).

and Chicago, shown in Figure 5.30a, with three distinct peaks around breakfast, lunch and happy hour (around 6 p.m.). This behavior is consistent with the general pattern observed for the country, shown in Figure 5.29a. However, Las Vegas is one exception, since there is an intense activity during the dawn, besides many other peaks of activities that do not occur

in other cities.

Turning our attention to eating habits on weekdays, Figure 5.30 shows that most cities in the USA present activity patterns very similar to the general pattern identified for the country, both in terms of Slow and Fast Food places. However, as observed for drinking patterns, there are exceptions, such as Las Vegas, which exhibits distinct trends that reflect inherent idiosyncrasies of this city. We also note relevant similarities and differences in eating habits of people from cities in different countries. For example, comparing Figures 5.30b and 5.30c with similar graphs produced for different Brazilian cities, we find that while all curves for the Fast Food class are very similar, the curves for Slow Food places are quite different, reflecting distinct habits for each country, as discussed previously.

The curves for weekends have very distinct peaks of activities from those of weekdays, both at the country and city levels. For instance, as shown in Figure 5.29, English people have a very distinct drinking pattern from Americans on weekends. Moreover, the differences among the countries in terms of preferences at Slow Food places are also clear on weekends: Brazilians tend to go to Slow Food places more often at lunch time, whereas Americans and English people do it more at dinner time.

We note that there is no clear (dominant) temporal check-in pattern for Fast Food places on weekends, when considering different cities of a country. However, we do note that most activities happen after noon, which was expected. In contrast, there is a dominant pattern for check-ins at Slow Food places on the weekends, and it is similar to the one observed on weekdays. This is possibly because such places (often restaurants) have well-defined opening hours, serving meals around lunch and dinner times only, which coincide with the times of check-in peaks (Figures 5.29c, 5.29f, 5.30c, and 5.30f). Assuming that the height of such peaks reflects the importance of that meal for a certain culture, we note once again a key distinction between Americans and Brazilians.

5.5.2.3 Discussion

In addition to temporal and spatial patterns of check-ins at different types of places, we also compute the Shannon's entropy [Shannon, 1948] of preferences for each venue subcategory among all considered areas. The goal is to analyze whether the check-ins at specific subcategories are more concentrated at specific areas (low entropy) or not (high entropy). We compute the entropy for subcategories of each class (Drink, Fast Food and Slow Food) at country and city levels. The average entropy for subcategories of the Drink class is 3.23 (standard deviation $\sigma = 0.93$) for countries and is 3.88 ($\sigma = 1.09$) for cities. Sake bar is one example with low entropy (1.13 for countries and 1.89 for cities), which indicates that this subcategory is popular on very few countries and cities. Surely Japan contributes con-

siderably to this result. On the other hand, the average entropy for subcategories of the Slow Food class is much larger, 2.63 ($\sigma = 0.78$). This higher entropy reflects the widespread popularization of various cuisines. For example, a check-in at an Italian restaurant does not necessarily mean that it represents a behavior of an Italian, since it is a very international type of restaurant, confirmed by the high entropy (3.63). Note, however, that if the check-in at an Italian restaurant is made at lunch time it could be more likely to represent a Brazilian behavior than American, since Brazilians have their main meal at lunch time, as presented in Section 5.5.2.2. Time plays an important role in this case.

Given these considerations and all the observations reported here, we propose the use of spatio-temporal correlations of check-ins as cultural signatures of regions.

5.5.3 Identifying Cultural Boundaries

5.5.3.1 Clustering Regions

In this section, we use the cultural signatures of regions described above to identify similar areas around the planet according to their cultural aspects, delineating their so-called “cultural boundaries”. To that end, we first represent each area a by a high dimensional preference vector composed of 808 features, namely the normalized number of check-ins at each of the 101 subcategories in four disjoint periods of the day, on weekdays and on the weekends. We then apply the Principal Component Analysis (PCA) [Jolliffe, 2002] technique to these vectors to obtain their principal components¹¹. Finally, we use the k -means algorithm, a widely used clustering technique, to group areas in the space defined by these principal components. We perform this analysis for areas defined at the country, city and neighborhood levels.

The score values for the first two principal components (P.C.) generated by the PCA for countries, cities, and regions are shown in Figures 5.31a, 5.31b, and 5.31c, respectively. The variance in the data explained by these first two components is shown in each figure. Each color/symbol in those figures indicates a cluster obtained by k -means, which used the p first principal components that explain 100% of the variation in the data ($p=15$ for countries, $p=26$ for cities and $p=22$ for regions). The k value in the k -means varied according to the characteristics of the considered areas. For countries, we set $k=7$ (same number of clusters used in [Inglehart and Welzel, 2010]). Following the same logic, we set $k=4$ for cities, since we considered cities from 4 different continents/countries, and $k=3$ for regions inside a city, because we considered 3 cities. We used the cosine similarity to compute the similarity between locations.

¹¹Alternative methods could be applied to reduce the dimensionality of these vectors. A comparison of these methods is out of the scope of the present work.

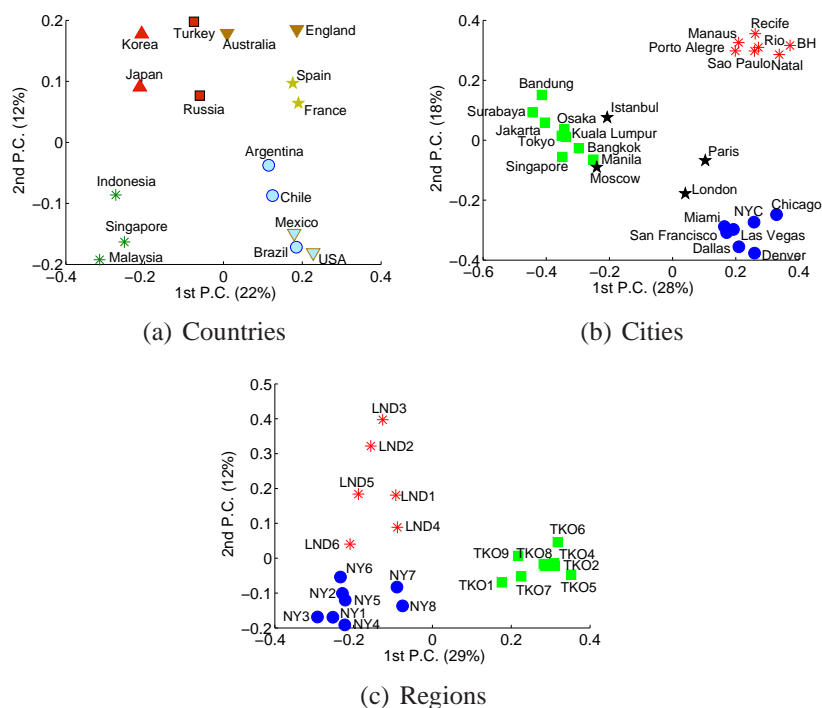


Figure 5.31: Clustering results for countries, cities, and regions inside cities.

It is possible to observe in Figure 5.31a that countries with closer geographical proximity are not necessarily associated with the same cluster. For example, Australia and Indonesia are *not* in the same cluster. Although they are geographically neighboring countries, they are culturally very distinct. When analyzing large cities from the considered countries, Figure 5.31b shows that they are well clustered by the geographical regions where they are located: Asia, Brazil, Europe and USA. Intuitively, this result makes sense, since, for instance, cosmopolitan European capitals tend to present more similar cultural habits among each other than among cities from different continents. Turning our attention to regions inside London, NY, and Tokyo, we observe in Figure 5.31c that all regions in the same city are in the same cluster. This result was also expected when considering all features. Besides that, when we analyze a subset of features, for example, drinking habits during weekends in all regions of London, NY, and Tokyo, Figure 5.32 shows this result, we find that some regions of London and NY are clustered together. This is corroborated by the results shown in Section 5.5.2: for certain categories, there are regions from different cities that are very similar and, thus, end up clustered together.

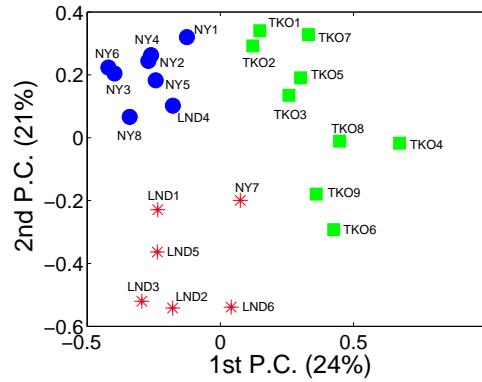


Figure 5.32: Clustering results for cities on weekend, considering only the Drink class.

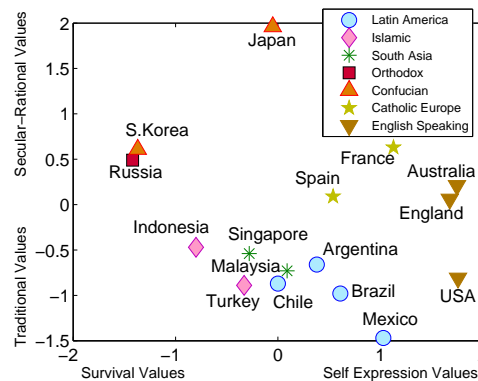


Figure 5.33: The cultural map of the World given by the World Values Survey [Inglehart and Welzel, 2010].

5.5.3.2 Comparing with Survey Data

Similarly to us, Ronald Inglehart and Christian Welzel proposed a cultural map of the world based on the World Values Surveys (WVS) data from 2005 to 2008 [Inglehart and Welzel, 2010]. This map is shown in Figure 5.33 and contains only the countries we analyze in this thesis. It reveals two major dimensions of cross-cultural variation: a traditional versus secular-rational values dimension and a survival versus self-expression values dimension. Moreover, it offers a division of the world into clusters, similarly to what we have done in the previous section. Comparing Figures 5.31a and 5.33, observe that the similarities are striking, with only two major differences. First, the “Islamic” cluster dissolved, with Turkey joining Russia and Indonesia joining Malaysia and Singapore. Second, USA and Mexico left the “English Speaking” and the “Latin America” clusters, respectively, and paired up to form a new one. Note, nevertheless, that these differences might not be surprising as these new boundaries.

We formally investigate the differences between boundaries given by the WVS study

Table 5.6: The Spearman’s rank correlation coefficient ρ (and its respective p-value) between the rank of similar countries generated from WVS and by our approach.

| Country | <i>dataset₁</i> | | <i>dataset₂</i> | |
|-----------|----------------------------|---------------|----------------------------|---------------|
| | ρ | p-value | ρ | p-value |
| - | | | | |
| Argentina | 0.56 | 0.03 | 0.77 | 0.0007 |
| Australia | 0.32 | 0.23 | 0.60 | 0.02 |
| Brazil | 0.48 | 0.06 | 0.81 | 0.0002 |
| Chile | 0.32 | 0.23 | 0.53 | 0.04 |
| England | 0.87 | 0 | 0.70 | 0.004 |
| France | 0.85 | 2e-06 | 0.61 | 0.01 |
| Indonesia | 0.84 | 4e-05 | 0.75 | 0.001 |
| Japan | 0.38 | 0.15 | 0.39 | 0.13 |
| Korea | 0.68 | 0.004 | 0.45 | 0.08 |
| Malaysia | -0.16 | 0.54 | 0.11 | 0.68 |
| Mexico | 0.55 | 0.03 | 0.71 | 0.003 |
| Russia | 0.78 | 0.0006 | 0.76 | 0.001 |
| Singapore | 0.34 | 0.20 | 0.65 | 0.008 |
| Spain | 0.78 | 0.0005 | 0.75 | 0.001 |
| Turkey | -0.18 | 0.50 | -0.31 | 0.24 |
| USA | 0.70 | 0.004 | 0.67 | 0.005 |

and by our approach. In order to do so we rank, for a given country, all the other countries according to their cosine similarity towards it. We compute the similarity using the dimensions produced by the WVS data [Inglehart and Welzel, 2010] and the dimensions computed by our approach. Then, we compute the Spearman’s rank correlation coefficient ρ between these two ranks to see, for instance, if the most similar (and distinct) countries to England using the WVS data are ranked similarly when we use our approach. In our approach, we use two different datasets. In *dataset₁*, we use the full set of features, as done so far. In *dataset₂*, we use solely the features extracted from the fast food check-ins performed during the weekends¹². Table 1 shows these results. We highlight in bold all the coefficients which are statistically significant, i.e., with a *p-value* < 0.05 . Observe that the correlation ρ is significant and positive for several countries. For *dataset₁* and *dataset₂*, 9 and 12 countries have similar ranks with the ones given by the WVS, respectively. This shows that our approach, which is based solely on one week of participatory data, has a clear potential to reproduce cultural studies performed using surveys, such as the ones relying on the WVS, which is based on 4 years of survey data.

We would also like to point out the reasons for the differences between our cultural map and the WVS map, as well as for the negative correlations seen in Table 1. First, the traits of each dataset are significantly different. While the WVS looked at several cultural dimensions, from religion to politics, from economics to lifestyle, we looked only at food and drink preferences. Second, the WVS data has a distance of 4 to 7 years to our data. During this time, significant cultural changes may have happened, given that the world is

¹²This particular set of features was chosen because it was the configuration which gave the best results.

getting more connected at every day. Third, the most significant differences are related to multi-ethnic, multicultural, and multilingual countries, such as Malaysia and Turkey. In these countries it is probably hard to find culturally homogeneous samples of individuals, which might be the cause of the discrepancies seen between our results and those described in [Inglehart and Welzel, 2010].

5.6 Discussion

In sum, the use of participatory sensor networks can help us better understand the dynamics of cities and urban social behavior, and from this we are able to offer smarter services to meet people's needs. We demonstrated this by proposing different techniques and methodologies to that end, including:

- a technique for summarizing the city dynamics based on transition graphs;
- a technique for identification of points of interest in the city;
- a new methodology for identifying cultural boundaries and similarities across populations;
- presentation of possibilities to better understand city dynamics through people movements.
- presentation of possibilities to the use of PSNs to the analysis of social and economic aspects of cities' inhabitants.

In Chapter 4 we discussed some possible limitations of our datasets. Here we summarize the performed procedures to tackle those limitations. One possible limitation of our Foursquare dataset is that it might be considered small, especially the one used to demonstrate the City Image technique (one week). To analyze to which extent this might impact the conclusions drawn from the City Images, we collected extra check-ins (for one extra week) and recalculated the City Images for each considered city using all the data available. We observe that the new City Images are very similar to the corresponding ones produced using our original one-week dataset, the changes, if observed, occur in some transitions classified in the indifference range.

Besides that, we also performed several statistical treatments in the datasets. Here we illustrate some examples. In the City Image and POIs identification techniques we created null models, identifying data that could be generated in a random fashion. This step is important because prevent us to use data that do not have relationship with the phenomena

| Our proposal | Verification |
|--|---|
| City Images | Clusterization agrees with common knowledge |
| Methodology for cultural boundaries identification | Very similar results with WSV |
| Sights identification | Compatibility with TripAdvisor's recommendation |

Table 5.7: Summary of the approaches applied to verify our results.

we are interested in. We also studied the ratio of check-ins per inhabitant, showing that it is similar among all the studied cities. We were also always concerned in normalizing the data before performing comparisons.

In addition, we always tried to verify the results obtained using the techniques and methods proposed here. Table 5.7 summarizes this discussion. In order to investigate if the City Images generated reflects the reality, we proposed a method for clustering them. The results are shown in the Appendix B. We could confirm that they are compatible with common knowledge, showing that the results does reflect typical transitions of performed by inhabitants of those cities. To investigate the accuracy of our method for cultural boundaries identification, we compared our results with the cultural map of the world based on the World Values Surveys (WVS), one of the most important studies of this area. We observe that the similarities of this study with our results are very good. Finally, we studied the identified sights by the POI identification technique, and we found that they cover seven out of all the eight Landmarks recommended by TripAdvisor as the most important cultural and leisure areas of the studied city.

As we can see, despite the possible bias and limitations of our data, the results we present in this work, as demonstrated, hold strong. Nevertheless, although these limitations do prevent us to make some general assertions, they do *not* invalidate our techniques and methodologies. We believe that applying the proposed techniques to a larger less biased datasets in future research may provide an even more accurate representation of the city dynamics and urban social behavior.

Chapter 6

Participatory Sensor Networks as Sensing Layers

Data from different PSNs are associated with a spatial-temporal context that can be correlated or not. In order to find out what is the case, a characterization study is needed. With that in mind, we characterize distinct PSNs in previous chapters. Particularly from the analysis performed in Section 4.4, we have found evidence that the studied PSNs are correlated and might complement each other. This result called our attention to the potential for joint use of data from these PSNs, considering each dataset from a PSN as a “sensing layer” (or just layer, for short).

This chapter is dedicated to discuss the concept of sensing layers, and it is organized as follows. Section 6.1 defines the concept of sensing layers and proposes a framework for working with sensing layers. Section 6.2 discusses how to process sensing layers, defining examples of operations that can be applied to them, as well as strategies of processing sensed data using the proposed operations. Section 6.3 presents some proposed applications that illustrate the potential of using sensing layers.

6.1 Sensing Layers

In this section, we first define the concept of sensing layers in Section 6.1.1, and then formalize a model for sensing layers in Section 6.1.3. Section 6.1.2 discusses the usefulness of sensing layers. Section 6.1.4 discusses some issues when dealing with data of distinct layers. Finally, Section 6.1.5 discusses the results of this chapter.

6.1.1 Basic Concepts

A sensing layer represents data, with the corresponding attributes, from a given source of data. The data represented by sensing layers have to come from a source that can be considered a sensor. Examples of data sources are: web services, such as weather condition provided by “The Weather Channel”¹; traditional wireless sensor networks; income census; and participatory sensor networks. In these examples the sensors are: webservice of The Weather Channel; physical sensor in a WSN; census of a city; and user & mobile device in a PSN. In this context, the applications or organizations provide a data stream, with very different throughputs. The census sensing, for instance, may be slow, e.g., data sharing every four years. These examples help to illustrate the ubiquity and diversity of data that may be available. This universe of “ubiquitous data” may be complex to understand and work with, opening opportunities for research studies. Given that, one essential step is a characterization study, since data from different sources can be very heterogeneous and not correlated.

We discuss the concept of sensing layers for participatory sensor networks, most of the time, because this is an emerging source of data with powerful characteristics, such as (near) real time and very large scalability, as shown in Chapter 4. Due to these special characteristics, the use of PSNs as sensing layers simultaneously with other layers, even derived from other sources, may bring new information about city dynamics and urban social behavior, which could enable the design of more sophisticate services (as discussed later). All the concepts discussed in this chapter can be used for other data sources associated to a predefined geographical region, sometimes with required adaptations.

Sensor nodes in a PSN are comprised of users, each one with a portable device. Thus, when we refer to a user, we are referring to a sensor node in a PSN. Note that the sensing activity depends on the willingness of each person to participate in the sensing process. Thus a user, as well as any kind of sensor, may or may not be a contributor during a certain time window.

Figure 6.1 illustrates the idea of sensing layers, showing four different layers for a city: **Traffic alerts** layer provides traffic conditions in certain locations, such as traffic jam or accident (obtained, for example, from Waze or Bing Maps); **Check-ins** layer provides category of a certain place, such as school or pub (obtained for example, from Foursquare); **Weather condition** layer provides climate conditions observed in a certain location, such as windy or rainy (obtained, for example, from Weddar or The Weather Channel); and **Pictures of places** layer provides photos of a certain place, such as a monument (obtained, for example, from Instagram). As illustrated, a plane represents a sensing layer, where the observations of a certain aspect of a predefined geographic region are disposed. Each observation (at each

¹<http://www.weather.com>.

layer) has the following attributes associated with it: time (when the observation occurred), space (geographic location), contributor sensor (e.g., user u) and specific data from a layer (*specialty data*).

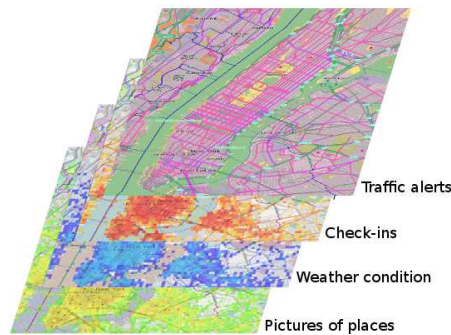


Figure 6.1: Sensing layers for a city. Each layer gives information about a specific aspect of the city.

Figure 6.2 expands the illustration of PSNs presented in the Chapter 3, now embedding the concept of sensing layers. Note that the figure illustrates three types of sensors: a traditional wireless sensor; companies providing data, such as “The Weather Channel”; and users sharing real-time data with their portable devices. This figure depicts three sensing layers, namely *pictures of places* (obtained from Instagram), *traffic alerts* (obtained from Waze) and *check-ins* (obtained from Foursquare). Other layers could be obtained by other types of data source, such as traffic condition provided by Google Maps, census data, or even be derived from one or more layers, as will be exemplified latter.

A sensing layer consists of data describing specific aspects of a geographical location. As shown in Figure 6.2 by a box labeled “big raw data”, these data should be collected (e.g., using an API) and processed, which also includes analysis and data standardization. The last step is the data storage. These steps do not include the extraction of context (or knowledge) from the obtained data, but organize them [Dey and Abowd, 2000]. However, data of sensing layers could be used for context inference, generating new information.

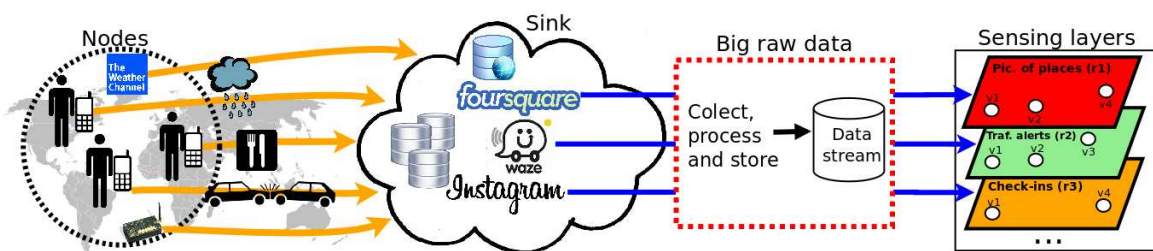


Figure 6.2: Overview of participatory sensor networks with the concept of sensing layers.

To illustrate these processes, consider data from a PSN derived from Foursquare. In Foursquare, users can, among other activities, perform check-ins at locations and leave tips on visited places. From these data, we can define at least two layers, namely: check-ins, containing the check-ins performed by users (check-ins can be used to discover popular places, for example), and tips of places, containing tips, such as “this restaurant has amazing food”, provided by users about certain places. The creation of layers, as shown in Figure 6.2, depends on specific operations for each system. In the case of Foursquare check-ins, a possible way to get them is through Twitter, as explained in Chapter 4. This means that we have to collect, analyze, and process tweets. The coding of tools to perform those steps varies according to the system or application. Next, we must define a structure to represent and store the data of interest associated with a given place where it was shared, thus representing a layer. Each data in a layer has the following attributes:

t: time interval when the data was created;

a: location (e.g., GPS coordinates, neighborhood area) where the data was generated. We represent all locations by an area²;

s: specialty data;

u: one or more IDs of user(s) who generated the data;

Each layer has also a variable *h*, which indicates the status of the layer, where $h = 0|1$, representing the inactive and active states, respectively. The list below represents some examples of layers that are currently available:

1. check-ins (example of source: Foursquare);
2. tips of locations (example of source: Foursquare);
3. traffic alerts (example of source: Waze);
4. pictures of places (example of source: Instagram);
5. average income per area (example of source: census);
6. weather condition (example of source: The Weather Channel);
7. noise level (example of source: Noise Tube³).

²Even if the data is referred by a GPS coordinate it is error prone. For this reason it is interesting to consider an area for this point, for example, a circle with radius *x*, from the GPS coordinate

³<http://noisetube.net>.

6.1.2 Usefulness of Sensing Layers

The processing of a set of sensing layers may enable a large-scale study of each monitored aspect in (near) real time, and provides historical data on patterns observed over long periods. Sensing layers can be applied to several contexts of urban computing, for example, helping to better understand the dynamics of cities and urban behavior in different regions of the world, and respond quickly to unexpected changes.

The use of sensing layers currently in the literature is commonly performed independently, i.e., there is no joint analysis. The individual use of a sensing layer can still be very useful. For instance, using a sensing layer containing traffic information may enable real-time identification of highways with accidents and potholes, whose detection is difficult with traditional sensors, but it becomes more feasible when users participate in the sensing process. Such detection opens opportunities for various services, such as assist smart cars in the correct identification of problems on the road.

Despite the usefulness of using single layers only, services based on just one layer might lack of complementary data. For example, Google Flu Trends⁴, a service based on Google queries, is a type of sensing layer. Very recently, a group of social scientists reported that Google Flu Trends not only wildly overestimated the number of flu cases in the U.S. in the 2012-13 flu season, but has also consistently overshoot in the last few years [Lazer et al., 2014]. According to them, the problem might be because Google Flu Trends is not using complementary information in their service. Indeed, the analysis reported in [Lazer et al., 2014] shows that combining Google Flu Trends with CDC⁵ data, works best. Seems that the way to save this interesting service is using multiple layers, even Matt Mohebbi, co-inventor of Google Flu Trends, agrees with that [Lohr, 2014].

The joint analysis of multiple sensing layers can also be extremely useful in building new applications. For example, we know that a common complaint of inhabitants of large cities is traffic jam. With this in mind, an application that naturally emerges is one that has the goal of inferring the causes of jam, an essential step for addressing the problem. This is not an easy task to accomplish, and the result may vary from place to place. However, the joint analysis of different sensing layers of the city could contribute to build a more robust application. For example, we could cross-check information provided by the following layers: traffic alerts, derived from Waze; check-ins, derived from Foursquare; and pictures of places, derived from Instagram. The first layer provides near real-time data about where traffic jams are occurring. The second one provides data about types of places located in the areas of jams. Having that, it is possible to better understand the areas of interest (for exam-

⁴<http://www.google.org/flutrends>.

⁵Centers for Disease Control and Prevention - <https://data.cdc.gov/>.

ple, identifying a commercial area). Finally, by analyzing the picture of places layer, we can get visual evidence of what is happening in almost real time near the areas of jams. When analyzing data from these three layers together, we can detect, for example, cars blocking intersections, and infer the possible causes of them. Obviously, other layers may also be used, such as the weather condition, layer derived from systems such as Weddar or other traffic condition layer provide, for instance, by Bing Maps⁶.

6.1.3 A Formal Model for Sensing Layers

Let $U = \{u_1, u_2, \dots, u_n\}$ represent a set of sensors (users & mobile device, WSN sensors, etc.), and let $P = \{p_1, p_2, \dots, p_n\}$ represent a set of sensing systems (E.g., WSNs or PSNs). Recall that for simplicity throughout the text the descriptions of concepts are mainly based on PSNs, but the concepts applies for other sensing processes as well. In fact, an application considering also other source of data, besides PSN, is illustrated in Section 6.3.2.

Each user $u_i \in U$ can share unlimited data on any PSN $p_k \in P$. Each j -th data shared $d_j^{p_k}$ into a PSN p_k has the form $d_j^{p_k} = \langle t, m \rangle$, where t refers to a timestamp when user u_i has shared data in p_k , and m is a tuple containing attributes of the shared data. The tuple m is composed of the attributes present in all sensing layers data, in this case $m = (a, u, s)$, where a is the area of the location where the data was shared, s is the specialty data, and u refers to the user $u_i \in U$ who shared the data.

The data shared in $p_k \in P$ can be viewed as a data stream B^{p_k} . We define that a data stream B^{p_k} consists of all n data shared by users U in a PSN p_k in a given time. Thus, $B^{p_k} = \langle d_1^{p_k}, d_2^{p_k}, \dots, d_n^{p_k} \rangle$, and B^{p_k} represents a sensing layer r_{p_k} . Table 6.1 shows examples of data present in sensing layers that have been shared in the three PSNs p_1 , p_2 , and p_3 , illustrated in Figure 6.3, which represents three users sharing data in different PSNs, p_1 (red cloud), p_2 (green cloud) and p_3 (orange cloud) at three different time intervals ($T1$, $T2$ and $T3$). Note that data in the same stream can have the same time⁷, since they may have been shared by multiple users simultaneously.

One way to work with sensing layers is to represent them in the same structure, what we call here *work plan*, containing one or more layers. This work plan represents the resulting plan composed by data combined after applying appropriate algorithms to the corresponding layers we are interested in. How to perform this combination depends on the functionality of the layer(s) that it captures. The abstraction used to represent a combination of data from one or more layers is a data dictionary M , which is a collection of pairs $\{key : value\}$. This structure was chosen because of its simplicity, which helps to ease the

⁶www.bing.com/maps.

⁷This model faces the clock synchronization problem. Therefore, “same time” means close times accepted to be considered equivalent.

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

(a) Foursquare PSN

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

(b) Waze PSN

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

(c) Instagram PSN

Table 6.1: Data stream describing users activity in three different PSNs: Foursquare, Waze, and Instagram.

concepts understanding. Keep in mind that other structures could be used, as long as they respect the principles represented here.

We define that the operation responsible for the work plan creation is called *COMBINATION*($\mathcal{F}, relation()$), where \mathcal{F} is a subset of $\mathcal{B} = \{B^{p_1}, B^{p_2}, \dots, B^{p_n}\}$, or $\mathcal{F} \subseteq \mathcal{B}$, and *relation*() is a function that defines the relationship between data from the streams B^{p_k} contained in \mathcal{F} . The function *relation*() defines the keys of the work plan M , and the data that these keys refer to, which are other observations of the data $d_i^{p_k}$ not used as key. The operation *COMBINATION* results in the work plan M .

To demonstrate the operation *COMBINATION*, we illustrate here two types of relations used to combine data: (1) by location and (2) by users (sensors). To demonstrate a work plan containing combined data by location, consider the activity shown in Figure 6.3. In this case, $\mathcal{F} = \{B^{p_1}, B^{p_2}, B^{p_3}\}$. The work plan M_1 represents this activity, and it is illustrated in Figure 6.4. Observe that the work plan represents data that have been shared across all considered layers. The color of the symbol representing a given data d_i' indicates from which layer it was extracted. The data shared in the same location are grouped and indexed by the key that represents the location. In the work plan M_1 , one key k_i is represented

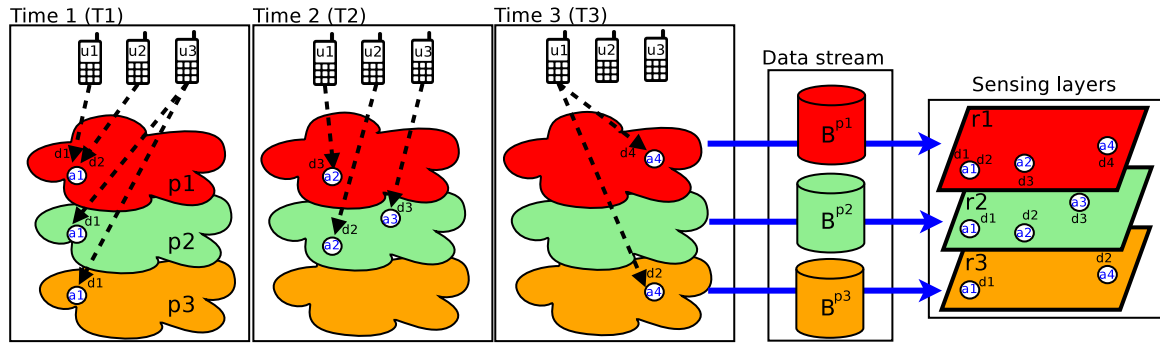


Figure 6.3: Illustration of sharing data in three PSNs throughout the time, resulting in layers.

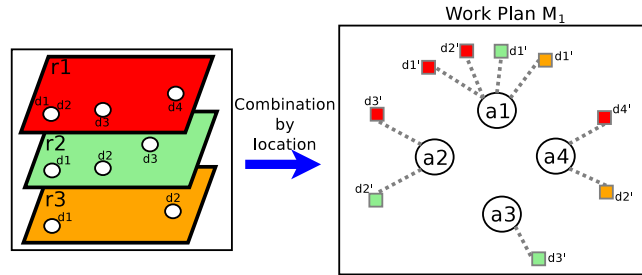


Figure 6.4: Combination by location.

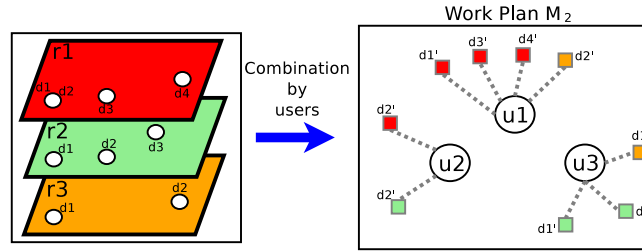


Figure 6.5: Combination by users.

by a_i , which is a unique area among all areas of all data shared in the considered layers: r_1 , r_2 , and r_3 . The $d_i^{r_j'}$ refers to the observations not used as key of the data d_i from the layer r_j , or $\langle t, u, s \rangle$. Thus, each unique areas become a key in work plan M_1 . Work plan M_1 , as described, presents the following structure: $M_1 = \{a_1 : \{d_1^{r1'}, d_2^{r1'}, d_1^{r2'}, d_1^{r3'}\}, a_2 : \{d_3^{r1'}, d_2^{r2'}\}, a_3 : \{d_3^{r2'}\}, a_4 : \{d_4^{r1'}, d_2^{r3'}\}\}$.

Figure 6.5 illustrates the combination by user. In this case, a work plan is build containing keys that represent user ids. The figure shows the work plan M_2 , which was created considering the activities shown in Figure 6.3. The content of the work plan is: $M_2 = \{u_1 : \{d_1^{r1'}, d_3^{r1'}, d_4^{r1'}, d_2^{r3'}\}, u_2 : \{d_2^{r1'}, d_2^{r2'}\}, u_3 : \{d_1^{r2'}, d_1^{r3'}, d_3^{r2'}\}\}$. As we can see, each unique user has become a key in M_2 . This work plan grouped all attributes by the same user in different layers.

6.1.4 Issues of Data from Multiple Layers

There are issues when dealing with data from several layers simultaneously. For instance, in order to perform data combination, such as by location or user, as exemplified, we have to make sure that the data is consistent in all layers. This is a mandatory condition for correct functioning.

Consider that we want to combine two layers A and B by locations. The format of data location in Layer A is expressed as latitude and longitude, and as street address in Layer B. One way to solve this inconsistency is performing a geocoding process, using, for example, the Yahoo! geocoding tool⁸. In this way the street address will be transformed in a geographic coordinate (latitude and longitude).

Another issue that might happen when combining data by location is regarding to areas that overlap each other. How to define a key in this case? One possibility is consider several keys, one for the intersection between those areas, and one or two⁹ as the area(s) not overlapped. Another option is to define just one key, this might be interesting when one area is inside another, so the key becomes the bigger area.

The combination by users is specially an issue when our sensor is an user, as in PSNs, because the same user may participate in different layers. Let's suppose we want to combine data by users using the check-ins layer (obtained from Foursquare) and the picture of places layer (obtained from Instagram). Since we are dealing with independent systems, users (sensors) have different identification. One way to try to bypass this issue is verifying other networks in order to match the user ID of one layer in another. For example, users of Foursquare and Instagram tend to be also users of Twitter [Duggan and Smith, 2014]. In this way, the key in the combination process could be the twitter ID.

Note that if the combination by user desired is between a PSN layer and other layers that doesn't have users as sensors, such as WSNs, the inconsistency does not exist, because every sensor has its unique ID. Although it is necessary to evaluate if a combination by users (sensors) between those layers lead to the desired information.

Another issue is that different layers might refer to data valid for different interval of times. This is natural because some data sources provides near real time data, others not. For example, an alert in a Waze PSN refers to a traffic situation that may not exist five minutes later. However, a census data usually is valid for a big interval of time, months or years, until the next census is released. We have to be aware of all those issues when designing new applications and define a way to treat them.

There might be other issues. For example, issues related to the volume of data. If

⁸<https://developer.yahoo.com/boss/geo>.

⁹If one area is not completely inside another.

we do not have significant data for a certain layer, its utilization may not lead to the correct information extraction. Different data sources may present different characteristics for this issue. For instance, in a PSN many factors influence the volume of data, for example, geographical, cultural and economic aspects. The granularity of areas may also influence other data sources. If we consider, for example, data from WSN as a layer we may not have data for an entire metropolis, because of size restrictions of those networks.

In summary, note the importance of a characterization process. We have to know the properties of the layers we want to use, in order to verify if their simultaneous use may lead to the intended information extraction. The *relation()* informed to the *COMBINATION* encapsulate the solution chosen for dealing with heterogeneous data, which is application dependent. If there is no solution to eliminate the inconsistency between data from two layers, then they cannot be used together.

6.1.5 Discussion

The use of layered (multi-layer) models to extract new information or design new applications is not new. Very recent studies focused on a particular type of multi-layer network, the multiplex, where each agent can be networked in different ways, and with different intensity, on several multiple layers simultaneously. This model is useful, for example, to study links that the same user has in different social networks (layers), for instance, to better understand the information spreading. Another example is the study of transportation in a city. The network of bus routes and stops (layer 1) is different from a subway network (layer 2) in the same city, but a user can use both networks to reach its destination [Domenico et al., 2013].

In the same direction, Xin et al. [2005] proposed a layered graph model to develop routing and interface assignment algorithms. Laura et al. [2002] proposed a layered model for the Web network, aiming to design a model that resembles better the complex nature of the Web. A GIS (geographic information system) is another example, because it often utilizes a layered model for characterizing and describing our world. It uses maps to visualize and work with geographic information in several layers [Chang, 2010]. GIS is related to the ideas proposed here, in fact, some GIS tools could be used to support the proposed framework, for example, in the combination process. Our proposal differ from a simple implementation of a GIS because it is not driven by jurisdictional (such as a city), purpose, or application requirements. We focus on the discussion of a sensing layer framework. Besides that we envision demonstrate the potential of simultaneous use of sensing layers derived from PSNs, for the extraction of new information related to the study of city dynamics and urban social behavior.

More close relate to our proposal, there are studies that consider different sources of

data simultaneously to better understand the dynamics of cities. For example, Bollen et al. [2011] investigated whether collective mood states derived from Twitter feeds are correlated to the value of the Dow Jones Industrial Average over time. Sagl et al. [2012] analyzed the collective human behavior based on mobile data, and correlated it with meteorological data from weather stations.

In sum, this work differs from all previous studies because: (i) define the concept of sensing layers; (ii) propose a framework that enables integration of the analysis and exploration of multiple layers simultaneously; and (iii) present applications that use the proposed framework and illustrate the potential of using multiple sensing layers.

6.2 Processing Sensing Layers

This section discusses how to process one or more sensing layers. To that end a number of example operations are proposed. Section 6.2.1 presents examples of such operations and Section 6.2.2 presents some strategies to process layers using the proposed operations.

6.2.1 Operations

In Section 6.1.3, we illustrate how to represent sensing layers in a work plan, for example, by location (M_1) or users (M_2). The general purpose of work plans is to be basic structures that can be easily manipulated. Recall that the structure chosen here to represent a work plan is a data dictionary. Having a work plan, as the M_1 or M_2 shown in Figures 6.4 and 6.5, we can apply operations to derive other structures and also extract new information. The list below provides examples of some generic operations:

- **dGRAPH** (directed graph): This operation is represented by the algorithm 2. It expects as input a work plan M , and the result is a directed graph $G = (V, E)$. This operation builds a directed graph $G = (V, E)$, where each key k_i in the work plan represents a node $v_i \in V$, and the data indexed by k_i are attributes of v_i . An edge $e = (v_i, v_j)$ is added depending on the desired analysis, which is expressed through some specific operations, as we describe below. Initially, $E = \emptyset$. All variables of the work plan are incorporated in the graph;
- **CNG** (change): This operation is represented by the Algorithm 3. It expects a work plan M , a layer identification (ID), and a status (0 or 1). It results in the alteration of the variable h of the informed layer, i.e., it changes the status of a layer through the variable h . If the informed status is 0, then $h = 0$ and the work plan are adjusted

Algorithm 2: Operation dGRAPH.

```

input : work plan  $M$ 
output: directed graph  $G$ 
1  $G \leftarrow \emptyset$  //  $G$  is directed graph
2 foreach  $key \in M$  do
3    $G.insertNode(key)$ ;
4    $G.insertAttributes(key.attributes)$ ; // insert attributes of the entry  $key$  in the
   last added node
5 end

```

accordingly, i.e., this particular layer of the work plan is disabled. The layer disabled can be enabled again with the same data it had previously at the disabling time;

Algorithm 3: Operation CNG.

```

input : work plan  $M$ , a layer  $ID$ , and an integer  $i$ 
output:  $M'$  with the modifications imposed by the change of  $h$ 
1  $hiddenLayers$ ; // structure to keep hidden layers
2 if  $h = 0$  then
3    $hiddenLayers.insert(ID)$ ;
4   foreach  $key \in M$  do
5     foreach  $data \in key$  do
6       if  $data$  is from layer  $ID$  then
7          $hiddenLayers.ID.insert(key \leftarrow data)$ ;
8          $M' \leftarrow M.remove(key \leftarrow data)$ ;
9       end
10    end
11    if  $size(key) = 0$  then
12       $M' \leftarrow M.remove(key)$ ;
13    end
14  end
15 else
16    $dataLayer \leftarrow hiddenLayers.ID$ ;
17   foreach  $key \in dataLayer$  do
18     foreach  $data \in key$  do
19        $M.insert(key \leftarrow data)$ ; //  $key$  entry is created if it doesn't exist
20     end
21   end
22 end

```

- **RESET**: This operation is represented by the Algorithm 4. It expects a directed or undirected G graph, and results in a work plan M . It is extracted all the necessary information from the graph to build a corresponding work plan. All variables of the graph are incorporated in the work plan;
- **dEDGE** (directed edges): This operation is represented by the Algorithm 5. It expects a directed graph G resulted from a work plan combined by locations, and results in a graph G' containing directed edges. This operation creates a directed edge from node v_i to node v_j if and only if at least one user shared data, in any layer, in the location represented by the node v_j right after sharing data, also in any layer, in a location represented by the node v_i . The weight of an edge represents the total number

Algorithm 4: Operation RESET.

```

input : graph  $G$  representing sensing layer(s)
output: work plan  $M$  representing the  $G$ 

1  $M \leftarrow \emptyset$ ;
2 foreach  $node \in G$  do
3    $allData \leftarrow \emptyset$ ;
4   foreach  $data \in node$  do
5      $allData.insert(data)$ ;
6   end
7    $key \leftarrow node$ ; // the node ID is the key of the work plan
8    $M.insert(key \leftarrow allData)$ ;
9 end

```

Algorithm 5: Operation dEDGE.

```

input : directed graph  $G$  resulted from a work plan combined by locations
output: graph  $G'$  containing directed edges

1  $allUsers \leftarrow \emptyset$ ;
2 foreach  $node \in G$  do
3   foreach  $data \in node$  do
4      $location \leftarrow nodeID$ ;
5     if  $data.user \notin allUsers$  then
6        $allUsers.insert(data.user)$ ;
7        $allUsers[data.user].insert([data.time, location])$ ;
8     else
9        $allUsers[data.user].insert([data.time, location])$ ;
10    end
11  end
12 end
13 sort the data of each key of  $allUsers$  by chronological order;
14 foreach  $User \in allUsers$  do
15   foreach  $userData \in User$  do
16      $loc \leftarrow userData[2]$ ;
17     if  $userData$  is not last data of  $User$  then
18        $nextLoc \leftarrow$  location of next  $userData$  of  $User$ ;
19        $G'.insertEdge(loc, nextLoc)$  // directed edge  $loc \rightarrow nextLoc$ 
20     end
21   end
22 end

```

of transitions performed from v_i to v_j considering transitions of all users. Note that it is possible to have more than one transition for the same user;

- **DEL** (delete): This operation is represented by the Algorithm 6. It expects a graph G and an integer t . It results in a subset graph G_{subset} derived from G . This operation deletes edges $e_i \in E$ (E is a set of edges of G), with weight $w_i < t$;
- **rdGRAPH** (random directed graph): This operation is represented by the Algorithm 7. It expects a directed graph $G(V, E)$, and results in a random directed graph $G_R(V, E_R)$. The random graph G_R is constructed keeping the same nodes of G and uses the same number of individual transitions of G . However, instead of considering the real transition $v_i \rightarrow v_j$ performed by an individual, the operation randomly chooses two nodes to replace v_i and v_j , simulating random transitions performed by users;

Algorithm 6: Operation DEL.

```

input : graph  $G$  representing sensing layer(s) and an integer  $t$ 
output: subset graph  $G_{subset}$  derived from  $G$ 

1 foreach  $edge \in G$  do
2   | if  $edge.weight < t$  then
3   |   |  $G_{subset} \leftarrow G.remove(edge);$ 
4   |   end
5 end

```

Algorithm 7: Operation rdGRAPH.

```

input : directed graph  $G(V, E)$  representing sensing layer(s)
output: random directed graph  $G_R(V, E_R)$ 

1  $numEdges \leftarrow numberEdges(G);$ 
2  $G_R \leftarrow \emptyset$  //  $G_R$  is directed graph
3 foreach  $node \in G$  do
4   |  $G_R.insertNode(node);$ 
5 end
6 for  $i \leftarrow 1$  to  $numNodes$  do
7   |  $rndNode1 \leftarrow randomNode(G);$  //  $randomNode()$  retrieves a random node
8   |  $rndNode2 \leftarrow randomNode(G);$ 
9   |  $G_R.inserEdge(rndNode1, rndNode2);$ 
10 end

```

- **MERGE**: This operation is represented by the Algorithm 8. It expects a work plan M_1 , a work plan M_2 , and a data relation $relation()$. M_1 and M_2 have to be produced following the same data relation, for example, by locations as explained above in the process **COMBINATION**. This operation results in a work plan $M_{merged}(V, E)$ representing the merge of the sensing layers represented by M_1 and M_2 . This operation merge information of M_1 and M_2 , respecting the data relation informed $relation()$.

Algorithm 8: Operation MERGE.

```

input : a work plan  $M_1$ , a work plan  $M_2$ ,  $relation()$ 
output: a graph  $M_{merged}$ 

1 foreach  $key1 \in M_1$  do
2   | foreach  $key2 \in M_2$  do
3   |   |  $allDataKey1 \leftarrow key1.retrieveAllData();$ 
4   |   |  $allDataKey2 \leftarrow key2.retrieveAllData();$ 
5   |   | if  $by\ relation() key1\ and\ key2\ should\ be\ merged$  then
6   |   |   |  $mergedKey \leftarrow key1\ and\ key2\ merged;$ 
7   |   |   |  $allData \leftarrow allDataKey1 \cup allDataKey2;$ 
8   |   |   |  $M_{merged}.insert(mergedKey \leftarrow allData);$ 
9   |   |   else
10  |   |   |  $M_{merged}.insert(key1 \leftarrow allDataKey1);$ 
11  |   |   |  $M_{merged}.insert(key2 \leftarrow allDataKey2);$ 
12  |   |   end
13  |   end
14 end

```

We can have also specific operations to produce new information (which could be represented in a new layer), using one or more existing layers, such as the following operations:

Algorithm 9: Operation fPOIS.

```

input : Work plan  $M$  representing a sensing layer
output: Work plan  $M_{POIs}$  containing points of interest
// identificaPOIs represents the algorithm 1 (Section 5.3)
1  $M_{POIs} \leftarrow \text{identificaPOIs}(M)$ ;
```

- **fPOIS** (find POIs): This operation is represented by the Algorithm 9. It expects a work plan M representing a layer such as *check-ins* and *pictures of places* combined by locations. Other layers might also be used, but previous verification of feasibility is needed, for example, data might not be available for the geographical region of interest. This operation results in a work plan of a new layer containing popular areas, or points of interest (POI), based on the number of activities performed on them. This operation identifies POIs applying the algorithm 1, specified in Section 5.3, to select geographic areas;
- **fSIGHTS** (find sights): This operation is represented by the Algorithm 10. It expects a work plan M^{POIs} containing POIs, and results is a graph G^{SIGHTS} containing sights. This operation identifies sights from a work plan M^{POIs} , where keys are the areas a of POIs identified in a particular pre-defined geographic region. This algorithm is described in Section 5.3. More details of this operation are presented in Section 6.3.1.

Algorithm 10: Operation fSIGHTS.

```

input : work plan  $M_{POIs}$  containing points of interest
output: graph  $G_{SIGHTs}$  containing sights
1  $G_{POIs} \leftarrow \text{dGRAPH}(M_{POIs})$ ;
2  $G_{POIs-FLow} \leftarrow \text{dEDGE}(G_{POIs})$ ;
//  $t$  is identified in the way described in Section 5.3
3  $G_{SIGHTs} \leftarrow \text{DEL}(G_{POIs-FLow}, t)$ ;
```

We chose specifically those operations because they are used in the applications presented in the next sections. Note that other operations can be proposed. For instance, another operation to create edges differently from *dEDGE*. This new operation, called for example *uEDGE*, could be suitable for a graph G produced from a work plan combined by users. The operation *uEDGE* could create an undirected edge between v_i and v_j , if and only if user u_i , represented by node v_i , shared data in the same location (layer independent) that user u_j , represented by node v_j . The weight of an edge represents the total number of locations that nodes v_i and v_j have in common. Other operations could be designed to add (directed or undirected) edges with different way to assign weights.

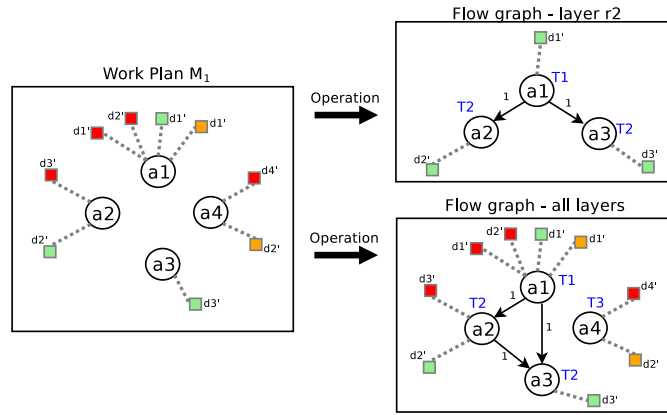


Figure 6.6: Illustration of flow graph creation from one single layer, and also from multiple layers.

6.2.2 Processing Strategies

As shown in the previous section, our framework provides several operations useful to process sensing layers in several manners. To give an example of the results we can obtain in processing sensing layers using those operations, we demonstrate how to obtain: flow graphs, graphs that map the locations where the same user shared data, thus capturing the movements or transitions in a geographical area; and also points of interest and sights. It is particularly interesting to illustrate the creation of flow graphs because it is a fundamental piece of some operations, for instance *fSIGHTS*.

Algorithm 11: Generation of flow graph for one single layer.

input : work plan M combined by locations
output: flow graph $G_{r_2}^{flow}$ that represents data from the layer r_2

- 1 $M \leftarrow M_1$; // M_1 is the work plan created previously
- 2 $M' \leftarrow CNG(M, r_1, 0)$;
- 3 $M'' \leftarrow CNG(M', r_3, 0)$;
- 4 $G \leftarrow dGRAPH(M'')$;
- 5 $G_{r_2}^{flow} \leftarrow dEDGE(G)$;

Consider the data sharing of the situation illustrated in Figure 6.3. After a certain time, we can process the data in order to extract knowledge in different ways. Take for instance the flow graph labeled “flow graph - layer r_2 ”, shown in Figure 6.6. The Algorithm 11 describe the steps necessary to generate this graph, referred to as $G_{r_2}^{flow}$ (built from layer r_2). In this algorithm we consider the work plan M_1 as explained above (combined by locations). We initially apply the operation *CNG* hiding layers r_1 and r_3 . After that, we have to generate a directed graph G using *dGRAPH* and apply the operation *dEDGE* in G , obtaining $G_{r_2}^{flow}$. In this case, we have a flow graph that represents data from a single layer. With this

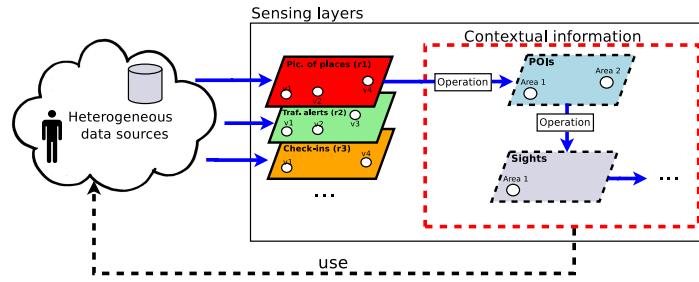


Figure 6.7: Illustration of new layers creation from the picture of places layer.

graph we can extract many valuable information, for example, regular trajectories in a city, as shown in Section 5.1.

Another possible analysis is to consider different layers simultaneously. In Figure 6.6, the part named the “flow graph – all layers” shows a graph, which we call G_{all}^{flow} . The Algorithm 12 describes the steps necessary to generate G_{all}^{flow} . This algorithm also consider the work plan M_1 created above. As we can see in the algorithm, in order to obtain G_{all}^{flow} we need to apply the operation $dEDGE$ in G . In the resulting graph, the nodes represent data shared in the same location at any layer. Edges connect nodes $v_i \rightarrow v_j$ if at least one user shared data in the location represented by node v_j , right after sharing a data in the location represented by the node v_i .

Algorithm 12: Generation of flow graph for multiple layers.

input : work plan M combined by locations

output: flow graph G_{all}^{flow} that represents multiple layers

- 1 $M \leftarrow M_1$; // M_1 is the work plan created previously
 - 2 $G \leftarrow dGRAPH(M)$;
 - 3 $G_{all}^{flow} = dEDGE(G)$;
-

New information could be obtained by processing data available from one or more sensing layers. Points of interest (POI) in a city, identified from data shared in Instagram and obtained using operation $fPOIS$, represent an example. To identify a sight it is necessary the POIs, according to the operation $fSIGHTS$. This is demonstrated in Figure 6.7. In this figure, the new information obtained is expressed as new layers. Note that these new layers are represented in the box labeled “Contextual information”, which had its meaning explained in Chapter 3. Basically, new information generated from other sensing layers are contextual information. Recall that contextual information might have the power to influence the data generation. For example, once users know where the points of interest are they may tend to share more data in those places instead of others. For more details in this direction see the discussion of Figure 3.2 in Section 3.1.

6.3 Applications Using the Sensing Layers Framework

In this section, we discuss two applications that illustrate the potential of the proposed framework for working with sensing layers. Those application are presented on Sections 6.3.1 and 6.3.2. Section 6.3.3 presents some final considerations about this section.

6.3.1 Identification of Sights

First, we discuss an application that identifies sights considering multiple layers simultaneously, highlighting the improvements on the strategy presented in Section 5.3, which considers only one layer. In this analysis, we consider the Instagram-New and Foursquare-New datasets, described in Section 4.4.

The picture of places layer (r_1) is represented by the dataset Instagram-New, and the layer check-ins (r_2) by Foursquare-New dataset. Our goal is to obtain results using both layers. To that end, we first combine the data by location, producing a work plan M_1 . First we want to identify sights for the layer r_1 . With that in mind, we disable layer r_2 from M_1 using the operation *CNG* obtaining M_{r_1} . After that, we apply *fPOIS* in M_{r_1} to generate $M_{r_1}^{pois}$, work plan containing the POIs. In the resulting work plan $M_{r_1}^{pois}$, the keys are the areas of the identified POIs. In the scenario illustrated in Figure 6.7, we have only two keys for a work plan representing POIs, represented by Area 1 and Area 2.

Algorithm 13: Identification of sights using sensing layers.

```

1  $M_1 \leftarrow r_1$  and  $r_2$  combined by locations;
   // identifying sights for layer  $r_1$ 
2  $M_{r_1} \leftarrow \text{CNG}(M_1, r_2, 0)$ ;
3  $M_{r_1}^{pois} \leftarrow \text{fPOIS}(M_{r_1})$ ;
4  $G_{r_1}^{sights} \leftarrow \text{fSIGHTS}(M_{r_1}^{pois})$ ;
   // identifying sights for layer  $r_2$ 
5  $M_1 \leftarrow \text{CNG}(M_1, r_2, 1)$ ;
6  $M_{r_2} \leftarrow \text{CNG}(M_1, r_1, 0)$ ;
7  $M_{r_2}^{pois} \leftarrow \text{fPOIS}(M_{r_2})$ ;
8  $G_{r_2}^{sights} \leftarrow \text{fSIGHTS}(M_{r_2}^{pois})$ ;
   // All sights of r1 and r2 layers
9  $M_{r_1}^{sights} \leftarrow \text{RESET}(G_{r_1}^{sights})$ ;
10  $M_{r_2}^{sights} \leftarrow \text{RESET}(G_{r_2}^{sights})$ ;
11  $M_{total}^{sights} \leftarrow \text{MERGE}(M_{r_1}^{sights}, M_{r_2}^{sights}, \text{relation}());$  // relation() by location

```

Each POI in $M_{r_1}^{pois}$ represents a popular area a in a given geographical region, e.g., a city. Popularity is identified through the volume of shared data made available by users u . That is, a POI represents the activity performed by a group of users u in a time interval t . Note that, the specialty data s , in this particular case, is the POI area itself. We use the work

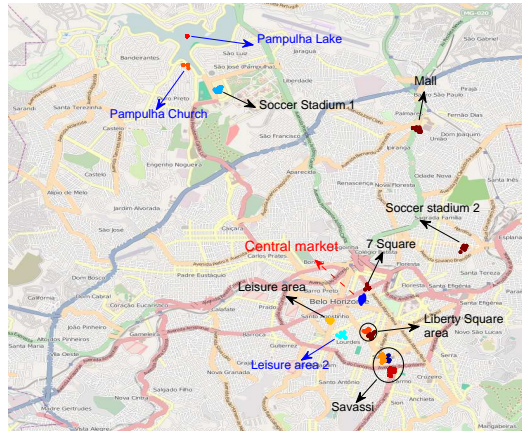


Figure 6.8: All identified sights with Foursquare and Instagram datasets.

plan $M_{r_1}^{pois}$ for the extraction of sights with the help of the operation $fSIGHTS$, which is represented by the Algorithm 10. First, the operation $fSIGHTS$ creates a directed graph (in the example $G_{r_1}^{pois}$), from the received work plan (in the example $M_{r_1}^{pois}$), using operation $dGRAPH$. Next it maps the flow of users performed between POIs. For this, it applies the operation $dEDGE$ in $G_{r_1}^{pois}$ obtaining $G_{r_1}^{pois-flow}$. After that, popular transitions that connect two nodes $v_i \rightarrow v_j$ are selected. For that, it uses the operation DEL in the graph $G_{r_1}^{pois-flow}$ using a parameter t . The parameter t in this case is calculated in the way presented in Section 5.3. According to the conjecture considered in the algorithm of the operation $fSIGHTS$, the popular transitions selected connects the sights, which are represented in the graph $G_{r_1}^{sights} = (V', E')$. In this graph, nodes $v_i \in V'$ are the areas a of the identified sights.

Next, we identify the sights $G_{r_2}^{sights}$ for r_2 . First, we enable r_2 and disable layer r_1 from M_1 . The next steps are performed similarly to the way it was presented for r_1 . After that we merge the contextual layers containing the sights for layers r_1 and r_2 , $M_{r_1}^{sights}$ and $M_{r_2}^{sights}$, respectively, in the work plan M_{total}^{sights} . This work plan contains all identified sights, which are shown by the Figure 6.8.

The sight indicated by a red arrow (Central Market) was identified only by Foursquare. Sights pointed by a blue arrow (Pampulha Church, Pampulha Lake, and leisure Area 2) were not identified with Foursquare. All sights are very relevant. It is important to observe the potential for complementary results using both layers.

6.3.2 Economic-Cultural Analysis of Regions

The application described in this section allows various economic-cultural analyses. In this document, we focus on two. The objective of the first analysis is to correlate the general

sentiment expressed in the tips for all locations in a given census tract a_i (geographic region defined for the purpose of taking a census), with the median income of the inhabitants of this tract. On the other hand, the aim of the second analysis is to study the movement of users in the considered tracts, taking into account the typical income of these tracts. This second analysis aims to identify possible social segregation in a city.

To illustrate this application, we consider two datasets derived from Foursquare and one derived from the census of NY. The first, named CHECKINS-NY, consists of 34,677 check-ins performed in New York City, in a week of April 2012. CHECKINS-NY is a subset of the dataset Foursquare-Crawled. The second dataset, named TIPS-NY, contains all the tips contributed by users up to January 2013 in all unique locations of the dataset CHECKINS-NY. The tips were collected through the Foursquare API. Each tip contains a location, a user ID, a time, and the textual content of the tip. We consider only tips in English. We define that a tip is in the English language if at least half of the words of the tip is listed in a dictionary containing key words in English. This resulted in 157,197 tips (2,531 discarded). The last dataset, named CENSUS-NY, contains information of the census of New York City, and it refer to the 2006-2010 American Community Survey. Figure 6.9 represents some examples of tracts considered in the CENSUS-NY. The area of each tract is pre-defined by the census of New York. It contains, among other information, the median income per tract (information we are interested here).

The TIPS-NY dataset is used to represent a sensing layer called tips of locations (r_1). The layer r_1 is composed of a data stream B_i . Each data stream has the form: $\langle t, (a, u, s) \rangle$. An example of the specialty data s of this layer is: “This place is awesome, I recommend the burger.”. The income layer (r_2), derived from CENSUS-NY, is composed of a dataset d_j for different tracts of New York. Each specialty data in d_j has the median income of the inhabitants of a particular tract. The form of d_1 is $t = 2006-2010$, $a = [\text{area of Tract 1}]$, $u = \text{“USA Census”}$, $s = \text{“median income in US\$ for the Tract 1”}$. Note that, this is an example of layer obtained from a different source than PSNs. This illustrates the use of other sources of data about predefined geographical regions.

For the first analysis we combine the data from the layers r_1 and r_2 . The chosen method is the combination by location, method described in Section 6.1.3. This combination process consider the keys as the areas of the tracts. Each key k_i combines, among other data, the tips of all the places that are located within the area of a tract, and the median income information of the tract. The combination process results in a work plan M_1 .

Thus, we use M_1 to calculate the general sentiment about all locations in each tract. For this analysis, we used the program SentiStrength [Thelwall et al., 2010]¹⁰, to classify the

¹⁰We used the tool IFeel [Gonçalves et al., 2013] to help in the selection of this sentiment analysis program.

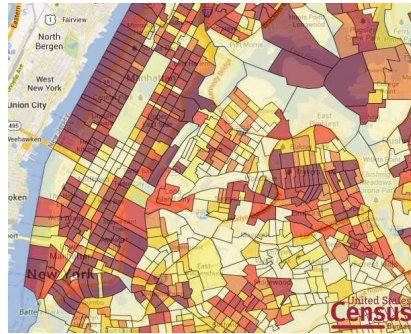


Figure 6.9: Examples of New York City census tracts.

sentiment expressed in the tips. SentiStrength computes the sentiment of a tip in a scale from -4 (strongly negative) to 4 (strongly positive), 0 indicates a neutral sentiment. This program is applied to each tip and then combined by location, and finally by tract.

Then we calculate the average sentiment for all locations in a given tract. Next, we group the tracts in five income groups: less than US\$25,000; between US\$25,000 and US\$50,000; between US\$50,000 and US\$75,000; between US\$75,000 and US\$100,000; and over US\$100,000. Finally, we calculate the average value of sentiment for each of the five income groups, considering all tracts that belong to each group.

Table 6.2 presents this result, and also for each income group, the percentage of average sentiment that falls in one of five range of sentiment: $(+3, +4)$, $(+1, +2)$, (0) , $(-1, -2)$, and $(-3, -4)$. As we can observe, the result suggests that poor tracts tend to have the worst sentiment expressed by users. This may be associated with low quality services in these tracts. With the tract income increasing, opinions tend to be more positive. Although the average sentiment for the richest tracts group (over US\$100,000) is slightly lower than the second richest (between US\$75,000 and US\$100,000), this group still has a larger number of positive tips compared to all other groups, and does not have negative tips. Note the potential of this analysis for social studies, e.g., for the study of inequalities in the quality of services in cities.

| Group | Mean Sent. (std) | $(+3,+4)\%$ | $(+1,+2)\%$ | $(0)\%$ | $(-1,-2)\%$ | $(-3,-4)\%$ |
|----------------------------|------------------|-------------|-------------|---------|-------------|-------------|
| <25000 | 0,46 (0,67) | 0 | 73,08 | 21,15 | 5,77 | 0 |
| ≥ 25000 and <50000 | 0,73 (0,63) | 1,23 | 84,31 | 12,92 | 1,23 | 0,31 |
| ≥ 50000 and <75000 | 0,81 (0,46) | 0,40 | 93,28 | 5,93 | 0,39 | 0 |
| ≥ 75000 and <100000 | 0,9 (0,36) | 0 | 96,97 | 3,03 | 0 | 0 |
| ≥ 100000 | 0,87 (0,28) | 0,96 | 98,08 | 0,96 | 0 | 0 |

Table 6.2: General sentiment per groups of tracts

For the second analysis, which has the steps represented in the Algorithm 14, the

dataset CHECKINS-NY is used to represent a sensing layer called check-ins (r_3). We combine layers r_2 (as defined above) and r_3 by location on the work plan M_2 . We then create a graph G_2 , and use it to generate a flow graph G_2^{flow} , where the edges are the transitions performed by the same user in different tracts (nodes in the graph). We exclude loops, i.e., visits from the same user on the same tract, generating then $G_2^{flow'}$. To gather evidence of the existence of segregation, we study the assortativity related to the median income by tract in $G_2^{flow'}$. This is a way to try to observe the existence of segregation.

Algorithm 14: Analysis 2.

```

1  $M_2 \leftarrow r_2$  and  $r_3$  combined by location;
2  $G_2 \leftarrow (M_2)$ ;
3  $G_2^{flow} \leftarrow dEDGE(G_2)$ ;
4  $G_2^{flow'} \leftarrow removeLoops(G_2^{flow})$ ;
5 foreach  $node \in G_2^{flow'}$  do
6   if  $node.income \leq US\$75,000$  then
7      $node.insert(class \leftarrow "A")$ ;
8   else
9      $node.insert(class \leftarrow "B")$ ;
10  end
11 end
12  $assortativity \leftarrow calcAssortativity(G_2^{flow'})$ ; // assortativity by attribute "class"
13  $assortRndGraph \leftarrow \emptyset$ ;
14 for  $i \leftarrow 1$  to 10 do
15    $G_{Ri} \leftarrow rdGRAPH(G_2^{flow'})$ ;
16   foreach  $node \in G_{Ri}$  do
17      $node.insert(choose\ random\ class\ (A\ or\ B))$ ;
18     // The number of A and B nodes respects the numbers observed in
19     //  $G_2^{flow'}$ 
18   end
19    $assortTemp \leftarrow calcAssortativity(G_{Ri})$ ;
20    $assortRndGraph.insert(assortTemp)$ ;
21 end
22 calculate i.c. of 95% for the average value of  $assortRndGraph$ ;

```

The assortativity measures the similarity of connections in the network relative to a particular attribute, and ranges from -1 to $+1$ [Newman, 2002]. In an *assortative network* (with positive assortativity), vertices with similar values for a given attribute (e.g., the same income) tend to be connected (be similar) to each other, whereas in a *disassortative network* (negative assortativity), the opposite happens. All tracts were associated with a class based on the median income of the tract: Class A for median incomes up to US\$75,000; and Class B for higher median incomes. The assortativity considering these two classes as attributes of $G_2^{flow'}$ is 0.14. Thus, the network for this attribute is assortative, indicating a trace of segregation, i.e., users tend to share content (or attend) in tracts that have the same class of income.

After that, we create ten random graphs $G_{Ri}(V, E_{Ri})$, where $i = 1, \dots, 10$, using the operation *rdGRAPH*. For each graph G_{Ri} is also randomly associated a class of a node,

A or B. The number of nodes of class A and B are also consistent with the one observed in G_2^{flow} . After that, we calculate the assortativity for all random graphs $G_{R1..10}$. The assortativity for all graphs, with 95% confidence level, are in the range is: $[-0.0084, -0.0014]$. As we can see, these random networks do not indicate segregation. Obviously, in order to draw any conclusion in this sense, a more detailed investigation is needed. However, this result shows the potential for joint analysis of multiple layers.

Note also the potential of considering the same layers to generate a work plan M_3 combined by users. Besides identifying users' preferences, we can also try to infer their social class studying the income of the tracts that the user visits. This can be useful for social studies, and for more effective advertising.

6.3.3 Discussion

In this section we demonstrated two applications that illustrate the potential of sensing layers. The first one consider two layers derived from PSNs. The other one considers also a layer not derived from PSN, to demonstrate the usefulness and flexibility of our proposal.

Obviously many other applications could be proposed. For example, in any city is likely to find many places where people perform more often a particular activity, for example an area of bars and restaurants where people meet to socialize. These locations could be identified with the help of the check-ins layer. The information provided by other layers could help users choose the best areas of interest at the moment. For example, a user could use the information provided by the traffic alerts layer to identify among all the options, the area with the lowest number of traffic problems at the time, and use the picture of places layer to view the style of the establishments in those areas and the people who frequent them.

Chapter 7

Conclusions and Future Work

This chapter summarizes this thesis and discusses directions for future research. Section 7.1 presents the conclusion of this thesis. Section 7.2 presents the future work. Finally, Section 7.3 lists and comments all the publications performed during the doctoral period.

7.1 Conclusions

Applications are becoming increasingly mobile, designed to infer user interests and location, and make different sorts of predictions. A mobile device is not just a better option, but may be the only option for many people. This is similar to another phenomenon that has happened for some years now: more and more people are ditching their landlines in favor of cellphones. When people get to that point, they tend to acquire not any cellphone but preferably the latest generation smartphone.

This document presented the dissertation entitled Large Scale Study of City Dynamics and Urban Social Behavior Using Participatory Sensor Networks. In this work we show that the use of participatory sensor networks can help us better understand the dynamics of cities and urban social behavior, and from this we are able to offer smarter services to meet people's needs.

Using several large scale datasets, we characterized and analyzed the main properties of PSN three different types of PSNs: location sharing services (namely Foursquare, Gowalla and Brightkite); photo sharing services (particularly Instagram); and traffic alert services (particularly Waze). We identified several properties of PSN, for instance, the planetary scale of those networks, as well as the highly unequal frequency of data sharing, both spatially and temporally, which is highly correlated with the typical routine of people. We also performed a comparison of different PSNs derived from Instagram and Foursquare aiming to understand whether data from one system could complement the other, or if they are

compatible regarding the study of city dynamics and urban social behavior. This analysis gave us insights about the potential for joint use of data from these applications, considering each PSN as a sensing layer. In general, our analysis pointed out several challenges of this emerging type of network, which may restrict its use, but also showed that there are good opportunities. In particular, we demonstrate a range of fruitful opportunities that emerge when using PSNs to the large scale study of city dynamics and urban social behavior.

In this direction, we presented a visualization technique called City Image, and illustrated its use in different cities around the world. This technique summarizes the city dynamics based on transition graphs that map the movements of individuals between different location categories in the PSN. We also showed the use of this technique for clustering cities based on their similarities in terms of movement patterns, which can be exploited to build city recommendation systems. Finally, we investigated the use of centrality metrics, computed on transition networks built at the granularity of specific venues, as a means to complement the City Image technique towards a deeper understanding of the city dynamics.

Next, we propose a technique for point of interest identification. The technique considers that each pair of coordinates (longitude, latitude) is associated with a point that represents a shared data, for instance a photo. We start by computing the geographic distance between each pair of points, and grouping together the points that are close to each other, e.g., those that have a distance smaller than a certain distance (dependent threshold). To capture the POIs, we exclude groups that may have been generated by random situations (i.e., random people movements), and thus do not reflect the dynamics of the city. To identify those groups, we analyze the number of data sharing in each group and adopt simple statistical methods. The technique is also able to extract sights out of the identified POIs, using the transition of people between POIs for that.

We also propose a new methodology for identifying cultural boundaries and similarities across populations using self-reported cultural preferences recorded in PSNs, such as Foursquare, which is the system we use to demonstrate the methodology. Besides being globally scalable, our methodology also allows the identification of cultural dynamics more quickly than traditional methods (e.g., surveys), e.g., one may observe how countries or cities are becoming more culturally similar or distinct over time.

In this document we also present the definition and applicability of the concept of sensing layers. The use of a set of sensing layers may be applied to several contexts of urban computing, for example, helping to better understand the dynamics of cities and urban behavior in different regions of the world, and respond quickly to unexpected changes. In this direction, we proposed a framework for integrating multiple sensing layers, which was illustrated in the construction of two applications using multiple sensing layers.

7.2 Future Work

In our study three types of systems, namely location, photo, and traffic alerts sharing services, have been explored. These systems fall into three different sensing layers - location categories, picture of places, and traffic alerts. We particularly have studied the time and location dimensions of our data from these layers. Besides that, we have explored also the specialty data provided by the location categories layer (e.g., in the City Image technique), but we have not explored the photos themselves provided by the picture of places layer nor the alerts from traffic alerts layer. Certainly, a range of fruitful opportunities may emerge when exploring the specialty data offered by each layer. For example, applying image processing techniques on the photos shared by people could potentially be useful in many cases, such as, a new way to capture people's sentiment about certain place, or use the photos to learn particular characteristics of different regions, in the direction of the study [Doersch et al., 2012].

Other possibility of future work include the development of other applications that exploit the proposed framework for integrating other sensing layers. For example, the traffic alerts layer derived from Waze and the weather condition layer derived from Weddar. With this we could build a more accurate mapping of the city dynamics. In the same direction, a future work is to investigate the interplay between data obtained from traditional wireless sensor networks and data obtained from PSNs. This step is fundamental to offer applications that is based on both source of information.

Another future work is build applications and services for smart cities exploring some of the opportunities presented in this document, such as traffic monitoring, information dissemination and recommendation systems. For instance, specifically about traffic related applications, there are several opportunities reported in Section 4.5.

From the methodology of cultural boundaries identification, one of the obvious directions for future work is to exploit the cultural criteria identified here, to perform social studies at large scale (e.g., study of global culture). Besides that, we also envision the development of recommendation mechanisms considering the cultural information of specific urban areas. This could be useful, for instance, for location-based social networks like Foursquare to improve their current recommendation systems. Another future work is to develop applications for companies that have businesses in one country and want to verify the compatibility of cultural preferences across different markets.

Future work specifically related to the City Image technique is to build new city recommendation services that explores the City Image technique and the proposed city clustering methodology. For that an essential step is extend our analyses to a very large number of cities in the world, or even all of them. Another possibility is to use the transition matrices

to complement a urban mobility model in different cities. This new model may improve the accuracy of a varied number of applications, such as traffic engineering in communication networks and transportation systems.

In our analysis we considered static graphs varying in the time, when needed, such as in the City Image technique. One concept that we could also explore is temporal graphs, a representation that encodes temporal data into graphs while fully retaining the temporal information of the original data. Temporal graphs enable analysis of the dynamic temporal properties of data by using existing graph algorithms (such as centrality algorithms), with no need for data-driven simulations. We believe that valuable information can be extracted using with temporal graphs from PSN data.

Quality control of PSN data is an important issue which cannot be neglected, as mentioned in Chapter 3. In this direction, another future work is to evaluate the quality of the data provided by PSNs. One of the possibilities is propose quality metrics for PSN data, considering, for example, amount of data per users and per area, data distribution, data entropy, and data trust. Defining quality metrics we can have a way to measure how accurate the result obtained is, and also have a parameter for deciding about the use of certain layer. Another possibility is to investigate spam or other malicious behavior in PSNs. If those malicious behaviors start to be significant in the system, this is a fundamental step in order to propose solutions.

Finally, is also a future work to evaluate other kind of sensing layers that could be extracted from the current technologies. In the same direction is also a future work compare different sensing layers with similar purposes. For example, is the traffic layer obtained from Waze better for traffic inference than the one obtained from Bing Maps? can they complement each other? If yes, how to fuse these information?

These are only some examples of future work that could be performed from this thesis. Certainly, a range of other possibilities can also be proposed.

7.3 Comments on Publications

Section 7.3.1 lists all the publications obtained direct from the results of this thesis. Section 7.3.2 presents other publications performed during the doctoral period.

7.3.1 Contributions from the Thesis

The list below contains all the publications derived directly from the thesis:

- [Silva et al., 2012b] published in ACM International Workshop on Hot Topics in Planet-scale Measurement (HotPlanet'12). This was our first work towards the understanding participatory sensor networks properties. The PSN analyzed was derived from location sharing services, particularly Gowalla and Brightkite;
- [Silva et al., 2013b] published in IEEE Symposium on Computers and Communications (ISCC'13). This extended the work [Silva et al., 2012b] analyzing also two different PSNs derived from Foursquare. Besides that, we also presented some challenges and opportunities to the study of city dynamics;
- [Silva et al., 2013c] (**2nd best paper award**) published in Brazilian Symposium on Computer Networks and Distributed Systems (SBRC'13). In this work we investigated properties of a PSN derived from Instagram, a photo sharing service;
- [Silva et al., 2013d] published in IEEE International Conference on Distributed Computing in Sensor Systems (DCOSS'13). This study extended the work [Silva et al., 2013c] in several ways, for instance, performing new analysis of users' contribution and the temporal photo sharing pattern in different cities and POIs. This study also propose a technique for point of interest identification, based on the popularity of areas where people shared pictures. The technique is also able to extract sights out of the identified POIs, using for that popular transitions of people between POIs;
- [Silva et al., 2013f] published in Springer International Conference on Social Informatics (SocInfo'13). In this work we studied properties of a PSN derived from Waze.
- [Silva et al., 2012d] (**Best paper award**) published in IEEE International Conference on Cyber, Physical and Social Computing. In this study we proposed a technique named City Image. This technique provides a visual summary of the city dynamics based on the movements of individuals. As demonstrated, this technique is promising way to better understand the city dynamics, helping us to visualize the common routines of their citizens;
- [Silva et al., 2013a] book chapter published in the book: Ubiquitous Social Media Analysis edited by Springer. In this work we survey models and approaches applied in PSNs to support different applications and techniques;
- [Silva et al., 2014c] under revision in ACM Transactions on Internet Technology (TOIT). This study is an extended version of [Silva et al., 2012d]. This study builds upon on [Silva et al., 2012d] by several ways: analyzing the proposed technique to a

much larger number of cities; showing how to use the technique to perform a quantitative comparison of multiple cities, illustrating it by clustering cities based on their similarity in terms of transitions; and including complementary analysis to the City Image technique that focus on transitions between specific locations in a city;

- [Silva et al., 2014b] accepted for publication in International AAAI Conference on Weblogs and Social Media (ICWSM'14). In this study we propose a new methodology for identifying cultural boundaries and similarities across populations using self-reported cultural preferences recorded in PSSs.
- [Silva et al., 2013e] published in ACM SIGKDD International Workshop on Urban Computing (UrbComp'13). In this study we perform a comparative study of different PSNs derived from Instagram and Foursquare. We analyze those PSNs to investigate whether we can observe the same users' movement pattern, the popularity of regions in cities, the activities of users who use those social networks, and how users share their content along the time;
- [Silva et al., 2014a] published in IEEE Wireless Communications Magazine. In this study we discuss the potential of location-based social media systems as sources of large scale participatory sensing from which valuable knowledge about city dynamics and urban social behavior can be drawn. We also discuss the technical challenges involved in building and deploying such methods. We also introduce the concept of sensing layers;
- [Silva et al., 2014d] accepted for publication in Brazilian Symposium on Computer Networks and Distributed Systems (SBRC'14). In this study we formalize the concept of sensing layers, presents a framework for working with multiple sensing layers, and also illustrates the potential of the joint use of multiply sensing layers through two applications;
- [Silva et al., 2014e] this work is under revision in ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems. This is an extended version of [Silva et al., 2014d] where we present more details about the proposed framework.

7.3.2 Other Publications

It is important to point out that the preliminary study towards the research topic of this thesis generated two other contributions:

- [Silva et al., 2012c] (**Best paper award**) published in Brazilian Symposium on Pervasive and Ubiquitous Computing (SBCUP'12). This work performs a study of the current state of research in ubiquitous computing. We collected information about all the papers published in the main ubicomp conferences (UbiComp, Pervasive, and PerCom) and performed a data mining process extracting statistics such as most productive authors and institutions. We also analyzed the collaboration among authors, identifying, for instance, communities' formation. Besides that, we analyzed all papers published in 2010 and 2011, creating a taxonomy of recent ubicomp research;
- [Silva et al., 2012a] published in Journal of Applied Computing Research. This work is an extension of the work performed in [Silva et al., 2012c], where a more detailed analysis of the collaboration network and the proposed taxonomy are presented.

In parallel with my thesis research topic, I have been participating in other studies related to Computer and Social Networks, one as first author and two as co-author:

- [Silva et al., 2011] published in Elsevier International Journal of Computer and Telecommunications Networking (Computer Networks). This study is an extension of the work I performed during my master on live streaming of user generated videos;
- [Maia et al., 2012] published in Brazilian Symposium on Computer Networks and Distributed Systems (SBRC'12). This work analyzes the SBRC authors' collaboration network;
- [Maia et al., 2013] published in Journal of the Brazilian Computer Society. This work is an extension of the study [Maia et al., 2012].

Bibliography

- Abowd, G. D. and Mynatt, E. D. (2000). Charting past, present, and future research in ubiquitous computing. *ACM Trans. Comput.-Hum. Interact.*, 7:29--58. ISSN 1073-0516.
- Abowd, G. D., Mynatt, E. D., and Rodden, T. (2002). The human experience. *IEEE Pervasive Computing*, 1:48--57. ISSN 1536-1268.
- Akyildiz, I., Su, W., Sankarasubramaniam, Y., and Cayirci, E. (2002). Wireless sensor networks: a survey. *Computer Networks*, 38(4):393 – 422. ISSN 1389-1286.
- Arrow, K. J. (1972). Gifts and Exchanges. *Philosophy and Public Affairs*, 1(4):343--362. ISSN 00483915.
- Axelrod, R. (1997). The dissemination of culture: A model with local convergence and global polarization. *J Conflict Resolut.*
- Barabási, A. (2005). The origin of bursts and heavy tails in human dynamics. *Nature*, 435:207--211.
- Barth, F. (1969). *Ethnic groups and boundaries: the social organization of culture difference*. Scandinavian university books. Little, Brown.
- Becker, H., Naaman, M., and Gravano, L. (2011). Beyond trending topics: Real-world event identification on twitter. In *Proc. of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM'11)*, Barcelona, Spain.
- Bell, G. and Dourish, P. (2007). Yesterday's tomorrows: notes on ubiquitous computing's dominant vision. *Personal Ubiquitous Comput.*, 11:133--143. ISSN 1617-4909.
- Benevenuto, F., Magno, G., Rodrigues, T., and Almeida, V. (2010). Detecting spammers on twitter. In *Proc. of the 7th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)*, Redmond, USA.

- Bilandzic, M. and Foth, M. (2012). A review of locative media, mobile and embodied spatial interaction. *International Journal of Human-Computer Studies*, 70(1):66--71. ISSN 1071-5819.
- Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993--1022.
- Blossfeld, H.-P., Klijzing, E., Mills, M., and Kurz, K. (2005). *Globalization, uncertainty and youth in society*. Routledge, London.
- Bollen, J., Mao, H., and Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1--8.
- Bradley, M. M. and Lang, P. J. (1999). Affective norms for english words (anew): Instruction manual and affective ratings. Technical report, Technical Report C-1, The Center for Research in Psychophysiology, University of Florida.
- Brenner, J. and Smith, A. (2013). 72% of online adults are social networking site users. <http://goo.gl/HTgNy3>.
- Brockmann, D., Hufnagel, L., and Geisel, T. (2006). The scaling laws of human travel. *Nature*, 439(7075):462--465.
- Brush, A. B., Krumm, J., and Scott, J. (2010). Exploring end user preferences for location obfuscation, location-based services, and the value of location. In *Proceedings of the 12th ACM International Conference on Ubiquitous computing, Ubicomp '10*, pages 95--104, Copenhagen, Denmark. ACM.
- Burke, J., Estrin, D., Hansen, M., Parker, A., Ramanathan, N., Reddy, S., and Srivastava, M. B. (2006). Participatory sensing. In *In: Workshop on World-Sensor-Web (WSW06): Mobile Device Centric Sensor Networks and Applications*, pages 117--134.
- Burt, R. S. (1992). *Structural Holes: The Social Structure of Competition*. Harvard University Press.
- Carole, C. (1997). *Food And Culture: A Reader*. Routledge, 2nd edition. ISBN 0415977770.
- Chaey, C. (2012). Foursquare explore is now a search tool anyone can use, no check-ins required. <http://goo.gl/MQ22DU>.
- Chang, K.-t. (2010). *Introduction to geographic information systems*. McGraw-Hill New York.

- Cheng, Y. (1995). Mean shift, mode seeking, and clustering. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 17(8):790--799.
- Cheng, Z., Caverlee, J., Lee, K., and Sui, D. Z. (2011). Exploring Millions of Footprints in Location Sharing Services. In *Proc. of the Fifth Int'l Conf. on Weblogs and Social Media (ICWSM'11)*, Barcelona, Spain.
- Cho, E., Myers, S. A., and Leskovec, J. (2011). Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '11*, pages 1082--1090, San Diego, California, USA. ACM.
- Cochrane, R. and Bal, S. (1990). The drinking habits of sikh, hindu, muslim and white men in the west midlands: a community survey. *British Journal of Addiction*, 85(6):759--769. ISSN 1360-0443.
- Cranshaw, J., Schwartz, R., Hong, J. I., and Sadeh, N. (2012). The Livelihoods Project: Utilizing Social Media to Understand the Dynamics of a City. In *Proc. of the Sixth Int'l Conf. on Weblogs and Social Media*, Dublin, Ireland.
- Cranshaw, J., Toch, E., Hong, J., Kittur, A., and Sadeh, N. (2010). Bridging the gap between physical location and online social networks. In *Proceedings of the 12th ACM International Conference on Ubiquitous computing*, pages 119--128, Pittsburgh, USA. ACM.
- Dey, A. K. and Abowd, G. D. (2000). Towards a Better Understanding of Context and Context-Awareness. In *CHI 2000 Workshop on the What, Who, Where, When, and How of Context-Awareness*, The Hague, The Netherlands.
- Dey, A. K., Abowd, G. D., and Wood, A. (1998). Cyberdesk: a framework for providing self-integrating context-aware services. In *Proceedings of the 3rd international conference on Intelligent user interfaces, IUI '98*, pages 47--54, New York, NY, USA. ACM.
- Doersch, C., Singh, S., Gupta, A., Sivic, J., and Efros, A. A. (2012). What makes paris look like paris? *ACM Trans. Graph.*, 31(4):101:1--101:9. ISSN 0730-0301.
- Domenico, M. D., Sole-Ribalta, A., Cozzo, E., Kivela, M., Moreno, Y., Porter, M. A., Gomez, S., and Arenas, A. (2013). Mathematical formulation of multi-layer networks. *Physical Review X* 3, 041022.
- Duggan, M. and Smith, A. (2014). Social media update 2013. <http://goo.gl/JhuiOG>.

- Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)*, Portland, USA.
- Festinger, L. (1967). *Social pressures in informal groups: a study of human factors in housing*. Stanford University Press.
- Fisk, P. R. (1961). The graduation of income distributions. *Econometrica*, 29(2):171–185.
- Foursquare (2013). About foursquare. <https://foursquare.com/about>.
- Frias-Martinez, V., Soto, V., Hohwald, H., and Frias-Martinez, E. (2012). Characterizing urban landscapes using geolocated tweets. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pages 239--248, Amsterdam, The Netherlands.
- Gao, H., Tang, J., and Liu, H. (2012). gSCorr: modeling geo-social correlations for new check-ins on location-based social networks. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, pages 1582--1586, Maui, USA. ACM.
- Garcia-Gavilanes, R., Quercia, D., and Jaimes, A. (2013). Cultural dimensions in twitter: Time, individualism and power. In *Proc. of ICWSM'13*, Boston, USA. AAAI.
- Giannotti, F., Pedreschi, D., Pentland, A., Lukowicz, P., Kossmann, D., Crowley, J., and Helbing, D. (2012). A planetary nervous system for social mining and collective awareness. *The European Physical Journal Special Topics*, 214(1):49--75.
- Ginsberg, J., Mohebbi, M., Patel, R., Brammer, L., Smolinski, M., and Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457:1012–1014. doi:10.1038/nature07634.
- Goel, V. (2013). Maps that live and breathe with data. <http://goo.gl/fCWkpf>.
- Gomide, J., Veloso, A., Jr., W. M., Almeida, V., Benevenuto, F., Ferraz, F., and Teixeira, M. (2011). Dengue surveillance based on a computational model of spatio-temporal locality of twitter. In *ACM Web Science Conference (WebSci)*, Evanston, USA.
- Gonzalez, M. C., Hidalgo, C. A., and Barabasi, A.-L. (2008). Understanding individual human mobility patterns. *Nature*, 453(7196):779--782.

- Gonçalves, P., Araújo, M., Benevenuto, F., and Cha, M. (2013). Comparing and combining sentiment analysis methods. In *Proceedings of the 1st ACM Conference on Online Social Networks (COSN'13)*, Boston, USA.
- Inglehart, R. and Welzel, C. (2010). Changing Mass Priorities: The Link between Modernization and Democracy. *Perspectives on Politics*, 8(02):551--567.
- Instagram (2013). Instagram today: 150 million people. <http://blog.instagram.com/post/60694542173/150-million>.
- Instagram (2014). Instagram today: 200 million strong. <http://blog.instagram.com/post/80721172292/200m>.
- Java, A., Joshi, A., and Finin, T. (2008). Detecting communities via simultaneous clustering of graphs and folksonomies. In *Proceedings of WebKDD*, Las Vegas, USA.
- Ji, R., Xie, X., Yao, H., and Ma, W.-Y. (2009). Mining city landmarks from blogs by graph modeling. In *Proceedings of the 17th ACM International Conference on Multimedia*, pages 105--114, Beijing, China. ACM.
- Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer, second edition.
- Joseph, K., Tan, C. H., and Carley, K. M. (2012). Beyond local, categories and friends: clustering foursquare users with latent topics. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 919--926, Pittsburgh, USA. ACM.
- Kim, D. H., Han, K., and Estrin, D. (2011). Employing user feedback for semantic location services. In *Proc. of the 13th international conference on Ubiquitous computing, UbiComp '11*, pages 217--226, Beijing, China. ACM.
- Kisilevich, S., Krstajic, M., Keim, D., Andrienko, N., and Andrienko, G. (2010). Event-based analysis of people's activities and behavior using flickr and panoramio geotagged photo collections. In *14th International Conference on Information Visualisation*, pages 289--296, London, UK. IEEE.
- Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604--632.
- Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, 78(9):1464--1480.
- Kostakos, V. et al. (2009). Understanding and Measuring the Urban Pervasive infrastructure. *Personal and Ubiquitous Computing*, 13(5):355--364. ISSN 1617-4909.

- Krumm, J. (2009). *Ubiquitous Computing Fundamentals*. Chapman & Hall/CRC, 1st edition. ISBN 1420093606, 9781420093605.
- Lane, N., Miluzzo, E., Lu, H., Peebles, D., Choudhury, T., and Campbell, A. (2010). A Survey of Mobile Phone Sensing. *IEEE Communications Magazine*, 48(9):140–150. ISSN 0163-6804.
- Laura, L., Leonardi, S., Caldarelli, G., De, P., and Rios, P. D. L. (2002). A multi-layer model for the web graph. In *Proceedings of the 2nd International Workshop on Web Dynamics*, Honolulu, USA.
- Lazer, D., Kennedy, R., King, G., and Vespignani, A. (2014). Google flu trends still appears sick: An evaluation of the 2013-2014 flu season. <http://j.mp/1m6JBX6>.
- Lee, R. and Sumiya, K. (2010). Measuring geographical regularities of crowd behaviors for twitter-based geo-social event detection. In *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks*, pages 1--10, San Jose, USA. ACM.
- Lohr, S. (2014). Google flu trends: The limits of big data. <http://nyti.ms/1g6KiHU>.
- Long, X., Jin, L., and Joshi, J. (2012). Exploring trajectory-driven local geographic topics in foursquare. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing, UbiComp '12*, pages 927--934, Pittsburgh, Pennsylvania. ACM.
- Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395--416. ISSN 0960-3174.
- Maia, G., Guidoni, D. L., Silva, T. H., Souza, F. S., de Melo, P. O. V., Soares, C. A., Almeida, J. M., and Loureiro, A. A. (2012). Análise da rede de colaboração do simpósio brasileiro de redes de computadores e sistemas distribuídos: As primeiras 30 edições. In *Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, Ouro Preto, MG, Brazil.
- Maia, G., Vaz de Melo, P., Guidoni, D., Souza, F., Silva, T., Almeida, J., and Loureiro, A. (2013). On the analysis of the collaboration network of the brazilian symposium on computer networks and distributed systems. *Journal of the Brazilian Computer Society*, 19(3):361–382. ISSN 0104-6500.
- Malmgren, R. D., Stouffer, D. B., Motter, A. E., and Amaral, L. A. N. (2008). A poissonian explanation for heavy tails in e-mail communication. *Proceedings of the National Academy of Sciences*, 105(47):18153--18158.

- Mashhadi, A. J. and Capra, L. (2011). Quality Control for Real-time Ubiquitous Crowdsourcing. In *Proc. of the 2nd Int'l Workshop on Ubiquitous Crowdsourcing (UbiCrowd'11)*, pages 5--8, Beijing, China.
- McGill, R., Tukey, J. W., and Larsen, W. A. (1978). Variations of box plots. *The American Statistician*, 32(1):12--16.
- McPherson, M., Smith-Lovin, L., and Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415--444.
- Murdock, G. (1949). *Social Structure*. Macmillan.
- Naaman, M., Zhang, A. X., Brody, S., and Lotan, G. (2012). On the study of diurnal urban routines on twitter. In *International Conference on Weblogs and Social Media*, Dublin, Ireland.
- Newman, M. (2010). *Networks: an introduction*. Oxford University Press, Inc.
- Newman, M. E. (2002). Assortative mixing in networks. *Phy. rev. let.*, 89(20):208701.
- Ng, A. Y., Jordan, M. I., Weiss, Y., et al. (2002). On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 2:849--856.
- Nguyen, T. and Szymanski, B. K. (2012). Using location-based social networks to validate human mobility and relationships models. *arXiv preprint arXiv:1208.3653*.
- Nister, D. and Stewenius, H. (2006). Scalable recognition with a vocabulary tree. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2161--2168. IEEE.
- Noulas, A., Scellato, S., Mascolo, C., and Pontil, M. (2011a). An Empirical Study of Geographic User Activity Patterns in Foursquare. In *Proc. of the Fifth Int'l Conf. on Weblogs and Social Media (ICWSM'11)*, Barcelona, Spain. AAAI.
- Noulas, A., Scellato, S., Mascolo, C., and Pontil, M. (2011b). Exploiting Semantic Annotations for Clustering Geographic Areas and Users in Location-based Social Networks. In *Proc. of the Fifth Int'l Conf. on Weblogs and Social Media (ICWSM'11)*, Barcelona, Spain. AAAI.
- Pascoe, M. J. (1998). Adding generic contextual capabilities to wearable computers. In *Proceedings of the 2nd IEEE International Symposium on Wearable Computers, ISWC '98*, pages 92--99, Washington, DC, USA. IEEE Computer Society.

- Pentland, A. (2010). To signal is human. *American scientist*, 98(3):204--210.
- Poblete, B., Garcia, R., Mendoza, M., and Jaimes, A. (2011). Do all birds tweet the same?: characterizing twitter around the world. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 1025--1030, Glasgow, UK. ACM.
- Pontes, T., Magno, G., Vasconcelos, M., Gupta, A., Almeida, J., Kumaraguru, P., and Almeida, V. (2012). Beware of what you share: Inferring home location in social networks. In *International Conference on Data Mining Workshops(ICDMW)*, pages 571--578, Brussels, Belgium.
- Quercia, D., Capra, L., and Crowcroft, J. (2012). The social world of twitter: Topics, geography, and emotions. In *The 6th international AAAI Conference on weblogs and social media, Dublin*.
- Reddy, S., Estrin, D., Hansen, M., and Srivastava, M. (2010). Examining micro-payments for participatory sensing data collections. In *Proc. of the 12th ACM International Conference on Ubiquitous Computing, Ubicomp'10*, pages 33--36, Copenhagen, Denmark. ACM.
- Redondo, J., Fraga, I., Padrón, I., and Comesaña, M. (2007). The spanish adaptation of anew (affective norms for english words). *Behavior research methods*, 39(3):600--605.
- Sadilek, A., Kautz, H., and Bigham, J. P. (2012). Finding your friends and following them to where you are. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 723--732, Seattle, USA. ACM.
- Sagl, G., Blaschke, T., Beinat, E., and Resch, B. (2012). Ubiquitous geo-sensing for context-aware analysis: Exploring relationships between environmental and human dynamics. *Sensors*, 12(7):9800--9822. ISSN 1424-8220.
- Sakaki, T., Okazaki, M., and Matsuo, Y. (2010). Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web, WWW '10*, pages 851--860, Raleigh, North Carolina, USA. ACM.
- Saroiu, S. and Wolman, A. (2010). I am a sensor, and i approve this message. In *Proc. of the Eleventh Workshop on Mobile Computing Systems and Applications, HotMobile '10*, pages 37--42, Annapolis, Maryland. ACM.
- Satyanarayanan, M. (1996). Fundamental challenges in mobile computing. In *Proceedings of the fifteenth annual ACM symposium on Principles of distributed computing, PODC '96*, pages 1--7, Philadelphia, Pennsylvania, USA. ACM.

- Scellato, S., Noulas, A., Lambiotte, R., and Mascolo, C. (2011). Socio-spatial Properties of Online Location-based Social Networks. In *Proc. of the Fifth Int'l Conf. on Weblogs and Social Media (ICWSM'11)*, Barcelona, Spain. AAAI.
- Schilit, B., Adams, N., and Want, R. (1994). Context-aware computing applications. In *Proceedings of the 1994 First Workshop on Mobile Computing Systems and Applications*, pages 85--90, Washington, DC, USA. IEEE Computer Society.
- Schmitt, H., Scholz, E., Leim, I., and Moschner, M. (2005). *The Mannheim Eurobarometer Trendfile 1970-2002. Data Set Edition 2.00: Appendix*. Zentralarchiv für Empirische Sozial.
- Scott, J. P. and Carrington, P. J. (2011). *The SAGE Handbook of Social Network Analysis*. Sage Publications Ltd. ISBN 1847873952, 9781847873958.
- Shankar, P., Huang, Y.-W., Castro, P., Nath, B., and Iftode, L. (2012). Crowds replace experts: Building better location-based services using mobile social network interactions. In *Int. Conf. on Perv. Comp. and Comm. (Percom'12)*, pages 20--29, Lugano, Switzerland.
- Shannon, C. E. (1948). A mathematical theory of communication. *Bell system tech. jour.*, 27.
- Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):888--905.
- Silva, T., Almeida, J. M., and Guedes, D. (2011). Live streaming of user generated videos: Workload characterization and content delivery architectures. *Comput. Netw.*, 55(18):4055--4068. ISSN 1389-1286.
- Silva, T., Vaz De Melo, P., Almeida, J., and Loureiro, A. (2014a). Large-scale study of city dynamics and urban social behavior using participatory sensing. *Wireless Communications, IEEE*, 21(1):42--51.
- Silva, T. H., Celes, C. S. F. d. S., Mota, V. F. S., and Loureiro, A. A. F. (2012a). A picture of present ubicomp research exploring publications from important events in the field. *Journal of Applied Computing Research*, 2(1):32--49.
- Silva, T. H., Vaz de Melo, P. O. S., Almeida, J., and Loureiro, A. A. F. (2012b). Uncovering properties in participatory sensor networks. In *Proc. 4th ACM International Workshop on Hot Topics in Planet-scale Measurement*, pages 33--38, Low Wood Bay, Lake District, UK.

- Silva, T. H., Vaz de Melo, P. O. S., Almeida, J. M., and Loureiro, A. A. (2013a). Social media as a source of sensing to study city dynamics and urban social behavior: Approaches, models, and opportunities. In *Ubiquitous Social Media Analysis*, volume 8329, pages 63–87. Springer Berlin Heidelberg.
- Silva, T. H., Vaz de Melo, P. O. S., Almeida, J. M., and Loureiro, A. A. F. (2012c). Challenges and opportunities on the large scale study of city dynamics using participatory sensing. In *Brazilian Symposium on Pervasive and Ubiquitous Computing (SBCUP'12)*, Curitiba, Brazil.
- Silva, T. H., Vaz de Melo, P. O. S., Almeida, J. M., and Loureiro, A. A. F. (2012d). Visualizing the invisible image of cities. In *Proc. IEEE International Conference on Cyber, Physical and Social Computing*, pages 382--389, Besancon, France.
- Silva, T. H., Vaz de Melo, P. O. S., Almeida, J. M., and Loureiro, A. A. F. (2013b). Challenges and opportunities on the large scale study of city dynamics using participatory sensing. In *18th IEEE Symposium on Computers and Communications (ISCC'13)*, pages 528--534, Split, Croatia.
- Silva, T. H., Vaz de Melo, P. O. S., Almeida, J. M., and Loureiro, A. A. F. (2013c). Uma Fotografia do Instagram: Caracterização e Aplicação. In *Brazilian Symposium on Computer Networks and Distributed Systems (SBRC'13)*, Brasília, Brazil.
- Silva, T. H., Vaz de Melo, P. O. S., Almeida, J. M., Musolesi, M., and Loureiro, A. A. F. (2014b). You are What you Eat (and Drink): Identifying Cultural Boundaries by Analyzing Food & Drink Habits in Foursquare. In *Proc. of International AAAI Conference on Weblogs and Social Media*, Ann Arbor, MI, USA.
- Silva, T. H., Vaz de Melo, P. O. S., Almeida, J. M., Salles, J., and Loureiro, A. A. F. (2013d). A picture of Instagram is worth more than a thousand words: Workload characterization and application. In *Proc. of the IEEE International Conference on Distributed Computing in Sensor Systems (DCOSS'13)*, pages 123--132, Cambridge, USA.
- Silva, T. H., Vaz de Melo, P. O. S., Almeida, J. M., Salles, J., and Loureiro, A. A. F. (2013e). A comparison of foursquare and instagram to the study of city dynamics and urban social behavior. In *Proc. ACM SIGKDD Int. Workshop on Urban Computing (UrbComp'13)*, pages 1--8, Chicago, USA.
- Silva, T. H., Vaz de Melo, P. O. S., Almeida, J. M., Salles, J., and Loureiro, A. A. F. (2014c). Revealing the city that we cannot see. *ACM Transactions on Internet Technology (TOIT)*, 0(0):00–00.

- Silva, T. H., Vaz de Melo, P. O. S., Almeida, J. M., Salles, J., Viana, A., and Loureiro, A. A. F. (2014d). Definição, Modelagem e Aplicações de Camadas de Sensoriamento Participativo. In *Brazilian Symposium on Computer Networks and Distributed Systems (SBRC'14)*, Florianópolis, Brazil.
- Silva, T. H., Vaz de Melo, P. O. S., Almeida, J. M., Salles, J., Viana, A., and Loureiro, A. A. F. (2014e). Participatory sensor networks as sensing layers. *ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*.
- Silva, T. H., Vaz de Melo, P. O. S., Viana, A., Almeida, J. M., Salles, J., and Loureiro, A. A. F. (2013f). Traffic Condition is more than Colored Lines on a Map: Characterization of Waze Alerts. In *Proc. of the Int. Conference on Social Informatics (SocInfo'13)*, pages 309--318, Kyoto, Japan.
- Sinnott, R. W. (1984). Virtues of the Haversine. *Sky and Telescope*, 68(2):159+.
- Smith, I., Consolvo, S., Lamarca, A., Hightower, J., Scott, J., Sohon, T., Hughes, J., Iachello, G., and Abowd, G. D. (2005). Social Disclosure of Place: From Location Technology to Communication Practices. In *Proc. of the 3rd Int. Conf on Perv. Comp. (Pervasive '05)*, pages 134--151, Munich, Germany.
- Song, C., Qu, Z., Blumm, N., and Barabási, A.-L. (2010). Limits of predictability in human mobility. *Science*, 327(5968):1018--1021.
- Sørensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *Biologiske Skrifter*, 5(4).
- Srivastava, M., Abdelzaher, T., and Szymanski, B. (2012). Human-centric sensing. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 370(1958):176--197. ISSN 1471-2962.
- TheEconomist (2012). *It's a hit: Google casts its web across the continent. Any complaints?* The Economist, Nairobi.
- Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., and Kappas, A. (2010). Sentiment in short strength detection informal text. *J. Am. Soc. Inf. Sci. Technol.*, 61(12):2544--2558. ISSN 1532-2882.
- Toch, E., Cranshaw, J., Drielsma, P. H., Tsai, J. Y., Kelley, P. G., Springfield, J., Cranor, L., Hong, J., and Sadeh, N. (2010). Empirical models of privacy in location sharing. In *Proc.*

- of the 12th ACM International Conference on Ubiquitous Computing, Ubicomp '10*, pages 129--138, Copenhagen, Denmark. ACM.
- Valori, L., Picciolo, F., Allansdottir, A., and Garlaschelli, D. (2012). Reconciling long-term cultural diversity and short-term collective social behavior. *Proc. of Nat. Acad. of Sci.*, 109(4):1068--1073.
- Vasconcelos, M. A., Ricci, S., Almeida, J., Benevenuto, F., and Almeida, V. (2012). Tips, dones and todos: uncovering user profiles in foursquare. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, pages 653--662, Seattle, Washington, USA. ACM.
- Vaz de Melo, P. O. S., Faloutsos, C., and Loureiro, A. A. (2011). Human dynamics in large communication networks. In *SIAM International Conference on Data Mining*, Mesa, AZ, USA.
- Wang, C., Xie, X., Wang, L., Lu, Y., and Ma, W.-Y. (2005). Detecting geographic locations from web resources. In *Proceedings of the 2005 workshop on Geographic information retrieval*, pages 17--24. ACM.
- Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236--244.
- Watson, J. (2006). *Golden arches east: McDonald's in East Asia*. Stanford University Press.
- Weiser, M. (1991). The Computer in the 21st Century. *Scientific American*, 265(3):94--104.
- Weiser, M. (1993a). Some computer science issues in ubiquitous computing. *Commun. ACM*, 36:75--84. ISSN 0001-0782.
- Weiser, M. (1993b). Ubiquitous computing. *Computer*, 26:71--72. ISSN 0018-9162.
- Weiser, M. (1994). Keynote: Building invisible interfaces. In *7th annual ACM symposium on User interface software and technology*.
- Weiser, M. (1996). Weiser's website about ubicomp. <http://www.ubiq.com/weiser/weiser.html>. Website accessed for the last time in May of 2013.
- Weiser, M. and Brown, J. S. (1996). The coming age of calm technology. Technical report, Xerox PARC.

- Weiser, M., Gold, R., and Brown, J. S. (1999). The origins of ubiquitous computing research at parc in the late 1980s. *IBM Syst. J.*, 38:693--696. ISSN 0018-8670.
- Xin, C., Xie, B., and Shen, C.-C. (2005). A novel layered graph model for topology formation and routing in dynamic spectrum access networks. In *International Symposium on New Frontiers in Dynamic Spectrum Access Networks*, pages 308–317, Baltimore, USA.
- Yardi, S., Romero, D., Schoenebeck, G., et al. (2009). Detecting spam in a twitter network. *First Monday*, 15(1).
- Yick, J., Mukherjee, B., and Ghosal, D. (2008). Wireless sensor network survey. *Computer Networks*, 52(12):2292 – 2330. ISSN 1389-1286.
- Yu, H., Sun, G., and Lv, M. (2012). Users sleeping time analysis based on micro-blogging data. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 964–968, Pittsburgh, USA. ACM.
- Zheng, Y., Zhang, L., Xie, X., and Ma, W.-Y. (2009). Mining interesting locations and travel sequences from gps trajectories. In *Proceedings of the 18th International Conference on World Wide Web*, pages 791--800, Madrid, Spain. ACM.
- Zheng, Y.-T., Zha, Z.-J., and Chua, T.-S. (2012). Mining travel patterns from geotagged photos. *ACM Trans. Intell. Syst. Technol.*, 3(3):56:1--56:18. ISSN 2157-6904.

Appendix A

General City Images

Figures A.1 and A.2 show the City Image, for all analyzed cities in Section 5.1, built using aggregated data across all time periods. These images provide a general picture of each city, and serve to illustrate broad differences across cities.

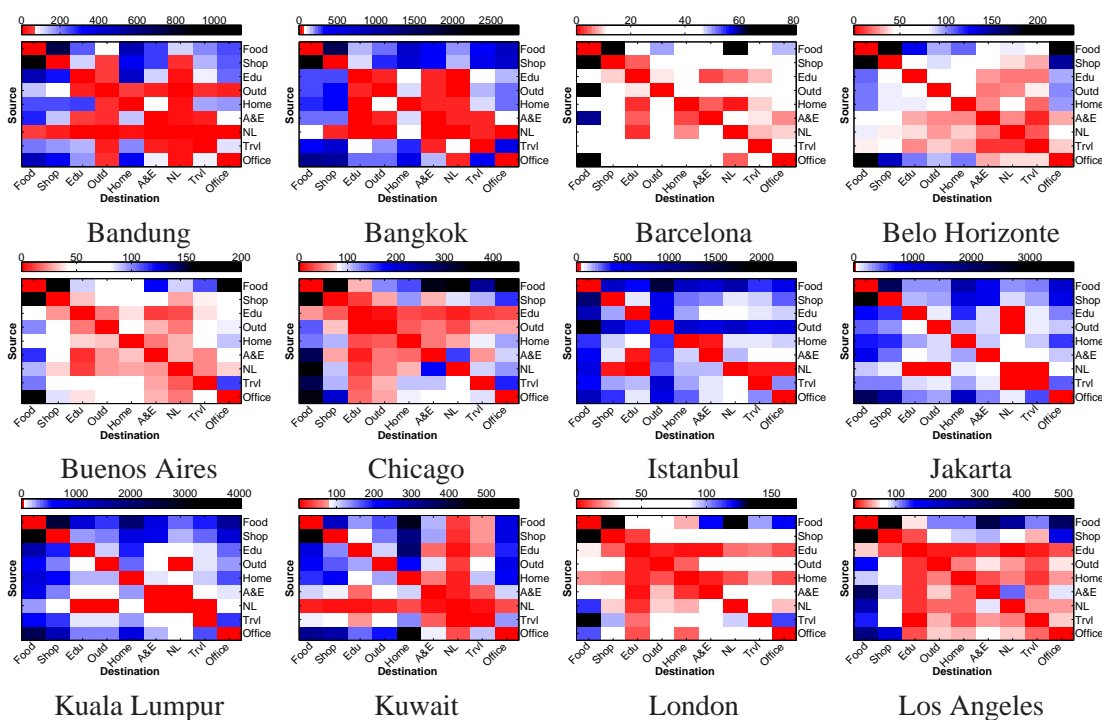


Figure A.1: The general City Image, which does not consider different periods separately of all cities.

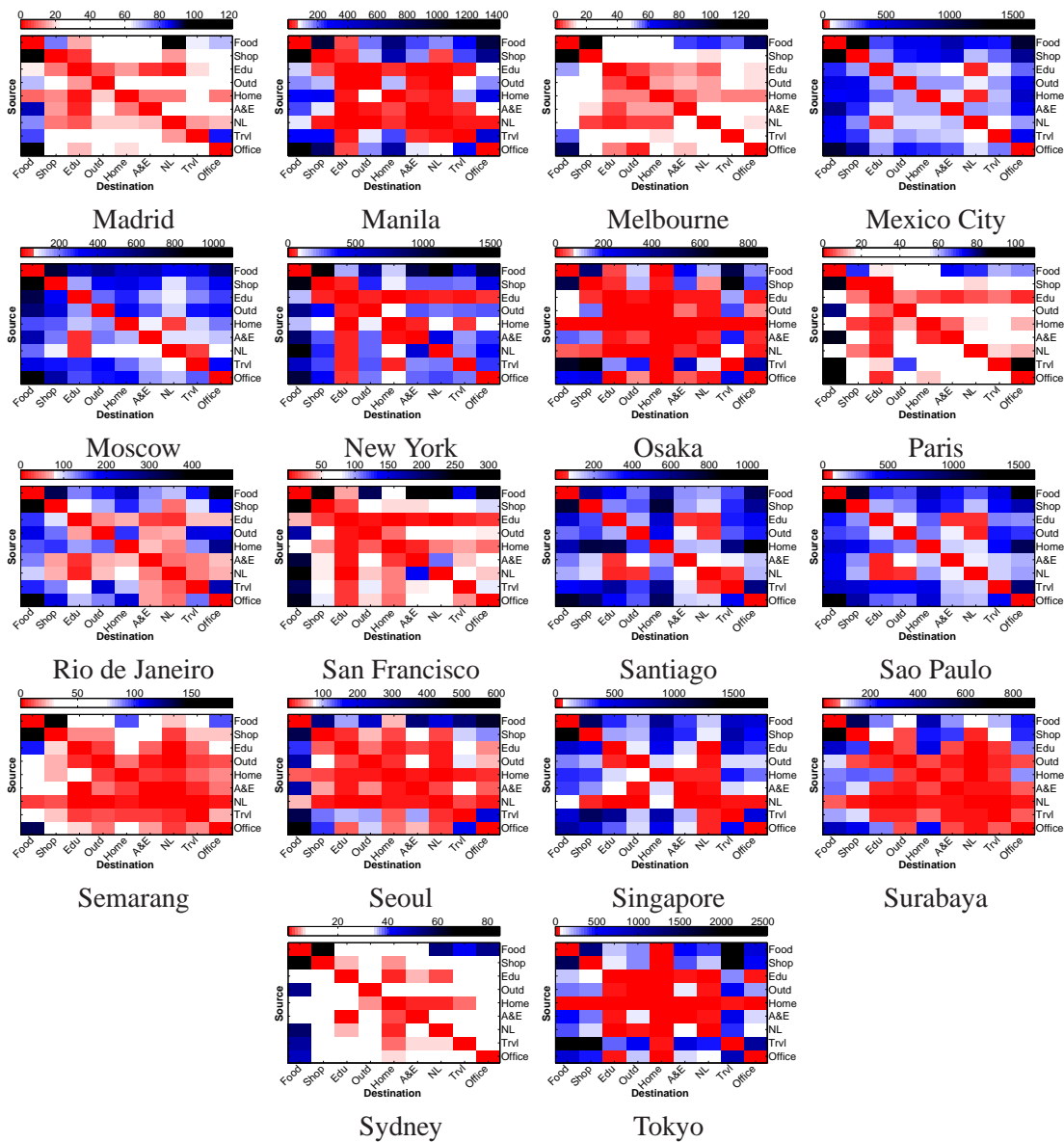


Figure A.2: The general City Image, which does not consider different periods separately of all cities.

Appendix B

Quantitative Comparison of Cities

In this appendix we propose an application that use the City Image technique for numerical comparison of different cities, by exploiting the values in each square matrix given by the technique. Specifically, we propose to compare two cities i and j by following the steps:

1. For each city i , the weight of each transition t of its City Image is normalized by the maximum weight of all transitions in this particular City Image. We refer to this normalized value as t'_i . As a result, we produce a vector $T_i = (t'_{i,1}, t'_{i,2}, \dots, t'_{i,81})$ containing all normalized transitions (total of 81, as there are 9 location categories) for a specific City Image;
2. We then compute the Euclidean distance $d_{i,j}$ between each pair of vectors (T_i, T_j) of cities i and j . By doing so we are calculating the distance between each considered city for all transitions.

More generally, the comparison of multiple cities produces a vector D containing the distance between each pair of cities. Vector D could then be used in several ways. For example, it could be exploited to cluster cities by similarity (in terms of movement patterns), as shown in the following steps:

1. Build a hierarchical cluster tree for the cities based on the distances in vector D using, for example, the Ward's method Ward Jr [1963]. This is a general agglomerative hierarchical clustering procedure, where the criterion for choosing the pair of clusters to merge at each step is based on the optimal value of an objective function. In our case, this objective function is the minimum total intracluster variance, which is computed based on the distances D ;
2. Determine the number of clusters c to be generated by visually inspecting the hierarchical cluster tree created, using, for example, a dendrogram plot of the tree;

3. Prune the tree created in step 1 in order to have c clusters.

We applied this procedure to compare and cluster the 30 cities analyzed in Section 5.1.3, considering two different time periods: weekdays during the day, to study the typical time when users perform their main routines; and weekend during the night, to study the typical period when people perform leisure activities. Figure B.1 shows the dendrograms built for each period. The red lines (dashed ones) indicate the cuts used to define the number of clusters c in each case. We defined c equal to 9 clusters for weekdays during the day and 7 clusters for weekend during the night.

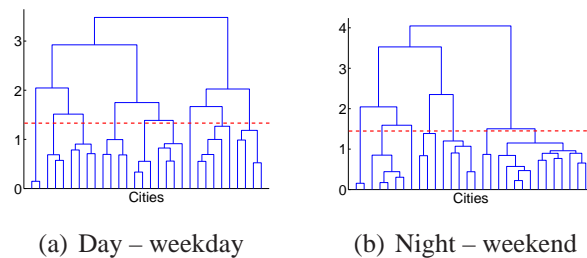


Figure B.1: Dendrogram plots for the binary cluster tree of 30 different cities, in two different time periods.

Tables B.1 and B.2 show the clustering results for weekdays during the day and weekends during the night, respectively. Note that, in general, cities from the same country or that are geographically close to each other were grouped together. The geographical proximity, which may reflect, to some extent, cultural similarity, is favorable to produce a similar behavior between the inhabitants from those cities, and might be the explanation to the clustering results. However, there are exceptions. For example, for weekdays during the day, San Francisco was grouped apart from other American cities, whereas Bangkok, far away from USA, was grouped in the same cluster as some American cities. Thus, the inhabitants of cities of the same country do not necessarily have similar behavior, reflecting heterogeneous patterns which are natural to occur in large countries, such as USA. Conversely, large geographical distances also do not necessarily imply large differences in people's habits. For instance, cities with good transportation system or many options for outdoor activities, such as beaches and parks, tend to favor transitions containing *travel* and *outdoor*, regardless of their particular geographical location, and tend to differ from other cities, even cities in the same country, that do not have such facilities.

We note that the proposed city clustering procedure and the city distance metric could be applied to a much larger number of cities in the world, with several potential applications. One example is a personalized city recommendation system for supporting tourism-oriented

| Cluster | Cities |
|---------|---|
| 1 | Bandung, Semarang, Surabaya |
| 2 | London, Paris, Madrid |
| 3 | Kuwait, Singapore, Moscow, Santiago |
| 4 | Sydney, Melbourne, Seoul, San Francisco |
| 5 | Rio, Belo Horizonte, Sao Paulo, Barcelona, Buenos Aires |
| 6 | Jakarta, Kuala Lumpur, Manila, Mexico City |
| 7 | Los Angeles, Chicago, New York, Bangkok |
| 8 | Tokyo, Osaka |
| 9 | Istanbul |

Table B.1: Clustering results for weekday during the day.

| Cluster | Cities |
|---------|---|
| 1 | Kuwait, Singapore, Kuala Lumpur, Manila, Bangkok |
| 2 | Tokyo, Osaka |
| 3 | Seoul, Jakarta, Bandung, Semarang, Surabaya |
| 4 | Rio, Belo Horizonte, Sao Paulo |
| 5 | Istanbul, Moscow |
| 6 | Santiago |
| 7 | Los Angeles, Chicago, San Francisco, New York, Melbourne, Sydney, Paris, Madrid, London, Barcelona, Buenos Aires, Mexico City |

Table B.2: Clustering results for weekend during the night.

applications. Such application could explore the proposed city clustering strategy to suggest new cities that the user might like, based on the user's interests (which could be inferred from prior user's interactions in the system). For example, by learning that a user liked Bandung during the day, the application might suggest Surabaya as a city to visit, as the two cities are grouped in the same cluster and thus have similarities. Location-based social media (like Foursquare) could benefit from this strategy to improve their current recommendation systems, by introducing the City Image as a new criteria.

Appendix C

Cultural Analysis of Individuals

In this Appendix, we use the map of preferences presented in Section 5.5.1.3 to analyze the individual preferences of users, showing, among other results, that food and drink preferences are good indicators of cultural similarities.

In order to assess the cultural similarities among users, we construct a similarity network $G_s = (V_s, E_s)$, where s is a similarity threshold used to build the network, vertices V_s represent the set of users, and an edge (v_i, v_j) exists in E_s if users v_i and v_j have a similarity score above s . The similarity score $s_{i,j}$ between two users v_i and v_j is the Jaccard index (JI) between their preference vectors¹ multiplied by 100. In this way, $s_{i,j}$ varies from 0 to 100 and measures the percentage of preferences shared by the users v_i and v_j . For example, considering a similarity threshold $s = 65$ (or 65%-network²), there is an edge between vertices v_1 and v_2 if the corresponding users have, at least, 65% of preferences in common. We have built two similarities networks: G_s^1 ; and G_s^2 . The network G_s^1 considers only food and drink preferences, i.e., only check-ins at food and drink places. On the other hand, G_s^2 consider all preferences, i.e., all Foursquare subcategories, including food and drink venues. To build both networks we consider only the users who performed at least 7 check-ins in the dataset (i.e., at least one check-in per day on average). In total, 28,038 users were considered in G_s^1 and 194,902 in G_s^2 . Moreover, isolated nodes were disregarded. We here consider the following values of $s \in \{65, 70, 75, 80, 85, 90, 95, 100\}$. Note that G_s^1 and G_s^2 are undirected unweight and symmetric graphs.

We first analyze relevant properties of G_s^1 and G_s^2 . Figure C.1a shows the percentage of vertices (i.e., users) in the two largest components of the network G_s^1 , for various values of s (figure omitted for the network G_s^2 due to space limitations). Figure C.1a shows that the largest component of the 65%-network practically contains all nodes. The percentage of

¹The Jaccard index of sets A and B is computed as $\frac{A \cap B}{A \cup B}$.

²Network created with a threshold s is referred to as s -network.

users in the largest component slowly decreases as the similarity threshold increases, until s reaches 85. For larger values of s , the number of users in the largest component drops sharply, becoming comparable to the size of the second largest component. This is explained by observing networks built using large values for s , such as the 100%-network, where every component is composed of very similar users. Since users with very similar preferences are rare, the largest components tend not to have very large differences in size. We note that the results for the network G_s^2 are similar to those observed for the network G_s^1 , for example, the largest component of the 65%-network also contains practically all nodes.

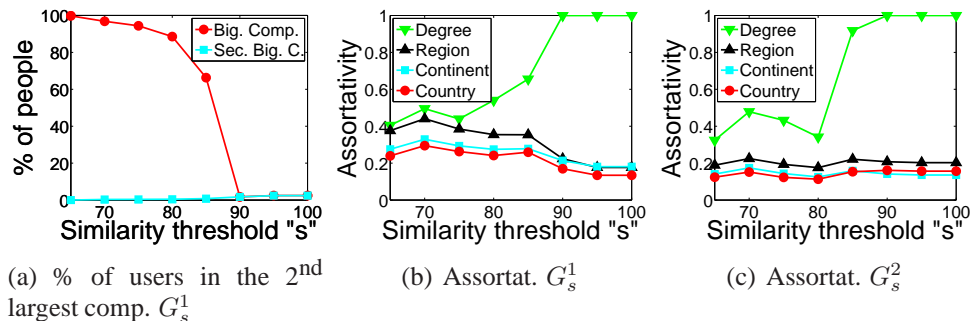


Figure C.1: General metrics for all similarity networks.

In order to verify the tendency of users from the same region to be connected, we calculate the assortativity of the similarity networks. Assortativity measures the similarity of connections in the network with respect to a given attribute, and varies from -1 to $+1$ [Newman, 2002]. In an *assortative network* (with positive assortativity), vertices with similar values of the given attribute (e.g., same country) tend to connect with (be similar to) each other, whereas in a *disassortative network* (with negative assortativity), the opposite happens. The assortativity analysis for the networks G_s^1 and G_s^2 formed from various values of s are shown in Figures C.1b and C.1c, respectively. Note that the assortativity for the network G_s^1 with respect to the geographical attributes (region Western/Eastern, continent, and country) decreases with the similarity threshold. This happens because most of the edges in the networks, formed from similarity threshold $s \geq 90$, connect users who have preference vectors with a few positive features (as defined in Section 5.5.1.3). This also helps to explain why, in both figures, the degree assortativity increases with the similarity threshold: considering only very particular tastes, the network tends to be composed mostly of cliques, making the degree assortativity very close to 1.

On the other hand, if we vary the value of s in the network G_s^2 , the assortativity for geographical attributes remains roughly the same. It is possible to explain this behavior by looking at the size of the preference vector F for the network G_s^1 , which is much smaller compared to that for the network G_s^2 (101 against 435). Since the preferences are distributed

over almost all the categories, a larger preference vector implies a lower probability of having preferences in common between two users, and, consequently, fewer edges in a similarity network, even for lower values of s . Note also that, in both Figures C.1b and C.1c, all similarity networks we take into consideration are assortative. However, the assortativity values of the geographical attributes for G_s^1 are most of the time higher compared to those obtained for G_s^2 . When considering all preferences/features we also increase the number of features that do not discriminate cultural differences sufficiently well (e.g., venues like homes, hotels, student centers, and shoe stores), since they are essentially present in all the cities and countries in the world. This suggests that, in this case, a similarity network considering only food and drink preferences might provide better insights in the study of cultural differences.

