Ph.D. Thesis

Human Mobility, Social Networks and Economic Development: a Data Science perspective

Luca Pappalardo

Supervisor
Dino Pedreschi

Supervisor
Fosca Giannotti
Art is anything you can do well. Anything you can do with Quality.

R. M. Pirsig
Abstract

Understanding human social behavior is a longstanding dream of mankind, a really profound point from both pragmatic and philosophical perspectives. The ability of drawing a comprehensive picture of human behavior and dynamics is helpful in many problems, which characterize our modern and complex society: the prevention of devastating pandemic diseases; the diffusion of new ideas or technologies over a social network; the patterns of success in different spheres of our activities. Big Data are nowadays a powerful social microscope which paves the road to realize the dream, allowing to “photograph” the main aspects of the society and to create a comprehensive picture of human behavior.

This thesis proposes to study human behavior and dynamics through a combination of techniques from network science and data mining. In the context of human mobility, we use mobile phone data and GPS trajectories from vehicles to show that people can be profiled into two distinct categories, namely returners and explorers, according to their recurrent mobility patterns. We construct a new mobility model that can reproduce the observed dichotomy and show that returners and explorers play a distinct quantifiable role in spreading phenomena. We also investigate the issue of activity recognition from human movements by presenting a classification model to recognize the activity performed by an individual by observing some characteristics of her movements.

We then move from individuals to connections, entering the domain of social network analysis. We investigate the challenging problem of community detection in dynamic social networks presenting Tiles, an innovative algorithm able to track the history of social communities in a streaming fashion. We also address the fascinating problem of the information diffusion over a social network, studying the spreading of musical tastes over a music social media. We show that certain individuals act as musical leader or innovators and that they can generate three different patterns of diffusion.

Finally, we investigate the potentiality of Big Data in providing estimate for the socioeconomic development of a territory. We use mobile phone data and GPS trajectories from vehicles to show that human mobility, and mobility diversity in particular, is highly correlated to wellbeing at municipality and province level. Individuals’ movements and quality of life are linked aspects of society, opening the scenario for the definition of new statistical index that rely on Big Data to monitor the economic health of a territory.

We conclude the thesis by revising the most promising research directions which open up from the results summarized in the thesis and introducing other interesting aspects related to the data-driven study of human behavior and dynamics.
Acknowledgments

PhD is not a journey you can do alone. Rather than a period of study, it’s a period of change. It is a complex mosaic made of all the friends, acquaintances, colleagues, random encounters, temporary loves, cities, rooms, houses, books, failures and successes I experienced during these four adventurous, hard and wonderful years.

First of all, thanks to Valentina who accompanied me during the first and most difficult year of PhD. Involuntarily, she taught me that crisis and change are the best blessing that can happen to people and that creativity is born from the distress. Thanks also to Clara, who is on my mind while I am writing these words, just for being as special as she is.

Thanks to my best friends from Salerno for loving me regardless the geographical distance: Flavio, Enrico, Nicola, Grazia, Alessandro, Francesco R., Francesco C., Roberto, Mattia F., Valentina, Angela, Sarah, Gemma, Brunella. Thanks also to the rockers Antonio, Giulio, Giancarlo, Luigi, Paolo for the beautiful rock nights in Salerno, and to Andrea for the crazy ones in the USA. Thanks to my Boston friends Georgios, Gabriele, Mario and all the guys of the Barabási lab, thanks to Michele&Clara and Fabrizio&Monica for the hospitality. Thanks to my family who often supported me morally and economically. A great thank to my Pisa friends, for the profound thoughts and the stupid ones, for the suggestions and the criticism, for their humor and rare intelligence: Matteo, Lillo, Gianluca, Francesco, Davide B., Damiano, Sabino, Davide P., Alex, Luigi, Michele, Jacopo, Antonio, Manuela, Antonella, Roberto, and especially the brazilians Igo, Vinicius and Amilcar. Thanks also to my “Hungarian” bro, Michele, for hosting me during the period in Budapest and especially for the unforgettable endless night conversations.

I owe much to my colleagues at KDDLab for stimulating the passion in the beautiful discipline of data science, for the great collaborations and suggestions, and for the funny work days: Giulio, Diego, Paolo, Lorenzo, Barbara, Chiara F., Chiara R., Roberto, Mirco, Vittorio, Anna, Letizia, Francesca, Riccardo. A very special thank to Salvo for his teachings and for his cheerfulness that made my labor entertaining and informative. Thank you to Giovanni and Pierpaolo for the intense nights working together eating and drinking.

A very special thanks to my master and friend Filippo for his competence, sympathy, humility and patience, to János Kertész for his kindness and professionalism, and to László Barabási for his time and his precious suggestions.

“Fortune is more important than talent”, Woody Allen is used to say. I agree with him. Opportunities determine the important changes of our lives, and opportunities generally happen by chance. My big chance was to meet my supervisors, Dino and Fosca,
who believed in me, gave me the opportunity to travel, meet new people, study, make errors, fail, learn, success and then produce this thesis. For all these reasons and for their invaluable contribution, this thesis is dedicated to them.
## Contents

1 Introduction

I Setting the stage

2 State of the art
   2.1 Human Mobility
      2.1.1 Motion Models: pollen and animals
      2.1.2 Human Mobility Patterns
      2.1.3 Predictability of Human Mobility
      2.1.4 Models of Human Mobility
      2.1.5 Mobility Data Mining
   2.2 Network Science
      2.2.1 Erdős-Rényi model
      2.2.2 Small worlds
      2.2.3 Scale-free networks
      2.2.4 Community Structure
   2.3 Human Mobility and Social Networks
   2.4 Data Mining and Network Science: a vision of convergence

3 The Social Microscope
   3.1 GPS data
   3.2 GSM data
   3.3 Properties of Mobility Data
   3.4 LastFM data

II Understanding Human Mobility

4 The patterns of car travel
   4.1 Understanding human mobility by car
   4.2 Inferring the traffic count by GPS data
   4.3 Conclusions


5 Returners and Explorers: dichotomy in Human Mobility
   5.1 Measuring Human Mobility ........................................... 72
   5.2 Returners and Explorers ............................................. 73
   5.3 Mobility Models ..................................................... 77
   5.4 The impact of returners and explorers on spreading phenomena. ........ 78
   5.5 Discussion .......................................................... 81
   5.6 Technical details and Supplementary Information ....................... 82
      5.6.1 Data Filtering ................................................... 82
      5.6.2 Mobility Measures .............................................. 83
      5.6.3 The patterns of returners and explorers ......................... 83
      5.6.4 Split methods .................................................. 86
      5.6.5 Mobility Models ................................................. 87

6 Learning Activities from Individual Mobility Networks ................... 95
   6.1 Related Works ....................................................... 96
   6.2 Individual Mobility Networks ...................................... 97
   6.3 Individual Activity Recognition .................................. 99
      6.3.1 Individual Mobility features .................................. 99
      6.3.2 Activity-Based Cascade Classification ......................... 102
   6.4 Experiments and Evaluations ..................................... 103
      6.4.1 Competing classification approaches .......................... 104
      6.4.2 Accuracy evaluation ........................................... 105
      6.4.3 Results ........................................................ 106
   6.5 Conclusions ......................................................... 107

III Understanding Social Networks ........................................ 111

7 Community Discovery in Dynamic Social Networks ....................... 113
   7.1 Evolutionary Community Discovery .................................. 115
   7.2 TILES algorithm ..................................................... 115
      7.2.1 TILES properties ................................................ 118
   7.3 Experiments ........................................................ 119
      7.3.1 Evaluation on Synthetic networks ............................... 119
      7.3.2 Datasets ........................................................ 121
      7.3.3 TILES communities characterization .......................... 121
   7.4 Conclusions ......................................................... 124

8 The local diffusion patterns of musical listenings ....................... 127
   8.1 Leader Detection ..................................................... 129
   8.2 Measures .......................................................... 131
   8.3 Experiments ........................................................ 133
      8.3.1 Data .......................................................... 133
8.3.2 Characterization of the Measures ........................................ 134
8.3.3 Local spreading patterns of musical genres ............................... 136
8.4 Related Works ........................................................................ 140
8.5 Conclusions ........................................................................... 140

IV Understanding the interplay between Human Behavior and Economic Development 143

9 Human Mobility, Social Networks, and Economic Development 145
  9.1 Measuring Human Behavior .................................................... 146
     9.1.1 Socio-Economic Data ....................................................... 146
     9.1.2 CDR data .................................................................... 147
     9.1.3 Social and Mobility measures based on CDR data ............... 149
  9.2 Results ................................................................................ 152
  9.3 Discussion ........................................................................... 154
  9.4 Conclusions ......................................................................... 155

10 Small area estimators using Big Data sources 163
  10.1 Big Data in small area estimation .......................................... 164
  10.2 Big Data and small area estimation methods ........................... 165
  10.3 Big Data as covariates in area level models ........................... 167
  10.4 Discussion ......................................................................... 172

11 Conclusions and Future Works 175

Bibliography 179
List of Figures

2.1 Some examples of Brownian motions. .............................. 27
2.2 Brownian motion and Lévy-flights ............................... 28
2.3 Short time trajectories of dollar bills in the United States. .... 29
2.4 Distribution of radius of gyration. ................................. 30
2.5 Distribution of mobility entropies. ............................... 31
2.6 The Erdős-Rényi model. ........................................... 38
2.7 The Small-World model ........................................... 39
2.8 The Kleinberg model ............................................. 40
2.9 The Barabási-Albert model ....................................... 41

3.1 Spatial distribution of GPS trajectories in the Octo dataset. .... 52

4.1 The general patterns of car travels ............................... 61
4.2 Spatial distribution of centers of mass in central Italy. ......... 62
4.3 Spatial distribution of most frequent locations in central Italy. .. 64
4.4 Visited locations over time and visitation frequency ............. 65
4.5 Distribution of fraction of time spent ............................ 66
4.6 Traffic sensed by VMP and GPS in Pisa .......................... 67
4.7 Wavelet decomposition for VMP and GPS .......................... 68
4.8 Real traffic versus the scaled GPS signal ........................ 68
4.9 Real traffic versus models ....................................... 70

5.1 A visualization of the complexity of the explored mobility patterns .... 72
5.2 The distributions of k-radii and total radii ........................ 74
5.3 The correlation between recurrent and overall mobility .......... 75
5.4 The ratio between recurrent and overall mobility ................ 76
5.5 The individual mobility networks of returners and explorers .... 77
5.6 EPR model predictions ........................................... 79
5.7 Temporal evolution of geographic spread for returners and explorers ... 80
5.8 The impact of returners and explorers on spreading phenomena .... 81
5.9 Mean frequency/time spent by users in the most frequent locations .... 84
5.10 Scatterplots of total $g$ versus the distance between the two most frequent locations ........................................ 84
5.11 Scatterplots of total $g$ versus the cluster-$g^{(k)}$ .................. 85
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.12</td>
<td>Split methods for returners and explorers</td>
<td>86</td>
</tr>
<tr>
<td>5.13</td>
<td>Number of returners and explorers with $k$</td>
<td>88</td>
</tr>
<tr>
<td>5.14</td>
<td>Values of the Pearson coefficient of the correlation between total $r_g$ and the distance between the two most frequent locations</td>
<td>88</td>
</tr>
<tr>
<td>5.15</td>
<td>Correlation between total $r_g$ and $r^{(k)}_g$ for the EPR datasets</td>
<td>90</td>
</tr>
<tr>
<td>5.16</td>
<td>Distribution of the ratio $s_k$ for the EPR models</td>
<td>91</td>
</tr>
<tr>
<td>5.17</td>
<td>Correlation between total $r_g$ and the distance between the two most frequent locations for the EPR models</td>
<td>92</td>
</tr>
<tr>
<td>5.18</td>
<td>Distribution of total $r_g$ for the EPR models</td>
<td>93</td>
</tr>
<tr>
<td>6.1</td>
<td>Example of Individual Mobility Network</td>
<td>98</td>
</tr>
<tr>
<td>6.2</td>
<td>Activity distribution for the trips of the volunteers in our dataset</td>
<td>104</td>
</tr>
<tr>
<td>6.3</td>
<td>Accuracy comparison between CRF and RF</td>
<td>106</td>
</tr>
<tr>
<td>6.4</td>
<td>Distribution of precision for 10k executions of the ABC classifier</td>
<td>109</td>
</tr>
<tr>
<td>6.5</td>
<td>Distribution of activities per hour</td>
<td>109</td>
</tr>
<tr>
<td>6.6</td>
<td>Distribution of activities per hour excluding “going home” activity</td>
<td>110</td>
</tr>
<tr>
<td>7.1</td>
<td>Communities identified by a static community detection algorithm</td>
<td>114</td>
</tr>
<tr>
<td>7.2</td>
<td>Appearance of a new edge</td>
<td>116</td>
</tr>
<tr>
<td>7.3</td>
<td>Community growth in Tiles</td>
<td>118</td>
</tr>
<tr>
<td>7.4</td>
<td>Comparison between Tiles and other community detection algorithms</td>
<td>120</td>
</tr>
<tr>
<td>7.5</td>
<td>Distributions of size, community per node, and clustering coefficient of communities</td>
<td>122</td>
</tr>
<tr>
<td>7.6</td>
<td>Average community similarity, community coverage, and distribution of transition time</td>
<td>123</td>
</tr>
<tr>
<td>8.1</td>
<td>Example of leader detection</td>
<td>129</td>
</tr>
<tr>
<td>8.2</td>
<td>Distributions of nodes’ degree and listeners per artist</td>
<td>133</td>
</tr>
<tr>
<td>8.3</td>
<td>RCA scores of genres in clusters</td>
<td>136</td>
</tr>
<tr>
<td>8.4</td>
<td>Distribution of objects and genres per leader</td>
<td>138</td>
</tr>
<tr>
<td>8.5</td>
<td>Distributions of leaders’ degree, centrality and clustering</td>
<td>138</td>
</tr>
<tr>
<td>9.1</td>
<td>Distributions of the economic measures</td>
<td>148</td>
</tr>
<tr>
<td>9.2</td>
<td>Distributions of mobility and social measures</td>
<td>151</td>
</tr>
<tr>
<td>9.3</td>
<td>Spatial distribution of Orange users over the French municipalities</td>
<td>152</td>
</tr>
<tr>
<td>9.4</td>
<td>Relation between diversity measures and economic measures</td>
<td>157</td>
</tr>
<tr>
<td>9.5</td>
<td>Distributions of mobility entropy and D-social in the deciles of the EDI</td>
<td>158</td>
</tr>
<tr>
<td>9.6</td>
<td>Relation between diversity measures and economic measures for null model NM1</td>
<td>159</td>
</tr>
<tr>
<td>9.7</td>
<td>Relation between the diversity measures and the economic measures for null model NM2</td>
<td>160</td>
</tr>
<tr>
<td>9.8</td>
<td>The mean values of mobility entropy and D-social for groups of municipalities based on the deciles of the economic indicators</td>
<td>161</td>
</tr>
</tbody>
</table>
0.0. LIST OF FIGURES

10.1 Map of HCR and $S_{E(d)}$ for Tuscany provinces .................................. 168
10.2 Estimates of the HEI and HCR measures ................................................. 171
10.3 Model estimates versus direct estimates ..................................................... 172
## List of Tables

2.1 Average path length and clustering coefficient of real networks. . . . . . . . 38  
3.1 Dataset used in the thesis. . . . . . . . . . . . . . . . . . . . . . . . . . . . . 50  
3.2 Example of GSM phone records. . . . . . . . . . . . . . . . . . . . . . . . . 53  
5.1 Example of GPS records . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 83  
5.2 Summary of properties of the datasets. . . . . . . . . . . . . . . . . . . . . 83  
6.1 Features summarizing the main aspects of individual mobility. . . . . . . . 101  
6.2 Examples of features extracted from a IMN . . . . . . . . . . . . . . . . . . . 102  
6.3 Precision and Recall of the RF and ABC classifiers . . . . . . . . . . . . . 108  
7.1 Datasets statistics used in Tiles experiments . . . . . . . . . . . . . . . . . . . 121  
8.1 Pearson correlation coefficient between width, depth, strength and other  
   network measures . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 134  
8.2 Partial correlation and p-value of clustering and closeness centrality with  
   width . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 135  
8.3 Presence of different diffusion patterns per genre. . . . . . . . . . . . . . . 139  
9.1 Socio-economic indicators derived from the INSEE . . . . . . . . . . . . . . 147  
10.1 Variables used as covariates . . . . . . . . . . . . . . . . . . . . . . . . . . . 170
Chapter 1

Introduction

Is it possible to observe, understand, describe and predict human activities? Answering this critical question is a longstanding dream of mankind, a really profound point from both a pragmatic and a philosophical perspective. The ability of drawing a comprehensive picture of human behavior means realizing the idea the we can know the present and the future, and profit from this knowledge by adjusting our policy to it. The laws of our movements could help design new smart sustainable cities or prevent the outbreak of devastating pandemic diseases; the patterns of shopping behavior would be valuable for suggesting to customers products according to their disparate tastes; the mechanism of social interactions could be exploited to speed up the diffusion of new ideas or technologies; the patterns of success might be used to improve our performance in every sphere of our activities. In the same way we rely on the laws of physics to manipulate the environment and build new technologies, we could take advantage of the laws of human behavior to control our society and bring it to a better future.

In 1959, however, the illustrious philosopher Karl Popper denied the possibility of a historicists’ doctrine: when humans are involved, prediction is impossible because history is not repetitive and neither are our motivations and desires [1]. So, in contrast with many natural phenomena like lunar eclipses or weather conditions, we cannot understand human actions so deeply to predict future movements, interactions, wars, revolutions. Popper’s negative answer, due his authority, set the agenda of social sciences for the decades to come. At that time, the shared sensation that we cannot lay bare the laws of our complex society was presumably driven by the lack of a suitable tool to capture what is going on in our society. Observation, and the increasingly sophisticated tools that allow it to be performed, are the first and fundamental part in the scientific process of discovery: Pre-Socratic philosophers began to speculate on the structure of the world by examining the stars through their own eyes; astronomers observe and understand the origin and evolution of the universe through the powerful lenses of their huge telescopes; biologists reveal the structure of cells and explain the birth of life by their advanced microscopes. The history of science has often been driven through new instruments of observation, which parted the doors to worlds unknown before. Popper and his contemporaries thought that human behavior was not scientifically understandable because at that time society was
not observable. In fact, they could not imagine that in the future a cheap, abundant, growing, big, valuable tool would be available to track people's movements, interactions, sentiments, opinions, transactions, and even eating habits. Fifty years ago, Popper could not imagine that the scientists of the new millennium would be able to observe humans like stars or cells, under the powerful lenses of Big Data.

Every day we produce huge loads of data about ourselves simply by living in our modern technological world: we click Web pages, post videos on Facebook and thoughts on Twitter, make calls on our cell phones, communicate through emails, shop with credit cards, publish our sports performance, diet and nightlife through a plethora of different online social networks. When it comes to producing data, we are really prolific: the world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s [2]. Such enormous and precious amount of data represent the digital trace of our everyday actions, usually referred as Big Data due their size and complexity [3]. Big Data are greedily collected by institutions and companies which spy us with the purpose of creating an ever-changing, up-to-the-minute mosaic of human behavior. We are just at the beginning of a data revolution, which will profoundly impact all aspects of the society: government, business, science and entertainment [4, 5, 6]. Big Data, indeed, are finally the tool of observation that was missing at the time of Popper, a powerful social microscope which allows scientists to observe, understand, describe and even predict many aspects of human behavior. Such a social microscope allows us to “photograph” the main aspects of the society, and to use the picture to understand more complex and often hardly observable features, such as the socio-economic status of a territory. The quantity and the quality of our actions, indeed, clearly depend on and are influenced by the economic condition of the territory where they take place. The monitoring of the economic status of a territory is responsibility of research and statistics institutes, which collect economic data by means of surveys and censuses. Censuses are complex and expensive to carry out, so surveys represent the feasible way to collect data. However, surveys refer to a tiny sample of the population and due their complexity are updated after months, preventing an effective “nowcasting” of the current economic health of a territory. Big Data pave the road to overcome such shortcomings. The understanding of the relationships between human activities and economic well-being is a key step toward the creation of a sociometer: a virtual instrument to measure in real-time the health condition of a population.

The interest around the analysis of Big Data and the possibility to compile them into a comprehensive picture of human behavior have infected all branches of human knowledge, from sports to economy. However, two aspects in particular attracted the interest of scientists in the last decade, due the striking abundance of data in those contexts: human mobility and social networks.

The last decade has witnessed an abundance of novel insights on the underlying patterns of human mobility triggered by the emergence of big mobility data, i.e. massive datasets of digital traces of human whereabouts which portray mobile activity at unprecedented scale and detail. Examples include the satellite-enabled Global Positioning Systems (GPS) and the mobile phone networks, that allow for sensing and collecting
society-wide proxies of human mobility. This new social microscope has attracted scientists from diverse disciplines, especially from network science [7, 8, 9, 10] and data mining [11, 12, 13, 14], and has fueled advances from public health to transportation engineering, urban planning, and the design of smart cities. Across all studies, there is a consensus that a stunning heterogeneity of the individual travel patterns coexists with a high degree of predictability. Humans exhibit a broad diversity of mobility behaviors combined with repeating regular patterns, dictated by daily routines. Although these discoveries have shed light on interesting aspects of human mobility, many key aspects are still unclear and open for future research: Which are the factors that shape the observed heterogeneity and predictability? Can we predict the type of movements individuals perform by observing just their movements?

Social networks are the other hot topic which recently attracted the attention of scholars ranging from sociology to statistical physics, partly due to the public interest in the last generation social media like Facebook and Twitter. The analysis of social interactions data provided by emails, the call graph, Facebook and Twitter and so on revealed for the first time the complexity underlying the social networks. Hubs exist in our social networks, who strongly contribute to the “six degrees of separation”: in average, an individual is separated from anyone else on the earth by six other individuals. Another characteristic is the community structure, the tendency of the social network to partition into social clusters of densely connected sets of individuals. In literature, there is no unique formal definition of a network community since the problem of community detection is still open. In the context of social network analysis, another hot problem regards information diffusion: How do diseases, information, ideas, innovations and technologies spread over a social network? Can we model such diffusion to make realistic simulations?

In the present thesis we aim to exploit the lenses of Big Data to answer some of the most challenging and critical open questions in the understanding of human mobility. Relying on big mobility data about car travels and phone calls, we go deeply into the complexity of the mobility sphere and investigate the impact of recurrent movements on the overall mobility of individuals. We will show that people can be profiled into two distinct categories, namely “returners” and “explorers”, according to their recurrent mobility patterns. We will show and discuss the inadequacy of state-of-the-art models and the importance of the discovery in the context of outbreak prevention. Moreover, we will also address the issue of activity recognition from human movements, providing a semantic amplifier for Big Data: a model able to recognize the activity performed by an individual by observing some characteristics of her movements.

We then move our attention from individuals to interactions, entering the domain of social network analysis. Here, we address two hot problems that characterize the study of social networks. First, we investigate the challenging problem of community detection in dynamic social networks, i.e. social networks evolving over time. State-of-the-art algorithms generally treat the community detection problem in a static way, without taking into account the fact that networks and communities evolve over time. Social interactions do not occur with a rigid temporal discretization, but they flow in streaming as time goes
by. Consequently, the social communities also have to change fluidly over time. We will present Tiles, an innovative algorithm able to track the history of social communities in a streaming fashion. Second, we address the fascinating problem of the information diffusion over a social network, studying the spreading of musical tastes over a music social media. Here, we show that certain individuals act as musical “leaders” or innovators, and that they can generate three different patterns of diffusion, providing a characterization of musical genres based on the discovered patterns.

Finally, we leverage on the acquired knowledge about human mobility and social networks to study their relations with the economic wellbeing. How are our mobility sphere and our social world connected to the economic development of the territory where they take place? We will see that by suitably adjusting the lenses of our social microscope we can uncover interesting relations, especially when we consider the diversification of our social and mobile activities at a city level. We will discover that we are part of a complex organism where our mobile and social behavior feeds, and is fed by, the economic development of the territories. In this context, Big Data have a high potential in providing representative, relatively inexpensive and readily available measures as proxies of socio-economic indicators of poverty, well-being and progress.

The methodology used during our research adventure constitutes the main objective of the present thesis. We advocate the need of a convergence of network/complexity science and data mining research and a progressive merge of the two scientific methodologies. Statistical physics and network science aim at discovering the global models of complex social phenomena, by means of statistical macro-laws governing basic quantities. On the other hand, data mining aims at discovering local patterns of complex social phenomena, by means of micro-laws governing behavioral similarity or regularities in sub-populations. The need of such dualistic approach is clear in the case of human mobility. In the overall set of individual trajectories across a large city researchers observed a huge diversity: while most travels are short, a small but significant fragment of travels are extraordinarily long. Despite this complexity represented in the data, mobility data mining can automatically discover travel patterns corresponding to set of travelers with similar mobility. The above dual scenario of global diversity and local regularity is perceived today as the signature of social phenomena, and represents a foundational tenet of computational social sciences. Although network science and data mining emerged from different scientific communities using largely different tools, we need to reconcile the macro/global approach of the first with the micro/local approach of the second within a unifying theoretical framework, because each can benefit from the other and together have the potential to support realistic and accurate models for simulation and what-if reasoning of social phenomena.

The thesis is organized in four parts. The first part, Setting the Stage, is devoted to the issue related to the understanding of human behavior through the analysis of Big Data. We firstly summarize in Chapter 2 the history of the study of human behavior, both in the mobility and in the social context. Here, we revise the main discoveries and contributions from network science and data mining. We will then present in Chapter 3 the social microscope we used in our studies, i.e. the Big Data we exploited, together with
some issue related to the analysis of the different kinds of data.

In the second part, Understanding Human Mobility, we tackle some challenging open questions in human mobility. Chapter 4 is devoted to the analysis of car travels as prelude for Chapter 5, where we investigate the existence of a dichotomy in human mobility. Chapter 6 is dedicated to the definition and implementation of the ABC activity classification model as semantic amplifier for Big Data. The third part, Understanding Social Networks, is dedicated to two interesting problems in social network analysis: community detection in dynamic social networks (Chapter 7), and the patterns of diffusion of musical tastes (Chapter 8).

The fourth part investigates the interplay between the different aspects of human behavior. Chapter 9 and Chapter 10 analyze human mobility and social interactions in two different territories, France and Tuscany, and show that the way we move and interact is related to the economic environment where our activities take place. Finally, Chapter 11 concludes the thesis by presenting possible future research directions in the study of human behavior.
Part I

Setting the stage
Chapter 2

State of the art

Fueled by big data collected by a wide range of high-throughput tools and technologies, a new wave of data-driven and interdisciplinary science have rapidly proliferated during the past decade, impacting a wide array of disciplines, from physics and computer science to cell biology and economics [5]. In particular, the ICT’s are inundating us with huge amounts of information about human activities, offering access to observing and measuring human behavior at an unprecedented level of details. These large-scale datasets offer objective description on human activity patterns and have started to reshape our discussions on quantifying and understanding human behavior. An impressive shift has been witnessed in statistical physics, complex system theory and data mining since the beginning of the new millennium, when the possibility of analyzing large datasets of human activities has boosted a renewed interest in the study of human mobility and social networks.

The understanding of how humans move has attracted particular interest in recent years, due to the data availability and to the relevance of the topic in various domains, from urban planning and virus spreading to emergency response. The study of social networks has its roots in the graph solution of the famous “Königsberg bridges” problem proposed by Euler in the 18th century. In the last decade, motivated by the ubiquity of networks in our contemporary society (Internet, the Web, Facebook, Twitter and so on), scientists from many disciplines dusted off the seminal studies of the past century and started to unveil the structure of complex networks, and social networks in particular.

The contribution of this chapter is to provide a brief account of this body of research, with a focus on recent results on the empirical laws that govern the movements and interactions of individuals. We first discuss how the key variables of people’s travels can be described by universal laws, validated against different datasets of real observations. We also discuss how predictable people’s movements are, and present the main mobility models able to describe in a realistic manner salient aspects of human mobility. Next, we move from individuals to interactions among individuals, and enter the domain of social network analysis. An extraordinary effort has been devoted to understand the interconnectedness of individuals, i.e. the structure of the social networks we inhabit, and how this structure influences social phenomena, such as the diffusion of information or
the formation of communities. We provide a brief account of the key findings of network science so far, for the purpose of discussing the recent results on how human mobility shapes and impacts social relations, and the other way around. Again, empirical laws were found that offer quantitative accounts of the intuition that people from the same social circles tend to be co-located in space and time more than people that are far apart in the social network. Building on this relation among social and mobility variables, it is possible to shed more light on how social and mobile behavior evolve over time.

The results surveyed in this chapter are the basic tools for research in human dynamics, and are at the convergence between data mining research and network science research. Such convergence is the starting point of the present thesis, aimed at combining the analytical power of statistical physics and knowledge discovery to deeply understand human behavior dynamics.

2.1 Human Mobility

We live in an era in which understanding individual mobility patterns is of fundamental importance for epidemic preventions and urban planning. Human movements are inherently massive, dynamical, and complex. Aided by modern transportation technologies, we can now travel to any place on the globe in just one day or two. On the other hand, while the mobility of our fellow species is mainly governed by mating needs and food resources, human mobility is fundamentally driven by ourselves, from job-imposed restrictions and family related programs to involvement in routine and social activities. Therefore, quantifying the regularities and singularities behind human movements had remained as an often elusive goal. Thanks to the availability of large-scale datasets generated by various domains of modern technologies, ranging from registration of dollar bills to mobile phone services and GPS devices to location based websites, we have witnessed a proliferation of studies on human mobility. In this section, we will start from the most fundamental models for motions, rooting back to the 19th century. We will then move to the present describing several recent empirical observations of human mobility and the new generation of mobility models provided in the context of two main scientific fields: network science and data mining.

2.1.1 Motion Models: pollen and animals

In 1827, while he was studying sexual relations of plants, botanist Robert Brown noticed that granules contained in grains of pollen were in constant motion, and that this motion was not caused by currents in the fluid or evaporation [15]. He thought at first that they were jiggling around because they were alive or because of the organic nature of the matter. So, he did the same experiment with dead organic and inorganic matter finding there was just as much jiggling. The movement evidently had nothing to do with the substance ever being alive or dead, and this left him and his contemporaries with a puzzling question: What is this mysterious perpetuum motion that keeps the pollen moving? A possible
2.1. HUMAN MOBILITY

explanation for the so-called Brownian motion\(^1\) is that all the molecules in the fluid are in vigorous motion, and these tiny granules are moved around by this constant battering from all sides as the fluid molecules bounced off. The particle of pollen behaves like a really huge balloon in the midst of a dense crowd: as the individuals move around they push the balloon from all directions; sometimes the balloon will move to the left, occasionally to the right, overall displaying a random, jittery motion like paths in Figure 2.1.

\[ \text{\textbf{Figure 2.1: Some examples of Brownian motions.}} \]

Such atomic-molecular thesis was guessed by Einstein, who in 1905 published a theoretical analysis of Brownian motion and showed that the mean distance reached by particles from the first collision point must grow with the square root of time \([16]\). It means, for example, that after 4 seconds, the distance is only twice \((\sqrt{4} = 2)\) the one found after a second, and not four times as insight would suggest. Einstein’s calculations were confirmed experimentally in 1908 by physicist Jean Baptiste Perrin, who convinced even the most skeptical about the validity of atomic-molecular hypothesis \([17]\). Before Einstein, Louis Bachelier derived independently several mathematical properties of Brownian motion \([18]\), including the equation for the probability \(P(x, t)\) for the position \(x\) of a Brownian random walker at time \(t\), when the walker starts as the origin at time \(t = 0\). The equation for \(P(x, t)\) is given by the diffusion equation, with a Gaussian solution. Therefore, a Brownian motion is basically a random walk with a normal distribution for the position of the random walker after a time \(t\), with the variance proportional to \(t\). It means that random walkers tend to travel roughly the same distance between sightings.

Measurements on albatrosses, monkeys and marine predators \([19, 20, 21]\) suggested that animal trajectories are characterized by jumps at very large distances and can be approximated by the so-called Lévy-flight, a generalization of Brownian motions. As Figure 2.2 suggests, the length of the jumps in a Brownian motions are very similar producing a peaked distribution of jumps’ length. In contrast, for Lévy-flights the length of the jumps is highly variable, producing a power law distribution of jumps’ length.

---

\(^1\)The first observation of Brownian motion was reported in 1785 by the Dutch physician Jan Ingenhaysz. However, Brown was the first to discover the ubiquity of the phenomenon.
2.1.2 Human Mobility Patterns

Are human movements governed by Lévy-flight, like marine predators and monkeys, or do they follow their own laws? To answer the above questions, humans need to be observed under a "microscope", like Perrin observed atoms and was able to experimentally confirm Einstein’s theory. The technological era, at last, allows us to track human mobility and to test models, thanks to the exploding prevalence of mobile phones, GPS, and other handheld devices. Such devices are our social microscopes.

In 2006, Dirk Brockmann and his colleagues used the geographic circulation of bank notes in the United States as a proxy for human traffic, assuming that individuals transport money as they travel [7]. They analyzed data collected at the largest online bill-tracking Website www.wheresgeorge.com, and found that most bills remain in the vicinity of their initial entry, yet a small but a significant number have traversed distances of the order of the size of USA (Figure 2.3), consistent with the intuitive notion that short trips occur more frequently than long ones. The researchers calculated that the probability $P(r)$ of a bank note traversing a distance $r$ in two weeks period over a range of distances between 10km and 3,500km follows a power law: $P(r) \sim r^{-(1+\beta)}$, with an exponent $\beta \approx 0.6$. Moreover, they found that the typical distance $X(t)$ from the initial starting point as a function of time is a power law: $X(t) \propto t^{1/\beta}$.

As we know, for Brownian motion the distance $X(t)$ scales according to the square-root law. For a power law distribution the variance diverges for exponents $\beta < 2$, implying that bank note dispersal lacks a typical length scale, resembling Lévy-flights. Lévy-flights are superdiffusive: they disperse faster than ordinary random walks. This discovery was a major breakthrough in understanding human mobility on global scales. In the light of this discovery, in dispersal humans are similar to animals. However, our intuition suggests that we do not move in a completely random way. There are regularities in our lives: most of us have a home, a work, a hobby. These activities necessarily shape our trajectories. Instead, if we do follow a pure Lévy-flight we rarely find our way back home, but our position increasingly moves away from the initial one.

To further investigate human mobility patterns, starting from 2008 Barabási and his
2.1. HUMAN MOBILITY

Figure 2.3: Short time trajectories of dollar bills in the United States from www.wheresgeorge.com. Lines connect origin and destination locations of bank notes that traveled for less than a week. Figure from [7].

Team analyzed the trajectories of 1 million anonymized mobile phone users whose position was tracked for a six-month period [8, 9, 10, 22]. Contrary to bills, mobile phones are carried by the same individual during her daily routine, offering the best proxy to capture individual human trajectories. Each time an individual makes a call the mobile phone operator registers the coordinates of the cell towers communicating with the phones, effectively tracking her locations. The time-ordered list of cell phone towers from which the individual made the calls forms her global mobility trajectory. A first result from the analysis of mobile phone records apparently confirmed in a certain way observations on bank notes: the distribution of displacements $\Delta r$ between user’s positions at consecutive calls is well approximated by a truncated power law: $P(\Delta r) = (\Delta r + \Delta r_0)^{-\beta} \exp(-\Delta r/\kappa)$, with exponent $\beta = 1.75 \pm 0.15$, $\Delta r_0 = 1.5$km and $\kappa = 400$km [8]. Even if such equation suggests that human motion follows a truncated Lévy-flight, differences from randomness emerge from other measures. The distribution $P(r_g)$ of radius of gyration $r_g$, the characteristic distance traveled by a user when observed up to time $t$, also follows a power law, in contrast with random walks which show peaked distributions (Figure 2.4, left) [8]. So, humans are different from bank notes and animals since a great heterogeneity characterizes our mobility: most people usually travel in close vicinity to their home location, while a few frequently make long journeys.

González et al. [8] also detected the tendency of people to return to previously visited location with a frequency proportional to the ranking in popularity of the location with respect to other locations. They found that the probability $F_{pt}(t)$ that a user returns to the position where he was first observed after $t$ hours shows several peaks at 24h, 48h and 72h (Figure 2.4, right), capturing the recurrence and temporal periodicity inherent to human mobility [8]. Song et al. [10] extended the experiment to a larger dataset and measured the distribution of the visiting time (i.e. the time interval $\Delta t$ a user spends at
one location). The resulting curve is well approximated by a truncated power law with an exponent $\beta = 0.8 \pm 0.1$ and a cutoff $\Delta t = 17h$, which the authors connected with the typical awake period of humans. Moreover, they also found that the frequency $f_k$ at which a user visits its $k$th most visited location follows a Zipf’s law ($f_k \sim k^{-\xi}$) with parameter $\xi \approx 1.2 \pm 0.1$. This also suggests that the user visitation frequency follows $P(f) \sim f^{-(1+1/\xi)}$. Individuals display significant regularity, returning to a few highly frequented locations, such as home or work. This regularity does not apply to the bank notes: a bill always follows the trajectory of its current owner; that is, dollar bills diffuse, but humans do not.

![Figure 2.4](image)

Figure 2.4: The distribution $P(r_g)$ of the radius of gyration measured for the users. The solid line represents a truncated power-law fit. The dotted, dashed and dot-dashed curves show $P(r_g)$ obtained from random walk, pure and truncated Lévy-flights models. The picture on the right shows that the prominent peaks capture the tendency of humans to return regularly to the locations they visited before, in contrast with the smooth asymptotic behavior (solid line), predicted for random walks. Figure from [8].

### 2.1.3 Predictability of Human Mobility

What is the role of randomness in human behavior and to what degree is human behavior predictable? This question is crucial, because the quantification of the interplay between the predictable and the unforeseeable is very important in many applications. From predicting the spread of human diseases to city planning and resource management in mobile communications, our ability to foresee the whereabouts and mobility of individuals can help us to improve or save human lives.

Eagle and Pentland first studied human predictability by tracking the movements of one hundred MIT students during nine months [23, 24]. By offering them free smart phones, they collected approximately 450,000 hours of information about users location, communication and device usage behavior. Eagle and Pentland arranged the whereabouts of students into three groups (home, work and elsewhere), and attempted to identify the amount of predictable structure in a student’s life using the Shannon entropy [25]. People with high entropy are more variable and harder to predict, while low-entropy individuals
are characterized by strong patterns across all time scales. They discovered a striking level of predictability in students’ routines, reaching in some cases about the 96% accuracy in predicting future whereabouts. Even the most unpredictable moments of students were by no means random: they were indeed concentrated at Friday and Saturday night, which are typical party times in students’ life.

In 2009, Song et al. [9] also exploited the concept of entropy to provide a quantitative evaluation of the limits in predictability for human walks, using a 3-month-long mobile phone dataset of about 50,000 individuals. The authors computed for each user three entropy measures: the random entropy $S_{\text{rand}}$ in the case of location visited with equal probability; the entropy $S_{\text{unc}}$ that depends only on frequencies of visits; and the real entropy $S$ that considers the probability of finding particular time-ordered subsequences in the trajectory. As shown in Figure 2.5A, the distribution $P(S)$ of the real entropy has a peak in $S = 0.8$ indicating that the real uncertainty in a typical user’s whereabouts is $2^{0.8} \approx 1.74$. It means that a user who chooses randomly her next location could be found on average in two locations. A big difference emerges in respect to the random entropy, for which the peak at $S_{\text{rand}} = 6$ implies $2^6 = 64$ locations.

Figure 2.5: (A) The distribution of the entropies $S$, $S_{\text{rand}}$ and $S_{\text{unc}}$ across 45,000 users. (B) The distribution of $\Pi_{\text{max}}$, $\Pi_{\text{rand}}$ and $\Pi_{\text{unc}}$ across all users. (C). The dependence of $\Pi_{\text{max}}$ on the user’s radius of gyration $r_g$. For $r_g > 10\text{km}$, $\Pi_{\text{max}}$ is largely independent of $r_g$. (D). The fraction of time a user spends in the top $n$ most visited locations, the resulting measure $\tilde{\Pi}$ representing an upper bound of predictability $\Pi_{\text{max}}$. Figure from [9].

To represent the fundamental limit for each individual’s predictability, Song et al. [9] defined the probability $\Pi$ that an appropriate algorithm can predict correctly the user’s future whereabouts. If a user with entropy $S$ moves between $N$ locations, then her
predictability is bounded by the maximal predictability $\Pi_{\text{max}}(S_N)$. For a user with $\Pi_{\text{max}} = 0.2$, this means that, no matter how good the predictive algorithm is, only in the 20% of the time can we hope to predict her whereabouts. They determined $\Pi_{\text{max}}$ separately for each user and found that the distribution $P(\Pi_{\text{max}})$ is peaked around $\Pi_{\text{max}} \approx 0.93$. Figure 2.5B highlights that $\Pi_{\text{rand}}$ and $\Pi_{\text{unc}}$ are instead ineffective predictive tools.

In a historical record of the daily mobility pattern of the users there is a potential 93% average predictability in user mobility, an exceptionally high value rooted in the inherent regularity of human behavior. The most surprising is the lack of variability in predictability across the population, obtained by explored impact of home, language groups, population density and rural versus urban environment [9]. Although the population has an inherent heterogeneity, the maximal predictability varies very little, there are no users whose predictability would be under 80%. Knowing the history of a person’s movements, we can find patterns and regularities in her mobility, and to foresee her current location with extremely high success probability.

The substantial predictability of individuals’ movements also emerges from the work by Schneider et al. [26], where they investigated the mechanism responsible for the daily mobility patterns of individuals. Schneider et al. analyze the spatio-temporal trajectories of thousands of persons as individual networks, discovering that nearly the entire population can be described with a few unique daily motifs. They also analyzed how more or less likely a motif is found under the condition that the individual has a given motif on another day, finding that the probability to find the same daily motif during another day is significantly larger compared to a randomized dataset. People, hence, tend to repeat their schematic mobility behavior in time. Their observation imply that each person has a characteristic daily motif: individuals have personal number of preferred places on a daily basis, which are most likely visited in a specific sequence given by its characteristic motif.

2.1.4 Models of Human Mobility

High heterogeneity, high regularity, high predictability: these are the main patterns which characterizes the mobility of humans. How to combine such ingredients to create a realistic model which captures the salient aspects of how people move? In recent years, many approaches have been proposed to answer this challenging question [27].

Song et al. [10] proposed the Exploration and Preferential Return (EPR) model, which does not fix the set of preferred locations but allows them to emerge naturally during the evolution of the mobility process. They introduce two basic mechanisms that together describe human mobility: exploration and preferential return. Exploration is a random walk process with truncated power law jump size distribution. Preferential return reproduces the propensity of humans to return to the locations they visited frequently before. An agent in the model selects between the two modes: with probability $P_{\text{new}} = \rho S^{-\gamma}$ (where $S$ is the number of sites visited so far by the agent, $\rho$ and $\gamma$ are two model parameters) the individual moves to a new location, whose distance from the current one is chosen from the known power law distribution of displacements. With complementary probability $P_{\text{new}} = 1 - \rho S^{-\gamma}$, she returns to one of the $S$ previously visited places (with
2.1. HUMAN MOBILITY

the preference for a location proportional to the frequency of visits). As a result, the model has a warmup period of greedy exploration, while in the long run users mainly move around a set of previously visited places.

The Orbit model, first proposed in [28], generates the whole sets of locations at the beginning of the simulation. Each agent then selects a subset of these locations and moves between them based on a predefined customizable behavior. Movements are described by a Markov chain where each state represents a specific location in the scenario. A realistic distribution of reappearance frequencies is then achieved indirectly by defining probabilities of transitions between places.

In the Time-Variant Community (TVC) model [29], movements are split into time slot, during which different reference locations are associated with users. Within each time slot, a user can either move in a restricted area or freely in the whole simulation scenario. This approach is extended by the SLAW model [30], which can also reproduce the preferences for shorter trips. This is achieved by exploring two strategies: first, the locations are distributed on the area so that the distance among them follows a heavy-tailed distribution; second, the daily routes of the users are formed from the jumps between a randomly selected subset of these locations minimizing the overall trip length. Since part of this subset remains fixed for each user, the SLAW model is also able to capture regular returns to the same location.

The Swim model [31] assigns to each agent a home location, which is a randomly and uniformly chosen point in the plane. The model captures the power-law distribution of displacements allowing agents to select the destination for next moves depending on the weight of each site, which grows with the popularity of the place and decreases with the distance from the home. The popularity of a location depends on overall preferences and it is calculated as the number of other people encountered the last time the agent visited the place.

Gravity models have a long history of use in describing and forecasting the movements of individuals, goods and services. This class of models characterize the distribution of trips between locations, based on the populations of origin and destination and the distance between them. The seminal work by Zipf [32] provides a theoretical motivation for movement between two cities \(a\) and \(b\) being governed by a \(\frac{P_a P_b}{d}\) relationship, where \(P\) is the city population and \(d\) is separation distance. In other words, he argued that movements follow a sort of “gravity law”: the probability that an individual or a group of individuals move between two locations is inversely proportional to the separation distance of the two locations. Gravity models have been applied in many areas of application, such as analyzing and forecasting the demand for goods and services in populations [33] and the spread of biological agents or human diseases [34, 35, 36].

Simini et al. [22] extend the gravity model introducing the radiation model, a stochastic process able to capture the local mobility decisions of individuals, analytically deriving mobility fluxes that require as input only information on the population distribution. The radiation model predicts mobility patterns in good agreement with mobility and transport patterns observed in a wide range of phenomena, from long-term migration patterns to
communication volume between different regions. Given its parameter-free nature, the model can be applied in areas where mobility measurements are not available, significantly improving the predictive accuracy of most of the phenomena affected by mobility and transport processes.

2.1.5 Mobility Data Mining

A different perspective on the study of human mobility is provided by the data mining community, whose interest on the analysis of human movements, in form of trajectories, generated the new sub-field called mobility data mining [37]. While statistical physics is aimed at discovering the global models of human mobility, by means of statistical macro-laws governing basic quantities, mobility data mining is aimed at discovering local mobility patterns, by means of micro-laws governing behavioral similarity or regularities in sub-populations.

Thanks to the advances in mobile communication and positioning technology, large amounts of moving objects data can be collected in form of trajectories. The extraction of spatio-temporal patterns in trajectories, which represent movement patterns of individuals, provides useful information in contexts like traffic flow control or location-aware advertising [38]. The Data Mining analysis step of the Knowledge Discovery process in Databases (or KDD) [39], a relatively young and interdisciplinary field of computer science, is the most intuitive and attractive approach to describe structure of trajectories and to extract frequent spatio-temporal patterns. Mining spatio-temporal patterns means searching for concise representations of interesting behaviors of single or groups of moving objects. An example of this process is the mining of frequent spatio-temporal patterns, a problem commonly addressed by a feature extraction solution: sets of features are derived from the data, yielding spatio-temporal predicates that describe each trajectory, and then generic mining algorithms are applied on the new feature-based representation of data, extracting frequent sets, association rules or frequent sequences of features. Clearly, the choice of the attributes to extract is a crucial aspect of the process. A basic family of trajectory features consists of individual-based features, such as spatial and temporal aggregates (length of the path covered, amount of time spent in a place, min/max/average velocity, most frequent direction followed, etc.); spatial events (visiting some pre-defined spatial regions or visiting twice the same place); and spatio-temporal events (temporally localized manoeuvres like performing U-turns, abrupt stops, sudden accelerations etc.).

The opposite alternative to the feature-based approach is the direct analysis of trajectories. In this case no pre-processing on the trajectories is performed, making it almost impossible to see a configuration occurring more than once perfectly in the same way, thus some kind of tolerance to small perturbations is needed. To address such problems, in [40] a trajectory is approximated by means of a sequence of spatial segments obtained through a simplification step. Then patterns are extracted in the form of sequences of contiguous spatial segments. Frequent sequences are represented as sequences of rectangles such that their width quantifies both the average distance between each segment and the points in the trajectory it covers. In another work [41] authors define the spatio-temporal periodic
2.1. HUMAN MOBILITY

pattern mining problem, and propose a mining algorithm for retrieving maximal periodic patterns. The time is simply assumed to be discrete, and spatial locations are discretized dynamically through a density-based clustering algorithm. Each time a periodic pattern is generated in the form of a sequence of spatial regions, a check is performed to ensure that all regions in the pattern are dense. A variant of the problems mentioned above has been proposed in [13]. In this work, the notion of T-pattern is introduced as a sequence of points in space with transition times, which express the time taken to move from each point to the next one in the sequence. A pattern of $n$ elements is any sub-sequence of $n$ points of a trajectory, such that each point in the sub-sequence falls within a spatial neighborhood of the corresponding point in the pattern, and transition times are approximately the same as in the pattern up to a given time tolerance threshold. Then, frequent T-patterns are extracted, by heuristically grouping close spatial points into rectangular regions, and representing sets of similar transition times through intervals.

A completely different issue in analyzing large quantities of mobility data is to divide the dataset into logically distinct groups, such that the objects in the same group are more similar that the others outside. In data mining this issue is addressed by clustering [39], consisting essentially in trying to outline groups of individuals that show similar behaviors. There are two main families of approaches to the clustering problem: applying generic clustering algorithms by defining some distance function between trajectories, or defining ad hoc notions and algorithms tailored around the specific data type [37]. In the former case, two individuals are considered similar if they follow approximately the same spatio-temporal trajectory: at each time instant (with some tolerance) they are approximately in the same place. Imposing both temporal and spatial constraints is sometimes too restrictive. Hence, the temporal constraint is often removed defining two objects to be similar if they simply follow the same route, at any moment in time. As a further simplification step, we can group objects that simply perform similar movements, like going in the same direction or performing the same turns. This is equivalent to finding groups of objects that perform similar sequences of changes (or non-changes) in their direction.

Ad-hoc approaches aim to find groups of individuals which follow a common trajectory, allowing a limited amount of random noise. Some techniques group together people that are likely to be generated from a common core trajectory by adding Gaussian noise, while other techniques use Hidden Markov models to model clusters, and adopt a mixture model approach for the parameter estimation task [37]. Alternative methods find groups of objects that move together within some time interval of minimum size. For example, trajectories are clustered using a density-based algorithm where the adopted distance is the average spatial distance between the trajectories within a given time interval. Then, for each time interval, the algorithm can be run focusing on the trajectory segments laying within the interval. The final objective is to discover which time interval results in the clusters of best quality and then return these clusters together with the interval [37].

In addition to descriptive methods, data mining offers a wide perspective of predictive tasks, which include prediction of locations, trajectories and events, as well as the classifi-
cation of trajectories. Location prediction consists in forecasting the future position of an individual, given the current location and the velocity vector. In the wireless communication networks context, where predicting the motion of users in the future is of fundamental importance, several algorithms have been investigated to predict user locations. Some of them train neural networks based on the location area, or apply Gauss-Markov models based on the location and velocity. Others analyze historic trajectories, derive predominant patterns and apply the most similar pattern to the trajectory in question [37]. Since people generally follow daily or weekly routines, it is possible not only to predict positions, but also to anticipate their most likely route and destination. In [42], authors adapt a transition matrix to personal preferences to predict the most likely route and destination of a single person within a given time frame, whereas in [43] Laasonen improves the model by incorporating information about residence times.

Giannotti et al. [11] illustrate how a knowledge discovery process based on massive collections of trajectory data can unveil the complexity of human mobility and answer some fundamental questions on mobility: what are the frequent patterns of people’s travels? How do big attractors and extraordinary events influence mobility? How to predict areas of dense traffic in the near future? How to characterize traffic jams and congestions? Firstly, authors analyzed the detailed trajectories from tens of thousands private cars with on-board GPS receivers in the areas of Milan and Pisa, and found a heavy-tailed distribution in both length and duration of trips. In order to assess the significance of datasets, they compared the Milan dataset with survey data collected by the Milan municipality, finding a significant match in movement and presence distributions. Then, to answer the above questions, researchers used M-Atlas as a complete querying, analysis and mining system centered onto the concept of trajectory. M-Atlas [44] manages three main object types: data, M-model, and M-pattern. A M-pattern represents the common behavior of a group of trajectories, and comprise trajectory clustering, descriptions of frequent behaviors, spatio-temporal coincidence of moving points, and flows of trajectories which move from a region to another. M-models are instead the global models extracted by a data mining algorithm, like histograms of distances between trajectories, compact representations of a set of T-Patterns, and origin-destination matrices. Such models and patterns can be constructed by M-Atlas through the execution of data mining methods in its library. Thanks to M-Atlas mining quering language, Giannotti et al. [11] characterized the main flows from the city center toward the suburbs of the city of Milan, grouping them through proper clustering algorithms. Furthermore, to understand how users access big mobility attractors, authors constructed an origin-destination matrix between the entire city as origin and the individual parking lots as destinations, and the Linate airport parking lot emerges as the top destination. To detect frequent segments of trips that are followed by a significant volume of vehicles, researchers used a model constructor generating T-Patterns, finding that northern travels tend to concentrate on the outer ring earlier than the southern travels, which instead use a small segment of the ring.
2.2 Network Science

Network science is a scientific discipline that examines the interconnections among diverse physical, engineered, information, biological, cognitive, semantic and social networks. A network has a very flexible definition: it is a set of nodes connected by links. In mathematical terms a network is represented by a graph, a pair of sets $G = (V, E)$, where $V$ is a set of nodes and $E$ is a set of edges that connects elements of $V$. According to the definition, any system with coupled elements can be represented as a network, so that our world is full of networks.

The scientific study of networks has its roots in the eighteenth century when the legendary mathematician Leonhard Euler solved the famous “Seven Bridges of Königsberg” problem [45]. In his short paper of 1736, he inadvertently started the immense branch of graph theory, the basis for our thinking about networks. The field of graph theory continued to develop, and specific networks like regular and disordered lattices were main objects of study in physics and natural sciences up to the end of the 20th century. However, in the last decade network scientists showed that most natural and artificial networks do not resemble lattices and started to uncover the complexity behind several networks, from the Internet to the biological and social networks. These recent discoveries have excited both the scientific community and the public audience. The success of network science is accompanied by several general audience books [46, 47, 48, 49], annual symposium series that brings together artists and scientists [50], network science research inspired art-projects, documentaries, movies, series of novels and short stories, from science fiction to literary novels. In this section we provide a brief account of the main discoveries and researches in network science, from the early models by Erdős and Rényi to the algorithms developed in the last few years.

2.2.1 Erdős-Rényi model

In the twentieth century the theory of random graphs has been an important breakthrough, introduced by Solomonoff and Rapoport [51] and independently by Erdős and Rényi [52]. In their seminal paper, the latter defined a random graph $G_{n,m}$ as $n$ labeled nodes connected by $m$ edges, chosen randomly from the $n(n-1)/2$ possible edges. All possible graphs with $n$ nodes and $m$ edges form a probability space in which every realization is equiprobable. In an alternative and basically equivalent definition of a random graph $G_{n,p}$ we start with $n$ nodes, every pair of nodes being connected with probability $p$. Figure 2.6 shows three random graphs from three different spaces. Studies of several mathematicians in random graph theory showed some major outcomes [53, 54, 55]: the degree distribution follows a Poisson distribution; at $\langle k \rangle = pn = 1$ there is a phase transition meaning that for $\langle k \rangle < 1$ the random graph is composed typically of isolated trees, while for $\langle k \rangle > 1$ a giant component emerges; the diameter of the giant component is similar to the average path length; both the giant component and the diameter scale with the logarithm of the size of the system (small world effect). The random graph theory has dominated scientific thinking about networks since 1959 and it was the first attempt to describe networks
in communication and life sciences. Erdős and Rényi described networks as completely uncorrelated. However, our insight suggests that behind a complex system there is an underlying network with non-random topology. Measurements on real world networks at the end of last century confirmed this intuition and showed that real networks differ from uncorrelated random graphs in several characteristics.

2.2.2 Small worlds

In 1998, Watts and Strogatz [56] analyzed the neural network of the worm *C. elegans*, the collaboration graphs of movie actors and the western U.S. power grid, and found that these systems do resemble random graph in having small characteristic path lengths, but do not in having a high clustering coefficient, a measure that indicates how strong is the presence of social triangles measuring a “all-my-friends-know-each-other” property. The clustering coefficient is very low in the Erdős-Rényi network but typically large in regular lattices (Table 2.1). They called the three analyzed systems “small-world” networks, in analogy with the famous small-world phenomenon discovered in the sixties by sociologist Stanley Milgram [57, 58], a phenomenon also popularly known as *six degrees of separation* thanks to a play by John Guare [59]. In a primary paper of 1967, Milgram decided to see whether in the “small-world experience” (an unknown person we meet knows a person we know) there was more to the phenomenon than just anecdotes. He described an experiment in which he asked Midwestern volunteers to send packages to strangers in Boston. The midwesterns were not allowed to mail the packages directly but had

<table>
<thead>
<tr>
<th></th>
<th>$L_{\text{actual}}$</th>
<th>$L_{\text{random}}$</th>
<th>$C_{\text{actual}}$</th>
<th>$C_{\text{random}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Film actors</td>
<td>3.65</td>
<td>2.99</td>
<td>0.79</td>
<td>0.00027</td>
</tr>
<tr>
<td>Power grid</td>
<td>18.7</td>
<td>12.4</td>
<td>0.080</td>
<td>0.005</td>
</tr>
<tr>
<td><em>C. elegans</em></td>
<td>2.65</td>
<td>2.25</td>
<td>0.28</td>
<td>0.05</td>
</tr>
</tbody>
</table>

Table 2.1: Characteristic path length $L$ and clustering coefficient $C$ for three real networks, compared to Erdős-Rényi networks with the same number of vertices ($n$) and average number of edges per vertex ($k$)
to relay them through personal contacts. Milgram reported that the average number of intermediaries for completed chains was surprisingly 5.5, which made for a six-linked chain.

With the aim of creating a model that fits the observed characteristics, Watts and Strogatz defined the small-world model [56]: a procedure to interpolate between regular and random networks. The intuition behind the Watts-Strogatz model is to start from a regular network, for instance a ring, in which each node is connected to its $k$ nearest neighbors. The small-world model is then created by taking each edge in turn and, with probability $p$, moving one end of that edge to a new location chosen uniformly at random from the lattice, except that no double edges or self-edges are ever created (in a variation of this model [60], there is no rewiring, but a few extra link are added, connecting randomly selected nodes). When $p = 0$, we have a regular lattice, with a very high clustering coefficient and high geodesic distances. When $p = 1$, every edge is rewired and we have a random graph, with small-world effect, and very low clustering. Between these two extremes $0 < p < 1$, a wide region exists for which the model has both low path lengths and high clustering (Figure 2.7). Watts and Strogatz quantified structural properties of small-world networks by their characteristic path length $L(p)$ and clustering coefficient $C(p)$. They found that there is a broad interval of $p$ over which $L(p)$ is almost as small as the characteristic path length $L_{\text{random}}$ of random networks, and yet $C(p) \gg C_{\text{random}}$, the clustering coefficient of random networks. The introduction of shortcuts connects nodes that would otherwise be much farther apart than $L_{\text{random}}$. For small $p$, each shortcut contracts the distance not just between the pair of nodes that it connects, but between their immediate neighborhoods, neighborhoods of neighborhoods and so on. By contrast, when an edge is removed from a clustered neighborhood $C(p)$ remains practically unchanged for small $p$ even though $L(p)$ drops rapidly. The important implication here is that at the local level the transition to a small world is almost undetectable.

The small-world model attracted a lot of interest around science of networks. However, it could explain how Milgram’s experiment got near their targets, but not how they ultimately found them. To investigate conditions in which individuals find small chains,
Jon Kleinberg [61] considered a model based on the small-world network and on a decentralized algorithm inspired by Milgram’s experiment. However, rather than using a ring as the basic structure, Kleinberg began from a two-dimensional grid and allowed edges to be directed. In the grid, nodes have $p \geq 1$ local contacts within a lattice distance as well as long-range contacts to $q$ other nodes. The probability of a long-range edge is proportional to $d(u, v)^{-r}$, where $d()$ is the lattice distance which separates the nodes, and $r$ a parameter of the algorithm. To send a message we start with two arbitrary nodes in the network, source node $s$ and target node $t$. The goal is to transmit the message from $s$ to $t$ with as few steps as possible. The algorithm is decentralized, that is the current message holder $u$ knows only the set of local contacts among all nodes, the position of target node $t$ on the grid, and the locations and long-range contacts of all nodes that have come in contact with the message. Kleinberg interestingly found that $r = 2$ is the only possible parameter for $r$ in the two-dimensional grid where a decentralized algorithm is able to perform the transmission task in expected $O(\log n)$ steps. He generalized these results for multi-dimensional spaces: for any $k$-dimensional space, a decentralized algorithm can construct paths of length polynomial in $O(\log n)$ if and only if $r = k$.

Besides the small world theory, another important contribution of sociology to network science was the “strength of weak ties hypothesis” guessed by Mark Granovetter [62]. In the seventies, he argued that the network behind our society consists of small, strongly connected circles of friends. Weak ties connect the members of these circles to their acquaintances, who have strong ties to their own friends. Granovetter discovered that weak ties play a crucial role in finding a new job, since they are the bridge to the outside social world.

Recently, thanks to availability of network data, both the small-world theory and the “strength of weak ties” hypothesis have been confirmed in a lot of natural and artificial networks like the WWW [63, 64], the Internet [65, 66], food webs [67], metabolic networks [68], scientific collaboration networks [69, 70], movie actor networks [71], in the neural network of worm *C. elegans* [56, 72], in email contacts network [73], in the Messenger contacts network [74] and in mobile phone networks [75]. These findings suggest that pioneering results of Milgram and Granovetter are pervasive in networks, in nature and

![Figure 2.8: The Kleinberg model: a two-dimensional grid network with $n = 6$, $p = 1$, and $q = 2$. $v$ and $w$ are the two long-range contacts.](image)
technology. Using data on about 70 billions links among \( \approx 720 \) million Facebook users, Boldi et al. [76] discovered that the average number of acquaintances separating any two people in the United States is 4.37, and that the number separating any two people in the world is 4.74. So, the Facebook world is even smaller than the one depicted by Milgram: in average, only four steps separate us from any other Facebook user on the planet.

2.2.3 Scale-free networks

In many real networks, some nodes - the hubs - are more highly connected than others. Think about Yahoo! for the Web, ATP protein for metabolic networks, Heathrow for air traffic network, an important person for social networks. To quantify this effect, we use the degree distribution \( P(k) \), the fraction of nodes that have degree \( k \). The simple Erdős-Rényi model predicts a Poisson distribution for \( P(k) \), but for many real networks \( P(k) \) is highly skewed and decays much more slowly than a Poisson, following a power law \( P(k) \sim k^{-\gamma} \). These networks are called scale-free, because they do not have a scale, that is a characteristic degree. Although this exponent is not universal [77], it has been observed in several artificial and natural networks, such as the WWW [63], the Internet backbone [78], metabolic reaction networks [68], phone call graphs [75], co-authorship networks of scientists [69,70] and movie actors [71]. Nevertheless, the scale-free case has stimulated a great deal of theorizing, in particular to explain why hubs and power law emerge in complex networks. In 1999, Barabási and Albert [79] showed that the heavy-tailed degree distribution emerges because networks expand continuously by the addition of new nodes. New nodes attach preferentially to sites that are already well connected, thus richly connected nodes get richer. They indeed defined a “scale-free model”, also known as Barabási-Albert model, based on two ingredients: growth and preferential attachment. We start from a small number of nodes \( m_0 \), and at every time step we add a new node with links to \( m \) different already existing nodes. The probability that a new node will be connected to node \( i \) depends on the connectivity \( k_i \) of that node. After \( t \) time steps the model leads to a random network with \( t + m_0 \) nodes and \( mt \) edges. This network evolves into a scale-invariant state with the probability that a node has \( k \) edges following a power law with exponent \( \gamma_{model} = 3.0 \).

![Figure 2.9: The Barabási-Albert model. Starting from two connected nodes (top left), in each panel a new node (the white one) is added to the network. When deciding where to link, new nodes prefer to attach to the more connected ones. A few highly connected hubs emerge.](image)
The Barabási-Albert model predicts that the oldest node always has the most links, a principle often called the “the rich get richer”. According to the model, late nodes can never turn into the largest hubs. In reality a node’s degree does not depend on the node’s age only. Instead webpages, companies, or actors have intrinsic qualities that influence the rate at which they acquire links. In other words, nodes have a “fitness” property, an intrinsic value related to some capacities such as management quality in companies, or the content offered by a website. The “fitness model”, also called Bianconi-Barabási model [80, 81], assumes that preferential attachment is driven by the product of a node’s fitness and its degree. A fitness value is associated to each new node in the evolving network, and it is linked to pre-existing nodes with a probability proportional to the product of nodes degree and fitness. Hence, the model assures that even a relatively young node with initially only a few links can acquire links rapidly if it has larger fitness than the rest of the nodes.

2.2.4 Community Structure

A remarkably attractive research issue in network science is community detection, i.e. the problem of finding natural groups of nodes in a network and connections between them, to uncover and understand the coarse structure of a network.

There is no unique definition of a network community. In many networks, the communities are natural, non-overlapping modules of networks. In other context, communities can overlap. In a third type of study, the communities are hierarchically embedded one into another. Intuitively a community is usually considered to be a set of entities where each entity is closer to the other entities within the community than to the entities outside it. They may correspond to groups of pages of the World Wide Web dealing with related topics [82], to functional modules such as cycles and pathways in metabolic networks [83], to groups of related individuals in social networks [84] and so on. When networks are small, it is easy to identify communities by analyzing the network visually and indicate its dense, closely connected parts. When a network is large, however, the manual detection is impossible and an algorithmic solution is needed. Unfortunately, communities are poorly defined and often hardly distinguishable, so it is very hard to develop a universally efficient numerical algorithm for the problem [85]. Moreover, communities are very dynamic objects, which change according to the evolution of the underlying network.

Two surveys by Fortunato [86] and Coscia et al. [87] explores the most popular community detection techniques and try to classify algorithms given the typology of the extracted communities. One of the most adopted definitions of community is based on the modularity concept [88, 89], a quality function of a partition which scores high values for partitions whose internal cluster density is higher than the external density. The seminal algorithm proposed by Girvan and Newman [88, 90] iteratively removes links based on the value of their betweenness, i.e. the number of shortest paths that pass through the link. The procedure of link removal ends when the modularity of the resulting partition reaches a maximum. The method introduced by Clauset et al. [89] is essentially a fast implementation of a previous technique proposed by Newman [88]. A fast and efficient
2.3. HUMAN MOBILITY AND SOCIAL NETWORKS

The greedy algorithm, Modularity Unfolding, has been successfully applied to the analysis of huge subset of the WWW [91]. Modularity is not the only key concept that has been used for community detection: an alternative approach is the application of information theory techniques, as for example in INFOMAP [92]. An interesting property for community discovery is the ability to return overlapping sub-structures which allow nodes to be part of more than one community. A wide set of algorithms were developed over this property, such as cFINDER [93], and DEMON [94].

A different category of community detection algorithms aim to deal with dynamic networks [95]. Communities, indeed, are affected by changes in network topology: as time goes by the rise and fall of nodes and edges determines the rise and fall of social clusters that a static community discovery algorithm often is unable to detect. In order to understand how communities evolve, many approaches have been followed so far. Some strategies track the evolution of communities by identifying key events in their life (birth, death, merge, split). The works in [96, 97] propose an extended life-cycle model to track the evolution of communities, with a two-step procedure: (i) the graph is divided in \( n \) temporal snapshots and for each snapshot a set of communities is extracted; (ii) for each community an evolutionary chain is built by observing its evolution through temporal adjacent sets. Takaffoli et al. introduced MODEC [98], a framework to detect the evolution of communities obtained at different snapshots in a dynamic social network. The problem of detecting the transition of communities is solved by identifying events that characterize the changes of the communities across time.

A different methodology to detect community in an evolutionary scenario is to design a procedure in which each community identified at time \( t \) is influenced by the ones detected at time \( t - 1 \) (avoiding the need to match communities) [99, 100, 101]. Although those approaches reduce the complexity of the matching phase, they are based on a static temporal partition of the complete temporal network. Some evolutionary algorithms do not partition the full temporal annotated graph, but try to build (and maintain) communities in an online fashion following the rising of new nodes and edges. Qi et al. [102] propose the IC-DRF model to dynamically maintain a community partition of moving objects based on trajectory information up to the current timestamp. However, due to the information used to update the community membership, the approach is suitable only for a specific kind of networked data. Lin et al. [103] propose an algorithm that extracts communities taking care of the topology of the graph at the specific time frame \( t \) as well as the historical evolutive patterns of previously computed communities. Cazabet introduces iLCD [104], an overlapping online approach to community detection which re-evaluates communities at each new interaction according to the path lengths between each node and its surrounding communities.

2.3 Human Mobility and Social Networks

In the previous sections we presented the evolution of the study on human mobility and network science, describing the main patterns and models that characterize the mobility
and social behavior of individuals. Here, we take a step further in our journey of understanding human behavior by focusing on the interplay between human mobility and social networks, with the purpose of highlighting to what extent human movements affect social dynamics, and how social interactions influence the way people move.

Recent advances on human mobility and social networks have turned the interplay between these two aspects into a crucial missing chapter in our understanding of human behavior. To make progress along this direction requires large-scale data that simultaneously capture the dynamical information on individual movements and social interactions. Thanks to the increasing availability of mobile phone datasets and location-based online social networks (LBSN), scientists start to look into challenging questions: at what extent human mobility patterns shape and impact our social ties, and how do our social surroundings affect where we go? The central hypothesis here is that social interactions increase with physical proximity. Indeed, social links are often driven by spatial proximity, from job- and family-imposed shared programs to joint involvement in various social activities. These shared social foci and face-to-face interactions, represented as overlap in individuals trajectories, are expected to have significant impact on the structure of social networks. There are three lines of inquiry in current literature: (1) geographic propinquity yields higher probability of forming a tie; (2) overlap in trajectories predicts tie formation; (3) social environment affects individual mobility.

The considerable influence of the geographic distance on the formation, the evolution and the strength of friendships is probably rooted in the very nature of our social brain. According to the anthropologist Robin Dunbar, there is a physical cognitive limit in the number of strong ties our brain is able to manage, partly because they must be powered by a form of social grooming, a time-consuming activity mainly based on geographical proximity and face-to-face contacts [105, 106, 107]. Recent analysis on Facebook and email data [108, 109] confirmed Dunbar’s intuition, showing that the volume of communications is inversely proportional to geographic distance and that the probability \( P(d) \) of having a friend at a certain distance decrease following a sort of “gravity law”. Although in the last decades technology has contributed to reduce distances, proximity is still important for the establishment of relevant relationships, breaking down the illusion of living in a “global village”: a small world in which physical and cultural distances vanish and where lifestyle become homogeneous [110, 111].

In studying the social vs geography problem, data from Location Based Social Networks (LBSNs) proved to be very useful. In LBSNs, users share their current location by checking-in on websites such as Foursquare, Facebook, Brightkite or Gowalla. In addition to provide the explicit social network of users, LBSNs give high resolution location data, as one can distinguish between a checkin to different floor of the same building. Scellato et al. [112] used information from both the social and location components of several LBSNs to identify the relation between friendship and geographic distance. They noticed a weak positive correlation between the number of friends and their average distance and observed that the socio-spatial structure of the users can not be explained by taking into account separately geographic factors and social mechanisms. Cranshaw et al. [113] studied the
entropy related to LBSNs locations in order to understand how it affects the underlying social network. They found two main results: (i) co-locations at high entropy locations are much more likely to be random occurrences than co-locations at low entropy locations; (ii) users who visit high entropy locations tend to be more social, having more ties in the social network than users who visit less diverse locations.

Given that two persons have been on multiple occasions in the same geographic place at the same time, how likely are they to know each other? This is another interesting and open problem about the interplay between sociality and mobility, regarding to which extent social ties between people can be inferred from co-occurrence in time and space. Crandall et al. [114] studied such problem by analyzing a huge dataset from the popular photo sharing site Flickr, reaching interesting and striking conclusions. They inferred a spatiotemporal co-occurrence between two Flickr users if they both took photos at approximately the same place and at approximately the same time. Rather surprisingly, they found that even a very small number of co-occurrences can lead to orders-of-magnitude greater probabilities of a social tie. Indeed two users have nearly 5,000 times the baseline probability of having a social tie on Flickr when they have just five co-occurrences in a day in a 80km range of distance. With the aim of a deeper understanding of the underlying phenomenon, they developed a mathematical model in which the probabilities of friendship as a function of co-occurrence qualitatively approximate the distributions they observed in the Flickr data.

Wang et al. [115] presented a data mining approach to the question: to what extent individual mobility patterns shape and impact the social network? Following the trajectories and communication patterns of approximately 6 Million mobile phone users over three months, they defined three group of similarity measures: mobile-homophily (similarity in trajectories), network proximity (distance in the call graph) and tie strength (number of calls between two users). Exploring the correlation between these measures, researchers discovered that they strongly correlate with each other. The more similar two users’ mobility patterns are, the higher is the chance that they have close proximity in the social network, as well as the higher is the intensity of their interactions. Starting from these results, they designed a link prediction experiment [116] by constructing the entire repertoire of both supervised and unsupervised classifiers, based either on network and/or mobility quantities. Results showed that mobility on their own carry high predictive power, comparable to that of network proximity measures. By combining both mobility and network measures, in the supervised case authors obtained that only one third of the actual links were missed by the predictor. The results of this study suggest that Granovetter’s theory should be integrated with a mobility dimension: the strength of a tie is correlated not only to social proximity (the extent to which people share the same community) but also to their mobility behavior (the overlapping of their spatio-temporal trajectories).

Cho et al. [117] investigated the interaction of the person’s social network structure and their mobility using datasets that capture human movements from Gowalla, Brightkite and phone location trace data. Since they uncovered a surprising increase of the effect of distant
friends on an individual’s mobility, they tried to understand if friendships influence where people travel, or if it is more traveling that influences and shapes social networks. In order to measure the degree of causality in each direction, they downloaded the Gowalla social network at two different time points $t_1$ and $t_2$, three months apart. Considering friendships at time $t_1$, they calculated a set of checkins $C_a$ that occurred after time $t_1$ and quantified the influence of sociality on future movements by measuring what fraction of them occurred within the vicinity of friend’s homes. Similarly, researchers examined the influence of mobility on creating new social ties by examining a set of checkins $C_b$ before time $t_1$ and counted the fractions of checkins led to creation of new friendships. They found that whereas there is, on average, a 61% probability that a user will visit a home of an existing friend, the probability that a checkin will lead to a new friendship is only 24%. Such results were confirmed in phone call data, with the influence of friendship on individual’s mobility about 2.5 times greater than the influence of mobility on creating friendships. Moreover, data also display a strong dependency between probability of friendship and trajectory similarity, suggesting the there is a strong presence of social and geographical homophily.

The most interesting aspect of such main findings is that they can be used to develop a model of human mobility dynamics combining the periodic daily movement patterns with the social movement effects coming from the friendship network.

2.4 Data Mining and Network Science: a vision of convergence

We have discussed in this chapter how the tools of statistical physics and complexity science have been applied to the study of human mobility, both focusing on individual movements and considering also the social relations among individuals. We have observed how, in both cases, general laws can be devised and empirically validated based on the newly available mobility data, shedding new light on the underlying mechanisms behind phenomena that, at first sight, seem to be governed by chaos. We conclude with an observation that spontaneously emerges from the current trend of research, as presented here: there is an evident push towards the convergence of network/complexity science and data mining research, a progressive merge of the two scientific communities that is only beginning today, but it is steadily increasing due to the advantages of combining the complementary strengths of the two approaches. Why is this merge convenient?

We learned in this chapter that statistical physics and network science are aimed at discovering the global models of complex social phenomena, by means of statistical macro-laws governing basic quantities; the ubiquitous presence of power laws and other long tailed distributions witnesses the behavioral diversity in society at large, such as the huge variability and individual differences of human movements. On the other hand, data mining is aimed at discovering local patterns of complex social phenomena, by means of micro-laws governing behavioral similarity or regularities in sub-populations, such as the mobility patterns and clusters discussed in Section 2.1.5. In the overall set of individ-
ual trajectories across a large city we observe a huge diversity: while most travels are short, a small but significant fragment of travels are extraordinarily long; therefore, we observe a long-tailed, scale-free distribution of quantities such as the travel length and the users’ radius of gyration. Despite this complexity represented in the data, mobility data mining can automatically discover travel patterns corresponding to set of travelers with similar mobility; in such sub-populations the global diversity vanishes and similar behavior emerges. The above dual scenario of global diversity (whose manifestation is the emergence of scale-free distributions) and local regularity (within clusters, or behavioral profiles) is perceived today as the signature of social phenomena, and seems to represent a foundational tenet of computational social sciences. Although network science and data mining emerged from different scientific communities using largely different tools, we need to reconcile the macro/global approach of the first with the micro/local approach of the second within a unifying theoretical framework, because each can benefit from the other and together have the potential to support realistic and accurate models for simulation and what-if reasoning of social phenomena.
Chapter 3

The Social Microscope

“The telegraph, the telephone, the radio, and especially the computer made possible to track anyone, anywhere in the world”, writes American journalist Steven Levy in his book Crypto [118]. In fact, almost everything we do nowadays requires the use of some digital device: from communications to travels, every human action is digitalized in some form. Take my daily routine, for example. I wake up early in the morning and generally start my day with a run in the park, tracking my ride with a smart phone app which summarizes and stores my performance. Back home, I have a regenerative breakfast and mark the food I consume through another app on my phone, with the purpose of controlling my fitness. Finally, I open my laptop and connect my mind to the Internet: I exchange emails with colleagues, chat with friends on Facebook, tweet short sentences on Twitter, call my parents via Skype, surf the Web and google topics of interest, buy goods on Amazon, use Youtube or Spotify to provide a soundtrack to my job. Your routine, presumably, is not so different from mine: you use everyday technological devices to accomplish your activities. All these digital actions are not lost in a black hole, but collected by institutions and companies which spy many aspects of our life, from movements to tastes. Simply by using our technological devices you and I are becoming objects under a social microscope: as biologists observe micro-organisms under their microscopes, data scientists observe our actions under the powerful lenses of Big Data. The term “Big Data” has not become a universally shared definition, but generally refers to any collection of datasets so large and complex that it becomes difficult to process using traditional data tools. The social network of Facebook is Big Data, the search logs stored by Google are Big Data, as well as other kinds of data greedily stored by companies around the globe.

Mobile phone operators, for instance, store in their huge databases information about all the calls you make: they know whom you called yesterday, when you called her, and where you performed the call. In other words, they track in detail your social network, your mobility, and the timing of your activities. The social microscope of Facebook has an even more powerful lens. It knows our social network, how and when we interact with our friends through likes, comments, chat conversations, audio/video calls. Moreover, Facebook collects the events we attend, our geographic positions through geotagged posts, and even our favourite movies, books, places and so forth. Dozens of different companies
and institutions make the same tracking with other aspects of your private life. How do they use such a broad social microscope? What do they produce from data?

Google is a master at creating data products. Think about spell checking, for which Google built a dictionary of misspellings and corrections. Or think about speech recognition, for which Google has made huge strides by using the voice data they have collected. Tracking the progress of the influenza epidemic by following searches for flu-related topics is another example of data product. Google is not the only company that knows how to use data. Facebook, Twitter and LinkedIn use patterns of friendships to suggest other people you may know. Amazon correlates your searches with those of other users to create surprisingly appropriate recommendations. Both friendships or items recommendations are “data products” that help to drive the companies’ more traditional retail business. Data collected from users provides added value. Whether data are search terms, voice samples, or product reviews, users are in a feedback loop in which they contribute to the products they use.

While companies are interested in creating data products, researchers and academics are more interested in exploiting the social microscope for producing scientific laws: mechanisms and patterns which can explain how some aspects of our complex society work and evolve. Producing patterns of human behavior and dynamics from the analysis of Big Data is exactly the purpose of this thesis. To understand these phenomena we need to observe movements and interactions under a proper social microscope: for our work, indeed, we rely on the analysis of mobility data and social network data from various sources. In this chapter we describe the datasets we used to analyze, understand, and describe the mobile and social aspects of human behavior. In Section 3.1 we describe the Octo dataset, which stores the trajectories of thousands of vehicles traveling in Italy, and in Section 3.2 present the two GSM datasets we exploited in our research, which convey information about both movements and social interactions. The two kinds of mobility datasets are different from many perspective, and each one has its own advantages and disadvantages. Section 3.3 addresses such issue highlighting the major differences. Section 3.4 describes the dataset we used in our social investigation on musical tastes. Table 3.1 summarizes the characteristics of the datasets used in the works of the present thesis.

<table>
<thead>
<tr>
<th>dataset</th>
<th>type</th>
<th>volume</th>
<th>period</th>
</tr>
</thead>
<tbody>
<tr>
<td>EC</td>
<td>mobility&amp;sociality</td>
<td>3M users</td>
<td>6 months</td>
</tr>
<tr>
<td>Orange</td>
<td>mobility&amp;sociality</td>
<td>20M users</td>
<td>45 days</td>
</tr>
<tr>
<td>Octo</td>
<td>mobility</td>
<td>150,000 users</td>
<td>1 month</td>
</tr>
<tr>
<td>LastFM</td>
<td>sociality</td>
<td>75,969 users</td>
<td>1 year</td>
</tr>
</tbody>
</table>

Table 3.1: Dataset used in the thesis.
3.1 GPS data

The GPS (Global Positioning System) is a satellite navigation system that utilizes more than two dozen satellites. It broadcasts precise timing signals by radio to GPS receivers, allowing them to accurately determine their location (longitude, latitude, and altitude) in any weather, day or night, anywhere on Earth. A GPS receiver calculates its position by precisely timing the signals sent by GPS satellites high above the Earth. Each satellite continually transmits messages which specify the time the message was transmitted and the precise positioning information. The receiver computes the distance to each satellite by determining the transit time of each message it receives. These distances along with the satellites’ locations are used to compute the position of the receiver, in form of latitude, longitude and other information like elevation, direction and speed. GPS-enabled devices provide us with all the required information for trajectory tracking, giving access to accurate, time-stamped locations of each tracked moving point. Since nowadays GPS receivers are very cheap, they are embedded in many devices we use everyday like smartphones and vehicles, allowing to track human mobility. Vehicle tracking systems, person tracking systems and pet tracking systems use GPS to locate a vehicle, person or pet. The GPS devices are attached to the vehicle, person or the pet collar, providing continuous tracking and mobile updates.

In the present thesis we use a massive real-life GPS dataset, namely the Octo dataset, obtained from tens of thousands private vehicles with on-board GPS receivers. The owners of these cars are subscribers of a pay-as-you-drive car insurance contract, under which the tracked trajectories of each vehicle are periodically sent to a central server for antifraud and anti-theft purposes. This data set has been donated for research purposes by Octo Telematics Italia S.r.l. [119], the leader for this sector in Europe. The market penetration of this service is variable on the territory, but in general covers around 2% of the total registered vehicles. The Octo dataset stores information of approximately 9.8 Million different car travels from 159,000 cars tracked during one month (May 2011) in an area corresponding to central Italy (a 250km×250km square, Figure 3.1). The GPS device automatically turns on when the car starts, and the sequence of GPS points that the device transmits every 30 seconds to the server forms the global trajectory of a vehicle. When the vehicle stops no points are logged nor sent. We exploit these stops to split the global trajectory into several sub-trajectories, that correspond to the travels performed by the vehicle. Clearly, the vehicle may have stops of different duration, corresponding to different activities. To ignore small stops like gas stations, traffic lights, bring and get activities and so on, we choose a stop duration threshold of 20 minutes: if the time interval between two consecutive observations of the vehicle is longer than 20 minutes, the first observation is considered as the end of a trip and the second observation is considered as the start of another trip. We also perform the extraction of the trips by using different stop duration thresholds (5, 10, 15, 20, 30, 40 minutes), without finding significant differences in the sample of short trips and in the statistical analysis we present in the current thesis.
CHAPTER 3. THE SOCIAL MICROSCOPE

Figure 3.1: Spatial distribution of GPS trajectories in the Octo dataset. The trajectories correspond to car travels performed by vehicles passed through an area corresponding to central Italy in May 2011.

3.2 GSM data

Mobile phones are nowadays very common technological devices carried by individuals in their daily routine, offering a good proxy to study structure and dynamics of human social behavior. Indeed, phone records capture information about both social links and human displacements: each time we make a call a social relationship of some kind is expressed, and the tower that communicates with our phone is recorded by the carrier, effectively tracking our location.

The GSM (Global System of Mobile Communications) is the most popular standard for mobile phones in the world, nowadays used by more than 1.5 billion people across more than 210 countries and territories. The ubiquity of GSM standard make international roaming very common between mobile phone operators, enabling subscribers to use their phones in many parts of the worlds. GSM networks consist of a number of base stations, each responsible for a particular spatial area (known as “cell” or “tower”). Hence, for each GSM-enabled device we can collect information about the base stations it was served by at different timestamps, and as such, assume its movement. A GSM-enabled device can be tracked by collecting all the communication signals transmitted (cell, signal strength) between this device and the networks infrastructure or by studying the log of the outgoing calls (UserID, data and time of the call, duration of the call, the cell where the call began,
the cell where the call ended). However, in both levels the accuracy of trajectories that can be collected is very low since the most detailed level of available information is the network cell and not a spatial point. Data about calls are generally stored in form of CDRs (Call Detail Records), describing each phone call performed by the users. Each call is represented by a tuple with timestamp, caller and callee identifiers, duration of the call, and the geographical coordinates of the tower serving the call. Table 3.2 provides an example of the structure of CDRs. The time-ordered list of towers from which a user performed the calls compose a mobile trajectory, which describes her movements during the period of observation. Such mobility traces are more accurate in densely populated areas, where much more phone towers are installed to carry the heavier load. This means that in rural areas, where a single tower usually covers several square kilometers, short movements are not collected. Besides the mobility dimension, GSM data also provide us the social dimension, since we can reconstruct from the calls a weighted directed call graph, where nodes are users and edges are interactions between users. The weight of an edge can be either the total number of calls or the total duration of calls between the users during the period of observation.

In our research we exploited the access to two GSM datasets. One is the EC dataset, collected by a European carrier for billing and operational purposes and describing each phone call performed by 3 millions users in a period of 6 months. The other one is collected by French Telecom provider Orange [120], the biggest French carrier covering approximately the 30% of the mobile phone market in France. Here, CDRs store information of about 200 millions phone calls sent or received by about 20 millions Orange users in a period of 45 days (from 01/09/2007 to 15/10/2007).

### 3.3 Properties of Mobility Data

Due to the manifold of techniques to record mobility information, mobility datasets can differ in several aspects. Each mobility dataset, indeed, inevitably presents some limitations with respect to its observation space, sampling coverage and resolution.

The observation space delimits the spatial, temporal, population and movement dimensions of the mobility data. The spatial dimension, i.e. *where* movements take place, refers to the region where individuals are observed. Typically, this is a city, community or even large administrative areas. In the temporal dimension, i.e. *when* movements take place, the observation space defines the time moment, time interval or time cycle of the movements (e.g. January 2014). The population observation space, i.e. *who* performs the

<table>
<thead>
<tr>
<th>timestamp</th>
<th>coordinates</th>
<th>caller</th>
<th>callee</th>
<th>type</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008/04/01 23:45:00</td>
<td>(132.567, 23.642)</td>
<td>A45J23</td>
<td>F45J23</td>
<td>SMS</td>
</tr>
<tr>
<td>2008/04/02 06:02:10</td>
<td>(143.282, 54.221)</td>
<td>K65232</td>
<td>V56YT4</td>
<td>Call</td>
</tr>
</tbody>
</table>

Table 3.2: Example of GSM phone records.
travels, defines the set of people whose movements are studied. As it is generally impossible to monitor the complete observation space, mobility datasets are often samples of the spatial, temporal and population dimensions. Regarding the observation space, the Octo GPS dataset refers to trips performed during one month (May 2011) in an area corresponding to a region in central Italy (a $250\text{km} \times 250\text{km}$ square). In contrast, the EC dataset and the Orange dataset cover two big European countries and a period of observation of 6 months and 45 days respectively. Finally, Octo represents the 2% sample of the overall population of vehicles in Italy, while the mobile phone datasets cover users of two majors European operators, with 1,000,000 users (EC dataset) and 20 million users (Orange dataset).

Resolution, i.e. the level of aggregation of the amount of detail in a dataset, is another interesting aspect of mobility data. In the population dimension the smallest possible measurements unit is a single person, which corresponds to the natural resolution. The spatial resolution of a dataset can vary between a few centimeters and several kilometers depending on the monitoring technology used to collect the data. The spatial resolution of GPS data is very high, providing very detailed information about the spatio-temporal position of users. Conversely, information provided by the GSM data is not very accurate in terms of space and time because an individual may be anywhere within a tower’s reception area, which can span up to tens of square kilometers.

The temporal resolution of a dataset corresponds to the time span of a single measurement. In this case, the GPS data have a incredibly high resolution with a sampling rate of a few seconds. For GSM data the situation is more complex, due the bursty nature of human activities [121]. For most of the time, indeed, people are inactive (do not perform calls) not allowing the observation of their current position. The temporal resolution of GSM data, therefore, is not terribly accurate since we can track people’s position only for a limited period of time.

The fact that one dataset contains aspects missing in the other dataset makes the two types of data suitable for an external validation of patterns emerging from human mobility behavior. We do not expect to observe exactly the same behavior in both datasets, but the same tendencies and laws behind the movement patterns.

### 3.4 LastFM data

LastFm [122] is a popular online social network, where people can discover, share and listen to new songs, artists and genres basing on what they like. A subscribed user can start listening the LastFm personalized Radio, or send data about her own online listenings. She can also express her preference and attach tags to each songs, describing the genre of the musical piece. As in other online social networks like Facebook, in LastFM users can add friends (undirected connections, the friendship request must be confirmed) and search for “neighbors”, i.e. other users with similar musical tastes. A user can see, in her homepage, her friends’ activities.

Using LastFm APIs, we downloaded a sample of the United Kingdom users graph,
starting from a set of nodes and implementing a breadth-first approach [123]. The breadth-
first procedure begins at a root node and inspects its neighboring nodes and, for each of
them, it inspects neighbors that have not yet been visited, and so on. Starting from
the initial set of nodes, we explored snowball samples up to the fifth neighbors. Thanks
to this procedure, for each user we retrieved her friendships and the number of single
listenings of a given artist in the time window from January 10 to December 11. As
an example, in the week between April 11, 2010 and April 18, 2010 the user 1234 has
listened 66 songs by the Metallica band. We also collected information about each artist
listened by at least one user in the period of observation, downloading the list of tags with
their counter, representing the number of users that assigned that tag to that artist. For
example, Metallica have 4 tags: metal with counter 50670, hard rock with 23405, punk
with 10500 and adrenaline with 670. We split all tags, associating the counter to each
single word, and filtered the words referring to a musical genre. In the last example we
obtained (metal, 50670), (punk, 10500), (hard, 23405), (rock, 23405), (adrenaline, 670);
hard and adrenaline have been discarded since they do not refer to major established
musical genres. Finally, we assigned a musical genre to an artist if the survived tag with
the greater counter had the relative rate $\geq 0.5$. The band Metallica is definitely metal,
since they have a rate $r_{metal}(Metallica) = \frac{50670}{50670 + 10500 + 23405} = 0.6$.

After the crawl and cleaning stages, we built the social graph $G$ as follows: nodes
correspond to users, and edges are generated generated by looking at the user’s friends
in the social media platform. The total amount of nodes is 75,969, with 389,639 edges
connecting them.
Part II

Understanding Human Mobility
Chapter 4

The patterns of car travel

Cars are the most popular private means for transportation in current society, diffused nowadays in every corner of the planet. For Italians in particular, cars also represent a symbol of freedom and achieved well-being, since the “fabulous sixties” where the Italian economic and industrial development began to change every aspect of culture and society. This is reflected in a natural tendency to privilege private vehicles in people’s travels, both at small and large distances: in 2011 more than 40 millions vehicles were circulating in Italy (for a population of almost 60 millions people), about 625 cars per 1000 inhabitants. It goes without saying that the Octo dataset described in Section 3.1, which stores the detailed spatio-temporal trajectories of approximately 10 million car travels in Italy, is a wonderful social microscope to observe people’s movements, extract mobility laws, and understand the mechanisms underlying human mobility.

Starting from the Octo dataset, previous works validated many mobility data mining tools and associated analytical models, including the discovery of access patterns to a city [11], individual profiles or travel routines [124], geographical borders of human mobility [125], and clusters of similar mobility behaviors [126]. Here we consider the Exploration and Preferential Return (EPR) model for individual human mobility [8, 10] introduced in Section 2.1.4, which explains the patterns and laws governing the key physical quantities of human movements, including the length/duration of travels, the characteristic distance traveled by individuals, the number and frequency of the locations visited by travelers. We address the following question: does this model apply to car travel?

It should be noted that the EPR model [8, 10] is developed with reference to GSM data, which have two main differences compared with the GPS trajectories in the Octo dataset: (i) mobile phone data pertain to general mobility while GPS data pertain only to cars; (ii) they are much less detailed than GPS trajectories, the latter providing for the precise spatio-temporal records of each travel with high exact geo-location and high sampling rate. In fact, each trajectory in the Octo dataset is a time-ordered sequence of position records $<id, x, y, t>$ where $id$ is the anonymized car identifier, $(x, y)$ are the lat-long coordinates, and $t$ is the time of the position. According to GPS standards, positioning accuracy is within a 15 meters bound in absence of obstacles and malfunctions. Sampling time, i.e. time between two observations, is 30 seconds in average when the car is in motion.
CHAPTER 4. THE PATTERNS OF CAR TRAVEL

It is therefore legitimate to investigate to what extent the previous models apply to GPS data, which deviations we can observe, and which new analytical opportunities we can provide by the finer spatio-temporal granularity. On the other hand, it is compulsory to investigate to what extent the GPS data are representative of the overall vehicular mobility, in order to generalize the validity of our findings. To this purpose, we use independent ground-truth measurements of global traffic volumes obtained by sensors placed in a set of locations during the same observational period of our GPS data. By leveraging on these ground-truth data, we show that the GPS data are an extremely accurate estimation of the overall volumes in each location, a good social microscope for car travels even if GPS data pertain to a 2% sample. Our model is based on standard machine learning techniques, tailored to the peculiarity of trajectory data. In conclusion, we obtain two intertwined results: first, the known human mobility models can be refined to deal with car mobility, and second, the available GPS data can indeed be used as a faithful proxy of car mobility. The work in this chapter is based on two papers [14, 127] published in 2013.

4.1 Understanding human mobility by car

Many studies on human mobility are based on the analysis of GSM phone data [8, 9, 10]. Even though mobile phones are carried by the same person during the daily routine offering a good proxy to capture individual trajectories, as stated in Section 3.3 they do not provide an accurate spatial information. Indeed, we know individuals’ positions only when they perform a call, and we only know the position of the tower managing the area the individual is within (an area of about 3km$^2$ in average), and not the individual’s actual geographic location. On the contrary, GPS traces provide high resolution location data, storing the geodetic coordinates with an average sampling rate of few seconds. These features are ideal in making a refined statistical analysis of human mobility patterns. However, in literature relatively few works are based on GPS data, mainly because it is difficult to obtain complete traces covering movements along the whole day. In this section, we present a statistical validation of human mobility patterns, by using the Octo dataset introduced in Section 3.1, which captures car mobility by means of tens of thousands private vehicles with on-board GPS receivers.

Our first measurement is the distribution of travel length, from which two different regimes clearly emerge (Figure 2, left). The first one (from 1km to about 20km) corresponds to low range travels, mainly located within the cities, and is characterized by an exponential distribution. The second regime corresponds to inter-city travels (i.e. travels connecting different urban areas), and is described by a power law distribution with exponent $\beta = 2.53$. Here, the observed scaling exponent is different from the one observed for GSM data ($\beta = 1.75$) [8] and for bank note dispersal ($\beta = 1.59$) [7], since it is influenced by a reduced range of distances in the GPS dataset. Indeed, both geographical limits (the region considered has a length of about 500 km) and physical limitations (travels longer than 500km are rare) provide a limitation to the range of possible distances.

In order to characterize human mobility patterns emerging from available trajectories,
4.1. UNDERSTANDING HUMAN MOBILITY BY CAR

we use the \textit{radius of gyration} as the characteristic travel distance covered by each vehicle, a measure of how far a vehicle is from its center of mass (mean location) \cite{low2}. Formally, we define radius of gyration of a vehicle as

\begin{equation}
    r_g = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (r_i - r_{cm})^2},
\end{equation}

where \( r_i \) represents the \( i = 1, \ldots, n \) positions recorded for the vehicle during the period of observation, and \( r_{cm} = \frac{1}{n} \sum_{i=1}^{n} r_i \) is the center of mass of the trajectory. For each vehicle in the Octo dataset, we compute the radius of gyration by taking all points composing its sub-trajectories as the \( n \) recorded positions. Then, we plot the distribution of \( r_g \), observing a power law with an exponential cutoff, \( P(r_g) \sim (r_g + r_0)^{-\beta} \exp(r_g/\tau) \) where \( r_0 = 5.54, \beta = 1.13 \) and \( \tau = 39.76 \) (Figure 4.1, center). The shape of the curve agrees with the previous results found on GSM data \cite{low2}, confirming that the majority of vehicles travel within a small distance, but some of them carry out long journeys. The difference between the predicted distribution and the observed behavior for people with low \( r_g \) (up to 5km) is presumably due to the tendency of covering small distances by foot, bike, or bus, resulting in a low probability to find such travels in the Octo dataset.

Figure 4.2 shows the spatial distribution of the centers of mass, with the color representing the value of relative \( r_g \). People with high \( r_g \) concentrate their centers of mass in the countryside, in the mountains (Apennines) and on the coast, whereas those with lower \( r_g \) are mainly located in urban areas. While the center of mass can be interpreted as the center position of a vehicle when it is moving, another interesting characteristic of individual mobility is the most frequent location \( L_1 \), i.e. the zone where a vehicle can be located with highest probability when it is stationary, most likely her home or work. To estimate \( L_1 \) for a vehicle \( u \), we compute all the locations where it goes by extracting origin and destination points of her sub-trajectories, without taking into consideration the
Figure 4.2: Spatial distribution of vehicles’ center of mass on the map of central Italy. The Figure shows that vehicles having small radius of gyration (red points) tend to concentrate their center of mass in the main urban centers of central Italy: Carrara, Pisa, Livorno, Pistoia, Empoli, Siena, Grosseto, Arezzo and the pool of towns composing the conurbation of Florence (Firenze, Prato, Sesto Fiorentino, Scandicci). Conversely, vehicles characterized by high radius of gyration (green points) are distributed in the countryside and on the coast.
4.1. UNDERSTANDING HUMAN MOBILITY BY CAR

time spent in each location by the vehicles. Then, we apply on these points the Bisecting K-means clustering algorithm [39]. It is an extension of the basic K-means algorithm that splits the set of all points in two clusters, dividing recursively the obtained clusters until they have a radius smaller than or equal to a threshold, set in our experiments to 250 meters. The centroid of the cluster with the highest frequency is chosen as $L_1$ for the vehicle $u$.

The most frequent location does not necessarily coincide with the center of mass, and the distance $d(L_1, c_m)$ tends to grow with the radius of gyration (Figure 4.1, right). The strong correlation shown in Figure 2 (right) is not obvious, and it is presumably due to the systematic nature of human motion. If a person travels arbitrarily on any direction from and to the same preferential location, then the distance between $c_m$ and $L_1$ would tend to zero, and the $r_g$ would have no relation with such distance. On the contrary, since each vehicle follows systematic travels among few preferred places (see Figure 4.4, right), the center of mass is pulled by these trips towards the mean point of the frequent locations. Therefore, the more a vehicle travels away from its $L_1$, the more the center of mass tends to be distant from the most frequent location. Figure 4.1 (right) also suggests that people with low $r_g$ have a larger probability to be located near the place where they live or work. On the contrary, vehicles traveling at large distances tend to be located in distant places, depending on the fact that they are moving or not. This phenomenon is confirmed by plotting on the map $L_1$ instead of $r_{cm}$, and noting that points corresponding to vehicles with high radius of gyration move towards urban centers, showing a power of attraction of cities on mobility by car (Figure 4.3). So, the majority of individuals tend to live or work within urban centers, but those characterized by high $r_g$ tend to stay far from these places when they move.

To estimate the trend of people to visit new distinct locations, we extract the number of clusters $S(t)$ visited by a vehicle, finding a power law $S(t) \sim t^{\mu}$. For vehicles having a small $r_g$ (within 1km), the exponent of the power law is $\mu \approx 0.3$, while it grows for vehicles traveling at large distances from the center of mass, $\mu = 0.65 \pm 0.03$ (Figure 4.4, left). In both cases, $\mu < 1$ indicates a decreasing tendency of vehicles to visit previously unvisited locations. Moreover, the visitation frequencies of individuals, that measures to what extent individuals return to the same place over and over, follow a Zipf’s like distribution $f_k \sim k^{-1.2}$ (Figure 4.4, right), confirming the pattern found in [10]. Clearly, the Zipf’s regime is more evident for vehicles with larger number of visited locations (above 80 visited location), whereas the frequency distribution of vehicles with less destinations does not allow us to state that it is compliant with a Zipf’s law, since it does not even cover an order of magnitude.

The time spent across the visited locations is another interesting mobility aspect it is worth investigating. Figure 4.5 (left) shows the GSM distribution of time spent for the five most frequent locations $L_1, \ldots, L_5$, computed on the EC mobile phone dataset introduced in Section 3.2. As we can see, the time spent is clearly unbalanced in favor of the most important location $L_1$. This is reasonable, because the most frequent location usually corresponds to an individual’s home or work place, which are the locations where
Figure 4.3: Spatial distribution of vehicles’ most frequent locations on the map of central Italy. The Figure shows that the most frequent location of the vehicles, regardless of their radius of gyration, tends to concentrate in the main urban areas of central Italy, corresponding to the cities of Carrara, Pisa, Livorno, Pistoia, Empoli, Siena, Grosseto, Arezzo and the pool of towns composing the conurbation of Florence (Firenze, Prato, Sesto Fiorentino, Scandicci).
4.2 INFERRING THE TRAFFIC COUNT BY GPS DATA

Figure 4.4: (Left) The number the visited distinct location $S(t)$ over time for different $r_g$ groups. $S(t)$ grows as $t^\mu$, with $\mu \approx 0.3$ for $r_g \leq 1$km and $\mu \approx 0.65 \pm 0.03$ otherwise. (Right) The visitation frequency $f_k$ of the $k$th most visited location, for vehicles that have been observed to visit $s = 40, 60, 80, 100$ and $120$ different locations. Empirical data are well approximated by a Zipf’s law, $f_k \sim k^{-1.2}$.

an individual spends most of the time. The plot also suggests that time is proportional to frequency: the more a vehicle visit a location, the more time she spends there. We observe the same pattern on GPS data, although the difference between $L_1$ and the other locations is less sharp (Figure 4.5, right). In both plots $L_1$ intersects the other curves approximately at the same points. This is interesting because, regardless the geographical scale and the portion of mobility considered, beyond a certain fraction of time is much more likely for a vehicle to be located in the $L_1$ than all the other locations.

The results of our analysis substantially confirm and refine the mobility patterns found on GSM data [8], with a difference in the population of very short range travelers which is underrepresented with respect to the prediction, and has a much slower rate of visiting new places. This suggests, on the one hand, that movements by car represents a significant portion of human travels, serving as a good social microscope that enables us to observe habits, trends and patterns in human mobility behavior, and on the other hand that it is worth investigating the particular tendency of very short range travelers.

4.2 Inferring the traffic count by GPS data

GPS trajectories of vehicles traveling within an urban territory could be very useful to address urban traffic monitoring and prediction, provided that such data are a trustable proxy of ground-truth. This is also important to assess the generality of our analytical results: to what extent our 2% sample of tracked vehicles is representative of the overall
mobility? Fortunately, there are many sensing technologies for proving traffic counts on streets, which can be used to evaluate the ability of GPS data to model local traffic, based on ground-truth. In this study, we use a dataset composed by logs collected in May 2011 from twelve Variable Message Panels (VMP). VMPs are devices situated in the outer belt of the city of Pisa with the purpose of counting all the vehicles entering the entry gates of the city. Therefore, we can assume that they capture the real number of vehicles passing from the corresponding roads. Each VMP stores information about the total count of vehicles per hour passing under that device. Exploiting the spatial precision of GPS data, we simulate the number of GPS vehicles crossing a VMP gate, by considering a spatial buffer of 30 meters radius around the position of the road sensors, and by aggregating hour by hour the number of GPS vehicles crossing those areas. Clearly, the GPS flow on a road arc can be measured both ways. We considered only the incoming flux to the city, since the VMP do not measure the outgoing traffic.

As the snapshot in Figure 4.6 suggests, there is a good match between the curves, which essentially differ by a scaling factor. To estimate the scaling, we consider the hourly vehicle count as a discrete signal, i.e. a time series, and analyze it through a discrete wavelet transform (DWT) [128]. DWT is a mathematical tool that projects a time series onto a collection of orthonormal basis functions and produces a set of coefficients, capturing information from the time series at different frequencies and distinct times. For a square-summable sequence $\mathbf{x} \in \mathbb{R}^n$, the decomposition has the following general form:

$$\mathbf{x} = \sum_{k \in \mathbb{Z}} (\varphi_k, \mathbf{x}) \varphi_k = \sum_{k \in \mathbb{Z}} c_k \varphi_k$$

(4.2)

where $c_k = (\varphi_k, \mathbf{x}) = \sum_{j=1}^n \varphi_k(j)x(j)$ are the coefficients of the transform ($j$ is a time index denoting the instants of the time series), and $\varphi_k \in \mathbb{R}^n$ the basis functions satisfying the orthonormality constraint.

For each of the twelve entry gates of Pisa, we apply a Daubechies 5 algorithm (db5) [129] to decompose in four layers the signal $\mathbf{r}_i \in \mathbb{R}^n$ ($n = 168$, the number of hours in a week)
4.2. INFERRING THE TRAFFIC COUNT BY GPS DATA

Figure 4.6: Traffic sensed by a VMP device and GPS traffic volume in one of the twelve entry gates of Pisa. The curves are very similar, suggesting that GPS data are a good approximation of real data.

representing the first week of the \( i \)-th VMP time series,

\[
\mathbf{r}_i = \sum_{k=1}^{m} \langle \phi_k, \mathbf{r}_i \rangle \phi_k = \sum_{k=1}^{m} d_k \phi_k, \quad (4.3)
\]

where \( i = 1, \ldots, 12 \) and \( m = 4 \). We repeat the same procedure on the twelve time series \( \mathbf{g}_i \in \mathbb{R}^n \) representing the signals of the corresponding GPS locations around Pisa,

\[
\mathbf{g}_i = \sum_{k=1}^{m} \langle \phi_k, \mathbf{g}_i \rangle \phi_k = \sum_{k=1}^{m} h_k \phi_k. \quad (4.4)
\]

Figure 4.7 shows two wavelet decompositions corresponding respectively to the real (left) and sampled (right) traffic count in one of the twelve entry gates of Pisa.

For each location \( i \), we divide layer by layer the relative coefficients, obtaining four scaling factors for each of the twelve locations of interest:

\[
s_{i,k} = \frac{d_k}{h_k}. \quad (4.5)
\]

We use the scaling factors to scale a GPS signal to estimate the real traffic volume (i.e. an estimation of the corresponding VMP signal). To this purpose, we multiply the decomposition of a GPS signal by the scaling factors as formalized by the following equation:

\[
Y_i = \sum_{k=1}^{m} s_{i,k} h_k \phi_k. \quad (4.6)
\]

To show the validity of this approach we measure the error with respect to the observed VMP traffic counts in all locations. In Figure 4.8 we plot the real VMP series, the scaled GPS signal, and the measured relative error at a selected VMP location. We note that the error is low when the GPS traffic is high. During the night hours the relative error tends to grow since there are too few circulating GPS vehicles, but the absolute error is still negligible. For a traffic manager it is crucial to have a precise estimation during the rush
Figure 4.7: A graphical representation of two wavelet decompositions corresponding respectively to the VMP (Left) and GPS (Right) traffic count in one of the entry gates of Pisa.

Figure 4.8: Comparison between the real traffic volume and the scaled GPS signal at a single location. The plot at the bottom shows the relative error with respect to the observed VMP traffic.
4.3. CONCLUSIONS

hours in order to design ad hoc intervention to avoid congestions, a situation for which our reconstruction provides a very high precision.

These results suggest that our sample is highly significative and represents a good approximation of real traffic. We can generalize the approach to scale the GPS traffic observed in a single VMP location to the total traffic entering the city, i.e. the total traffic measured by all the VMP sensors. This approach enables a real time traffic estimation based on the observation of the GPS vehicles alone, reducing the need for ad hoc installation of new panels. To generalize the inference to the real time scenario, we need a model trained on historic data capable of giving precise estimates for the traffic situations. To this aim, we create the signal $\vec{v}$, that represents the total real traffic obtained by summing hour by hour the traffic volume of all VMP devices. Then, we divide the signal $\vec{v}$ into two sub-signals: $\vec{v}_{TR}$ is the training set, used to learn the model; $\vec{v}_{TS}$ is the test set used to evaluate the performance of the extracted model. As training set we use the sub-signal corresponding to the first week of the Octo dataset, using the remaining three weeks as test set. We learn the model by extracting the scaling factors with equation (4.5) from the signal $\vec{v}_{TR}$, and use equation (4.6) to estimate the signal of the unseen traffic. The resulting series is then compared with the real signal $\vec{v}_{TS}$. We evaluate the accuracy of the model against two other approaches. A first approach is based on a Backpropagation Multilayer Feedforward Neural Network (ANN). A second approach is a naive predictor based on a single day model computed by averaging, for each hour, the values observed during the training week. Clearly, this trivial model is not robust against local variations and we use this as a baseline, and to show the adaptability of our approach to the fluctuation of traffic flows. Figure 4.9 compares the estimations produced by the three approaches. The DWT method maintains the general shape of the curve but overestimates the real traffic, in particular during the daylight hours. The ANN approach, on the contrary, provides volumes that are comparable with the VMP observations but it does not preserve the general shape of traffic. Finally, the naive approach shows how the phenomenon can not be captured by a static model. Although the shape of the naive curve is similar to the VMP curve, it tends to underestimate the real traffic and in particular the first peak in the morning, corresponding to people going to work. Furthermore, since it is a simple daily mean of the traffic in the observation period, it is not able to discriminate between working and nonworking days.

Of the three approaches, ours gives the best approximation of evolution of traffic during the week capturing the crucial peaks during the rush hours. The slight overestimation is acceptable in a monitoring scenario where, for example, an alert is raised when a given threshold is crossed, because the threshold can be set according to the expected overestimation.

4.3 Conclusions

In this chapter we showed the patterns of human mobility by car using the Octo dataset introduced in Section 3.1. Since the GPS data pertain only to private cars movements, we
used our data to assess the validity of the general laws of mobility derived from individual movements observed by means of GSM data. Moreover, we focused on the analysis of local behavior and validity of the dataset by comparing the observations with the ground-truth provided by real-traffic sensors. The experimental settings showed a close correlation between the real traffic volume and the scaled GPS flows obtained by means of a machine learning approach. We finally introduced a method, based on historic data analysis, to monitor real time traffic.
Chapter 5

Returners and Explorers: dichotomy in Human Mobility

The availability of massive digital traces of human whereabouts has offered a series of novel insights on the quantitative patterns characterizing human mobility. Indeed, satellite-enabled Global Positioning Systems (GPS) and mobile phone networks allow for sensing and collecting society-wide proxies of human mobility, like the GPS trajectories from vehicles and call detail records from mobile phones. This broad social microscope has attracted scientists from diverse disciplines, from physics and network science [7, 8, 9, 10] to data mining [13, 14, 12, 11], and has fueled advances from public health [130, 131, 132, 133, 134, 135] to transportation engineering [136, 137, 138], urban planning [139, 140, 141, 142], and the design of smart cities [143, 144, 145]. All these studies document a stunning heterogeneity of human travel patterns that coexists with a high degree of predictability: individuals exhibit a broad spectrum of mobility ranges while repeating daily schedules dictated by routine.

In this chapter we show that this seemingly conflicting coexistence of heterogeneity and predictability can be understood by quantifying the impact of recurring movements on mobility. To be specific we analyze mobile call records and GPS tracks of cars, allowing us to compare the overall mobility of an individual with her recurrent, or systematic, mobility. Two distinct mobility profiles emerge in both datasets: returners and explorers. The characteristic distance travelled by returners, estimated by their radius of gyration \( r_g \) [8, 14], is dominated by their recurrent movement between a few preferred locations. In contrast, recurrent mobility has only a vanishing contribution to the overall mobility of explorers, who have a tendency to wander between a varying number of different locations. We find that these two profiles are well separated: individuals persistently belong to one or the other of these two classes. We use individual mobility networks to show that both returners and explorers visit locations grouped in a few well-separated spatial clusters.

Finally, we show that current models of human mobility [10] cannot account for these two classes of individuals. We therefore propose an improved model, that takes into account geographical constraints and population density information and can reproduce the mobility patterns of returners and explorers. Finally, we show that the returner/explorer
dichotomy has important repercussions on spreading processes: we find that the global invasion threshold [146], which determines the diffusion rate of an epidemic, is significantly higher for explorers than returners. In other words, it is easier for explorers to induce a global epidemic.

The work in this chapter is based on the paper [147], submitted for publication to Nature Communications.

5.1 Measuring Human Mobility

Our first data source consists of Call Detail Records (CDR) containing the calls of 67,000 individuals selected from the EC mobile phone dataset described in Section 3.1, provided that they visit more than two locations during the observational period and that their average call frequency $f$ is $\geq 0.5 \text{ hour}^{-1}$ (see Section 5.6.1 for details). We reconstruct a user’s movements based on the time-ordered list of cell phone towers from which a user made her calls [8]. Our second data source is a GPS dataset that stores information about the trips of $\sim 46,000$ vehicles extracted from the Octo dataset described in Section 3.2. The visualization of the recorded trajectories demonstrates the complexity of explored mobility patterns (Figure 5.1). We assign each origin and destination point of the obtained sub-trajectories to the corresponding Italian census cell, using information provided by the Italian National Institute of Statistics (ISTAT) (see Section 5.6.1). We describe the movements of a vehicle by the time-ordered list of census cells where the vehicle stopped.

Figure 5.1: A visualization of the complexity of the explored mobility patterns. A fragment of the GPS trajectories used in our study, displaying trips originating in the metropolitan areas of Pisa (in red) and Florence (green). Although a plain geo-referenced visualization of experimental data, it reveals the confrontation of two “competing” metropolitan areas. It also demonstrates the ability of Big Data to portray social complexity.
We use the total radius of gyration \( r_g \) defined as \( [8, 14] \):

\[
r_g = \sqrt{\frac{1}{N} \sum_{i \in L} n_i (\vec{r}_i - \vec{r}_{cm})^2},
\]

(5.1)

to characterize the typical distance traveled by an individual. Here \( L \) is the set of locations visited by the individual, \( \vec{r}_i \) is a two-dimensional vector describing the geographic coordinates of location \( i \); \( n_i \) is the visitation frequency or the total time spent by the individual in location \( i \); \( N = \sum_{i \in L} n_i \) is the total number of visits or time spent, and \( \vec{r}_{cm} \) is the center of mass of the individual.

The most frequented location \( L_1 \) is the place where an individual is found with the highest probability when stationary, most likely her home. In general, the importance of each location \( L_k \) to an individual is defined by its rank, where \( L_k \) is the \( k \)-th most frequented location.

To understand how the \( k \) most frequented locations of an individual determine the characteristic distance traveled by her, we define the \( k \)-radius of gyration \( r_g^{(k)} \)

\[
r_g^{(k)} = \sqrt{\frac{1}{N_k} \sum_{i=1}^{k} n_i (\vec{r}_i^{(k)} - \vec{r}_{cm}^{(k)})^2}
\]

(5.2)

as the radius of gyration computed over the \( k \) most frequented locations \( L_1, \ldots, L_k \); \( N_k \) is the sum of the weights assigned to the \( k \) most frequented locations, \( \vec{r}_{cm}^{(k)} \) is the center of mass computed on the \( k \) most frequented locations (\( r_g^{(k)} = r_g \) if \( k \geq N \)). Thus, \( r_g^{(k)} \) represents the mobility range restricted to the \( k \) most frequented locations. For example, if an individual’s \( r_g^{(2)} \approx r_g \), then her characteristic traveled distance is dominated by the two most frequented locations. Conversely, if the \( r_g^{(2)} \) is much smaller than total \( r_g \) the two most frequented locations do not offer an accurate characterization of the individual’s travel pattern and we need to consider more locations.

### 5.2 Returners and Explorers

To investigate the role of the \( k \) most frequented locations for an individual’s mobility pattern we compare the probability distributions of total \( r_g \) and \( r_g^{(k)} \) for \( k = 2, \ldots, 10 \), for the GSM and the GPS data (Figure 9.2). All curves are long-tailed, indicating that most individuals cover small distances, but a few travel regularly over hundreds of kilometers (heterogeneity). We fit the distributions using the truncated power law \( [8, 14] \) \( P(r_g) = (r_g + r_0)^{-\alpha} \exp^{-r_g/r_{cut}} \) (Figure 9.2), finding two significant differences. First, the exponent \( \alpha \) of the distribution of \( k \)-radii is significantly higher that the exponent of total \( r_g \) (in GSM data for example \( \alpha_{r_g^{(2)}} = 1.89 \pm 0.09 \) and \( \alpha_{r_g} = 1.6 \pm 0.05 \)). Second, the exponential cutoff parameter \( r_{cut} \) is larger for small \( k \) (in GSM data \( r_{cut}^{r_g^{(2)}} \approx 691 \) and \( r_{cut}^{r_g} \approx 250 \)). Obviously, as \( k \) increases the \( r_g^{(k)} \) curve approaches the total \( r_g \) distribution.

The correlation between total radius and \( k \)-radius of gyration allows us to quantify the degree of similarity between overall and recurrent mobility. Figure 5.3 compares
Figure 5.2: The distributions of k-radii and total radii. The distributions of total $r_g$ and $r_g^{(k)}$, with $k = 2, 6, 10$, for the users in the GSM (a, b, c) and the GPS datasets (d, e, f). Black symbols indicate the total $r_g$, while red and blue triangles indicate the $r_g^{(k)}$ for the GSM and GPS data respectively. The dashed black line represents a truncated power-law fit of the total $r_g$, the red and blue solid lines represent a truncated power-law fit for the GSM and GPS data respectively.

The separation between the two classes is especially clear for high total radii of gyration, as for high total $r_g$ we find very few points between the diagonal and the abscissa in Figure 5.3. Yet, as the insets shows, the split into the two classes is valid for smaller total $r_g$ as well. The number of $k$-returners increases with $k$, and when $k$ equals the total number of visited locations each individual becomes a returner. The two profiles are observed in both the GSM and GPS data.
5.2. RETURNERS AND EXPLORERS

Figure 5.3: The correlation between recurrent and overall mobility. The scatter plots represent the correlation between total $r_g$ and $r_g^{(k)}$ for $k = 2, 4, 8$ in the GSM dataset (a, b, c) and the GPS dataset (d, e, f). Each point is colored from blue to red indicating the density of points in the corresponding region. Most of the points gather around the x-axis, the diagonal, and the origin. The insets magnifies the origin of the plot to [0, 100km] for GSM and [0, 16km] for GPS, demonstrating that the split emerges for smaller total radii as well. As total $r_g$ increases explorers points become returners. This transition is faster in the GPS case, consistent with the fact the vehicle mobility represents a subset of trips and visited locations.

We develop three algorithms to split the population into $k$-returners and $k$-explorers: the bisector method classifies an individual as a $k$-returner if $r_g^{(k)} > r_g/2$ or a $k$-explorer otherwise; a Support Vector Machine (SVM) classifier and the Expectation-Maximization (EM) clustering algorithm [39] extract the two patterns from the population (see Section 5.6.4 for details). The three methods produce similar results, indicating that the two classes are clearly separated and well defined. Consequently, subsequently we use the simpler bisector method to split the population into $k$-returners and $k$-explorers.

The ratio $s_k = r_g^{(k)}/r_g$ measures the impact of an individual’s recurrent mobility on her overall mobility: the higher the ratio the higher is the weight of the top $k$ locations in the trajectories of an individual. Figure 5.4 shows the probability distribution of the $s_k$ ratio for different $k$. We observe two peaks: the peak located at $s_k = 0$ corresponds to $k$-explorers, whose $k$-radius is significantly smaller than the total $r_g$; the peak at $s_k = 1$ corresponds to the $k$-returners, individuals whose $r_g^{(k)}$ is very similar to the total $r_g$. By increasing $k$ the $k$-explorers gradually become $k$-returners, causing the explorers and
returners peaks to decrease and increase respectively. The population reaches a balance of \( k \)-returners and \( k \)-explorers for \( k = 4 \) for GSM. In the GPS data regardless of \( k \) we always have more \( k \)-returners than \( k \)-explorers.

---

**Figure 5.4:** The ratio between recurrent and overall mobility. The distribution of the ratio \( s_k \) measured on the GSM data for \( k = 2, 4, 8 \) (a, b, c). We used 30 equal sized bins. The peak at \( s_k = 0 \) corresponds to explorers, while the \( s_k = 1 \) peak corresponds to returners. For small \( k \) in the GSM data explorers are more numerous than returners. As \( k \) increases the number of returners increases and overcomes the number of explorers. A balance in the population is reached at \( k = 4 \). (d, e, f) The \( P(s_k) \) for the GPS data. We again observe two peaks, but the returners peak \( s_k = 1 \) dominates starting from \( k = 2 \).

Returners and explorers are also characterized by a different spatial distribution of the visited locations. Figure 5.5 shows some representative examples of individual mobility networks [9, 26, 148] of 2-returners and 2-explorers with different total \( r_g \). For both profiles the visited locations tend to group in dense clusters with few outliers (see Section 5.6.3 for details). For 2-returners the two most frequent locations tend to separate, increasing total \( r_g \); while for 2-explorers they remain close and the total \( r_g \) grows because other groups of locations emerge far from the center of mass. Indeed, the distance between the \( k \) most important locations grows with total \( r_g \) more rapidly for returners than explorers (see Section 5.6.3). The higher the total radius of gyration, the more obvious is the difference between the two profiles.
5.3 Mobility Models

We compare our findings with the results produced by the Exploration and Preferential Return (EPR) individual mobility model [10], a state-of-the-art model that accurately captures the visitation frequency of locations, the distribution of the total radius of gyration across the population and its growth with time (ultraslow diffusion). The model incorporates two competing mechanisms, the exploration of new locations and the return to previously visited locations. We used the EPR model to simulate the mobility of 67,000 synthetic individuals (see Section 5.6.5 for details) and computed for each synthetic individual total $r_g$ and $r_g^{(k)}$. As shown in Figure 5.6a for $k = 2$ although there is a weak tendency for points to gather around the diagonal, the empirically observed split into returners and explorers is absent from the model trajectories. The difference between empirical and synthetic data is especially clear when we explore $P(s_k)$ (Figure 5.6b vs Figure

Figure 5.5: The individual mobility networks of returners and explorers. The mobility networks of returners and explorers, for $k = 2$. Nodes represent the locations visited by the individual, and each link denotes a travel observed between two locations. When total $r_g$ is small the two most important locations (red and blue) are close to each other for both explorers and returners. As the total radius increases the behavior of 2-returners and 2-explorers begins to differ: for 2-returners the two most important locations move away from each other; for 2-explorers they stay close and other clusters of locations emerge far from the center of mass.
5.4. In the model for small $k$ the $k$-explorers (with ratio $s_k \approx 0$) dominate the population. For $k \approx 60$ we have the perfect balance between $k$-returners and $k$-explorers, as observed for the GSM dataset for $k = 4$ (Figure 5.4b). Thus, the EPR model overestimates by more than one order of magnitude the number of locations needed to accurately estimate the total radius of gyration. Contrarily to the empirical results, in the EPR model there is no significant correlation between total $r_g$ and the sum of the distances of the $k$ most visited locations (Pearson coefficient is close to zero), neither for $k$-returners nor for $k$-explorers (see Section 5.6.5).

The observed discrepancies between the empirical data and the EPR model could arise from the fact that in the model individuals can travel arbitrarily large distances, increasing their total $r_g$ with each jump. To correct for this limitation we propose the $d$-EPR model, in which an individual selects a new location to visit depending on both its distance from the current position, as well as its relevance measured as the overall number of calls placed by all individuals from that location. We use the gravity model [32, 149] to assign the probability of a trip between any two locations, which automatically constrains individuals within the country’s boundaries (see Section 5.6.5 for implementation details). This modification is justified by the accuracy of the gravity model to estimate origin-destination matrices at the country level [150, 151, 152]. The obtained $d$-EPR model generates trajectories that are in much better agreement with the empirical data: the balance between $k$-returners and $k$-explorers in the population is reached at $k \approx 9$, in contrast with $k \approx 60$ in the original EPR model (Figure 5.6f), closer to $k = 4$ in GSM and $k = 2$ in GPS (Figure 5.3). Consequently the correlation plot of $r_g^{(k)}$ vs total $r_g$ displays the empirically observed split into returners and explorers (Figure 5.3 even at $k = 2$, Figure 5.6d). The correlation between total $r_g$ and the distance between the most visited locations is much higher than in the original EPR model and closer to the values of GSM and GPS data (see Section 5.6.3).

Hence, the $d$-EPR model of human mobility reproduces the key features of the aggregated mobility patterns in a confined geographical space, accounting for the two classes of individuals, returners and explorers. Further investigation is needed to fully understand the mechanism behind the discontinuous transition of individuals between the two classes as $k$ increases.

5.4 The impact of returners and explorers on spreading phenomena.

Our findings are particularly relevant to the geographical spreading of an epidemic, which are known to be a direct consequence of individuals’ movements [153, 154, 133, 130]. From “patient zero” (i.e. the first infected individual) the virus is passed on to individuals who come into contact with him, contributing to the rapid growth of the epidemic. Obviously, the wider the range of mobility, the faster will the virus diffuse over the population. The question is, how does the presence of the two mobility profiles uncovered above affect the spreading pattern.
5.4. THE IMPACT OF RETURNERS AND EXPLORERS ON SPREADING PHENOMENA.

Figure 5.6: EPR model predictions. (a, b, c) The prediction of the EPR model for $k = 2$. We find that 2-explorers dominate the population of synthetic individuals, and the balance in the population is reached only for $k = 60$, in contrast with $k = 4$ in the empirical data. (d, e, f) The results of the $\delta$-EPR model for $k = 2$. In this case the 2-explorers continue to dominate the population, although the balance is reached at lower values of $k = 9$, coming closer to empirical data. The insets in a, d magnifies the plot at smaller values of the total radius of gyration.

To test this, we split the mobility history of an individual into time periods, and capture the trajectory’s reach up to time $t$ using three measures: (i) the number of locations visited; (ii) the area covered; and (iii) the total radius of gyration $r_g(t)$. We observe that the trajectory of explorers is distributed over a larger territory, as they visit more locations, cover a larger geographic area, and have a higher $r_g(t)$ with respect to returners. This pattern applies both for GSM and GPS data (Figure 5.7).

To assess the different role the returners and explorers play in diffusion and spreading processes we consider the global mobility networks generated by individual mobility. The global mobility network is a graph whose nodes are locations and edges indicate the existence of at least one trip between two locations. To be specific, we focus on Tuscany, estimating the mobility of each individual through the GPS data and the number of residents in the locations through the official census cells provided by the Italian National Institute of Statistics (ISTAT). We build two separate global mobility networks, one considering only the trips of 10,000 randomly selected 2-returners, and the other considering only the trips of 10,000 randomly selected 2-explorers. For each network we compute the
mean degree $\langle k \rangle$, the mean square degree $\langle k^2 \rangle$, and the mean number of residents in the each location $\bar{N}$. We use these values to determine the global invasion threshold \[ R_s = \bar{N} \cdot C(\langle k^2 \rangle - \langle k \rangle)/\langle k \rangle^2, \] under the assumption of a diffusion dynamics with large subpopulations and a low reproductive number (i.e. close to the subpopulation epidemic threshold). Here the constant $C$ depends on the disease model and the mobility parameters (e.g. the reproductive number and the mobility rate), which are the same for the two classes of user profiles. In a metapopulation network an epidemic can spread and invade the system only if $R_s > 1$, and this global invasion threshold is affected by the topological fluctuations of the network’s degree: the larger is the degree heterogeneity, the higher is $R_s$ and therefore the higher is the chance that the epidemic will globally invade the metapopulation. We repeat the experiment 1,000 times, randomly choosing 10,000 2-returners and 10,000 2-explorers, and obtaining 1,000 values for the invasion threshold (Figure 5.8). The distributions of invasion thresholds for Tuscany for the 2-returners and the 2-explorers are peaked and well separated, finding that the average invasion threshold of 2-explorers is 11% higher than that of 2-returners. This means that for a disease model with the same parameter values, an epidemic diffusing on the 2-explorers’ mobility network
is more likely to invade the system than on the 2-returns’ mobility network. In other words, explorers are more effective at spreading a pathogen than returners. Although more refined metapopulation infection models are needed to provide accurate estimates of invasion probabilities, our analysis reveals a clear distinction between the diffusion properties of the returners and explorers’ mobility networks.

Figure 5.8: The impact of returners and explorers on spreading phenomena. (a) The distribution of the diffusion invasion threshold of 2-returners and 2-explorers in Tuscany over 1,000 experiments. In each experiment we choose 10,000 2-returners and 10,000 2-explorers randomly from the population. The global invasion threshold of 2-explorers is higher, hence it is easier for them to spread an epidemic. (b, c) The 100 most populated nodes of the global mobility networks of 2-returners (b) and explorers (c). The size of nodes and edges are proportional to the residents in the corresponding locations and to the number of individuals who traveled between the endpoints of the edge, respectively. The $<k>$ indicates the mean degree, density the network density (actual edges over possible edges), and cc indicates the average clustering coefficient of the network.

5.5 Discussion

In this chapter we report the existence of two distinct profiles characterizing human mobility: returners and explorers. Returners limit much of their mobility to a few locations, hence their recurrent and overall mobility are comparable. In contrast the mobility of explorers cannot be reduced to few locations. These patterns cannot be explained by the Exploration and Preferential Return (EPR) model of human mobility, unable to distin-
guish returners from explorers. We show that by incorporating a gravity model into EPR we can recover the two classes, the obtained extended model coming closer to the empirical observations characterizing the two profiles. The returner/explorer dichotomy has a strong impact on the spreading patterns: we show that explorers are the key actors in the spreading of diseases, ideas and information. The emerging profiles of returners and explorers offer another step towards deriving accurate models of human mobility, capable of generating realistic simulations, predictions and what-if reasoning.

5.6 Technical details and Supplementary Information

5.6.1 Data Filtering

We apply several filters to the GSM data. Firstly, for each user $u$ we discard all the locations with a visitation frequency $f = n_i/N \leq 0.005$, where $n_i$ is the number of calls performed by $u$ in location $i$ and $N$ the total number of calls performed by $u$ during the period of observation. This condition checks whether the location is relevant with respect to the specific call volume of the user. Since it is meaningless to analyze the mobility of individuals who do not move, all the users with only one location after the previous filter are discarded.

We select only active users with a call frequency threshold of $f = N/(24 \times 91) \geq 0.5$, where $N$ is the total number of calls made by $u$, $24$ is the hours in a day and $91$ the days in our period of observation. Finally, to exclude abnormally active users like line testers and alarm managers we discard the users with a huge number of calls $N > k \times 91$, where $k = 300$. Starting from $\approx 3$ millions users, the filtering results in 67,049 active mobile phone users.

Since GPS data do not provide explicit information about visited locations, we assigned each origin and destination point of the obtained sub-trajectories to the corresponding census cell, according to the information provided by the Italian National Institute of Statistics (ISTAT). As for the GSM data, we describe the movements of a vehicle by the time-ordered list of census cells where the vehicle stopped (Table 5.1). We filter the data by focusing only on trips performed within a single region (Tuscany), and by discarding all the vehicles with only one visited location or with less than one trip per day on average during the period of observation. This filtering results in a dataset of 46,121 cars.

Table 5.2 summarizes some characteristics of the datasets. The GSM and the GPS datasets differ in several aspects [14, 127]. The GPS data refers to trips performed during one month (May 2011) in an area corresponding to a single Italian region, while the mobile phone data cover an entire European country and a period of observation of three months. The GPS data represents a 2% sample of the population of cars in Italy [14], while the mobile phone dataset covers users of a major European operator, about the 25% of the country’s adult population. The trajectories described by mobile phone data include all possible means of transportation. In contrast, the GPS data refers to vehicle displacements only. The fact that one dataset contains aspect missing in the other dataset makes the two types of data suitable for an independent validation of the universality of the patterns.
5.6. TECHNICAL DETAILS AND SUPPLEMENTARY INFORMATION

<table>
<thead>
<tr>
<th>timestamp</th>
<th>Origin</th>
<th>Destination</th>
<th>Car</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011/05/12 08:31:20</td>
<td>(32.567, −2.546)</td>
<td>(32.7, −2.511)</td>
<td>F45J23</td>
</tr>
<tr>
<td>2011/05/24 17:53:08</td>
<td>(32.1982, −2.333)</td>
<td>(33.123, −2.31)</td>
<td>H2705L</td>
</tr>
</tbody>
</table>

Table 5.1: **Example of GPS records.** The GPS device automatically turns on when the car starts, and the global trajectory of a vehicle is formed by the sequence of GPS points that the device transmits each 30 seconds to the server via a GPRS connection. We exploit the stops to split the global trajectory into several sub-trajectories, that correspond to the single trips performed by the vehicle.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Volume</th>
<th>Space</th>
<th>Time</th>
<th>Conveyance</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPS</td>
<td>46,121 users</td>
<td>region</td>
<td>1 month</td>
<td>cars</td>
</tr>
<tr>
<td>GSM</td>
<td>67,000 users</td>
<td>country</td>
<td>3 months</td>
<td>many</td>
</tr>
</tbody>
</table>

Table 5.2: Summary of properties of the datasets.

emerging from human mobility behavior. The works in [14, 127] summarizes the main differences between the general mobility describe by GSM data and the vehicle mobility described by GPS data.

5.6.2 Mobility Measures

We evaluate the importance of a GSM location, i.e. its weight, using the visitation frequency as suggested in the seminal paper by González et al. [8]. For GPS data, we measure the weight using the dwell time of a vehicle in a certain census cell. As Figure 5.9 suggests, while in the GSM data there is no significant difference between the frequency and time distribution, for the GPS data the time spent in a location is much more discriminant of the importance of a location.

5.6.3 The patterns of returners and explorers

Returners and explorers are characterized by different spatial distribution of the visited locations. In both cases locations tend to group in dense clusters, however for returners the two most frequent locations tend to separate with increasing total $r_g$.

For explorers they remain close and total $r_g$ grows because other groups of locations emerge far from the center of mass. The higher the total radius of gyration, the stronger is the difference between the two classes. These observations suggest the distance between the $k$ most important locations grows with total $r_g$ more rapidly for returners than explorers. We measure the correlation between total $r_g$ and the sum of the distances between the $k$ most frequented geographic locations $\sum_{i,j\in\{1,\ldots,k\}} dist(L_i, L_j)$, both for GSM and GPS data. The correlation is significantly stronger for $k$-returners than $k$-explorers (0.96 vs 0.22 in GSM data, 0.98 vs 0.46 in GPS data, $k = 2$, Figure 5.10).

To study the tendency of locations to group in dense clusters, we define the cluster $k$-radius $c\cdot r_g^{(k)}$ as the radius of gyration computed on the $k$ most frequented geographic
CHAPTER 5. RETURNERS AND EXPLORERS: DICHOTOMY IN HUMAN MOBILITY

Figure 5.9: Bar charts of the mean frequency and mean time spent (in hours) by all the users in the ten most frequent locations, for the GSM (left) and the GPS (right) datasets. Whiskers indicate the standard deviation of frequencies and times.

Figure 5.10: Scatterplots of total $r_g$ versus the distance between the two most frequent locations $dist(L_1, L_2)$ for 2-returners and 2-explorers, for GSM data (a, b) and GPS data (c, d). The correlation is much stronger for $k$-returners than $k$-explorers, as we can see from the values of the Pearson correlation coefficient.
clusters. A geographic cluster of an individual is a dense group of locations representing a geographical unit of individual’s mobility. An individual that commutes weekly between two homes in two different cities has (at least) two different geographic clusters. The $c-r_g^{(k)}$ is computed on the $k$ most frequented clusters, considering the most frequent location of each cluster only. In the above cited example, the cluster radius of an individual corresponds to the radius of gyration computed on the two different homes. We compute the geographical clusters through the DBSCAN algorithm [39], which extracts dense groups of points according to two input parameters: $eps$, the maximum search radius; and $minPts$, the minimum number of points (locations) to form a cluster. We set $minPts = 2$ and $eps = 5, 10, 50, 100$km. The split into returners and explorers emerges also at cluster level, and it is clear until $eps = 10$km, where the clusters have the size of a medium sized city. For high values of $eps = 50, 100$km the number of computed clusters is small (mainly 2), penalizing the presence of explorers (Figure 5.11). The presence of returners and explorers in the population, hence, is independent of the spatial granularity of individuals’ location: it appears for GSM towers as well as when we take districts or entire cities as individual locations.

Figure 5.11: Scatterplots of total $r_g$ versus the cluster-$r_g^{(k)}$, for $k = 2$, and GSM data. The geographical clusters are computed through the DBSCAN algorithm with parameters $eps = 5, 10, 50, 100$km and $minPts = 2$. The returner/explorer dichotomy appears again and it is clear until $eps = 10$km (b), where clusters have the size of medium-sized city.
5.6.4 Split methods

We develop three methods to split the population into returners and explorers. The bisector method uses a curve bisecting the plane to detect the subpopulation of \(k\)-returners. A Support Vector Machine (SVM) and the Expectation-Maximization (EM) clustering algorithm extract the two patterns from the population by means of data mining techniques.

The bisector method uses the curve \(r_g^{(k)} - r_g/2 = 0\) to bisect the plane, defining all the users above the curve as \(k\)-returners. Figure 5.13 shows how the number of \(k\)-returners varies with the number of locations \(k\) considered into the \(k\)-radius. Figure 5.12a and d shows the split of the population according to the bisector method.

Support Vector Machines (SVM) [39] are supervised learning models that analyze data and recognize patterns. We first build the SVM classifier providing a set of training examples to the SVM learning algorithm, and then used the built model to classify individuals as \(k\)-returners or \(k\)-explorers. We describe each individual as a pair \((r_g^{(k)}, r_g)\). As training examples, we select the individual falling exactly on the diagonal (\(k\)-returners) or the abscissa (\(k\)-returners) of the total \(r_g\) versus \(r_g^{(k)}\) plot. Precisely, \(k\)-returners examples are all the individual for which \(r_g^{(k)} = r_g\), while \(k\)-explorers examples are all the individual for...
which \( r_y^{(k)} = 0 \). Figure 5.12b, e shows the split of the population according to the SVM method.

The Expectation-Maximization (EM) algorithm [39] is an iterative method for finding maximum likelihood of parameters in statistical models. It alternates between an expectation (E) step, which creates a function for the expectation of the log-likelihood based on the current estimate for the parameters, and a maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the E step. These parameter-estimates are then used to determine the distribution of the latent variables in the next E step. The EM algorithm outputs a pair of values for each individual, representing the probability to be a \( k \)-returner and a \( k \)-explorer. We assign each individual to the category with the highest probability. Figure 5.12c, f shows the split of the population according to the EM method.

The three methods produce similar trends of variation with \( k \) (Figure 5.13). For GSM data, \( k \)-returners are initially the minority in the population. They start outnumbering the \( k \)-explorers from \( k = 4 \). In the GPS case \( k \)-returners are immediately the majority, and the gap increases with the value of \( k \). Since the methods produce similar results, we focus on the simplest, the bisector method.

Figure 5.14 shows how the Pearson coefficient of the correlation coefficient varies with the number of locations \( k \). Two different behaviors emerge for GSM and GPS data. In the former case, returners’ correlation is strong and stable, explorers’ correlation is weak and slight increases. In the latter case, the returners’ correlation rapidly falls down, while the explorers’ one rapidly increases. The two curves crosses at \( k = 8 \).

5.6.5 Mobility Models

The Exploration and Preferential Return (EPR) individual mobility model [10] is a state-of-the-art model that accurately captures the visitation frequency of locations, the distribution of the radius of gyration across the population and its growth with time. The model does not fix the set of preferred locations but allows them to emerge naturally during the evolution of the mobility process. It incorporates two competing mechanisms: exploration and preferential return. Exploration is a random walk process with truncated power law jump size distribution. Preferential return reproduces the propensity of humans to return to the locations they visited frequently before. An agent in the model selects between the two modes: with probability \( P_{\text{new}} = \rho S^{-\gamma} \) (where \( S \) is the number of locations visited so far by the agent, \( \rho \) and \( \gamma \) are two model parameters), the individual moves to a new location, whose distance from the current one is chosen from the known power law distribution of displacements. With complementary probability \( P_{\text{new}} = 1 - \rho S^{-\gamma} \), the agent returns to one of the \( S \) previously visited places (with the preference for a location proportional to the frequency of visits). As a result, the model has a warmup period of greedy exploration, while in the long run agents mainly move around a set of previously visited places.

We implemented the original version of the EPR model, along with two improved versions: the \( s \)-EPR model, where agents are constrained within a limited geographical
CHAPTER 5. RETURNERS AND EXPLORERS: DICHOTOMY IN HUMAN MOBILITY

Figure 5.13: Number of returners and explorers in the population with $k = 2, 10$, for GSM data (a, b, c) and GPS data (d, e, f), according to the three split methods described above. In GSM data an balance of the two profiles is reached at $k = 4$, while in GPS data returners are immediately more numerous than explorers.

Figure 5.14: Values of the Pearson coefficient of the correlation between total $r_g$ and $\text{dist}(L_1, L_2)$ with $k = 2, 10$, for GSM data (a) and GPS data (b). For returners the correlation is much stronger. In GPS the two Pearson coefficients come close to each other as $k$ increases and the curves cross at $k \sim 7$. 
space; and the \( d \)-EPR model, in which an individual selects a new location depending on both its distance from the current position and its relevance measures ad the overall number of calls places by all users from that location. We use the gravity model to assign the probability of a trip between any two locations automatically constraining individuals within the country’s boundaries.

Figures 5.15, 5.16 and 5.17 show the patterns of returners and explorers emerging from the three versions of the EPR model.

**Implementation of the original EPR model**

We generate an initial (home) location for each of the 67,000 synthetic individuals by randomly selecting a point on a square of size 100 \( \times \) 100. We then repeat the following steps 1,000 times for each individual:

1. We extract a waiting time \( \Delta t \) from the distribution \( P(\Delta t) \sim \Delta t^{-1-\beta} \exp(-\Delta t/\tau) \), with \( \beta = 0.8 \) and \( \tau = 17 \) hours as measured in [10].

2. With probability \( P_{\text{new}} = \rho S^{-\gamma} \), where \( S \) is the number of distinct locations previously visited and \( \rho = 0.6 \) and \( \gamma = 0.21 \) [10], the individual visits a new location (step 3), otherwise she returns to a previously visited location (step 4).

3. If the individual explores a new location, a distance \( \Delta r \) is extracted from the distribution \( P(\Delta r) = \Delta r^{-1-\alpha} \) with \( \alpha = 0.55 \) as in [10], and the individual moves to a randomly selected location on the circle of radius \( \Delta r \) centered on her current location. The number of distinct locations visited, \( S \), is increased by one. The new locations can be outside the initial 100 \( \times \) 100 square.

4. If the individual returns to a previously visited location, it is chosen with probability proportional to the number of visits to that location.

**Implementation of the \( s \)-EPR model**

We place each of the 67,000 GSM users in her most visited location (GSM cell phone towers). For each individual we repeat the following steps:

1. *Same as the original model.*

2. *Same as the original model.*

3. If the individual explores a new location, a distance \( \Delta r \) is extracted from the distribution \( P(\Delta r) = \Delta r^{-1-\alpha} \) with \( \alpha = 0.55 \) as in [10], and an angle \( \theta \) between 0 and \( 2\pi \) is extracted with uniform probability. If the location at distance \( \Delta r \) and angle \( \theta \) from the current location is not in the country’s boundaries a new distance and a new angle are extracted until this condition is satisfied. When the new location is found the number of distinct locations visited, \( S \), is increased by one.

4. *Same as the original model.*
Figure 5.15: The correlations between total $r_g$ and $r_g^{(k)}$ for $k = 4, 10, 60$ for the s-EPR dataset (a, b, c) and the d-EPR dataset (d, e, f).
Figure 5.16: Distributions of the ratio $s_k$ with $k = 4, 10, 60$ for the original EPR model (a, b, c), the $s$-EPR model (d, e, f) and the $d$-EPR model (g, h, i). We used 30 equal sized bins.
Figure 5.17: Correlation between total $r_g$ and the distance between the two most frequent locations $\text{dist}(L_1, L_2)$ of returners and explorers, with $k = 2$, for the original EPR model (a, b), the $s$-EPR model (c, d) and the $d$-EPR model (e, f). In all the cases the correlation is much stronger for returners.
5.6. TECHNICAL DETAILS AND SUPPLEMENTARY INFORMATION

Figure 5.18: The distribution of total $r_g$ for the original EPR model (a), the s-EPR model (b), and the d-EPR model (c).

Implementation of the d-EPR model

We place each of the 67,000 GSM users in their most visited location (GSM cell phone towers). For each individual we repeat the following steps:

1. *Same as the original model.*

2. *Same as the original model.*

3. If the individual who is currently in location $i$ explores a new location, then the new location $j \neq i$ is selected according to the gravity model [32, 149] with probability $p_{ij} = \frac{1}{N} \frac{n_i n_j}{r_{ij}^2}$, where $n_{i(j)}$ is the total number of calls placed by all users from location $i(j)$ representing its relevance, $r_{ij}$ is the geographic distance between $i$ and $j$, and $N = \sum_{i,j \neq i} p_{ij}$ is a normalisation constant. The number of distinct locations visited, $S$, is increased by one.

4. *Same as the original model.*
Chapter 6

Learning Activities from Individual Mobility Networks

Human mobility is driven by our daily activities, such as going to work or school, shopping, transporting kids, and so on. The digital mobility traces collected through a variety of technologies, from navigation devices to smart phones, allow us to understand people’s movements in great detail. However, they generally fail to capture the purpose of such movements, i.e. the kind of activity behind each travel. This deficiency is a hard obstacle to the deployment of Big Data in many domains such as urban planning, traffic management, intelligent transportation systems, socio-demographic simulation and nowcasting, and emergency management [143]. For all such applications, information on why people move is crucial. The availability of society-wide data about mobility and activity of people would be a driver for a better comprehension of our complex society, and for smarter knowledge services for the individual and the collective sphere.

It is not a surprise, hence, that several researchers tackled the problem of activity recognition, i.e. how to infer the kind of activity associated to a travel on the only basis of the observed mobility patterns [155]. Show me how you move, I’ll tell what you do. The rationale behind such research follows a two steps method. First, use a small training mobility dataset annotated with activity information, obtained for instance by surveying some volunteers, to learn a classifier that maps mobility-related features into the different kinds of activities. Second, apply the classifier to unlabeled Big Data, to obtain large-scale mobility data annotated with activity information — the activity classifier acts as a semantic amplifier of Big Data. It goes without saying that this second step can be successful only if the predictive accuracy of the classifier is extremely high. Unfortunately, none of the methods for activity recognition proposed so far reach adequate performance for semantic amplification. Some existing methods, such as Conditional Random Fields (CRF) [156], obtain very accurate learners that, given a past history of activity-labelled movements of individuals, predict the activity associated to future unlabeled trips of the same individuals. As we discuss in this chapter, such learners exhibit poor performance when used to predict the activity associated to the movements of other individuals, whose data were not used in the learning process; which is the situation we face in semantic
amplification of unlabeled Big Data.

In this chapter, we describe a model for activity recognition targeted explicitly at the semantic amplification of big mobility data, called Activity-Based Cascading (ABC) classification. ABC departs completely from probabilistic approaches for two main reasons: (i) it exploits a set of structural features extracted from the Individual Mobility Network (IMN), a model able to capture the salient aspects of individual mobility; (ii) it uses a cascading classification as a way to tackle the skewed frequency of activity classes (home and work are generally very frequent compared to shopping and leisure). We leverage on a dataset of approximately 7,000 activity-annotated trips obtained from GPS receivers on board of private cars, and show how ABC classification reaches high precision (up to 0.98) and outperforms both state-of-the-art probabilistic methods (CRF) and decision tree classifiers. In summary, the novel contributions of the chapter are the following:

- Summarization of individual mobility by introducing a novel model based on a graph-based representation: Individual Mobility Networks (Section 6.2);
- Selection of predictive features extracted from the IMNs to improve decision tree classification (Section 4-A);
- Enhancing Cascade Classification with label propagation through successive steps (Section 4-B).
- Comparison with state-of-the-art methods to asses validity of the approach (Section 5);

The work in this chapter is based on the paper in [148], published in the proceedings of the 2014 International Conference on Data Science and Advanced Analytics (DSAA’14).

6.1 Related Works

The task of activity recognition consists in assigning a label to a movement according to its relevant characteristics. The vast literature on the subject may be organized according to the type of movement observed. Many works focus on the movement of individuals to recognize gestures, indoor activities, physical activity levels, surveillance, outlier and intrusion detection [157, 158]. We are mainly interested in those works considering the movement as the physical change in position of the individual, thus leaving a geographical place to reach another one. We can identify two large groups of inference methods: supervised and unsupervised.

Among the supervised approaches, some methods try to infer the mode of transportation [159, 160], the activity performed in a specific location [161, 162] and a combination of the two [155]. The methods that deals with the transportation mode try to infer if the individual is moving by foot, by car, by bike or by public transportation. This annotation exploits several features of the movements, speed, acceleration and, when available, other context data like accelerometer measurements. The learning approaches are based on discriminative methods, like decision trees [160] and conditional random fields.
6.2. INDIVIDUAL MOBILITY NETWORKS

( CRF) [155, 163]; or on generative methods, like Hidden Markov Models (HMM). We are not interested in the transportation mode, since we focus on the prediction of the activity at destination. When considering the activity from movement we can distinguish two main approaches: sequence learning approaches consider the activity of an individual as a sequence in a fixed temporal period (usually one day) and try to predict the labels for the whole sequence [155]; episode learning approaches try to label each single movement episode independently from the others [160].

Unsupervised methods are mainly based on clustering techniques [164, 165] or dimensionality reduction [164, 166]. Jiang et al. [164] analyze an activity-based travel survey conducted in the Chicago metropolitan area with the aim of exploring the daily activity structure of people. They describe how the considered population can be clustered into eight (weekdays) or seven (weekends) representative groups according to the activities performed by the individuals.

6.2 Individual Mobility Networks

An Individual Mobility Network (IMN) describes the individual mobility of a person through a graph representation of her locations and movements, grasping the relevant properties of individual mobility and removing unnecessary details.

Definition 6.2.1 An Individual Mobility Network (IMN) of an individual $u$ is a directed graph $G_u = (V, E)$, where $V$ is the set of nodes and $E$ is the set of edges. On nodes and edges the following functions are defined:

- $\omega : E \rightarrow \mathbb{N}$ returns the weight of an edge (i.e. the number of travels performed by $u$ on that edge);
- $\tau : V \rightarrow \mathbb{N}$ returns the time spent by the user in a given location;
- $p_e : E \times T \rightarrow [0, 1]$ estimates the probability $p_e(e, t)$ of observing the user $u$ moving on edge $e$ at time $t$;
- $p_l : V \times T \rightarrow [0, 1]$ estimates the probability $p_l(v, t)$ of observing the user $u$ at location $v$ at time $t$.

Nodes represent locations and edges represent movements between locations. We attach to both nodes and edges statistical information by means of structural annotations: edges provide information about the frequency of movements through the $\omega$ function; nodes provide an estimation of the time spent in each location through the $\tau$ function.

To clarify the concept of IMN, let us consider the network in Figure 6.1. It describes the IMN extracted from the mobility of an individual who visited 19 distinct locations. Location $a$ has been visited a total of 18 time units (days in the example), since $\tau(a) = 18$. The edge $e = (a, b)$ has weight $\omega(e) = \omega(a, b) = 20$, indicating that the individual moved twenty times from location $a$ to location $b$. 
The IMN of an individual is an abstraction of her mobility behavior. A location is an abstract entity without any reference to the geographic space. It can be interpreted as a subjective point of interest, a place around which the mobility of that individual gravitates. This allows the modeling of locations that are meaningful only for that individual, like his home or work place, etc. Accordingly, given the IMNs of two distinct individuals we are not able to determine whether they have visited the same location. This limitation, on the other hand, allows us to hide the actual places visited by the individual, providing a protection layer of sensitive information.

The computation of a IMN starts from the ordered sequence of an individual’s trajectories. It works in a “streaming” fashion: every time the individual performs a new trip, her IMN is updated according to the new trajectory. The origin and destination points of the new trajectory are mapped to locations in the IMN. The locations are obtained by aggregating all the origin and destination points of the past trajectories within a spatial threshold $\delta$.

**Definition 6.2.2** Let $P_u$ be the set of origin and destination points of an individual $u$, and $\delta$ a distance threshold. A location $L$ of $u$ is the aggregation of points in $P_L = \{ p \in P_u | \forall p_i, p_j \in P_L : d(p_i, p_j) \leq \delta \}$, where $d(p_i, p_j)$ is the Euclidean distance between $p_i$ and $p_j$. 

---

**Figure 6.1:** The IMN extracted from the mobility of an individual. Edges represent the existence of a route between locations. The function $\omega(e)$ indicates the number of trips performed on the edge $e$, while $\tau(x)$ the total time spent in a location $x$. 

- $\tau(n) = 3$
- $\tau(q) = 8$
- $\tau(c) = 12$
- $\tau(a) = 18$
- $\omega(b, a) = 2$
- $\omega(b, c) = 1$
- $\omega(a, b) = 10$
- $\omega(a, c) = 5$
- $\omega(a, i) = 6$
- $\omega(a, s) = 2$
- $\omega(a, t) = 4$
- $\omega(a, u) = 3$
- $\omega(a, m) = 2$
- $\omega(a, o) = 1$
- $\omega(a, l) = 1$
- $\omega(a, g) = 1$
- $\omega(a, d) = 1$
- $\omega(a, h) = 1$
- $\omega(a, r) = 1$
- $\omega(a, p) = 1$

This diagram illustrates the IMN extracted from the mobility of an individual. The locations and edges show the routes and the number of trips to each location.
6.3. INDIVIDUAL ACTIVITY RECOGNITION

Algorithm 1 Aggregate(\( \mathcal{G}, t, \delta \))

<table>
<thead>
<tr>
<th>Line</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:</td>
<td>( SP = \text{getStartingPoint}(t) )</td>
</tr>
<tr>
<td>2:</td>
<td>( EP = \text{getEndingPoint}(t) )</td>
</tr>
<tr>
<td>3:</td>
<td>( SL = \text{getNearestLocation}(SP, \delta) )</td>
</tr>
<tr>
<td>4:</td>
<td>( EL = \text{getNearestLocation}(EP, \delta) )</td>
</tr>
<tr>
<td>5:</td>
<td>( \text{update}(\tau(SL)) )</td>
</tr>
<tr>
<td>6:</td>
<td>( \text{update}(\tau(EL)) )</td>
</tr>
<tr>
<td>7:</td>
<td>( \text{update}(\omega(SL, EL)) )</td>
</tr>
</tbody>
</table>

When a new trajectory is produced, we search for an existing location, i.e. a place in space within a spatial threshold distance \( \delta \) from the current point. If such location exists the point is associated with such location; otherwise a new location is created. Algorithm 1 shows the detail of the incremental process. Since each update is performed in constant time, the aggregation procedure is very efficient.

6.3 Individual Activity Recognition

The task of activity recognition consists in assigning an activity to an individual according to her mobility habits. In this chapter we consider the specific issue of detecting the activity related to the destination of a trip, by looking at the characteristics of movements and visited locations. We tackle the problem into two steps: (i) we extracted a set of features from the IMN of each individual; (ii) we learn a classifier on the set of extracted features. To learn the classifier, we consider an extended version of the IMNs where the attributes of each trip include a label representing the activity performed at the destination node. In the learning phase, two are the critical choices to analyze: the set of features to consider (Section 6.3.1) and the learning algorithm to adopt (Section 6.3.2).

6.3.1 Individual Mobility features

We consider two distinct sets of features. The first set, named trip features, regards the characteristics of each single trip performed by an individual. As a second set of features, we introduce a new set of attributes, named network features, which regard the topological structure of each IMN.

Network features

We identify four classes of network features to capture the salient aspects of a IMN: centrality, predictability, hubbiness and volume.

Centrality measures the degree of connectivity of a location. Given a location, we want to state whether it is at the center of the individual’s mobility. For example in the network of Figure 6.1 location \( n \) cannot be considered central, since it is connected to just two edges and visited just two times from the same location \( b \). On the contrary, location \( a \) is a central place the individual visits starting from different other locations, presumably her home. The clustering coefficient [55] of a node captures these aspects by
estimating the probability that its neighbors are connected to each other. In other terms, it measures the probability that a triad of locations forms a triangle with their edges. The value of the clustering coefficient varies from zero to one: it is zero when all the triads of nodes are not closed, it is one when all the neighboring triads are connected, representing a dense mobility around the considered locations. The clustering coefficient of location $a$ in Figure 6.1 equals 0.12, while location $n$ has a clustering coefficient of 0. Another feature of centrality is the average path length \([55]\) of a location $x$: the average number of edges to traverse to reach the location $x$ from any other node in the network. A location that is at the periphery of the network is likely to have a high average path length. In the example of Figure 6.1, location $o$ is located at the border of the individual mobility network, having a high average path length. Location $a$ is at the center of the network, resulting in a quite low average path length.

Predictability measures the degree of uncertainty of an individual’s movements. In details, given a location $x$ we want to measure the accuracy of predicting the next location visited from $x$. Consider for example the central location $a$ in Figure 6.1, from which the individual visits uniformly many locations. Without other background information it is hard to estimate the probability of the next location. On the contrary, from location $n$ the individual visits always the same location $b$. To model these cases we adopt the concept of Shannon entropy \([167]\) to measure the distribution of outgoing trips from a node. Formally, the entropy of a location $x$ is given by the formula:

$$E(x) = -\frac{\sum_{y \in V} p(x,y) \log p(x,y)}{\log N},$$

where $N$ is the number of locations visited by the individual and $p(x,y)$ represents the probability of observing a travel from $x$ to $y$, i.e. $p(x,y) = \omega(x,y)/\sum_{e \in E} \omega(e)$. The entropy is zero when all the outgoing trips are concentrated in a single edge, while it is one when the trips are distributed uniformly over all the available edges. Another aspect to consider when studying the mobility uncertainty of a node is the number of neighbors in the network. Clearly, when a node has a few neighbors, the possible destinations are limited. On the IMN we can measure the degree of a location $x$ as the number of edges entering or leaving $x$ (the sum of in-degree and out-degree of the node).

Hubbiness measures the relevance of a location in the IMN. We can capture the relevance of a node $x$ by counting the different origin locations that have a trip ending in $x$ and, analogously, the number of distinct destination locations for the trips leaving $x$. These two aspects can be summarized by the in-degree and the out-degree \([55]\), i.e. the number of incoming and outgoing edges of a location. The relevance of a node (edge) in a graph is also proportional to number of path traversing that node (edge). Such property is synthesized by edge- and node-betweenness \([55]\). This measure computes, for each node (edge), the number of minimum paths connecting any two pairs of nodes in the network that pass through the considered node (edge). The value of betweenness is low when few paths traverse that portion of the network, making the node (edge) marginal. It is high when the number of passing paths is high, making the node (edge) relevant for the network.
Volume describes quantitatively the amount of mobility observed for each node (edge). In particular, we exploit the weight function $\omega$ of each edge to quantitatively measure its volume. The flow per location is measured by summing the weights of incoming (outgoing) edges. Similarly, we state the relevance of a location by considering the value of the function $\tau$ representing the time spent in that location.

**Trip features**

In the literature, the features used for the learning phase vary according to the application scenario. However, we can identify a subset of features that are shared by many approaches [163, 160, 168]. We start from this set and select the following features: length, duration, time interval (four distinct daily time intervals: 00-06, 06-12, 12-18, 18-00), and average speed.

Each family of features combines both topological properties of the graph and mobility specific measures. Moreover, each feature can be computed directly from the IMN without referring to the original trajectories used to learn the model. Table 6.1 summarizes the features used for the learning phase. From each IMN we derive a dataset where each row describes a single trip of the individual: some of the attributes of the row depend on the specific trip; the remaining features depend on the topology of the edge. Network features are the same for all the trips belonging to the same network edge. Since each edge refers to a pair of locations, the features of the nodes are computed for the outgoing node and the ingoing node and stored in distinct attributes of the table, using the suffix *-from* or *-to* to distinguish between them.

Table 6.2 displays some examples extracted from the IMN of Figure 6.1. Given an edge, the number of examples equals the value of the edge’s weight. Edge $e = (a, b)$, for instance, has weight $\omega(a, b) = 20$ and is represented by 20 rows in Table 6.2. The trip features (such as length and duration) are inferred from the distributions related to nodes and edges.
CHAPTER 6. LEARNING ACTIVITIES FROM INDIVIDUAL MOBILITY NETWORKS

<table>
<thead>
<tr>
<th>From</th>
<th>To</th>
<th>ccFrom</th>
<th>ccTo</th>
<th>weight</th>
<th>...</th>
<th>duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>b</td>
<td>0.12</td>
<td>0.22</td>
<td>20</td>
<td>...</td>
<td>11min</td>
</tr>
<tr>
<td>a</td>
<td>b</td>
<td>0.12</td>
<td>0.22</td>
<td>20</td>
<td>...</td>
<td>8min</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>b</td>
<td>n</td>
<td>0.22</td>
<td>0.0</td>
<td>2</td>
<td>...</td>
<td>3min</td>
</tr>
<tr>
<td>b</td>
<td>n</td>
<td>0.22</td>
<td>0.0</td>
<td>2</td>
<td>...</td>
<td>2min</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 6.2: Some examples of features extracted from the IMN in Figure 6.1.

6.3.2 Activity-Based Cascade Classification

We propose to re-structure the classification task into a step-wise process, which has some similarities with nested cascade classification [169]. For this reason we call it *Activity-Based Cascade (ABC)* classification.

The idea is to reduce the original multi-class classification problem to several binary classification problems, each trying to discriminate between one class and all the others. The classifiers are learnt in cascade according to ascending frequency of the classes, as for rule-based classifiers [170]. At each step of the cascade chain, the classifier tries to distinguish the instances of its class among those which were rejected from all the previous ones. For instance, if “going home” is the first class in the cascade, the first binary classifier will discriminate between “going home” and “all other classes”. Then, the second classifier will discriminate for another class without the instances of “going home”. By eliminating their instances we can simplify the job of discriminating the instances of the second type, belonging to more difficult classes.

We can further improve the method by using the classification results obtained at each step as source for later classifications. Indeed, the first classifier of the cascade allows to recognize all the trips/locations of its class (for instance, all the trips of type “going home”). We can use the information from the previous step as contextual information by computing some measures for all other trips/locations. For instance, we can compute the number of network hops between any trip/location and its closest “going home” trip/location. The result is that the number of attributes describing the instances grows at each step of the cascading process, and the later classes will benefit from a larger set of context information.

Clearly, the same step-wise process is adopted when we apply the classifier to a dataset of unlabelled IMNs, in order to perform the semantic amplification. The first step of the cascade is applied to all the trips of the IMNs to classify those ones belonging to the first class (in our example above, “going home”). Then, all the remaining trips are enriched with a context feature, e.g. their network distance from the closest trip labeled as “going home”. The process continues with the next step in the cascade. At the end of the cascade process, if some instances of the test set are classified as “all other classes”, they remain *unclassified*, meaning that the overall classifier was not able to properly classify it.
ABC Evaluation Criteria

To evaluate the accuracy of each classifier, we split the annotated dataset into a training set (used for the learning phase) and a test set (used for measuring the accuracy). Following the approach presented in [163], the split is done according to two strategies: leave-one-week-out (WO) and leave-one-user-out (UO). The first strategy extracts the trajectories of one week of each user, using them as test set. The second strategy extracts a single user from the dataset, using her whole mobility as test set.

The two accuracy measures drive the classification models towards two distinct goals. The WO strategy emphasizes what is learned from the past to predict the future. This approach is usually suited for recommendation-like services, where the previous examples are used to classify the new ones. However, it has a tendency to produce overfitted models that are hardly adaptable to new users. The UO strategy, instead, learns from the crowd to predict the activities of new users. It is more robust to overfitting since it generalizes more the learned model. In our experiments we used a generalization of this two approaches by splitting the annotated dataset into two parts corresponding to 60% and 40% of total size. Since the available dataset covers a period of one month, we generalize the WO strategy in a Temporal Split (TS) where three weeks are used for training and the last week is used as test set. The corresponding splitting strategy for UO is called User Split (US), where the 60% of the users compose the training set, and the remaining 40% compose the test set.

6.4 Experiments and Evaluations

In our experiments we used the Octo GPS dataset described in Section 3.1. A portion of the movements have been annotated by volunteers to reconstruct their activities during May 2011. In particular, volunteers annotated 6,953 distinct trips performed by 65 vehicles. The annotation was performed by using 13 distinct activities: going home, working, daily shopping, shopping, social activities, leisure, services, education and training, bring and get, touring, other. We introduced the activity type none for the activities that did not receive any tagging. Such none activities, which correspond to about the 5% of the training set, have been removed before the learning phase.

Figure 6.2 shows the distribution of activities for all the movements performed by the volunteers in the dataset. Clearly, “going home” and “working” activities have the highest frequencies in the dataset: this highlights the fact that the majority of movements are performed for systematic routines. A large part of the activities, however, is annotated with the “other” label generally due to the presence of activities that are difficult to be categorized by the user. For example, the “bring and get” activity is rarely annotated and often avoided, since it includes short stops.
6.4.1 Competing classification approaches

The task of classifying an individual’s activities by movements is usually approached with two main families of approaches. One group of methods [163, 160], based mainly on probabilistic approaches, consider the problem of tagging a whole sequence of activities of a user. They start from the assumption that an individual performs similar schedules in different days. The second group of methods, based on decision tree and Hidden Markov Models, tries to tag each single episode.

Since a IMN hides the specific sequence of movements performed by the individual, we cannot use directly sequence-based learning methods. However, we will show that the topological features of a IMN are capable of subsuming the probabilistic dependencies among the different activities of an individual. To this aim, we compare the performance of the ABC classifier introduced in Section 6.3.2 with two methods which represents the two families of approaches described above: Conditional Random Fields (CRF) and Random Forest (RF).

**Conditional Random Fields (CRF).** Conditional Random Fields (CRF) [156] are probabilistic graphical models to label sequence data. They are based on undirected graphical structures, where each link represents the conditional probability distribution over hidden states. Unlike Hidden Markov Models, there is no assumption of dependency in the structure. Liao et al. [163] show an instance of CRF for the activity recognition problem where the nodes represent sequences of observations. A trajectory is associated
with a node and all its properties are attached to the corresponding vertex, while hidden states represent the activities to be predicted. We used the network structure as presented in [163] with the implementation of the library CRF++\(^1\). Since the method depends on the sequential order of the movement episodes, we implemented this approach by considering directly the original trajectories of the users. The original article considers only trip features. We extended the implementation by including also the features derived from the corresponding IMN for each set of trajectories of the corresponding user. The efficiency of the learning phase is determined by the complexity of the model and, hence, by the number of features considered.

**Random Forest (RF).** The learning methods based on decision trees provide an efficient solution for classification problems. However, the learning phase and the classification step are usually limited to a single episode, with limited knowledge of contextual episodes (i.e. previous or successive movements). To overcome this problem, we extend the set of features for each movement to classify by coding the contextual information in a set of numeric attributes. This approach simplifies the search space compared with that structured in a CRF and focuses only on relevant properties of each movement. The procedure to extract these features from a IMN is described in Section 6.3.1. Moreover, the outcome of the learning phase of these methods is a tree-based model that allows the analyst to reason and interpret the outgoing model easily. The extended set of features is effective both for decision trees and CRFs. Among the available solutions for decision tree classifiers, in our experiments we chose Random Forest classifiers, using the implementation provided by the scikit-learn Python package.\(^2\)

### 6.4.2 Accuracy evaluation

How to compare the performances of the two learning approaches presented in Section 6.3.2? On one hand, the CRF classifier uses the raw trajectories of the training set. On the other hand, the Random Forest classifier is based on the IMNs extracted from the raw trajectories. The set of classes to predict, conversely, is fixed for both the methods. We compare the two approaches by means of two distinct accuracy methods, the TS and the US methods introduced in Section 6.3.2; using only trip features and both trip and network features.

Figure 6.3 shows the accuracy results for the two methods. Both methods have better performances when using the TS accuracy: they are more robust when we learn the model from the past and apply it to the future. Using the TS accuracy metric, the CRF classifier outperforms the RF, although the extended set of features (the ones including the network features) negatively influences the CRF results. In the RF, on the contrary, the extended network features are effective and contribute to increase the accuracy of the classification. The TS accuracy measure, however, has a marginal significance for our problem. Our goal is to perform a semantic amplification: extend the annotation performed on a small

---

\(^1\)https://code.google.com/p/crfpp/

\(^2\)http://scikit-learn.org/
portion of the population to a larger set of moving users. Therefore, in our context the US accuracy is the crucial scenario.

Moving to the US accuracy, both models produce a significant drop in the accuracy. In particular the CRF method shows a significant difference w.r.t. the TS scenario, suggesting a low efficiency in generalizing the model to new behaviors. The RF method suffers from an analogous drop in accuracy, but reach better overall performance than CRF on the extended set of features.

The experiments show that the extended network features we propose are effective for both methods. Moreover, they are capable of improving prediction accuracy for both the TS and the US accuracy measures. When considering the US accuracy, however, their contribution is still more valuable, since they are capable of representing general attributes of individual mobility. Thus, the RF based approach we propose is proven to be more effective from two points of view: (i) we have a very compact and aggregated view over individual mobility; (ii) we are capable to better generalize mobility features to predict users’ activities.

6.4.3 Results

In this section we compare the results of the RF classifier and the ABC classifier introduced in Section 6.3.2, using the US accuracy measure. As discussed in Section 6.3.2, the ABC method might leave some output instances unclassified. In this context the accuracy
(which would simply consider unclassified instances as errors) would not be a sufficiently informative measure. For this reason we adopt the standard precision and recall measures, in order to better highlight the trade-off between quality of the classification versus completeness. Table 6.3 shows, for each class, the precision and the recall achieved by the two classifiers. Results for RF shows that some classes, e.g. “education”, “bring/get” and “touring”, have low precision and recall, presumably due the low frequency of such classes in the annotated dataset (see Figure 6.2).

The ABC classification presents several advantages with respect to the RF. First, it mitigates the differences in the support of some activities, such as the “bring and get” activity. Moreover, it produces specialized decision tree models, which help in the understanding of each activity. The ABC classification is presented with two different modes: (i) a step-wise cascading classification; (ii) a propagation of the classes assigned in the previous steps to the current one. In both cases, the accuracy of the ABC approach outperforms the RF approach presented above and provides better results (around 10% of increment of accuracy) with label propagation. The comparative results are shown in Table 6.3. The test set used to evaluate the performance contains 2,594 tuples. The table reports the support for each activity, i.e. the number of samples for which the ABC classifier succeeds in assigning a label. It is clear that the ABC classifier cannot to classify all the instances in the test set. In the last column of Table 6.3 we report the number of tuples labelled by the classifier. At the end of the cascading classification, the samples that are not classified by the last decision tree remain unclassified. This is due to the selectiveness of the decision tree in the cascading chain, which is emphasized in the label propagation mode, since the additional features contribute to better distinguish the current class from the remaining ones. As a consequence, ABC classifies a reduced number of examples, but with a higher confidence.

ABC classification uses a fixed order of the classes in the cascading chain: in our experiments we sorted in ascending order the classes according to the their support, following the strategy in [170]. To check the significance of such sorting with respect to the outcome of the classification, we performed a series of classification processes by considering different shuffling of the class order (Figure 6.4). It is clear that the ordering have little and statistically non-significant influence on the accuracy of the ABC classifier.

### 6.5 Conclusions

In this chapter we described a methodology to recognize individuals’ activities from the observation of their movements. The approach is based on the definition of an Individual Mobility Network (IMN), a graph-based model capable of grasping salient properties of individual mobility. The set of annotated movements are enriched with features extracted from the IMN, and used to train a classifier for mapping mobility properties to semantic activities. The model we proposed, Activity-Based Cascading (ABC) classification, departs completely from previous approaches and outperforms the state-of-the-art probabilistic methods.
Table 6.3: Precision and Recall of the RF classifier (left) and the ABC classifier (right) with extended features, label propagation, and US validation.

We applied the semantic amplifier to a large dataset of mobility data collected by users living in Tuscany. Figure 6.5 shows the distribution of activities during the day: the movements annotated as “going home” are still the majority with a peak in the late afternoon, typically the time people go back home. Figure 6.6 shows the same distribution after removing “going home” activities, in order to highlight the other distributions. As we see, the working activity shows a clear peak in the morning, while shopping movements are clearly constrained in the opening hours of shops. It is worth noting the increase of social activities in the late afternoon (after the working hours).

Since the features of IMNs are abstracted from real geography, we can exploit this methodology to transfer the model to a different territory. From the experiments we have evidence that the approach is valid at a regional level, since the outcomes of the annotation within the same region provide reasonable results. As a preliminary proof, we focused on two cities in Tuscany, restricting the training set to Pisa and the test set to Florence. The results are promising: high frequency activities (going home and working) show high precision (respectively 90% and 80%). Activities with lower support, instead, have lower accuracy because of the low number of samples.

An open line of investigation is the influence of the geographical scale on the learning step. Clearly, trip features contain relative distances between locations and the expected duration of the trips. If we learn the model an a small-sized city, may we transfer it to a large-sized city?
6.5. CONCLUSIONS

Figure 6.4: Distribution of precision for 10k executions of the ABC classifier, varying class order for progressive classification.

Figure 6.5: Distribution of activities per hour. Time is shown in GMT+0, but real time is in GMT+1.
Figure 6.6: Distribution of activities per hour excluding “going home” activity. Time is shown in GMT+0, but real time is in GMT+1.
Part III

Understanding Social Networks
Chapter 7

Community Discovery in Dynamic Social Networks

Social networks are dynamic objects. During the course of our lives, social ties change and we experience new and different types of bonds: interactions with classmates usually decrease after the school period; a new job brings new social connections; a migration completely revolutionizes our social circles. Dealing with social networks as static elements, therefore, prevents us from grasping the most important aspect of our social ties: their evolution in time. The dynamism of our interactions also affect our social communities, the densely connected circles which can correspond to different social contexts, from family to work and sports. When events generate modification in our social interactions, they necessarily produce updates in the structure of our social communities.

The concept of “community” intuitively depicts a set of individuals that are very similar, or close to each other more than to anybody else outside the community. However, in scientific literature there is no unique definition of a network community: in many networks the communities are natural and non-overlapping modules of networks; in other context they can overlap; in a third type of study the communities are hierarchically embedded one into another. When networks are small, it is easy to identify communities by analyzing the network visually and indicate its dense, closely connected parts. When a network is large, however, the manual detection is impossible and an algorithmic solution is needed. Unfortunately, communities are poorly defined and often hardly distinguishable, so it is very hard to develop a universally efficient numerical algorithm for the problem [85].

For this reason, community discovery in social networks has emerged as one of the most challenging problems in social network analysis. Classical community discovery algorithms have proposed to deal mainly with static networks, i.e. networks which do not change in time. This choice underlines the assumption that networks can be somehow frozen in time because mutation in their topology happens only in the long run. Although working with “flattened” networks simplifies the formulation of algorithms, the assumption can lead to biased results when highly dynamic contexts are the objects of investigation. Networks often describe rapid-scale human dynamics: social interactions, call graphs, buyer-seller
Figure 7.1: Communities identified by a static community detection algorithm. Numbers on the edges represent the interaction time: even if the interactions are distant in time all the nodes are grouped together. Conversely, an evolutionary algorithm can find several different communities if the network is observed during different stages of its evolution (e.g. if observed at $t = 2$ we have $C_1 = \{u, x, y\}$ while if $t = 3$ a new community $C_2 = \{v, j, z\}$ will appear and for $t = 5$ the whole network compose a single community).

scenarios are all examples of realities for which the static assumption leads to results that overestimate or underestimate the real connectivity. A community extracted in a static way from a social network, i.e. without taking into account the temporal ordering of interactions, can group together nodes that have been in contact rarely and whose interactions can be very distant in time one from the others (Figure 7.1). To overcome this limitation the classical community discovery problem has to be adapted to the dynamic scenario: as done for clustering approaches [99] we need to introduce an evolutionary variant able to deal with rapidly evolving networks.

In this chapter, we discuss a formalization of the evolutionary community discovery problem and propose TILES, an algorithm which solves it by tracking the evolution of communities through time. The algorithm has two main characteristics. First, it proceeds in a streaming fashion updating the observed communities whenever a new interaction is generated. As a fall of a domino tile, every time a new interaction emerges in the network, TILES exploits a label propagation procedure to propagate the changes to the node surroundings, adjusting the neighbors’ community memberships. Second, it deals with overlapping communities, allowing each individual to belong simultaneously to several different communities, which correspond to the several contexts where people express social interactions (family, work, friends, etc.).

We compare our algorithm with other dynamic and static community detection algorithms, using synthetic networks with ground truth communities. With respect to its competitors, TILES has comparable execution times but a better match with the ground truth communities. We also provide a characterization of the community produced by our algorithm using Big Data: a nation-wide call graph of one million users whose interactions are tracked for one month; and a Facebook interaction network which covers a period of 52 weeks.
7.1 Evolutionary Community Discovery

Online social interaction networks, call graphs, economic transactions are all sources of information that evolve over time. The rise of new nodes and edges can lead to deep mutations of network topology: in time, new paths become accessible connecting once disconnected components. An analysis that considers networks as static entities - frozen in time - necessarily introduces bias on its results. In dynamic contexts the problem of community discovery in social networks needs to be revised and its formulation extended.

**Definition 7.1.1 (Evolutionary Community Discovery)** Given an interaction streaming source $S$ and a graph $G = (V, E)$, where $V$ is the set of nodes and $E$ the set of timestamped edges ($e \in E$ is defined as a triple $(u, v, t)$ where $u, v \in V$ and $t \in \mathbb{N}$ is the time of the edge generation by $S$) the evolutionary community discovery (ECD) problem aims to identify and update the community structure that compose $G$ as new interactions are generated by $S$.

The ECD problem models the scenarios where interactions among entities do not occur with a rigid temporal discretization, but they flow “in streaming” as time goes by. After all, this is how our social interactions actually take place: phone calls, SMS messages, tweets, Facebook posts do not appear at predetermined time slots but they are produced in a fluid streaming fashion. Consequently, the social communities also have to change fluidly over time.

A valid algorithm for the ECD problem needs to be able to answer the following question: given a community $C$ at timestamp $t$ and a streaming source $S$, what will be its structure at an arbitrary time $t + \Delta$? An algorithm for the ECD problem needs to produce a series of observations of communities through time. The algorithm we propose, **Tiles**, solves efficiently the ECD problem tracking the evolution of communities through a domino tile strategy.

7.2 **Tiles** algorithm

Social interactions determine how communities form and evolve. The emergence of a new edge in the network leads to changes on the communities equilibrium. A common approach in literature is to split the network into temporal snapshots and repeat a static community detection for each snapshot, to study the variation of the results as time goes by. This approach, however, introduces an evident issue: which temporal threshold has to be chosen to partition the network? This problem, which is obviously context dependent, also introduces another issue: once the algorithm is performed on each snapshot how can we identify the same community in consecutive time slots?

To overcome those issues we have designed **Tiles**, an algorithm that does not impose fixed temporal thresholds for the partition of the network and the extraction of communities. It proceeds in a streaming fashion updating the observed communities whenever a new interaction is generated by the streaming source. As a fall of a domino tile, every time
a new interaction emerges in the network TILES exploits a label propagation procedure to propagate the changes to the node surroundings, adjusting the neighbors’ community memberships. According to TILES, the belonging of a node to a community can be of two types: (i) \textit{weak} membership, which identify nodes in the “periphery” of the community (peripheral nodes); or (ii) \textit{strong} membership, for nodes in the “core” of the community (core nodes). If a node is involved in at least one triangle with the others belonging to the same community is defined as a core node, otherwise it is a peripheral node. Only core nodes are allowed during the label propagation phase to spread community membership to their neighbors, which becomes peripheral nodes if they do not participate in any triangle within the core. TILES is an algorithm for identifying overlapping communities, i.e. each node can belong to different communities, which can represent the different spheres of the social world of an individual (friendship, co-workshop, etc.).

The algorithm takes in input three parameters: (i) the graph $G$, which is initially empty; (ii) an edge streaming source $S$; and (iii) $\tau$, a temporal observation threshold. TILES produces as output, for each node, a series of timestamped observations each one composed by two sets: the weak community memberships and the strong community memberships of the node. The temporal observation threshold $\tau$ does not affect the execution of the algorithm, it only allows us to customize the output of the algorithm. Communities are indeed updated every time a new interaction appear, $\tau$ only permits to decide at which granularity we want to observe the community membership evolution of a node (every day, every week, every month, etc.).

Algorithm 2 shows the behavior of TILES. First of all (line 3-6) the new edge $e = (u, v)$ generated by the source $S$ is added to the graph. Then the following scenarios are considered:
Algorithm 2 Tiles$(G, S, \tau)$

Require: $G$: undirected graph, $S$: streaming source, $\tau$: temporal observation threshold

1: actual$_t$ = 0
2: while $S$.isActive( ) do
3:   $e$←$S$.getNewInteraction( )
4:   if $e \not\in G$ then
5:     $G$.addEdge($e$)
6:   end if
7:   if $|\Gamma(e_u)| = 1 \& |\Gamma(e_v)| > 1$ then
8:     WeakPropagation($e_u, e_v$)
9:   else if $|\Gamma(e_u)| > 1 \& |\Gamma(e_v)| = 1$ then
10:      WeakPropagation($e_v, e_u$)
11:   else
12:      CN ←$\Gamma(e_u) \cap \Gamma(e_v)$
13:      if $|CN| == 0$ then
14:         WeakPropagation($e_u, e_v$)
15:      end if
16:   end if
17:  end if
18:  if actual$_t$ == $\tau$ then
19:     OutputCommunities($G$)
20:     actual$_t$ = $e_t$
21: end if
22: end while

1. both the nodes $u$ and $v$ appear for the first time in the graph. No other actions are performed until the next interaction is produced by the source (Figure 7.2, top left);

2. one node appears for the first time and the other is already existing but peripheral. Since peripheral nodes are not allowed to propagate the community membership, no action is performed until the next edge is produced by the source (Figure 7.2, top left). The same case applies when both nodes are existing but peripheral;

3. one node appears for the first time in $G$, while the other is an already existing core node (line 7-10). The new node inherits a weak community membership from the existing core node (Figure 7.2, top right);

4. both nodes are core nodes already existing in $G$ (line 12-17). In this case two sub-scenarios can emerge:

   (a) Nodes $u$ and $v$ do not have common neighbors (line 13-15): they propagate each other a weak community membership through the WeakPropagation procedure (Figure 7.2, bottom left).

   (b) Nodes $u$ and $v$ do have common neighbors (line 17): their community memberships are re-evaluated and the changes propagated to their surroundings by the StrongPropagation function (Figure 7.2, bottom left and right).

Tiles communities grow gradually expanding their core and their peripheries through the WeakPropagation and the StrongPropagation procedures. The WeakPropagation procedure regulates the events in which a new node becomes part of an already
Figure 7.3: Community growth: four consecutive updates extracted from a Facebook interaction network (FB07 network). Colors identify “core” communities. “Peripheral” nodes are identified by solid lines while nodes that are not involved in communities by dashed lines. The new interaction is in red.

established community. Since the newcomer is not involved in any triangle with other nodes of the community it becomes part of its periphery. The same function is performed when a new interaction connects existing nodes that does not share any neighbors. The StrongPropagation procedure assumes that the nodes \( u \) and \( v \) have at least one common neighbor \( z \). For each triple \( (u,v,z) \) if at least two nodes are core for the same community the third one becomes core as well (example in Figure 7.3c-d), otherwise a new community is created upon the new triangle (example in Figure 7.3a-b). Once the core nodes are established, they propagate a weak membership to their neighbors, if they are not already within the community.

Due to its streaming definition, the complexity of Tiles is mainly shaped by the edge insertion phase, which can cause perturbation on the network topology and induce updates on the community structure. In the worst case scenario, this step has complexity \( O(|V|) \) since the most costly rule is applied when the edge endpoints \( u \) and \( v \) share \(|V|\) neighbors. However, reaching such upper bound is unusual due to the power law degree distribution which characterizes real world interaction networks.

7.2.1 Tiles properties

Due to its streaming nature, Tiles shows two main properties: (i) it can be used incrementally on a precomputed community set; (ii) it can be parallelized if specific conditions are satisfied. Moreover, in presence of a deterministic interaction source \( S \), Tiles output is uniquely determined. Here we discuss and formalize such characteristics.

Incrementality. As specified above, Tiles is called on an initially empty graph. However, it also works when a non-empty graph and a set of precomputed communities are passed as parameters. Given a deterministic streaming source \( S_t \) at time \( t \) and a non-empty graph \( G_t \), whose nodes are assigned to a community set \( C_t = (c_1, c_2, \ldots, c_n) \), Tiles has the incrementality property:

\[
\text{Tiles}(G_t, S_t) = \text{Tiles}(G, S_0)
\]  
(7.1)
where $G$ is an empty graph, $C$ is an empty community set, $S_0$ is the streaming source $S$ at the initial time. Since Tiles is incremental, the final community evolutions produced starting with the source $S$ at time 0 or at time $t$ are identical, assuming that the streaming source is deterministic.

**Compositionality.** Tiles is parallelizable by identifying disjoint streams of edges produced by the deterministic streaming source $S$. Given a graph $G$, and two disjoint stream of edges $S^i, S^{ii}$ iff $\forall (u_1, v_1) \in S^i (u_2, v_2) \in S^{ii}$: $\left( c(u_1) \cup c(v_1) \right) \cap \left( c(u_2) \cup c(v_2) \right) = \emptyset$, where $c(\cdot)$ returns the set of communities the node is part of, then:

$$\text{TILES}(G, S) = \text{TILES}(G, S^i) \cup \text{TILES}(G, S^{ii})$$ (7.2)

Isolating interactions among nodes of different communities makes it possible to parallelize the algorithm, resulting in faster computation of the community evolution.

## 7.3 Experiments

Evaluating results provided by a community discovery algorithm is a complex task, since a shared and universally accepted definition of what a community is does not exist. In the literature, each algorithm proposes its own idea of a community and of the properties nodes should share in order to belong to the same partition of the network. Moreover, different community discovery algorithms are often designed to solve slightly different problems (i.e. overlapping and non-overlapping communities, static and dynamic communities, modularity-based and density-based communities).

In this section we propose a validation of Tiles against other algorithms on synthetic networks with ground truth communities (Section 7.3.1). Then we characterize the communities produced by Tiles on two real datasets of social interactions (Section 7.3.2).

### 7.3.1 Evaluation on Synthetic networks

A comparison with the communities produced by different algorithms is the most straightforward way to assess the strengths and weaknesses of Tiles. Hence, in this section, we compare our algorithm with other static and dynamic community detection algorithms. The plethora of community definitions introduced by different approaches makes it questionable to directly compare the outputs obtained by two algorithms on the same network when a ground is not provided. Unfortunately, datasets with ground truth, i.e. real partition of the network into communities, are hard to find, especially for large scale networks. For this reason, we perform the comparison over several synthetic datasets, estimating the resemblance of algorithms’ communities with the provided ground truth partition of the network. To compare the ground truth with the structure delivered by the algorithm we adopt the Normalized Mutual Information (NMI) score, a measure of similarity borrowed from information theory [171]:

$$\text{NMI}(X : Y) = \frac{H(X) + H(Y) - H(X,Y)}{(H(X) + H(Y))/2}$$
where $H(X)$ is the entropy of the random variable $X$ associated to the partition produced by the algorithm, $H(Y)$ is the entropy of the random variable $Y$ associated to the ground truth partition, whereas $H(X,Y)$ is the joint entropy. NMI ranges in $[0,1]$ and is maximized when the compared communities are identical.

To obtain a ground truth community partition of the networks we use the LFR benchmark [172], which generates synthetic networks along with ground truth communities, according to the following input parameters:

- $N$, the network size;
- $\mu$, the average ratio for each node between edges to its community and edges with the rest of the network;
- $C$, the network density.

We produce a total of 2,500 different static synthetic networks varying the $N$ parameter from 1k to 500k nodes; the $\mu$ parameter from 0 to 0.9 with steps of 0.1; and the $C$ parameter from 0 to 0.9 with steps of 0.1.

On the produced networks we apply TILES and other overlapping community detection algorithms. One algorithm, iLCD [104] is a dynamic algorithm which approaches the ECD problem exploiting topological information only. In particular, iLCD re-evaluates communities at each new interaction produced by a streaming source according to the path lengths between each node and its surrounding communities. The other two algorithms are static ones: (i) DEMON [94], which exploits a label propagation procedure to build communities starting from ego networks; (ii) cFINDER [93], which computes communities based on the clique percolation method (CPM). It is worth underlining that the LFR benchmark does not generate a timestamped stream of edges. For this reason, we imposed a random temporal order on the edges in order to simulate the streaming source $S$ needed to apply TILES and iLCD. Conversely, such ordering is not needed for the other two algorithms since they are static.

We observe that varying the parameter $\mu$, TILES produces communities whose NMI w.r.t. the ground truth is comparable to DEMON and cFINDER, but significantly out-
performs iLCD, its direct dynamic competitor (Figure 7.4, left). Yet, the NMI of the compared methods is stable till the density parameter $C \leq 0.5$, i.e. half of all the possible edges are present in the network (Figure 7.4, center). Such a high density values, however, is not unusual for real interaction networks [55, 54], where the density usually falls in the range $[0.1, 0.2]$. Figure 7.4 (right) shows that TILES (implemented in Python language) has an average execution time comparable to iLCD (implemented in the Java language). Our algorithm produces communities whose NMI w.r.t. a ground truth is significantly higher than the other dynamic community detection algorithm, with a similar execution time.

7.3.2 Datasets

<table>
<thead>
<tr>
<th>Network</th>
<th>Nodes</th>
<th>Edges</th>
<th>CC</th>
<th>#Observations ($\tau$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FB07</td>
<td>19,561</td>
<td>304,392</td>
<td>0.104</td>
<td>52 (1 week)</td>
</tr>
<tr>
<td>CG</td>
<td>1,007,567</td>
<td>16,276,618</td>
<td>0.067</td>
<td>10 (3 days)</td>
</tr>
</tbody>
</table>

Table 7.1: Datasets statistics. CC identify the network clustering coefficient.

We performed TILES on two large interaction networks. The FB07 interaction network is extracted from the WOSN2009 [173] dataset\footnote{http://socialnetworks.mpi-sws.org/data-wosn2009.html} and regards online interactions between users via the wall feature in the New Orleans Facebook regional network during 2007. We adopted an observation period $\tau$ of a week to format the output of the algorithm.

The second dataset is a call graph extracted from the EC mobile phone dataset described in Section 3.2. It contains date, time and coordinates of the phone tower routing the communication for each call and text message sent by 1,007,567 costumers, in a period of one month. We discarded all the calls to external operators. As $\tau$ we adopted a window of 3 days.

The datasets allow us to test the algorithm on two different grounds: a “virtual” context, in which people share thoughts and opinions via a social media platform, and a “real” one, in which people directly keep in touch through a mobile phone. The general statistics of the datasets are shown in Table 7.1.

7.3.3 TILES communities characterization

The size of TILES communities follows a fat tailed distribution, for both the Facebook network and the call graph (Figure 7.5, first row). This means that the vast majority of communities have a few nodes, and a small but significant portion of the communities reach a size of several thousands nodes. Such a great heterogeneity also characterizes the community overlap, i.e. how many different communities a node belongs to (Figure 7.5, second row). The majority of nodes belong to just one or two communities, while some nodes have a high degree of overlap belonging to thousands different communities.
Figure 7.5: (First row) Community size distribution for the call graph (left) and the Facebook network (right). (Second row) Community per node distribution for the call graph (left) and the Facebook network (right). (Third row) Distribution of average clustering coefficient per community for the call graph (left) and the Facebook network (right).

Figure 7.5 (third row) shows the average clustering coefficient of communities, computed over the core nodes of each community. The communities maintain high average clustering coefficients as time goes by, with minimum values of 0.6 for the Facebook network and 0.8 for the call graph. This values are very high compared to the overall clustering coefficients of the networks (see Table 7.1), highlighting how Tiles is capable of producing dense social communities, with high clustering structure.

To study how communities change in time, we computed the similarity between each community and the community itself after a time slot $\tau$ (set at 1 week for FB07 and 3
Figure 7.6:  
(First row) Average community similarity (Jaccard coefficient) among consecutive observations; and community appearance per observation for the call graph (left) and the Facebook network (right).  
(Second row) Tiles communities nodes coverage for the call graph (left) and the Facebook network (right).  
(Third row) Distribution of transition time from periphery to core and ratio of nodes that move from the periphery to the core across consecutive observations, for the call graph (left) and the Facebook network (right).

days for GC), through the Jaccard coefficient between the nodes:

\[ Jaccard(C_t, C_{t+1}) = \frac{|C_t \cap C_{t+1}|}{|C_t \cup C_{t+1}|} \]

where \( C_t \) identifies the community \( C \) observed at time \( t \). Figure 7.6 (first row) shows that the Jaccard coefficient (blue line) is stable and high during the period of observation. This means that communities do not change so much, and the set of community nodes remain
stable as time goes by. The rate at which communities arise in time is described in Figure 7.6 (first row, green line): we notice that more than the 80% of communities arise just in the first ten days/weeks of the period of observation, then the appearance rate starts to stabilize.

Due to its definition, TILES searches for precise patterns within a network structure: triangle based communities. The resulting nodes coverage, i.e. how many nodes are included into communities, is hence strictly related to the clustering coefficient of the analyzed network: the greater the clustering coefficient the higher is the nodes coverage. This tendency is depicted in Figure 7.6 (second row) where we report, both for CG and FB07, how the ratio of “core” nodes, “periphery” nodes and the sum of the two (total nodes coverage) changes in time during the period of observation. The FB07 network, which has a clustering coefficient of $0.104$, shows a high coverage: the 80% of nodes are included in some communities. Conversely, the CG network reaches a coverage of only 50%, due its lower clustering coefficient of 0.067.

A peculiarity of TILES is the concept of community periphery. As discussed in Section 7.2, peripheral nodes are not involved in triangles with other nodes of the community. Every community node is at beginning a peripheral node, then it becomes a core nodes once it is involved within a triangle. The time of transition between the periphery to the core of a community, as shown by Figure 7.6 (third row), is generally short: in CG the 40% of the nodes become core nodes in just 3 days; in FB07 the 15% of the nodes perform the transition during the first week. We also investigated how many nodes make this transition across consecutive community observations: given two observation of a community $C$, what is the ratio of core nodes in $C$ at $t + \Delta$ that where in the peripheral nodes at time $t$? Figure 7.6 (third row) shows that this ratio fluctuates a lot in time, and takes values between the 30% and 50% of the nodes. This means that almost the half of the peripheral nodes become core nodes in the subsequent time window for both CG and FB07. Such information can be very valuable to boost several network analysis tasks, such as Link Prediction and Information Diffusion.

7.4 Conclusions

TILES solves the problem of tracking the evolution of overlapping communities in interaction networks: following a domino approach, each new interaction determines the re-evaluation of community memberships for the endpoints and their neighborhoods. It defines two types of community membership: weak membership, describing nodes in the periphery of the community; and strong membership, for core nodes which are involved in at least a triangle within the community. Compared with another dynamic community detection algorithm on synthetic networks, TILES shows similar execution times but a higher correspondence with the ground truth communities.

Many research directions open for future works, such as an extension of the algorithm to model edges and nodes disappearance, and a parallel implementation which exploits the compositionality and incrementality properties to further speed up the execution time.
The transition from the *periphery* of a community to its *core* is another interesting aspect to investigate, both in the baseline scenario and in a scenario where the disappearance of edges and nodes is allowed.
Chapter 8

The local diffusion patterns of musical listenings

One of the most fascinating aspects which characterize our global and highly interconnected society is the capability to spread fads, ideas, innovations and even epidemics in a strikingly fast way. Think about the extraordinary outbreak of the AIDS in the eighties, which killed almost 20 million people all over the world, or the spreading of computer viruses through emails or smart phones, or the diffusion of “viral” Youtube videos posted by famous or anonymous people, which are able to gain millions of fans in just one day or two. These are all example of diffusion in networks, a complex phenomenon which is attracting more and more interest nowadays in every context. The laws of diffusion, indeed, are intensively studied by sociologists seeking to understand fads, trends and riots; by marketing specialists trying to figure out how to spread their product; by epidemiologists hoping to curb deadly viruses; by hackers and computer managers determined to destroy and protect computer systems respectively.

Addressing the problem of diffusion, however, is very challenging since detailed data about the real spreading of items are rarely available. The spreading of the influenza disease from an individual to another, for example, cannot be observed in details because of the lack of context information: we do not know who passed the disease to whom, when she passed it and where the infection took place. For this reason biological disease spreading is generally modeled through meta-population approaches [174], who simulate the spatio-temporal diffusion of the pathogen on social and transportation networks, and exploit some known properties of the virus, such as virulence, period of incubation and period of recovering.

Social media platforms like Facebook or Twitter, conversely, provide the social microscope to observe in a detailed way the diffusion dynamics of many items: thoughts, opinions, concepts. The spreading of a Twitter hashtag, for instance, can be tracked with high accuracy: starting from the user who first generated it, we can follow the chain of tweets and retweets, obtaining a complete temporal diffusion chain of that particular topic. Such information allows us to built agent-based models which describe the diffusion dynamics within the online social network [175].
In this chapter we address the problem of information diffusion on online social media. We aim to study how the listenings of musical artists diffuse starting from musical innovators over the LastFM social media, a platform which allows users to discover, share and listen to music songs. In the setting of studying information diffusion, most of the attention of researchers has been put on how to maximize the number of nodes subject to the spreading process. This is done by choosing appropriate seeds in critical parts of the network, such that their likelihood of being prominent users (nodes active on an innovation before all the others) is maximum, to possibly achieve larger cascades. Large cascades are obviously part of the overall aim, but we argue that it is not the unique dimension of this problem. Three other dimensions are relevant for the diffusion of an item: the width, the depth and the strength of the social prominence of any given node in a network. The width of a node is being prominent for its immediate neighbors; the depth is its ability to be the root of long cascades; the strength is being the root of an intense activity. We argue that real-world scenarios focus on specific diffusion patterns requiring a multidimensional understanding of the prominence mechanics at play, along the three mentioned dimensions.

To clarify the point, consider the following scenarios: (i) an analyst needs information from the personal acquaintances of a subject, the important aspect is that many subject’s direct connections respond, ignoring people two steps away or more; (ii) a person wants to find another person with a given object, the important aspect is that some people are able to pass her message through a chain pointing to the target; (iii) an artist wants to influence people in a social network to her art, the important aspect is that some people are influenced above the threshold that will make them aware of the art. In (i) we want a broad diffusion in the first degree of separation. In (ii) we require a targeted diffusion similar to a Depth First Search. In (iii) there is the need of a high-intensity diffusion. Different scenarios may require any combination of the three.

In this chapter we use three measures to capture the characteristics of these three scenarios: the width, depth and strength of social prominence. The width measures the ratio of the neighbors of a node that follows the node’s actions. The depth measures how many degrees of separation there are between a node and the other nodes that followed its actions. The strength measures the intensity of the action performed by some nodes after the leader. We study what the relations are between these three measures to understand if we are capturing three orthogonal dimensions of social prominence. We also study the relations between the width, depth and strength measures and different node properties, with the aim of predicting the diffusion patterns of different events, given the characteristics of the nodes that lead their diffusion.

We validate our concepts in a musical context, relying on the LastFM dataset described in Section 3.4. We provide an algorithm to detect musical leaders, i.e. the prominent users who start listening to an artist before any of their neighbors. We calculate for each prominent user her width, depth and strength, along with its network statistics such as the degree and the betweenness centrality, looking for associations between them. We then create a case study to understand what are the different dynamics in the spread of
8.1 Leader Detection

Each diffusion process has its starting points. Any idea, disease or trend is firstly adopted by particular kinds of actors. Such actors show an increased sensibility and a pronounced inclination to novelties, being the first to perform actions or to adopt new items in a given social context [177]. We call such actors prominent actors, innovators or leaders, because they are able to anticipate and influence the behavior of other actors. Given a graph, several interesting problems arise regarding how information spreads over its topology: Can we identify the leaders? Can we characterize them?

Our approach aims to detect leaders through the analysis of the topology of the social network and the set of actions performed by the nodes. When discussing the roles of those entities, we refer respectively to the following definitions:

**Definition 8.1.1 (Social Graph)** A social graph $\mathcal{G}$ is composed by a set of actors (nodes) $V$ connected by their social relationships (edges) $E$. Each undirected edge $e \in E$ is defined as a couple $(u, v)$ with $u, v \in V$. With $\Gamma(u)$ we denote the set of neighbors of node $u$.

**Definition 8.1.2 (Action)** An action $a_{u,\psi} = (w, t)$ defines the usage by an actor $u \in V$, at a certain time $t$, of a specific object (or service) $\psi$ with a weight $w \in \mathbb{R}$. The set of all the actions of nodes belonging to a social graph $\mathcal{G}$ will be identified by $\mathcal{A}$, while the object set will be called $\Psi$.

In a musical application scenario, an action is the listening of a song $\psi$ by a given individual $u$ at a given time $t$, while the weight $w$ denotes how many times the user listened to the song. We identify with $\mathcal{G}_\psi = (V_\psi, E_\psi)$, where $V_\psi \subset V$ and $E_\psi \subset E$, the induced subgraph on $\mathcal{G}$ representing respectively the set of all the actors that have performed an action on object $\psi$, and the edges connecting them. We depict an example of the social graph and the set of actions in Figure 8.1 (left), where the induced subgraph for the object $x$ is highlighted with a dashed line.

---

Figure 8.1: Toy Example. On the left the social graph $\mathcal{G}$ and action set $\mathcal{A}$, where $x, y \in \Psi$ are the objects of the actions; in the center the induced subgraph for the action $x$; on the right the diffusion tree for $x$. In red we highlighted the leader (root) for the given tree.

The work in this chapter is based on two published papers [176].
Given the nature of a diffusion process, each leader has to be a prominent node among its neighbors, being the root of a cascade event that follows some rigid temporal constraints. In other words, a leader must perform an action before all his social neighbors. We defined that a node $u$ precedes a neighbor $v$ if given $t_{u,\psi} \in a_{u,\psi}$ and $t_{v,\psi} \in a_{v,\psi}$ is verified that $t_{v,\psi} > t_{u,\psi}$ and $t_{v,\psi} - t_{u,\psi} \leq \delta$. Here, $\delta$ is a temporal resolution parameter that limits the cascade effect: if $t_{v,\psi} - t_{u,\psi} > \delta$, we say that $v$ executed action $a_{v,\psi}$ independently from $u$, as $u$’s prominence interval is over. For example if $u$ listens to a new song today and her neighbor $v$ listens to the same song a month after, we can assume that they listened the song independently, without any social influence.

To detect the leaders with respect to a given action, we first transform the undirected subgraph $G_\psi$ in a directed one imposing that the source node of an edge must have performed its action before the target node. After that, we label each edge $(u, v)$ with $\min(t_{u,\psi}, t_{v,\psi})$ to identify when the diffusion started going from one node to the other. The directed version of $G_\psi$ represents all the possible diffusion paths which connect leaders with their “tribes”, i.e. the nodes she is able to influence in the adoption of the object (Figure 8.1 (center) shows an example for the object $x \in \Psi$). Given an object $\psi$, the symbol $L_\psi$ refers to the leader set, i.e. the nodes who act as leaders for that object. When no action is specified the set $L$ is used to describe the union of all the $L_\psi$ for the graph $G$. Clearly, in the modified directed social graph $G_\psi$ a leader does not have any incoming edges, since she is the first to perform an action on the object and cannot act after another individual (she is, in her surrounding, an innovator). Given this definition, for each directed connected component $C_\psi \subset G_\psi$ multiple nodes can belong to $L_\psi$.

One caveat of our problem definition comes from the fact that a leader may be influenced by exogenous events. This is not a problem as we are not measuring a node’s influence, but a node’s prominence, i.e. its propensity to act faster than others to any kind of exogenous and/or endogenous influence.

To study the real path of diffusion given an action $a$ and a leader $l$, we start from the modified directed graph $G_\psi$ and build a minimum diffusion tree:

**Definition 8.1.3 (Leader’s Minimum Diffusion Tree)** Given an action $a_\psi$, a directed connected component $C_\psi$ and a leader $l \in L_\psi$, the minimum diffusion tree $T_{l,\psi} \subset C_\psi$ is the minimum spanning tree (MST) having its root in $l$ and built minimizing the temporal label assigned at the edges.

An example of minimum diffusion tree for the node 1 and object $x$ is shown in Figure 8.1 (right). For each object, the diffusion process on a given network is independent. Moreover, given temporal dependencies on its adoption (expressed through actions $a_{\ast,\psi} \in A$), it is possible to identify the origin points of the diffusion. The identified leaders will show different topological characteristics and will be prominent in their surroundings in different ways: our aim is to classify diffusion patterns from leaders characterizing some of their common traits.

Algorithm 3 reports the pseudocode of the leader extraction procedure $\text{ExtractLeaders}$, which computes the leader on a given social network $G$ with respect to a set of actions
Algorithm 3 The pseudo-code of ExtractLeaders.

Require: $\mathcal{G} = (V, E); \Psi, \delta$
Ensure: $L, T$

1: $\mathcal{T} \leftarrow \{\}$
2: $\mathcal{L} \leftarrow \{\}$
3: for all $\psi \in \Psi$ do
4:    $\mathcal{G}_\psi \leftarrow \text{InducedSubgraph}(\mathcal{G}, \psi, \delta)$
5:    $\mathcal{T}_\psi \leftarrow \{\}$
6:    $\mathcal{L}_\psi \leftarrow \{\}$
7:    for all $C_\psi \in \mathcal{G}_\psi$ do
8:        for all $l \in C_\psi$ do
9:            if $\text{InDegree}(C_\psi, l) == 0$ then
10:               $\mathcal{L}_\psi \leftarrow \mathcal{L}_\psi \cup l$
11:               $\mathcal{T}_{l, \psi} \leftarrow \text{MST}(C_\psi, l)$
12:               $\mathcal{T}_\psi \leftarrow \mathcal{T}_\psi \cup \mathcal{T}_{l, \psi}$
13:        end if
14:    end for
15: end for
16: $\mathcal{L} \leftarrow \mathcal{L} \cup \mathcal{L}_\psi$
17: $\mathcal{T} \leftarrow \mathcal{T} \cup \mathcal{T}_\psi$
18: end for
19: return $\mathcal{L}, \mathcal{T}$

Ψ and a time tolerance for the influence δ. For all objects $\psi \in \Psi$, we first extract the directed induced subgraph $\mathcal{G}_\psi$, through the procedure InducedSubgraph. Then, for each connected component $C_\psi \in \mathcal{G}_\psi$, we choose the leaders as the nodes without incoming edges (performed by InDegree). We add them to the leader set $\mathcal{L}_\psi$ and we store in $\mathcal{T}_\psi$ their minimum diffusion trees, calculating the minimum spanning tree MST on the nodes in $C_\psi$ through the Kruskal’s algorithm [178] with root in $l$. At the end, we return the union of $\mathcal{L}_\psi$ and $\mathcal{T}_\psi$.

The complexity of the algorithm ExtractLeaders is super-linear. First, we cycle over all the actions ($\mathcal{O}(|\Psi|)$). Then, we cycle over all the connected components of $\mathcal{G}_\psi$ and, for each one, we cycle over the nodes belonging to them; together the two cycles reach the complexity of $\mathcal{O}(|V_\psi|)$. Within the inner loop a minimum spanning tree is computed for each leader ($\mathcal{O}(\log |V|)$) with Kruskal’s algorithm. As a consequence, the final complexity of the algorithm ExtractLeaders is $\mathcal{O}(|\Psi| \times |V| \log |V|)$. For large networks, it is fair to assume that $|\Psi| << |V|$, so the complexity would be $\Theta(|V| \log |V|)$. Moreover, since each action is independent of the others, with $|\Psi|$ computer processors the exact complexity would be $\mathcal{O}(|V| \log |V|)$.

8.2 Measures

The diffusion of listenings from a prominent individual can follow three different dimensions: width, the ratio of neighbors mirroring an action after a node; depth, how many
degrees of separation are in between a node and the most distant of the nodes mirroring its actions; and strength, how strongly nodes are mirroring a node’s action. These three dimensions specify the three ways a prominent individual can “infect” the other, and diffuse the listening through its social network surroundings.

Given a leader, the width aims to capture the direct impact of her actions on her neighbors, i.e. the degree of importance that a leader has over her friends.

**Definition 8.2.1 (Width)** Let $G$ be a social graph, $\psi \in \Psi$ an object and $l \in \mathcal{L}_\psi \subset V$ a leader: the function $\text{width} : \mathcal{L}_\psi \rightarrow [0,1]$ is defined as:

$$
\text{width}(l, \psi) = \left| \left\{ u \in \Gamma(l) \land \exists a_{u,\psi} \in \mathcal{A} \right\} \right| / |\Gamma(l)|.
$$

(8.1)

The value returned is the ratio of all the neighbors that, after the action of the leader, have performed the same action.

The depth measure evaluates how much a leader can be prominent among other prominent leaders, which can be prominent on other leaders and so on.

**Definition 8.2.2 (Depth)** Let $T_{l,\psi}$ be a minimum diffusion tree for a leader $l \in \mathcal{L}_\psi$ and a given object $\psi \in \Psi$: the function $\text{depth} : T_{l,\psi} \rightarrow \mathbb{N}$ computes the length of the maximal path from $l$ to a node $u \in T_{l,\psi}$. The function $\text{depth}_{avg} : T_{l,\psi} \rightarrow \mathbb{R}$ computes the average length of paths from $l$ to any leaf of the tree.

The last proposed measure, the strength, tries to capture quantitatively the total weight of the usage of an object after the leader’s action. A leader is strongly prominent if the nodes among which she is prominent are very engaged in utilizing what she used. Direct prominence diminishes as new adopters become more distant, in the network sense, from the original innovator. Therefore, we decided to introduce a distance damping factor.

**Definition 8.2.3 (Strength)** Let $T_{l,\psi}$ be a minimum diffusion tree for a leader $l \in \mathcal{L}_\psi$ and an object $\psi \in \Psi$; $0 < \beta < 1$ a damping factor: the function $\text{strength} : T_{l,\psi} \times (0,1) \rightarrow \mathbb{R}$ is defined as:

$$
\text{strength}(T_{l,\psi}, \beta) = \sum_{i \in [0,\text{depth}(l)]} \beta^i L(T_{l,\psi}, i)
$$

(8.2)

where $L : T_{l,\psi} \times \mathbb{N} \rightarrow \mathbb{R}$ is defined as:

$$
L(T_{l,\psi}, i) = \sum_{\left\{ u \in T_{l,\psi} : \text{distance}(l, u) = i \right\}} \frac{w_{u,\psi}}{w_u}
$$

(8.3)

and represents the sum, over all the nodes $u$ at distance $i$ from $l$, of the ratio between the weight of action $\psi$ and the total weight of all the actions taken.

Given the example in Figure 8.1, we can easily characterize the three diffusion patterns generated by action $x$ from the leader (node 1). The leader has 4 neighbors, the nodes in the set $\Gamma(1) = \{2, 4, 7, 8\}$. Given that $\Gamma_x(1) = \{u | u \in \Gamma(1) \land \exists a_{u,x} \}$ = $\{2, 4\}$ (only nodes 2 and 4 perform the action $x$), we have $\text{width}(1, x) = \frac{|\Gamma_x(1)|}{|\Gamma(1)|} = 0.5$. The longest chain
8.3 Experiments

In this section we present the characterization of width, depth and strength measures on the LastFM data described in Section 3.4, by searching for associations with network topology measures. Finally, we analyze the prominence of different users for different musical genres.

8.3.1 Data

In the experiments we used the LastFM dataset introduced in Section 3.4, a graph $G$ of 75,969 nodes and 389,639 edges. Figure 8.2 (left) depicts the log-binned degree distribution of $G$, along with the best fit. As expected the distribution is a power law: most of the users have a few friends, but a small fraction of them act as hubs, having hundreds of friends. Each action in the data is one user listening to an artist $w$ times in week $t$. In Figure 8.2 (right) we depicted the log-binned distribution of the number of listeners per artist, along with the best fit. Also in this case a huge heterogeneity emerges: some artists are extremely famous while the majority have a very few listeners. Each artist belongs to a music genre, represented by its tag. We focus on the 9 most popular genres: dance, electronic, folk, jazz, metal, pop, punk, rap and rock. Only artists belonging to the
CHAPTER 8. THE LOCAL DIFFUSION PATTERNS OF MUSICAL LISTENINGS

9.3.2 Characterization of the Measures

For each action in the dataset, we computed the leaders on the social network through the ExtractLeader procedure described by Algorithm 3. We set the time tolerance parameter of the algorithm to $\delta = 3$, meaning that if a user listened to a particular artist three weeks or more after its neighbor then we do not consider her neighbor to be prominent for her for that action. For each leader obtained we computed her width, depth and strength, setting the strength damping factor $\beta = 0.5$. Besides the three measures defined above, we calculated also leaders’ degree (number of edges connected to the node), clustering coefficient (ratio of triangles over the possible triads centered on the node), neighbor degree (average degree of the neighbors of the node), betweenness (share of the shortest paths that pass through the node) and closeness centrality (inverse average distance between the node and all the other nodes of the network).

Table 8.1 shows the Pearson correlation coefficient $\rho$ between the computed network measures. We highlighted the correlations whose p-value was significant and whose absolute value was strong enough to draw some conclusions. For the significance of p-values, the traditional choice is to set the threshold at $p < 0.01$. However, given our number of observations, we decided to be more restrictive, setting our threshold at $p < 0.0005$. We also consider the correlation coefficient value $\rho$ significant if $|\rho| > 0.1$.

Table 8.1: Pearson correlation coefficient $\rho$ between width, depth, strength and other network statistics for leaders.

<table>
<thead>
<tr>
<th>average depth</th>
<th>width</th>
<th>strength</th>
<th>degree</th>
<th>clustering</th>
<th>neighbor degree</th>
<th>betweenness</th>
<th>closeness</th>
</tr>
</thead>
<tbody>
<tr>
<td>width</td>
<td>-0.03</td>
<td>-0.23</td>
<td>-0.08</td>
<td>0.05</td>
<td>-0.08</td>
<td>-0.02</td>
<td>-0.13</td>
</tr>
<tr>
<td>strength</td>
<td>-0.01</td>
<td>0.13</td>
<td>0.05</td>
<td>-0.07</td>
<td>-0.59</td>
<td></td>
<td></td>
</tr>
<tr>
<td>degree</td>
<td>-</td>
<td>-0.31</td>
<td>0.13</td>
<td>0.05</td>
<td>0.00</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>clustering</td>
<td>-</td>
<td>0.02</td>
<td>-0.02</td>
<td>0.03</td>
<td>-0.02</td>
<td>0.77</td>
<td>0.56</td>
</tr>
<tr>
<td>neighbor degree</td>
<td>-</td>
<td>-</td>
<td>-0.16</td>
<td>-0.02</td>
<td>0.77</td>
<td>0.66</td>
<td>-0.32</td>
</tr>
<tr>
<td>betweenness</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-0.32</td>
<td>-0.00</td>
<td>0.39</td>
<td>-0.22</td>
</tr>
</tbody>
</table>

1To assure experiment repeatability, we made our cleaned dataset and our code available at the page http://goo.gl/h53hS
8.3. EXPERIMENTS

<table>
<thead>
<tr>
<th>Partial $\rho$</th>
<th>clustering</th>
<th>closeness centrality</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.087216</td>
<td>-0.536861</td>
<td></td>
</tr>
<tr>
<td>p-value</td>
<td>$1.57 \times 10^{-14}$</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 8.2: Partial correlation and p-value of clustering and closeness centrality with width, controlling for degree values.

The depth measure is associated with a low closeness centrality: a deep prominence is associated to nodes at the margin of the network. It is expected that nodes with high closeness centrality have also low depth: being central, they cannot generate long chains of diffusion. The eccentricity of all the nodes of the network ranges from 6 to 10, meaning that some leaders cannot have a depth larger than 5. To make a fair comparison, we recalculate the depth value capping it at 5, meaning that any depth value larger than 5 is manually reduced to 5. Then, we recalculate the correlation $\rho$ between the depth capped to 5 and the closeness centrality obtaining as result $\rho = -0.1366$, with $p < 0.0005$. We can conclude that central nodes are not associated with deep spread of their prominence in a social network.

For the width measure, the anti-correlation with the degree is not meaningful, as the degree is in the denominator of Definition 8.2.1. On the other hand, we observe two interesting correlations: (i) a positive association with clustering, meaning that nodes could be prominent in a tightly connected community; (ii) a negative association with closeness centrality, meaning that central nodes could not spread a wide influence. Both associations could be explained with the negative correlation with degree. Therefore, for both measures we run a partial correlation, controlling for the degree. In practice, we calculate the correlation between width and clustering (or closeness centrality) by keeping the degree constant. Results are in Table 8.2: even if significant according to the p-value, the relationship between width and clustering is very weak and deserves further investigation. On the other hand, it is confirmed that central nodes are also associated with low width, regardless their degree.

The strength measure is not strongly correlated with traditional network statistics (Table 8.1). As a consequence, hubs that are associated with low depth and low width, do not have necessarily high strength, making their prominence in a network questionable. On the other hand, strength appears to be negatively associated with depth, suggesting a trade-off between how deeply a node can be prominent in a network and how strong this prominence is on the involved nodes. Such anti-correlation between the strength and the depth may be due to the damping factor $\beta$: from Definition 8.2.3 we see that $\beta$ decreases nodes’ contributions at each degree of separation (i.e. at increasing depths). As a consequence, nodes that are farther from the leader contribute less to its strength, i.e. the highest the depth the smallest are the contributions to the strength. We recalculated the strength values by setting $\beta = 1$, therefore ignoring any damping factor and nullifying this effect. We obtained as result $\rho = -0.4168$ and a significant p-value, therefore concluding that the damping factor $\beta$ is not causing the anti-correlation between depth and strength. The anti-correlation is hence a property of prominent users: a leader who strongly influence
nodes in her surroundings hardly reaches nodes at many degrees of separation.

To sum up, we summarize the associations as follows: (i) central nodes are not necessarily prominent in a social network (low width and depth), a result that confirms results in [179] and [180]; (ii) longer cascades (higher depths) are associated with a lower degree of engagement (lower strengths), a phenomenon possibly related to the role played by “weak ties” [62]; (iii) being prominent among neighbors is probably easier if the node is in a tightly connected community (high width and clustering).

### 8.3.3 Local spreading patterns of musical genres

Starting from the width, depth and strength measures computed for each leader, here we characterize the spreading of musical genres among the users of the service. The object set $\Psi$ is composed by 402 artists, each one having a tag corresponding to her main musical genre. To characterize the typical values of width, depth and strength of the 9 most popular musical genres, we group the leaders through a clustering algorithm, using as features their width, depth and strength values. As clustering algorithm we use the self-organizing map (SOM) method [181] for the following reasons: (i) SOM does not require to set the number of clusters $k$; (ii) the standard k-means clustering algorithm [39] outperforms SOM only if the number of resulting clusters is very small (less than 7) [182], but our study of the best $k$ to be used in k-means with the sum of squared errors (SSE) methodology resulted in a optimal number of clusters falling in a range between 9 and 13 (in fact, SOM returned 12 clusters); and (iii) SOM performs better if the data points are contained in a warped space [183], which is our case.

In Table 8.3(a), we report the expected number of leaders with the given genre in the cluster. This is a measure known as revealed comparative advantage: $RCA(i,j) = \frac{freq_{i,j}}{\sum_{i} freq_{i,j}}$, where $i$ is a genre, $j$ is a cluster, $freq_{i,j}$ is the number of leaders who spread an artist tagged with genre $i$ that is present in cluster $j$. In Table 8.3(a), for each cluster we highlighted what is the genre with the highest unexpected presence. The centroids of the SOM are depicted in Figure 8.3(b): depth on the x-axis, strength on the y-axis and
the width as the color (strength and width are in log scale). We can identify the clusters characterized by the highest and lowest strength (9 and 4 respectively); by the highest and lowest depth (2 and 9 respectively); and by the highest and lowest width (11 and 1 respectively). There are also clusters with relatively high combinations of two measures: cluster 10 with high strength and width or cluster 5 with high depth and width.

From Table 8.3(a) we obtain a description of what values of width, depth and strength are generally associated with each musical genre. For space constraints, we report only a handful of them for the clusters with extreme values. Jazz music dominates clusters 1 (with the lowest width) and 4 (with the lowest strength): this suggests that jazz is a genre for which it is hard to be a prominent user. Cluster 9, with the lowest depth but the highest strength, is dominated by pop music (which dominates also clusters 10 and 11, both with high strength but low depth). As a result, we can conclude that prominent leaders for pop artists are embedded in groups of users very engaged with the new artist. On the other hand, it is unlikely that these users will be prominent among their friends too. Pop music strongly affects the neighbors of a musical leader, but the produced influence chain is not very long. Finally, cluster 2 with the highest density has a large majority of punk leaders. From this evidence, we can conclude that punk is a genre that can achieve long cascades, exactly the opposite of the pop genre. Jazz, pop and punk music are hence the genres better characterized by the cluster structure, presenting well defined characteristics in the patterns of diffusion from leaders: jazz music does not diffuse with high strength; pop music in contrast has high strength but low depth, and does not generate long diffusion chains; punk music finally presents high depth but low strength and width: it goes far away in the network but with a weak power of infection.

A leader is not bounded to be leader just for one object $\psi$, but she is free to be prominent in many $\psi$. Therefore, one leader can be counted in more than one genre. To help understand the magnitude of the issue, we depicted in Figure 8.4 the number of leaders influencing their neighbors for a given amount of actions (left) and for a given amount of genres (right). The typical leader influences one neighbor for one artist. However, there is a certain amount of leaders expressing their leadership for at least 8 objects and 4 genres. Figure 8.5 shows the log-binned distributions, for the leaders of each genre, of four of the topological measures studied in Section 8.3.2: degree, closeness centrality, clustering and neighbor degree. We omit betweenness centrality for its very high correlation with degree. Overall, there is no significant distinction between the genres in the distributions of the topological features.

The most noticeable information is carried by the degree distributions (Figure 8.5, top left). Each distribution appears very different from the overall degree distribution (Figure 8.2, left). There are less leaders with low degree than expected, therefore it appears that a high degree increases the probability of being a leader. On the other hand, we know that central hubs have on average lower depth and width. As a consequence, it appears that the best leader candidates are the nodes with an average degree between 10 and 100 (Figure 8.5, top left).

Using our leaders' minimum diffusion trees, we can extract some patterns that help us
CHAPTER 8. THE LOCAL DIFFUSION PATTERNS OF MUSICAL LISTENINGS

Figure 8.4: Distribution of number of objects (left) and of genres (right) per leader.

Figure 8.5: Distribution of leaders’ Degree (top left), Closeness Centrality (top right), Clustering (bottom left) and Neighbor Degree (bottom right) per tag.
Table 8.3: Presence of different diffusion patterns per genre.

<table>
<thead>
<tr>
<th>Pattern</th>
<th>dance</th>
<th>electronic</th>
<th>folk</th>
<th>jazz</th>
<th>metal</th>
<th>pop</th>
<th>punk</th>
<th>rap</th>
<th>rock</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3.62% (35.42%)</td>
<td>3.04% (22.50%)</td>
<td>3.94% (30.30%)</td>
<td>7.25% (62.50%)</td>
<td>4.14% (23.08%)</td>
<td>3.69% (32.01%)</td>
<td>6.56% (27.59%)</td>
<td>4.01% (27.97%)</td>
<td>4.22% (30.43%)</td>
</tr>
<tr>
<td></td>
<td>2.55% (25.00%)</td>
<td>3.92% (29.00%)</td>
<td>3.15% (24.24%)</td>
<td>4.35% (37.50%)</td>
<td>4.83% (26.92%)</td>
<td>3.61% (31.29%)</td>
<td>10.66% (44.83%)</td>
<td>5.60% (38.98%)</td>
<td>4.12% (29.71%)</td>
</tr>
<tr>
<td></td>
<td>3.40% (33.33%)</td>
<td>3.79% (28.00%)</td>
<td>3.94% (30.30%)</td>
<td>4.35% (37.50%)</td>
<td>6.90% (38.46%)</td>
<td>4.73% (41.01%)</td>
<td>12.30% (51.72%)</td>
<td>4.99% (34.75%)</td>
<td>4.52% (32.61%)</td>
</tr>
</tbody>
</table>

obtaining a complementary point of view over the leader prominence for different music genres. We mine a graph dataset composed by all diffusion trees $T_{l,ψ}$ with the VF2 algorithm [184]. Suppose we are interested in counting how frequent is the following star pattern: a leader influences three of its neighbors in the diffusion trees of pop artists. In our data, we have 5,043 diffusion trees for pop artists, of which 581 have at least four nodes. Since the VF2 algorithm found the star pattern in 186 of these graphs, we say that it appears in 3.69% of the trees, or in 32.01% of the diffusion trees that have a sufficient number of nodes to contain it. In Table 8.3 we report the results of mining three patterns of four nodes: i) the star-like pattern described above; ii) a chain where each node is prominent for (at least) one neighbor; iii) a split where the leader is prominent for a node, which itself is prominent for two other neighbors. Two values are assigned to each pattern and genre pair: the relative overall frequency, and the relative frequency considering only the trees with at least four nodes (in parentheses).

There is no necessary relation between the patterns and width, depth and strength measures: a low depth does not imply the absence of the chain pattern, nor does a high width imply a high presence of the star pattern. However, the combination of the two measures may provide some insights. For instance, we saw in Table 8.3(a) that jazz leaders are concentrated in the lowest width cluster. However, many jazz leaders who affect at least three nodes tend to be prominent in their neighbors, much more than in any other genre (7.25% of all leaders, 62.5% of leaders who are prominent for at least three other nodes). Therefore, jazz leaders have low prominence among their friends, however they are likely to have at least three neighbors for which they are prominent.

The chain pattern is more commonly found in pop leaders than in folk ones, even though the clusters of their leaders described in Table 8.3(a) would suggest the opposite. It seems that pop leaders are not likely to be prominent for nodes any further than the third degree of separation, while folk leaders tend to generate longer cascade chains. Also in this case, punk leaders are commonly found in correspondence with chain patterns, just as Table 8.3(a) suggested.

Although pop leaders show a much greater strength value than metal ones (by confronting in Table 8.3(a) their presence in high strength clusters like 9 or 10 and low strength clusters like 8 and 0), the split pattern tends to be more frequent in the metal genre (6.90%
against 4.73% of the trees). This phenomenon suggests us that metal leaders tend to be prominent for nodes strongly devoted to metal, inducing them to spread the music to their neighbors. Pop leaders, on the other hand, affect more neighbors with higher width and strength, presumably flooding their ego networks with the songs they like.

8.4 Related Works

In the literature, two phenomena are tightly linked to the concept of diffusion: the spread of biological [174] or computer [154] viruses, and the spread of ideas and innovation through social networks, the so-called “social contagion” [185], [186]. In both cases, the patterns through which the spreading takes place are determined not just by the properties of the pathogen/idea, but also by the network structures of the population it is affecting.

Some models have been defined to understand the contagion dynamics: the SIR, SIS and SIRS models [187, 188]. The idea behind them is that each individual transits between some stages in the life cycle of a disease: from Susceptible (S) to Infected (I), and from Infected to either Recovered (R) or again Susceptible. The availability of Big Data conveying information about human interactions and movements encouraged the production of more accurate data-driven epidemic models. The work in [174], for example, takes into account the spatio-temporal dimension, while authors of [154] study the spreading patterns of a mobile virus outbreak. Christakis and Fowler investigated the role of social prominence in the spread of obesity, smoking and happiness [189, 190]. Their results suggest that these health conditions may exhibit some amount of “contagion” in a social sense: although the dynamics of diffusion are different from the biological virus case, they nonetheless can spread through the social network.

8.5 Conclusions

In this chapter, we presented a study on the propagation of music listenings in a social network. We studied the patterns of diffusion from leaders (musical innovators) investigating three different dimensions: the prominence of a leader on how many neighbors, on how distant nodes and on how engaged nodes. We characterized each of these concepts with a different measure: width, depth and strength. We defined a leader detection algorithm and applied it to the LastFM online social network introduced in Section 3.4, showing that: (i) central hubs are usually incapable to strongly influence the behavior of the entire network; (ii) a trade-off exists between the length of the cascade chains and and how engaged each element of the chain is; (iii) leaders in tightly connected communities seem to achieve maximum engagement, although for this last point we do not have conclusive evidence. We also included a case study in which we show how artists in different musical genres are spread through the network.

Many future developments are possible. The limited prominence that central hubs have on the overall network may be studied in conjunction with the problem of network controllability [191]. Alternative leader detection techniques, such as the ones presented
in [192], can be confronted with our proposed algorithm. Finally, a deeper analysis of
the properties of the width, depth and strength measures can be performed, using addi-
tional techniques and exploiting data from other social media services like Twitter and
Facebook.
Part IV

Understanding the interplay between Human Behavior and Economic Development
Chapter 9

Human Mobility, Social Networks, and Economic Development

Big Data have the potential capability of creating a digital nervous system of our society, enabling the measurement, monitoring and short-term prediction of relevant quantitative aspects of the socio-economic structure in quasi real time [6]. An intriguing question is whether and how measurements made on the basis of Big Data can yield us high-fidelity proxies of socio-economic development and well-being, which can take into account not only material living standards (income, consumption and wealth), but many other important factors such as health, education, governance, social relationships, environment and security. In this chapter, we make a step along this direction and show how a widely available source of Big Data, mobile phone call records, support measurements of human movements and social relations and provide surprisingly strong correlations with several dimension of economic development.

Current research offers two different paradigms to explain the economic development of individuals, territories and social communities. On one side, a line of research investigate the impact of social networks on economic opportunities. Theoretical works suggest that more heterogeneous social ties generate more individual economic opportunities expressed as access to jobs [193], promotions [194], or higher salaries [195]. The seminal work by Eagle et al. [196] investigates empirically these theories by analyzing a nationwide mobile phone dataset and shows that, in the UK, regional communication diversity is positively associated to a socio-economic municipality ranking. On the other side, the fields of transport studies and economic geography investigate the impact of human mobility on economic development. While traditional approaches, like the ones by Sinclair [197] and Hall et al. [198], model impacts of transport and mobility in terms of size, density and distances of flows, current methodologies are changing towards a more personal way of investigating mobility and the effect on economic development [199].

Both the social and mobility paradigms recognize that economic development is related to individual or collective human behavior: how people interact and move within a territory affects, and is affected by, the economic development of the territory. From a scientific point of view, the empirical investigation of these relationships allows us to
validate existing theories with empirical proof. From a policy point of view, the investiga-
tion encounters the increasing demand by policymakers for continuous and updated information on the geographic distribution of poverty, inequality or life conditions.

In this chapter, we investigate and quantify the relations between human mobility, social networks and economic development in France based on nationwide gathered mobile phone data and small level census data. We first construct four individual metrics which describe different aspects of individual human behavior: the volume of sociality, the diversity of sociality, the volume of mobility and the diversity of mobility. In a second stage, we explore the associations between these four metrics and several indicators of economic development on a municipality scale. We find that the aggregated mobility diversity of individuals resident in the same municipality varies with socio-economic indexes such as income, education level and a deprivation index, whereas the aggregated social diversity does not. Consequently, the correlation between social diversity and economic development, which was significant in Eagle et al. [196] at the level of British communities, are less apparent at French municipality level, whereas the correlation between mobility diversity and economic development is clear. Diversity is hence a key concept for the social ecosystems, made up today of highly interconnected people immersed in a plurality of information and communication technologies. We also find that a remarkable systematic variation of the predictability of individual movements exists across geographical units defined on socio-economic indicators, an important finding when compared to previous works such as [9] which states that mobile predictability is very stable across different subpopulations delineated by personal characteristics like gender or age group. Being confirmed by two different null models that produce no correlations, our results reveal the high potential of Big Data in providing representative, relatively inexpensive and readily available measures as proxies of economic development, opening an interesting perspective to study human behavior through mobile phone data, as new statistical indicators can be defined to describe and possibly “nowcast” the economic health of a territory.

This chapter is based on the paper [200] submitted to the International Journal of Geographical Information Science (IJGIS).

9.1 Measuring Human Behavior

We study the relationship between human mobility, social networks and economic development on a nation-wide scale exploiting the access to two data sources. We use the first one, provided by the French National Institute of Statistics and Economic Studies (INSEE), to derive socio-economic indicators from small level census data. Secondly, we describe mobile and social behavior of individuals in France through the Orange GSM dataset described in Section 3.2.

9.1.1 Socio-Economic Data

We derive measures describing the socio-economic situation of French municipalities from the indicators provided by the French National Institute of Statistics and Economic Studies
### 9.1. MEASURING HUMAN BEHAVIOR

#### Measures

<table>
<thead>
<tr>
<th>Measures</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>residents</td>
<td>the official number of residents in the municipality</td>
</tr>
<tr>
<td>per capita income</td>
<td>the total income of the municipality divided by the residents in the municipality</td>
</tr>
<tr>
<td>primary education rate</td>
<td>the ratio between people with primary education only and people older than 15 y.o.</td>
</tr>
<tr>
<td>EDI</td>
<td>the European Deprivation Index as defined by [201]</td>
</tr>
</tbody>
</table>

Table 9.1: The socio-economic indicators at municipality level derived from the data provided by the French National Institute of Statistics and Economic Studies (INSEE).

(INSEE). To be specific, for each of the 36,568 municipalities we collect data on the number of residents, the per capita income, the education level, and a deprivation index which provides a poverty indicator (see Table 9.1).

Figure 9.1(a) shows the distribution of the (logarithm of the) per capita income, a metric for the available monetary resources of people in each municipality. The peaked distribution suggests that the majority of municipalities have a per capita income in the order of $10^4$. The European Deprivation Index (EDI) is a composite measure of relative prosperity, based on variables like home overcrowding, unemployment, accessibility to cars, house ownership and the environmental quality of each municipality, as described by Pornet et al. [201]. The distribution of EDI among the French municipalities produces a peaked curve, highlighting the presence of a characteristic deprivation value in France. Measures on the education level provide useful information about the degree of development of a territory as basically the higher the education level, the higher is the richness of a country [202]. We describe the education level of municipalities as the ratio of people with primary education only. The measure shows a peaked distribution with a mean value 0.15, characteristic of developed countries (see Figure 9.1(c)). In more general terms, the average education level in France is good, since only a small percentage of the population has a low level of education.

#### 9.1.2 CDR data

Being gathered for billing purposes by mobile phone operators, Call Detail Records (CDR) collect geographical, temporal and interactional information on human mobile phone use. As penetration rates of mobile phone use have been steadily growing over the last decades [203], these data sources show an enormous potential to empirically investigate human dynamics on a society wide scale. We use here the Orange dataset described in Section 3.2, consisting of CDR data describing 215 millions phone calls performed by 21 million users in France during a period of 45 days (from September 1 to October 14, 2007). In order to overcome several limitations and shortcomings of the original data and to retain individuals with reliable statistics, we carry out some filtering phases on the original data source.

Firstly, for each individual we remove all the locations visited only once or having a
visitation frequency under a given threshold. We set the threshold to $f = n_i/N < 0.005$, where $n_i$ is the number of calls performed by the individual in location $i$ and $N$ the total number of calls performed by the individual during the period of observation. The filter thus evaluates whether the used locations are relevant with respect to the specific volume of calls of the individual, hereby assuming that locations that are not visited regularly are of limited meaning for our analysis. Since it is meaningless to analyze the mobility of individuals who do not move, we discard all the users with only one location after the previous filter.

In a second filter we select only individuals with a relative high call frequency, assuming that only these individuals would produce usable insights in our analysis. We set the call frequency threshold to $f = N/45 < 0.5$, where 45 days is the length of our period of observation. This means that we delete all the users with less than one call each two days (in average). Finally, we avoid the presence of abnormally active users like line testers and alarm managers by discarding the users with a huge number of calls $N > k \times 45$, where $k = 300$.

We finally construct a social network for the 45 days of call communication. Whereas a single call between two individuals may not carry much information, reciprocal calls of long duration between two users serve as a signature of more intense work-, family-, leisure-, or service-based relationships. We translate the phone log data into a social network representation by linking two individuals in the social network if at least one reciprocated pair of calls exists between them during the period of observation (i.e. A called B and B called A). We define the weight of the links between two individuals $\Lambda$ and $B$, $w_{AB}$ and $w_{BA}$, as the aggregated duration of the calls between them. This procedure eliminates a large number of one-way calls, most of which correspond to single events and do not represent meaningful communications. Ultimately, the resulting mobile call graph contains 6,289,865 active users and 33,071,948 edges.
9.1.3 Social and Mobility measures based on CDR data

Based on the filtered CDR data, we construct two metrics that capture aspects of individual sociality: the degree and the D-social. Within a social network, we can express the volume of social interactions of an individual by counting the amount of links she possesses with others. This simple measure of connectivity is widely used in network science and is called the degree of an individual. Here the degree of an individual is the number of different individuals that are in contact by mobile phone calls with her. We can therefore see the degree as a proxy for the volume of sociality for each individual. Social networks, as in the case of call graphs, are usually directed networks, meaning that nodes have two different degrees: the in-degree (the number of incoming links) and the out-degree (the number of outgoing links). We define the social degree of an individual as the sum of the in-degree and the out-degree:

\[
deg(u) = \deg_{\text{in}}(u) + \deg_{\text{out}}(u) \tag{9.1}
\]

It is known that social networks are highly heterogeneous with respect to the number of friendships [75, 74] which often results in power-law degree distributions. Due to such a variability the degree of an individual provides a useful metric to differentiate between individuals.

The social diversity measure D-social, first defined by Eagle et al. [196], quantifies the topological diversity in a social network as the Shannon entropy associated with an individual’s communication behavior:

\[
D_{\text{social}}(i) = -\sum_{j=1}^{k} p_{ij} \log \left( \frac{p_{ij}}{\log(k)} \right) \tag{9.2}
\]

where \( k \) is the out-degree of individual \( i \), \( p_{ij} = \frac{V_{ij}}{\sum_{j=1}^{k} V_{ij}} \) and \( V_{ij} \) is the call volume between individual \( i \) and individual \( j \). We can interpret the D-social as a measure for the social diversification of each individual according to its own interaction pattern. In a more general way, individuals who possess a fixed way of daily communications with their connections reveal a low social diversification resulting in lower values for D-social, whereas individuals who call in a more erratic and unpredictable way show high social diversification leading towards higher D-social values. From another perspective, the social diversity is also related to the predictability of individuals’ calls: the higher the social diversification, or thus the values for D-social, the lower is the social predictability of the individual.

Based on the location traces registered by the CDR data, we define two metrics to describe individual mobility: the radius of gyration and the mobility entropy. The radius of gyration \( r_g \), first introduced in the mobility context by González et al. [8], characterizes the spatial spread of the locations visited by an individual from the trajectories’ center of mass (i.e. the weighted mean point of the locations visited by an individual), defined as:

\[
r_g = \sqrt{\frac{1}{N} \sum_{i \in L} n_i (r_i - r_{\text{cm}})^2} \tag{9.3}
\]
where $L$ is the set of locations visited by the individual, $n_i$ is the location’s visitation frequency, $N = \sum_{i \in L} n_i$ is the sum of all the single frequencies, $r_i$ and $r_{cm}$ are the vector of coordinates of location $i$ and center of mass respectively. The radius of gyration provides with a measure of mobility volume, indicating the characteristic distance traveled by an individual. Besides the volume of individual mobility, we estimate the diversity of individual mobility by using the Shannon entropy:

\[
E(u) = -\frac{\sum_{e \in E} p(e) \log p(e)}{\log N}
\]  

(9.4)

where $e = (a, b)$ represents a pair of location, $p(e)$ is the probability of observing a movement between locations $a$ and $b$, and $N$ is the total number of individual’s trajectories. Similarly to the D-social measure, the mobility entropy is one when, starting from a given location, each outgoing trip is directed to a different location (indicating a high diversity in mobility undertakings), while it is zero when all the outgoing trips are directed to the same location (indicating a low diversity in mobility undertakings). Seen from another perspective, the mobility entropy of an individual also quantifies the possibility to predict individual’s future whereabouts. Individuals having a very regular movement pattern possess a mobility entropy close to zero and their mobility undertakings are rather predictable. Conversely, individuals who have a mobility entropy going to one have very unpredictable mobility undertakings.

We compute the four individual measures defined above on the filtered CDR data: radius of gyration, mobility entropy, social degree and D-social. The measures of volume (radius of gyration and social degree) show a great variability across the population, producing heavy-tail distributions (Figure 9.2(a), (b)). In the case of radius of gyration, most individuals cover small distances, but a few travel regularly over hundreds of kilometers. These results are consistent with previous studies on human mobility, performed on mobile phone data [8] and on GPS tracks of traveling vehicles [14]. A similar heterogeneity characterizes the volume of people’s social interactions. The majority of individuals in the social network possess only a few social links, while a few hubs have an enormous number of links, confirming previous results discovered on an Instant Messaging platform [74] and on a mobile phone network [75].

In contrast with the measures of volume, the measures of diversity (D-social and mobility entropy) produce peaked distributions (Figure 9.2(c), (d)). Regardless the characteristic traveled distance and the social degree each individual is hence as predictable as everyone else, a finding that has been stated before in the mobility context by Song et al. [9]. In general, the observed distributions allow to state that the volume of movements and social interactions observed in the CDR dataset is heterogeneous among the population, whereas in terms of diversity of movements and social interactions we find that a substantial homogeneity exists among the investigated population.

Such observations raise the question to which degree sociality and mobility might, or might not be associated to one another. We find no relationship between volume or diversity measures of sociality and mobility on an individual level: the correlation between radius of gyration and social degree, as well as the correlation between mobility entropy
and D-social, is zero. This suggests that, although their distributions are comparable, mobility and sociality measures capture different aspects of individual behaviour. It is not possible, thus, to accurately predict individual mobility features from individual sociality features and vice versa.

Figure 9.2: The distributions of the mobility and social measures computed over the Orange users in the filtered dataset. (a) Distribution of radius of gyration; (b) Distribution of social degree; (c) Distribution of entropy; (d) Distribution of D-social.
9.2 Results

We aggregate the individual measures at the municipality level by assigning to each individual a home location according to the algorithm proposed in [204], which sets an individual’s home as the phone tower where the individual performs the highest number of calls during nighttime (from 10 pm to 7 am). Based on these locations, we assign the individuals to the corresponding municipality. To improve the reliability of the analysis, we discard municipalities with less than 1,000 official residents or less than 100 Orange users. After the filtering phase, we obtain a set of 4,450 municipalities. The map in Figure 9.3 shows the spatial distribution of Orange users across the (filtered) French municipalities.

Figure 9.3: The spatial distribution of Orange users over the French municipalities. The color of municipalities, in a gradient from blue to red, indicates the number of Orange users assigned to that municipality. Municipalities deleted during the filtering phase are in white. We observe that the number of Orange users in the municipalities is not uniform, but varies according to the density of the municipality.

Now that the individual measures for mobility and sociality are aggregated at the
municipality level, we investigate their interplay with the socio-economic indicators that are also available at this level of analysis. We find two main results. First, the measures of volume for both mobility and sociality (radius of gyration and social degree) are not correlated with the socio-economic indicators ($\rho \approx 0$). Second, we find that the mobility entropy is a better predictor for economic development than social diversity. Figure 9.4 shows the relations between the diversity measures and three economic indexes: the European Deprivation Index, the per capita income and the education level. For mobility entropy, clear tendencies appear: as the mean mobility entropy of municipalities increases, the deprivation index and the rate of uneducated people tends to decrease, while the per capita income tends to increase. Social diversity, in contrast, has a weak predictive power with respect to the economic development, exhibiting weaker correlations than the mobility entropy.

Figure 9.5 provides another way to observe the relations between the diversity measures and the economic development. We split the municipalities in deciles based on the values of the deprivation index, and for each decile we compute the distributions of the mobility entropy and D-social at the municipality level. For mobility entropy, as the deciles of the economic values increase both the mean and standard deviation of the mobility entropy distribution change. This is consistent with the observations made in the plots of Figure 9.4(a). Conversely, we do not observe a significant change in the average and the standard deviation for the social diversity distribution. Figure 9.4 and Figure 9.5 suggest us that the mobility entropy aggregated at the municipality level is better associated with the socio-economic indicators than the aggregated D-social measure. The perceived relation between mobility diversity and the European Deprivation Index, for example, is stronger and more evenly distributed over the different levels of EDI values for municipalities.

We compare our findings with the results produced by two null models, in order to test the significance of the correlations observed on the empirical data. In the first null model NM1, we randomly distribute the Orange users over the French municipalities by extracting uniformly $N$ users from the CDR dataset and assigning them to a random municipality with a Orange population of $N$ users. We then aggregate the individual measures of the users assigned to the same municipality. We repeat this process 100 times and take the mean of the aggregated values of each municipality produced in the 100 experiments. In the second null model NM2, we randomly shuffle the values of the socio-economic indicators over the French municipalities. We perform this procedure 100 times and take, for each municipality, the mean value of the socio-economic indicators computed over the 100 produced values. In contrast with empirical data, we find no correlation in the null models between the diversity measures and the socio-economic indicators, neither for mobility nor for sociality (Figure 9.6 and 9.7). This suggests that a spatial homophily exists: individuals residing in the same territory exhibit a similar mobile behavior, which is linked to the economic development of the territory itself.

Figure 9.8 summarizes our findings. Here we organize municipalities by deciles based on the investigated socio-economic indicators. We then compute the mean values for both mobility entropy and D-social of these decile groups. The results confirm our initial inter-
The associations between socio-economic indicators and mobility diversity are more explicit than the associations between socio-economic indicators and social diversity. Generalizing, we conclude that human mobility proves to be a better predictor for the economic development of cities when compared to social behavior. This is also true for other socio-economic indicators: mobility entropy is stronger correlated with per capita income and education level than social diversity.

9.3 Discussion

The most remarkable result in this study is the observation that human mobility, and mobility diversity in particular, is more associated with socio-economic measures on a municipality scale than social diversity. To be specific, on a municipality level mobility entropy is positively correlated with the per capita income and the education level, while negatively correlated with the deprivation index (Figures 9.4 and 9.8). Such results are significant since two null models, obtained randomly distributing users or socio-economic indicators over the French municipalities, cannot account for the observed relations producing null correlations between the behavioral measures and the economic development.

Generalizing our empirical findings, we can state that a greater diversification of human mobility is linked to a higher overall wealth, to a more educated territory and to a lower poverty and deprivation index. Remarkable is that a systematic variation of the mobility entropy distribution exists across geographical units defined on socio-economic indicators (Figure 9.8), delineating subpopulations where a different distribution of entropy emerges based on the occurrence of socio-economic indicators. This is an important finding when compared to the seminal work on the predictability of human mobility by Song et al. [9], which states that mobility entropy is very stable across different subpopulations delineated by personal characteristics like gender or age group. The contrast between our findings and the result of Song et al. [9] suggests that socio-economic situations on a city scale are more related to individual mobility than individual characteristics or communication patterns. The observed variation also suggests a relation between economic development and predictability: people resident in more developed and richer territories tend to show a higher mobility entropy and hence more unpredictable mobility patterns.

Although the relations between mobility diversity and socio-economic indicators appear clearly, it is difficult to formulate a hypothesis to explain their connections. Without a doubt, the relation between socio-economic indicators and mobility diversity is two directed. It might be that a well-developed territory provides for a wide range of activities, an advanced network of public transportation, a higher availability and diversification of jobs, and other elements that foster mobility diversity. As well as it might be that a higher mobility diversification of individuals lead to a higher economic well-being as it could nourish economy, establishes economic opportunities and facilitate flows of people and goods. Interpretations of the relation between mobility diversity and economic development are not directly derivable from the empirical results and should therefore be combined with more thorough theoretical insights.
A second line of discussion arises from the remarkable differences between the results of our analysis and the results obtained by Eagle et al. [196]. Whereas, for the UK, Eagle et al. find a strong positive association between social diversity of local communities and a British deprivation index, the proposed case of France does not exhibit such a clear relationship. In this perspective it is worth highlighting three main differences between our study and the one by Eagle et al. First, while we compute the measures of social diversity (D-social) in the same way as the UK study, the two deprivation indices are different and calculated by two different institutions. In particular, the UK deprivation index also includes information about the income and the education level, which are separate variables in our study. Second, our spatial granularity refers to municipalities (basically, cities), while Eagle's granularity refers to census areas, which are more fine-grained with respect to municipalities. Third, the social network constructed in our research uses mobile phone data solely, while the social network extracted in Eagle et al. case refers to both mobile phones and residential landlines depicting a more elaborate and objective view on individuals' social network. Discussing the difference between mobility diversity and social diversity as a proxy for economic development, we could say that mobile phone call records offer a realistic and objective account of each user's personal mobility, when observed over a long time, while severely underestimating the social connections, due to the multiple means that people have to interact in the physical and digital world. Moreover while nowadays the new communication technologies makes the social interactions relatively inexpensive, movements and displacements requires a much higher effort in terms of costs and time. Accordingly, it is no surprise that mobility gives a stronger account of real social phenomena.

From a geographic point of view, our results suggest that scale plays an important role in the relationship between mobility diversity and social diversity. On an individual level, no significant correlations between the two measures exists while at the inter-municipality level a significant relation emerges. The reasons for this scale related emergence of relationships is difficult to assess, and cannot simply be retrieved from our analysis only. Nevertheless, the results suggest that although individual possess their own, unique way of communicating and moving, their environment on a city scale implies some sort of convergence towards the behavior of other individuals in their geographical proximity. The analysis of CDR data form a unique way to empirically describe such scale related processes and might give direction to other future research efforts.

9.4 Conclusions

We know that bio-diversity is crucial to the health of natural ecosystems and for the balance, or wellbeing, of plant and animal species that inhabit them. The story narrated in this chapter shows that diversity is also a key concept for the social ecosystems, which are made up today of highly interconnected people immersed in a plurality of information and communication technologies. It makes thus sense, a lot of sense, to speak of diversity in a social meaning, especially for data scientists and quantitative sociologists. The seminal
work by Eagle et al. [196] showed that wellbeing is correlated to social diversity intended as the diversification of phone calls over individuals’ contacts. Our work shows that mobility 


diversity is even more important: a greater diversification of individuals’ movements within a territory is linked to a lower value of various poverty indicators, and a larger per capita income.

These findings open an interesting perspective to study human behavior through mobile phone data, as new statistical indicators can be defined to describe and possibly “nowcast” the economic health of a territory. Next chapter will provide another evidence of the high potential of Big Data in providing representative, relatively inexpensive and readily available measures as proxies of socio-economic indicators of poverty, well-being and progress. This monitoring and short-term prediction of prosperity at a small geographical scale (like a metropolitan area) can form an important tool for policymakers for smart city planning, epidemiology, hazard resilience strategies, and so on.
9.4. CONCLUSIONS

Figure 9.4: The relation between the diversity measures and (a)-(b) European Deprivation Index, (c)-(d) per capita income and (e)-(f) primary education rate. The color of a point indicates, in a gradient from blue to red, the density of points around it. In order to highlight the presence of tendencies, we split the municipalities into ten equal-sized groups according to the deciles of the measure on the x axis. For each group, we compute the mean and the standard deviation of the measure on the y axis and plot them through the black error bars. \( \rho \) indicates the Pearson correlation coefficient between the two measures.
Figure 9.5: The distributions of mobility entropy (a) and D-social (b) in the different deciles of the European Deprivation Index (EDI). We split the municipalities into ten equal-sized groups computed according to the deciles of the EDI measure. For each group, we plot the distributions of the mobility entropy and D-social measures. The blue dashed curve represents a fit of the distribution, the red dashed line represents the mean of the distribution.
Figure 9.6: The relation between diversity measures computed on the null model NM1 and (a)-(b) European Deprivation Index, (c)-(d) per capita income and (e)-(f) primary education rate. The color of a point indicates, in a gradient from blue to red, the density of points around it. In order to highlight the presence of tendencies, we split the municipalities into ten equal-sized groups according to the deciles of the measure on the x axis. For each group, we compute the mean and the standard deviation of the measure on the y axis (the black error bars).
Figure 9.7: The relation between diversity measures computed on the null model NM2 and (a)-(b) European Deprivation Index, (c)-(d) per capita income and (e)-(f) primary education rate. The color of a point indicates, in a gradient from blue to red, the density of points around it. In order to highlight the presence of tendencies, we split the municipalities into ten equal-sized groups according to the deciles of the measure on the x axis. For each group, we compute the mean and the standard deviation of the measure on the y axis (the black error bars).
Figure 9.8: The mean values of mobility entropy (a) and D-social (b) for groups of municipalities based on the deciles of the economic indicators. We split municipalities into ten equal-sized groups according to the deciles of the economic measures. For each group, we compute the mean for the mobility entropy and the D-social.
Chapter 10

Small area estimators using Big Data sources

Nowadays there is a growing request by policymakers for more detailed and up-to-date information about geographic distribution of poverty, inequality and life condition indicators. Unfortunately, statistical agencies are not prepared to provide estimates for the socio-economic situation of a territory with the requested spatial and temporal detail. Many socio-economic indicators, such as the GDP, are computed once every three months or even once a year, making timely and effective policy decisions impossible. Once the new measurements are available, indeed, they depict a socio-economic situation that could not reflect the current economic situation of the country/territory. In a rapidly changing world where new markets rapidly raise while old ones fall down, policymakers need today up-to-date measurements on the socio-economic health of a given territory. In this perspective, many national statistical agencies and researchers are now promoting Big Data methodologies and best practices to develop, evaluate and implement poverty estimation and poverty mapping. For example the European Commission funded two projects, SAMPLE (Small Area Methods for Poverty and Living Condition Estimates) and AMELI (Advanced Methodology for European Laeken Indicators), related to this topic. More recently the Sheveningen Memorandum and the European Statistical System data event in Rome tracked how the Big Data could be of real interest for official statistics.

In this chapter we present a contribution to the use of small area estimation methods combined with Big Data aimed to improve our ability to measure, monitor and predict well-being, deprivation, poverty, exclusion and inequality at a fine-grained spatial and temporal scale. We identify three possible approaches to the use of Big Data in the small area estimation framework: (i) use Big Data sources to obtain local socio-economic indicators comparable with those obtained with small area estimation methods; (ii) use Big Data to set out new covariates for small area models; (iii) use survey data to check and remove the self-selection bias of the values of the indicators obtained using Big Data. In Section 10.1 we describe the first two approaches focusing on the study of well-being. In Sections

\[\text{http://www.cros-portal.eu/content/big-data-event-2014}\]
10.2 and 10.3 we present two applications of these approaches using survey produced by the European Union and Big Data on individuals’ mobility referring to Tuscany. Finally, in Section 10.4 we address the last approach and conclude with some final remarks on the combined use of Big Data and small area estimation in a statistical framework. This chapter is based on the paper [205] to appear in the Journal of Official Statistics.

10.1 Big Data in small area estimation

In the context of understanding social complexity, the connection of social mining with statistical data collection, analysis and modelling is fundamental. Generally, statistical data are collected by means of sample surveys or censuses. Censuses are complex and expensive to carry out, so sample surveys represent a common way to collect data. To measure social complexity with a focus on the identification and quantification of social exclusion and deprivation, there is a demand by policymakers for local level estimates of the most relevant poverty and well-being indicators. Generally the local level (local administrative area, zone of local governance) identifies a so-called unplanned domain of estimation in sample surveys. Oversampling to increase the sample size could be a solution to assess poverty and deprivation at a local level, say at Local Administrative Unit levels 1 and 2 (LAU 1 and LAU 2), as is often required by policymakers. However, the high cost in terms of time and financial resources makes the instrument impracticable for obtaining accurate estimates. Big Data can represent today an alternative source of data for the local level, usually reaching a very high level of geographical detail.

The first opportunity to reconcile data from the two independent sources - Big Data and sample surveys - is to compare local measures extracted from Big Data with measures on related aspects obtained from survey data. Measures from Big Data sources are usually obtained very quickly, but they could be affected by a serious self-selection bias. Conversely, small area estimates are methodologically sound though requiring timely survey and population data that can be difficult to obtain. Comparing the two alternative sets of measures referring to the same areas can give useful insights on the potential of Big Data to benchmark small area estimates. If we find accordance on the level of deprivation and poverty measured by Big Data and survey data, the evidence for analysts and policymakers can be considered very strong. However, if we find discrepancy between the results obtained from the two sources of data, then there is need of further investigation on those areas. This is the rationale underlying the application we show in Section 10.2, where Big Data on individuals’ mobility are compared with small area poverty estimates.

An alternative is to use Big Data directly as a covariate in a small area model. At its heart, poverty mapping is about combining survey data that measures poverty incidence with auxiliary information about the population of interest. On one hand we have survey data collected ‘ad hoc’ such as consumption and income, on the other hand we have auxiliary information obtained from other surveys, from population censuses or from administrative registers. We can use Big Data as common variables between survey and auxiliary information to improve the precision of the small area estimates. The extension
of the covariates to include Big Data variables such as GPS trajectories from vehicles or social media search loads presents difficulties and challenges. For example, due to technical problems and law restrictions it is unfeasible to link administrative archives or survey data with unit-level data. We can overcome this problem by using the so-called area level models, such as the Fay-Herriot model [206], where direct estimates obtained from survey data are modelled with area level auxiliary variables, i.e. a unique variable value for each area. Since in the Fay-Herriot model unit level is not required, it is relatively easy to aggregate Big Data in the areas of interest and use them in the model. In section 10.3 we present an application where elaborations from Big Data are used as covariates in a Fay-Herriot model to estimate poverty indicators, accounting for the presence of measurement error in the covariates [207].

A second alternative is to use Big Data directly to measure poverty and social exclusion, appropriately taking into account the self-selection problem. We envision that survey data could be used to check and remove this bias, provided that unit-level information from Big Data sources will be available. We discuss this problem in the final section of the chapter.

10.2 Big Data and small area estimation methods

There has been rising interest in research on poverty mapping over the last decade, with the European Union proposing a core of statistical poverty indicators commonly known as Laeken Indicators [208]. In each of the EU countries these indicators can be computed using data coming from sample surveys, such as the EU-SILC survey (European Union Statistics on Income and Living Conditions). The EU-SILC fulfills information on the household equivalentized income for each of the sampled households: this information is fundamental to compute monetary poverty indicators, such as the Head Count Ratio, for any area of interest.

The Head Count Ratio - also known as the At-Risk-of-Poverty rate - measures the incidence of poverty of an area. It is a special case of the generalized measures of poverty introduced by Foster et al. [209], hereafter FGT. Denoting by $t$ the poverty line, that is the level of income that defines the state of poverty, by $d$ the area of interest, and by $\alpha$ a sensitivity parameter, the class of FGT poverty measures for a given area $d$ is defined as:

$$F_{\alpha,d} = \frac{1}{N_d} \sum_{j=1}^{N_d} \left( \frac{t - y_{jd}}{t} \right)^\alpha I(y_{jd} \leq t).$$

(10.1)

Here $y$ is a measure of income for unit or household $j$, $N_d$ is the number of units or households in area $d$, $I$ is the indicator function that is equal to 1 when $y_{jd} \leq t$, 0 otherwise. When $\alpha = 0$, $F_{\alpha,d}$ is equal to the Head Count Ratio or HCR indicator, which is simply the proportion of units or households in the area with an income at the poverty line or below the poverty line. Since this index is easy and fast to compute it is widely used in poverty estimation. However, it should always be supplemented by other poverty

---

2a household consists of one or more individuals who live in the same dwelling and also share at meals or living accommodation, and may consist of a single family or some other group of people.
indicators referring to the same areas, such as the Poverty Gap, a measure of the severity of poverty, and the mean of the Household Equivalised Income (HEI).

The computation of these indicators using data coming from the EU-SILC survey results in accurate estimates only for planned domains, that is for the administrative areas corresponding to the Italian regions (NUTS 2 level). When the interest is in obtaining the estimates at different geographical levels (province, municipality, district, etc.), we need to use small area estimation techniques since the sample size in many areas is too small to obtain accurate direct estimates. The areas of interest can correspond to administrative areas that are at the base of important local policy decision processes, such as provinces and municipalities, or to areas that can be used to analyze the socio-economic structure of the territory, such as the Local Labour Systems (LLSs).

Here we compute the estimates of the HCR and the mean HEI for the ten provinces in Tuscany by applying the M-quantile estimators proposed by Tzavidis et al. [210] and Marchetti et al. [211] to data from the EU-SILC 2008 and the Population Census 2001 [212]. M-quantile models [213] relax the parametric assumption of random effect models traditionally used for small area estimation [214] representing an advantage in many real data applications [215, 216]. The covariates we include in the model for the mean HEI are the house ownership status, the age of the head of the household, the employment status of the head of the household, the gender of the head of the household, the years of education of the head of the household and the household size. Although the 2008 EU-SILC data were collected six years after the census, the 2001-2007 period was one of relatively slow growth and low inflation in Italy, so it is reasonable to assume that there was relatively little change. Moreover, using an enlarged sample of the EU-SILC 2008 survey for the province of Pisa, Giusti et al. [215] showed that M-quantile small area estimates of the HCR and of the HEI were coherent with EU-SILC 2008 direct estimates using the oversample reliable estimates.

The use of Big Data as a covariate in the M-quantile unit-level models raises some issues. First, the auxiliary variables used in the small area model should be measured at the unit level and they should be the same as those available from the survey: the problem is that Big Data with such characteristics are usually unavailable due to confidentiality reasons. Second, even if available, Big Data cannot be considered as covering the population of interest, due to the self-selection problem. For these reasons, we use Big Data on human mobility separately in this application to produce a measure of entropy of individuals’ movements for the provinces of Tuscany. Generalizing the approach of Eagle et al. [196] and following the approach of Chapter 9, we study the correlation between the level of poverty estimated by small area estimation methods and the mobility diversity of inhabitants in the provinces of Tuscany extracted from Big Data.

We use the Octo dataset (Section 3.1) and, following the strategy of Chapter 5, we map vehicles’ traces on the road network and associate their position during the stops with the census sectors provided by the Italian National Institute of Statistics (ISTAT). In this way, we describe each vehicle’s trip by a tuple composed of the timestamp and a pair of coordinates corresponding to the origin and destination of the trip. The mobility
of a vehicle $v$ is given by:

$$E(v) = -\sum_{d=1}^{D} p_v(d) \log p_v(d)$$  \hspace{1cm} (10.2)

a measure of entropy where $d = (a, b)$ represents a pair of an area, $p_v(d)$ is the probability of observing a movement of vehicle $v$ between the area $a$ and $b$. The probability $p_v(d)$ is given by the ratio between the number of trips of $v$ between $a$ and $b$ and the total number of trips of $v$. The mobility value is low when the vehicle $v$ visits few distinct locations, showing low mobility diversity. On the other hand, when the mobility measure increases the vehicle $v$ makes trips to several areas as destinations. We associate each vehicle to the area where its most frequent location is situated, and then define the mobility of an area $d$ as:

$$E(d) = \frac{1}{V_d} \sum_{v \in d} E(v)$$  \hspace{1cm} (10.3)

where $V_d$ is the number of vehicles associated to area $d$. We compute for each area a measure of mobility complexity, namely the standard deviation of the mobility $E(d)$. Denoting by $V_d$ the subset of vehicles associated to area $d$, for a given area we measure the standard deviation of mobility by:

$$S_{E(d)} = \sqrt{\sum_{v \in d} (E(v) - E(d))^2 / V_d}.$$  \hspace{1cm} (10.4)

Figure 10.1 shows the M-quantile estimates of the HCR (on the left) and the standard deviations of mobility entropy $S_{E(d)}$ (on the right), and their root mean squared error (in parenthesis) for the ten provinces of Tuscany. The linear correlation coefficient between the values of HCR and the $S_{E(d)}$ is $-0.74$, suggesting that higher levels of heterogeneity of mobility are in the provinces where poverty is at lower levels. In other words, the diversification of movements in each province is a proxy of the level of poverty. This result is similar to that obtained in previous chapter, where we found that the diversity of individuals’ movements is correlated with the economic development of French municipalities. The case of Tuscany, jointly with the example of France, shows how Big Data have the potential to mirror aspects of well-being and other socio-economic phenomena, supporting the evidence emerging from survey data.

### 10.3 Big Data as covariates in area level models

In this section we present an application of the Fay-Herriot model proposed by Ybarra and Lohr [207] to estimate poverty indicators for the Local Labour Systems (LLSs) of Tuscany, using Big Data as covariates. LLSs are areas in which most of the daily activity of the people who live and work in them takes place; their definition is similar to that of the travel-to-work-areas (TTWAs) widely used in US and UK territorial analyses. According to the official EU nomenclature of local units they are intermediate between LAU 1 and
LAU 2 levels. We can overcome the limitations of the use of Big Data as covariates in unit level models by using area level models. Fay and Herriot [206] proposed a famous model to reduce the variability of direct estimators based on survey data by using auxiliary information coming from other data sources. The Fay-Herriot model [207] and its spatial extension [217] have been used to produce small area estimates of poverty indicators, such as HCR and mean HEI, using EU-SILC income data and auxiliary data coming from the population censuses.

Here we show an application where we use the Fay-Herriot model to produce estimates for HCR and mean HEI in the LLs of Tuscany. The estimates are obtained through area-level data coming from the EU-SILC survey 2011 and from Big Data on human mobility as covariate information. We use a modified version of the Fay-Herriot model proposed by Ybarra e Lohr [207] to allow for measurement error in the auxiliary variables. Ybarra and Lohr [207] assume that for a direct estimator \( y_d \) of the target variable \( Y_d \), the expected value is \( E[y_d] = Y_d \). Moreover, they assume that the auxiliary data source provides an estimator \( \hat{X}_d \) of a vector \( X_d \) of population characteristics, where the estimator \( \hat{X}_d \) has mean squared error \( MSE(\hat{X}_d) = C_d \). Ybarra and Lohr [207] show that when the auxiliary variables are measured with error, the traditional Fay and Herriot [206] estimator is worse than the direct estimator in terms of precision, and the estimated mean squared error of the Fay and Herriot estimator gives a misleading notion of precision.

Suppose that \( X_d \) is the true value of the auxiliary variable in small area \( d \) available for small area estimation. Since \( X_d \) may be measured with error, we substitute an estimator \( \hat{X}_d \) to \( X_d \) and use the following model:

\[
y_d = \hat{X}_d^T \beta + r_d(\hat{X}_d, X_d) + e_d
\] (10.5)
where \( r_d(\hat{X}_d, X_d) = u_d + (X_d - \hat{X}_d)^T \beta \), with \( u_d \sim N(0, \sigma_d^2) \) and random error \( e_d \sim N(0, \phi_d^2) \), with known \( \phi_d \). Ybarra and Lohr [207] assume that \( u_d \) is independent of both \( e_d \) and \( \hat{X}_d \), and that random variables in different small areas are independent. They also assume that \( \hat{X}_d \) and \( y_d \) are independent for each area, that is the case when \( X_d \) and \( Y_d \) are estimated using different data sources. In our application this is the case of Big Data auxiliary variables, while for auxiliary variables coming from EU-SILC this hypothesis is violated. The resulting EBLUP (Empirical Best Linear Unbiased Predictor) estimator derived by Ybarra and Lohr [207] is

\[
\hat{Y}_{dME} = \hat{\gamma}_d y_d + (1 - \hat{\gamma}_d) \hat{X}_d^T \hat{\beta}
\]

where \( \hat{\gamma}_d = (\hat{\sigma}_d^2 + \hat{\beta}^T C_d \hat{\beta}) / (\hat{\sigma}_d^2 + \hat{\beta}^T C_d \hat{\beta} + \psi_d^2) \) and the regression vector \( \beta \) and the variance component \( \sigma_d^2 \) are estimated according to an iterative procedure for the modified least squares as in Cheng and Van Ness [218].

We compute direct estimates of HEI and HCR for the 57 Local Labour Systems (LLSs) in Tuscany, using data from the EU-SILC 2011 available at the LLS level. As covariate information we use data coming from the same survey and Big Data on individuals’ mobility, under the hypothesis that mobility data can be predictive of well-being measures as Chapter 9 suggests. Note that 24 out the 57 LLSs of Tuscany are ‘out of sample areas’ with zero sample size. We describe the mobility of area \( d \) by the mobility entropy \( E(d) \) and the radius of gyration \( r_g(d) \) defined as:

\[
r_g(d) = \frac{1}{V_d} \sum_{v \in d} r_g(v),
\]

where \( V_d \) is the set of vehicles associated to area \( d \) and \( r_g(v) \) is the radius of gyration of vehicle \( v \) defined as in Chapters 5 and 9.

We use two different models to estimate the small area HEIs and HCRs (Table 10.1). To estimate the mean HEI using estimator \( \hat{Y}_{dME} \), let \( X_d \) be the matrix of the auxiliary variables of area \( d \): it contains a constant term, the direct estimates of the proportion of male as the head of the household, the direct estimates of the mean of the squared meters of the house (both from the EU-SILC 2011 survey), and the values of the \( r_g(d) \) from Big Data of human mobility. Let \( C_d \) be the corresponding variance-covariance matrix of the auxiliary variables, with the covariances set to zero, and let \( y_d \) be the direct estimate in area \( d \) of the mean HEI and \( \psi_d \) its standard deviation. For the HCR, the auxiliary variables matrix \( \hat{X}_d \) contains a constant term, the direct estimates of the mean of the age of the head of households (from EU-SILC 2011) and the mobility entropy \( E(d) \) extract from Big Data. Here \( y_d \) is the direct estimate of the HCR in area \( d \). As variance-covariance matrix \( C_d \) we use the estimated variances of the auxiliary variables’ mean estimates, setting the covariances to zero. As we know, Big Data auxiliary variables can be self-selected. However, as shown by Bethlehem [219] the bias due to the self-selection process is related to the correlation between the target variable (mobility index) and the response behaviour (having or not having a GPS). Using the results shown in Pappalardo et al. [14] we argue that in this application this correlation coefficient can be considered very small, and hence the bias due to the self-selection process could be negligible. In fact, Pappalardo et al.


measure | covariates | source
--- | --- | ---
HEI | ratio of male household’s head | EU-SILC 2011
mean size of house (m$^2$) | EU-SILC 2011
radius of gyration $r_g(d)$ | Big Data

Table 10.1: Variables used as covariates

[14] show that the mobility index measured using the sample of cars with GPS is coherent with the mobility registered for all the vehicles in the municipality of Pisa (data derived from traffic sensors spread around the city). Given this evidence, it is reasonable to use the hypothesis of independence between the mobility index and having a GPS, so that we can handle these data as if they were a simple random sample from the population of vehicles. The variances in the $C_d$ matrix are then computed using a simple random sample design variance formula (for finite populations).

An important problem in small area estimation is the synthetic prediction for non-sample areas, i.e. areas where there are no sampled units, even if in those areas there are population units with the characteristics of interest. The conventional approach for estimating a small area characteristic, say the mean, is the synthetic estimation [214]:

$$
\hat{Y}_{d,\text{out}} = X_{d,\text{out}}^T \hat{\beta}
$$

where $X_{d,\text{out}}$ is the auxiliary information for the out of sample area $d$ and $\hat{\beta}$ is the vector of estimated coefficients under a small area model. In the application presented here the problem is serious, since there are 24 out of sample areas (42% of the total number of small areas). Moreover the predictor $\hat{Y}_{d,\text{out}} = X_{d,\text{out}}^T \hat{\beta}$ cannot be applied because the EU-SILC auxiliary variables selected in our models are not available for the out of sample areas. Big Data auxiliary variables are instead available for all the areas. A possible synthetic predictor is $\hat{Y}_{d,\text{out}} = \hat{X}_d^T \hat{\beta} + \hat{X}_d^{BD} \hat{\beta}_{BD}$, where $\hat{X}$ is the matrix of the direct estimators of the EU-SILC auxiliary variables at a regional level, $\hat{X}_d^{BD}$ is the value of the Big Data auxiliary information for area $d$ and finally $\hat{\beta}$ and $\hat{\beta}_{BD}$ are the estimated regression coefficients (see Giusti et al. [220] for an example). Thus, using Big Data it is possible to obtain area-specific synthetic estimates for the out of sample areas taking into account the variability between areas that cannot be specified by only basing predictions on the values of $\hat{X}$. This represents one of the major advantages in the use of Big Data sources in small area estimation.

Finally, to estimate the mean squared error of $\hat{Y}_{d,\text{ME}}$ for both sampled and out of sample areas we use a parametric bootstrap approach, since the jackknife approach described in Ybarra and Lohr [207] was too unstable with our data, often producing negative estimates of the mean squared error. In the parametric bootstrap we first estimate $\beta$ and $\sigma^2_\mu$ as shown in [207], then we parametrically generate the errors $u^*_d \sim N(0, \hat{\sigma}^2_\mu)$ and $e^*_d \sim N(0, \phi^2_d)$. Using these random errors and the matrix of auxiliary variables we generate the bootstrap true values $Y^*_d = \hat{X}_d^T \hat{\beta} + \mu^*_d$ and the bootstrap direct estimates $y^*_d = Y^*_d + e^*_d$. In the next step, we generate a bootstrap matrix of auxiliary variables with errors $X^*_d = \hat{X}_d + \epsilon_d$, where $\epsilon_d \sim N_p(0, C^*_d)$ with $N_p$ indicating a multivariate normal of dimension $p$. Using
10.3. BIG DATA AS COVARIATES IN AREA LEVEL MODELS

Figure 10.2: Estimates of the HEI (left) and of HCR (right) for the LLSs in Tuscany. Small area estimates based on EU-SILC 2011 and the Octo mobility dataset. Out of sample areas are estimated using a synthetic estimator.

equation 10.6 with \( \hat{X}_d^* \) and \( \hat{y}_d^* \) we obtain a bootstrap estimate \( \hat{Y}_{dME}^* \) of \( Y_d \). Repeating this process \( B \) times to obtain \( B \) values of \( \hat{Y}_{dME}^{*b} \) and \( Y_{d}^{*b} \), where \( b = 1, \ldots, B \), the bootstrap mean squared error estimator of \( \hat{Y}_{dME} \) is

\[
\sum_{b=1}^{B} (\hat{Y}_{dME}^{*b} - Y_{d}^{*b})^2.
\]

We checked the performance of this bootstrap mean squared error estimator with a small simulation following the setting used in Ybarra and Lohr [207] in their simulation study. The bootstrap scheme works properly, showing an expected slight underestimation of the real (i.e. Monte Carlo) mean squared error. As an alternative to the bootstrap, for the out of sample areas we predicted the \( \psi_d \) values using a linear model based on the same variables used in the estimation process [221]. This method is feasible given that data coming from Big Data sources are available for all the small areas. Results for the means of HEI and HCR are mapped in Figure 10.2, where the estimates refer to the LLSs showing interesting intra-regional differences that would be lost if the scope of the analysis was limited to the regional level.

What is even more important is that under both models we obtained a remarkable gain in terms of precision with respect to the direct estimates. Even if this gain is marginally overestimated because the bootstrap mean squared error of \( \hat{Y}_{dME} \) underestimates the real mean squared error, the gain in precision is anyway evident. In Figure 10.3 we show a comparison between the bootstrap mean squared error estimates of \( \hat{Y}_{dME} \) and the mean squared error estimates of the direct estimates \( y_d \), both for HEI and HCR. Since the mean squared errors for the direct estimators are available only for the sampled areas, we report these ratios for the 33 sampled areas only. The gain in precision is observed for all the areas. In most of the areas the gain in precision is about 5%-20% for HEI and
Figure 10.3: The plot on the right shows the mean squared error estimates of the small area versus the direct estimates of the mean HEI. The plot on the left shows the same for the HCR estimates. Results are reported for the 33 LLSs sampled in the EU-SILC 2011 survey for Tuscany.

10%-40% for HCR. In some areas the gain is more than 50%. However, the mean squared error estimator of the small area estimator $\hat{Y}_{dME}$ should be treated carefully, due to its observed underestimation. These first results, showing an evidence of the gain in precision obtained combining survey data with Big Data, encourage further research on this topic.

10.4 Discussion

In this chapter we used indicators from Big Data sources in the study of poverty and living conditions finding that they have much to offer when combined with small area methods. Big Data, indeed, can be much faster in providing auxiliary variables for small area models than official data sources. Even when individual level data are not available due to privacy restrictions, area level summaries can provide useful covariates for area level models. Instead of using a direct survey estimate, we can generalize the Fay-Herriot model [206] to include covariates and obtain an indirect estimate of the variable of interest. Covariates based on social media, GPS trajectories, mobile phone records or other Big Data sources may augment or replace traditional auxiliary information for a wide variety of poverty and living conditions indicators. The advantage of these types of covariates is that they are often readily available and provide significant information on several socio-demographic variables. The disadvantage is that they require a more complex small area estimation model which considers measurement error as a way to handle functional covariates.

Our evidence of correlation between local poverty estimates obtained by surveys and independent estimates from Big Data sources is encouraging. However, while data quality of survey estimates has been widely studied and discussed, the quality of Big Data sources
has not been thoroughly considered. The usage of Big Data could generate a self-selection problem in relation to the population units, which could have severe effects and lead to biased final estimates. A way to remove the self-selection bias is to use quality survey data to check the differences in the distributions of common variables between Big Data and survey data. These differences can be used to compute weights that allow the reduction of bias due to self-selection of the Big Data. The main issue here is the availability of unit-level information from Big Data sources, due to confidentiality problems. Further methodological and experimental work are required to provide a solution to the problem.

On the one hand Big Data represents an incredible and huge source of data on social complexity and human behaviour, even though they do not ensure representation of the population. On the other hand survey data are high quality sources of data representative of the population of interest, but they are also expensive in terms of money and time to be collected properly. The interaction and integration between these two sources of data is an important challenge for research in statistics and data science. It is also extremely important to develop sound and effective statistical methodology in order to accommodate this abundantly rich class of Big Data resources [222].

Provided that statistics and social mining mature the ability to discover knowledge from these data, we envision that scientific research will be revolutionised by this new wave. Big Data and social mining are providing statistical agencies with novel means for measuring and monitoring our societies more precisely, continuously, and everywhere. The Big Data source by its nature is hence candidate to be one pillar of a statistical system that produces data using social network measures, as proxies of socio-economic indicators of poverty, well-being and progress.
Chapter 11

Conclusions and Future Works

In this thesis, we have studied different dimensions of human behavior through a data-driven approach which combines techniques from network science and data mining. We first presented salient events in the history of such fields, and then described our social microscope, the Big Data we used to observe, understand and describe human behavior. Human mobility is the main topic we investigated, focusing on three aspects: the patterns of car travels and their difference with general mobility; the returners/explorers dichotomy and its consequences; the ABC classifier, a model able to recognize human activities based on people’s movements. We then moved our interest from individuals to interactions, entering the domain of social network analysis. Here, we approached the challenging problem of community detection in dynamic social networks and the fascinating process of musical tastes diffusion over an online social network. Finally, we studied the interplay between human behavior and economic development in France and Tuscany, uncovering interesting relations between metrics of mobility and sociality extracted from Big Data and several socio-economic indicators at municipality and province level. Many research directions open up from the results summarized in the present thesis. Here, we want to revise the most promising ones, tracking possible future works and scenarios.

As we saw in Chapter 5 despite the global heterogeneity observed in human mobility, people can be profiled into returners and explorers, according to their recurrent mobility patterns. Explorers’ mobility cannot be reduced to a few important locations, resulting in a higher invasion diffusion threshold: explorers are more effective at spreading viruses, ideas or information than returners. Due to such results, a question naturally arises: can we exploit the dichotomy returners/explorers to prevent or limit an outbreak? The preliminary results presented in Chapter 5 suggest that a viral disease spread faster when carried by an explorer. To further investigate this hypothesis we plan to implement a SIR-like model [187, 223, 224] using the standard compartmentalization in which individuals can exist only in one of some discrete states, which generally are susceptible (S), infected (I), and permanently recovered (R). The dynamics of individuals based on travels between locations can be described by the Octo dataset (Section 3.1) or the GSM datasets (Section 3.2). We plan to investigate and model three different diffusion scenarios, corresponding to three possible viral outbreaks. The first one is an “influenza virus” scenario, where
an individual can be infected and become infectious after a period of latency, recovering after a period of convalescence. This scenario models most of the biological diseases which diffuse every year all over the world generating epidemics or pandemics, including the influenza disease which has been deeply studied in the context of complex systems [174, 225]. The second scenario covers a “computer virus” outbreak where the smart phone of an individual, once infected by other phones, is immediately infectious, recovering after a period thanks to anti-virus programs. Due to the growing world-wide diffusion of last generation smart phones with operating systems, the diffusion of computer viruses has been studied in recent years [226, 227]. Finally, we will investigate a “zombie virus” scenario, modeling the apocalyptic scenario where an individual can be infected but no recovering is possible, as the horror and science fiction scenario depicted in zombie movies [228].

The patterns of diffusion for musical items can be further investigated as well. As we showed in Chapter 8, starting from musical innovators/leaders, the listenings diffuse in three possible ways: width, depth, strength. Such three dimensions, however, model the spreading of music at a local level, i.e. the level of nodes and their surroundings. What about the patterns of diffusion of a new song, album, artist or genre over the entire social network? Can we model the spreading of musical tastes at a global level? The diffusion speed of a song, the amount of people it reaches, and the role of tastes in the diffusion process are interesting aspects from the point of view, for example, of music bands or music producers who aim to diffuse the product over the population. Here, we plan to implement several increasingly sophisticated models able to capture more and more aspects about the musical diffusion process. In a very baseline model, we do not consider leaders’ features and musical genre preferences, assigning the same two probabilities to each node: a probability to be initially infected (a musical leader); and a probability to infect its neighbors. Starting from the baseline, we can implement other models by adjusting the parameters according to the local network structure of a node and its genre preferences. The probability of a node to be a musical leader will depend on its network characteristics, if it is found that leaders show typical network features. The virulence of a node, i.e. its ability to infect other nodes, will depend both on his genre preferences and on the musical genre of the song/album/artist. The diffusion speed and the size of the infected population generated by a song/album/artist for each model will be compared with the actual speed and coverage of the song/album/artist over the LastFM social network described in Section 3.4, in order to test the capacity of the models to capture and describe the actual diffusion of musical tastes.

Another interesting research direction emerges if we move our focus from individuals to communities, investigating the relation between human mobility, social communities and economic development. In Chapters 9 and 10 we saw that economic well-being is correlated with individual interactions and movements, aggregated at geographical level. Studying the role of social communities means aggregating individuals at a social level, grouping them in densely connected social clusters. Many aspects here are worth investigating: the description of the mobility of communities, such as the definition of measures
which describe the volume and diversity of movements performed by individuals within a community; the predictive power of communities’ network features with respect to economic well-being, with the definition of a sociometer defined over social groups instead of individuals; the influence of individual mobility in shaping the network structure of social communities.

Besides human mobility and social networks, the striking proliferation of Big Data is affecting many other interesting aspects of our society. Sports data, for example, are now attracting growing interest. The perspective of securing a competitive advantage versus their peers is driving major sports organization to collect and analyze more and more data on their athletes: individual player performance, coaching or managerial decisions, game-based events, and the list goes on. The recent emergence of the so called online social fitness constitutes a good proxy to study the patterns underlying success in sports. Indeed, through these platforms users can collect, monitor and share with friends their sports performance, diet, and even burned calories. Since such data are generally made available by public APIs, this allows researchers to download and analyze information about thousands of professional and amateur sportsmen, giving an unprecedented opportunity to answer very fascinating questions: What are the main factors that shape sports performance? What are the characteristics that distinguish successful sportsmen? In our lab, we conducted a preliminary study on a sample of 30,000 cyclists downloaded via APIs from the social fitness platform Strava.com [229], downloading information about each cycling workout: average speed, duration of ride, cyclists heart bit rate and power. From the data we derived two basic metrics: a measure of training effort, that is how much a cyclist struggled during the workout; and a measure of training performance indicating the results achieved during the training. Through a dynamic temporal analysis of workout characteristics, we split the dataset in three subpopulations of cyclists presenting similar training behaviors. By analyzing the way people trained in the clusters we discovered that athletes that better improve their performance follow precise training patterns, with alternation of high stress peaks and rest periods. This is a confirmation of the overcompensation theory, the main medical sports theory applied to aerobic sports. Starting from these preliminary results, we plan to address another fascinating issue: What is the role of social influence in the sports success? Online social fitness platforms also provide a social dimension, allowing users to interact through post or comments. As we discovered, the training strategy is a key aspect in the success. Interactions with successful sportsmen could provide precious information and help less good sportsmen to improve their performance: may the success in sports exhibit some amount of “contagion” in social sense?
Bibliography


11.0. BIBLIOGRAPHY


11.0. BIBLIOGRAPHY


