



Universidade Federal do Ceará
Departamento de Computação
Curso de Ciência da Computação

Igo Ramalho Brilhante

Mobility data analysis under a complex network perspective:
from interactions among trajectories to movements among points
of interest

Fortaleza, Ceará

2012

Igo Ramalho Brilhante

**Mobility data analysis under a complex network perspective:
from interactions among trajectories to movements among points
of interest**

Dissertação submetida à Coordenação do Curso de Pós-Graduação em Ciência da Computação da Universidade Federal do Ceará, como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Área: Ciência da Computação

Orientador: Prof. Dr. José Antonio Fernandes de Macêdo

Coorientadora: Profa. Dra. Chiara Renso

Fortaleza, Ceará

2012

A000z Brilhante, R. I.
 Mobility data analysis under a complex network perspective: from interactions among trajectories to movements among points of interest / Igo Ramalho Brilhante. 2012.
 104p.;il. color. enc.
 Orientador: Prof. Dr. José Antonio Fernandes de Macêdo
 Coorientadora: Profa. Dra. Chiara Renso
 Dissertação(Ciência da Computação) - Universidade Federal do Ceará, Departamento de Computação, Fortaleza, 2012.
 1. 2. 3. I. Prof. Dr. José Antonio Fernandes de Macêdo(Orient.) II. Universidade Federal do Ceará- Ciência da Computação(Mestrado) III. Mestre

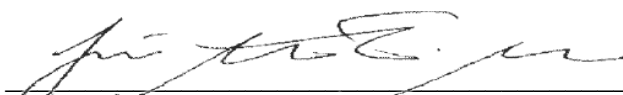
CDD:000.0

***MOBILITY DATA ANALYSIS UNDER A COMPLEX NETWORK
PERSPECTIVE: FROM INTERACTIONS AMONG TRAJECTORIES TO
MOVEMENTS AMONG POINTS OF INTEREST***

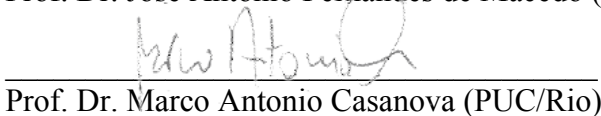
Igo Ramalho Brilhante

Dissertação apresentada ao Curso de Mestrado em Ciência da Computação da Universidade Federal do Ceará, como parte dos Requisitos para a obtenção do Grau de Mestre em Ciência da Computação do aluno Igo Ramalho Brilhante

Composição da Banca Examinadora:



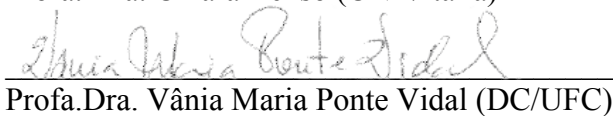
Prof. Dr. José Antônio Fernandes de Macedo (Presidente) (DC/UFC)



Prof. Dr. Marco Antonio Casanova (PUC/Rio)



Profa. Dra. Chiara Renso (CNR/Itália)



Profa. Dra. Vânia Maria Ponte Vidal (DC/UFC)

Aprovada em 10 de fevereiro de 2012

Aos meus Pais.

Acknowledgments

“O único lugar onde o sucesso vem antes do trabalho é no dicionário.”

(Albert Einstein)

Abstract

The explosion of personal positioning devices like GPS-enabled smartphones has enabled the collection and storage of a huge amount of positioning data in the form of *trajectories*. Thereby, trajectory data have brought many research challenges in the process of recovery, storage and knowledge discovery in mobility as well as new applications to support our society in mobility terms.

Other research area that has been receiving great attention nowadays is the area of complex network or *science of networks*. Complex network is the first approach to model complex system that are present in the real world, such as economic markets, the Internet, World Wide Web and disease spreading to name a few. It has been applied in different field, like Computer Science, Biology and Physics. Therefore, complex networks have demonstrated a great potential to investigate the behavior of complex systems through their entities and the relationships that exist among them.

The present dissertation, therefore, aims at exploiting approaches to analyze mobility data using a perspective of complex networks. The first exploited approach stands for the trajectories as the main entities of the networks connecting each other through a similarity function. The second, in turn, focuses on points of interest that are visited by people, which perform some activities in these points. In addition, this dissertation also exploits the proposed methodologies in order to develop a software tool to support users in mobility analysis using complex network techniques.

Keywords: Complex network. Trajectory. Mobility.

List of Figures

Figura 2.1 Example of a trajectory sample whose identifier is 223: (a) interpolation between the points; (b) “raw” data of the trajectory, the triples $\langle traj_{id}, x_i, y_i, t_i \rangle$ 25

Figura 2.2 (left) trajectories and (right) trajectories with geographic information [Alvares et al., 2007]

Figura 2.3 (a) a trajectory moving from left to right and (b) the same trajectory, but with semantic location associated with it. Besides of considering the passing through a location, the time spent on the place can be considered 26

Figura 2.4 Example of application with three candidate stops ($R_{C_1}, R_{C_2}, R_{C_3}$). Imagine a trajectory sample running through from left to right and t_0, \dots, t_{18} are the time points of T . First, T is outside any candidate stop, so it starts a move. Then T enters R_{C_1} at time t_1 such that the duration is long enough, $t_6 - t_1 \geq \Delta_C$, then (R_{C_1}, t_1, t_6) is the first stop. When the trajectory enters R_{C_2} , it do not spend time enough inside that candidate, so it is not a stop. We then have a move until T enters R_{C_3} , which fulfills the requests to be a stop, and so (R_{C_1}, t_1, t_6) is the second stop of T . The trajectory ends with a move [Alvares et al., 2007a] 29

Figura 3.4 Example of degree distribution: (a) node degrees; (b) degree distribution; (c) degree distribution plot 36

Figura 3.5	A random graph $\mathcal{G}_{n,p}$ generated with $n = 100$ and $p = 0.05$. Its average shortest path is 2.271 and its clustering coefficient is 0.099 (145 triangles)	40
Figura 3.6	The small world model varying p . This figure illustrates well the network structure by varying the probability p : $p = 0$ generates a regular network, while $p = 1$ generates a random network [Watts & Strogatz, 1998]	41
Figura 3.7	A WS network shaping a ring generated with $n = 100$ and $p = 0.2$ and $k = 10$. Its average shortest path is 2.548 and its clustering coefficient is 0.419 (624 triangles)	42
Figura 3.8	A BA network generated with $n = 100$ and $m_0 = 4$ and $k = 2$. The largest and red nodes represent the preferential attachments, highly connected nodes, which are more likely to establish new links with other nodes	43
Figura 3.9	An example of community discovery by performing the method proposed in [Blondel et al., 2008]. This algorithm is available in Gephi [Bastian et al., 2009]. This network was built from a user's profile in a social network and, thus, each community represented by a color shows a community of people. For instance, the blue community represents the friends from the university and the red one represents the family ties	44
Figura 4.1	A plotted complex network composed by 36,824 nodes and 306,572 edges generated in our experiments	50
Figura 4.3	Plot showing the number of trajectories for each day of the week	58
Figura 5.1	The building process of places network from one user history: From positional observations in (a) to the user history in (b), the candidates stops in (c). The trajectories set in shown in (d) where a move of duration of 8h30' (thus exceeding 4 hrs) splits the user history into two trajectories. The POI network is depicted in (e)	70
Figura 5.2	The plot of the POIs network generated from our experiments with 77 nodes and 677 edges: nodes represent the composite POIs; and edges represent the movement of users' trajectories between the nodes	73

Figura 5.3	Edge weight distribution of the network P_{oi}^N	74
Figura 5.4	The 109 communities discovered from the POIs network. The edge color identify the different communities	75
Figura 5.5	Community size distribution considering number of edges. Many communities are formed by a few edges, whereas a few communities are composed by a higher number of edges	76
Figura 5.6	Cumulative Distribution of <i>Compactness</i> . $P(Compactness \leq k)$ indicates the probability that <i>Compactness</i> takes on a value less than or equal to k	76
Figura 5.7	Correlation	77
Figura 5.8	The selected communities: the three largest communities by the number of edges are 72, 20, 25; and the two communities 76 and 63 act like a "bridge" between them characterizing the movement between two regions of the city	78
Figura 5.9	Temporal analysis of the network P_{oi}^N and communities 72, 20 and 25 showed in Figure 5.8	80
Figura 5.10	Communities to illustrate the measure <i>Compactness</i> considering different degrees of <i>compactness</i> : communities 104, 86 and 43 are less compact communities; 6 and 13 are more compact communities	81
Figura 6.2	Statistical analysis performed by M-Atlas: movement distribution (top), cumulative lengths distribution (left) and density of length over speed (right) [Trasarti et al., 2010]	87
Figura 6.3	Some functionality of Cytoscape. (a) Network layout algorithms. (b) Data attribute-to-visual mapping to control the appearance of their associated nodes and edges and data types as well. (c) Attribute of the selected nodes and edges. (d) Annotations are transferred to node and edge attributes by choosing the desired ontology and hierarchical level from a list of those available [Shannon et al., 2003]	88

Figura 6.4	NWB interface with the menu to compute network analysis on different types of networks, such as node degree, degree distribution, clustering coefficient and community detection	89
Figura 6.5	Interface of Gephi showing the <i>Data Laboratory</i> with <i>Data Table</i> , nodes and edges with their properties, and a graph visualization window [Bastian et al., 2009]	
Figura 6.6	MOBNET architecture consisting of three main layers: view engine, manager engine and storage	91
Figura 6.7	MOBNET interface: in the menu we can choose between <i>trajectory network</i> and <i>poi network</i> to build and visualize networks; <i>Network Manager</i> performs activities on the built networks	92
Figura 6.8	Building a <i>trajectory network</i> where the nodes are the trajectories from the dataset	93
Figura 6.9	Visualizing the nodes (trajectories) of a built <i>trajectory network</i> : this is a trajectory network with four nodes representing the four trajectories depicted on the map	93
Figura 6.10	Building a <i>poi network</i> where the nodes represent the points of interest, and the edges correspond to the movement of the trajectories between the points	94
Figura 6.11	Visualizing the discovered communities of a built <i>poi network</i>	95

List of Tables

Tabela 4.1	Information about trajectory dataset	54
Tabela 4.2	Four different parameter combinations	55
Tabela 4.3	Experiment 1: frequency of 3, spatial threshold of 0.3 km and temporal threshold of 30 minutes	56
Tabela 4.4	Experiment 2: frequency of 3, spatial threshold of 0.3 km and temporal threshold of 15 minutes	56
Tabela 4.5	Experiment 3: frequency of AVG, spatial threshold of 0.3 km and temporal threshold of 15 minutes	56
Tabela 4.6	Experiment 4: frequency of AVG, spatial threshold of 0.3 km and temporal threshold of 30 minutes	57
Tabela 4.7	Random Graph - Erdős and Rényi	57
Tabela 5.1	Degree correlation of P_{oi}^N	75
Tabela 5.2	Similarities between the communities in Figure 5.8	79

Tabela 5.3 Compactness of the some communities of Figures 5.8 and 5.10	82
--	-------	----

List of Algorithms

4.1	Compare two trajectories by their positions	51
4.2	Trajectory Network Generator	52
5.1	Points of Interest Network Builder	71

Contents

1	Introduction	18
1.1	Motivation.....	20
1.2	Contributions	21
1.3	Publications	21
1.4	Organization	22
2	Mobility Analysis	23
2.1	Preliminaries.....	24
2.1.1	Trajectory.....	24
2.1.2	Semantic Trajectory	25
2.2	Stops and Moves	27
2.3	Identifying Stops and Moves	27
2.3.1	SMoT (Stops and Moves of Trajectories)	28
2.3.2	CB-SMoT (Clustering-Based SMoT)	29
2.3.3	Stay Point Detection	30
2.4	Summary	31
3	Complex Network	32
3.1	Preliminaries.....	34

3.1.1	Concepts and Basic Definitions	34
3.1.2	Properties of Complex Networks	35
3.1.2.1	Degree and Degree Distribution	35
3.1.2.2	Clustering Coefficient or Transitivity	35
3.1.2.3	Shortest Path Length, Betweenness Centrality and Close- ness Centrality	37
3.1.2.4	Power Law Distribution	38
3.1.2.5	The Small-World Effect	39
3.2	Network Models	39
3.2.1	Random Graphs	40
3.2.2	Small-World Model	41
3.2.3	Models of Network Growth	41
3.3	Community Discovery	43
3.4	Summary	45
4	Trajectory Analysis using Complex Network	46
4.1	Basic Concepts and Related Work	47
4.1.1	Basic Definitions	47
4.1.2	Mobility Analysis	48
4.1.3	Complex Network Analysis	49
4.2	Complex Network and Trajectory	49
4.2.0.1	Step 1 - Build Trajectory Network	51
4.2.0.2	Step 2 - Analyze Trajectory Network Features	53
4.2.0.3	Step 3 - Identify relevant trajectories within trajectory network	53
4.3	Experiments	53
4.3.1	Experiments on the vehicles' movements in Milan city	54
4.3.2	Computed Trajectory Network Features	55
4.4	Conclusion	60
5	COMETOGETHER: discovering communities of places in mobility data	63
5.1	Related Work	64

5.2	Background	66
5.3	Problem Definition and Methodology	67
5.3.1	Building the Network	68
5.3.2	Communities of Points of Interests	70
5.4	Case Study	72
5.4.1	POI Network Characteristics	73
5.4.2	Communities Analysis	74
5.4.3	Large Communities	78
5.4.4	Compact Communities	81
5.5	Conclusion	82
6	MOBNET: a software tool to analyze mobility through complex network	83
6.1	Tools in Mobility Analysis	84
6.1.1	Weka-STPM	84
6.1.2	M-Atlas	86
6.2	Tools in Complex Networks	86
6.2.1	Cytoscape	87
6.2.2	Network Workbench	88
6.2.3	Gephi	89
6.3	MobNet	90
6.3.1	Overview	90
6.3.2	Trajectory Network	92
6.3.3	Points of Interest Network	94
6.4	Conclusion	95
7	Conclusions	96
7.1	Conclusion	96
7.2	Future Works	97
	References	99

CHAPTER 1

Introduction

The advent of tools for automated data collection tends to produce large amount of data stored in electronic format. This tremendous growth of data has opened the possibility of extracting useful information and knowledge from the data. In addition, the explosion of personal positioning devices like GPS-enabled smartphones has enabled the collection and storage of a huge amount of positioning data in the form of *trajectories*, i. e. spatio-temporal points identifying the positions of a *moving object*. Trajectory data have brought many research challenges in the process of recovery, storage and knowledge discovery in mobility area. Therefore, many opportunities in this area as well as new applications to support our society in mobility terms have arisen. Examples are numerous: tourist systems that offer meaningful information like recommending interesting places; systems to support the traffic managers in order to distribute the traffic flow more efficiently in the road network, thus prevent the unwanted traffic jams; techniques to study the migration of animals to find patterns of displacements, such as groups of animals that flock together; tools for companies that deliver goods or service to improve the care of their clients by avoiding and preventing possible delays.

Other challenges come from the moving object *interaction*, i. e., how and how much they interact to each other or how a moving object can influence a group (or vice-versa). For example, how can friends of mine influence the place I am used to visit? And what about our feelings? How do they characterize the way we drive our cars for instance? In addition, what about the places visited by people? How do they relate to each other by

considering the movements of people among them? People live in an environment where they move from one place to another, where “places” are not only “static geographical objects”, but they are also part of people’s lives. Thereby, a two-way relationship can be regarded between how the movements of people are affected by the location of places of interest, and how the places themselves are characterized and connected by the mobility of people. These are intriguing issues that give the intuition of the complexity in mobility data and the complexity of performing useful analysis on them, in which involve different entities and relations, creating complex systems of interactions that may be tough to be understood and analyzed.

Mobility, or trajectory data, however, are not the only complex systems in our life. In the reality, we are surrounded by complex systems: economic markets, the Internet, World Wide Web, disease spreading and human beings are complex systems. Because of this complexity and the fact that nothing happens in isolation, a new science called the *Science of Networks* has recently emerged. Most events and phenomena are connected, caused by, and interacting with a huge number of pieces of a complex universal puzzle. For instance, the spread of disease that could start in a city, but rapidly spread over the world causing a worldwide problem like the 2009 flu pandemic; unforeseeable combination of small mistakes in the power electric network that plugs a city into darkness, like happened to New York in 1977 leaving nine million of inhabitants in a mayhem of riots, plundering, and widespread panic. We have come to see that we live in a small world, where everything is linked to everything else and we are witnessing a revolution in the making as scientists from all different disciplines discover that complexity has a strict architecture. We have come to grasp the importance of the networks [Barabási, 2002].

Science of networks is the science of the real world: people, friendship, rumors, diseases, fad, firms, financial crisis. Two famous examples of networks are the Internet and the social networks. Studies on Internet network have focused on its largeness, its points of weaknesses or points susceptible to failure, and the understanding of how it evolves over time. Social networks in turn were firstly target of sociologists to study the relationships of small groups of people. The explosion of the Internet and social media (Facebook, Orkut, Youtube, etc) has enabled a number of analysis in social networks not only of small groups, but large groups in global scale. Sociologists are not the only ones interested in social networks since enterprises are too. Companies have studied how their employees are connected to each other forming a structure of network in which some employees carry important roles for the company, such as those employees that receive much confidence from the others, or those ones that can propagate new ideas or a new philosophy for the others. Therefore, networks model entities (people, web pages, cells, routers, etc) by some relationship among them (friendship, linkage, chemistry reactions, package transference, etc) in order to comprehend how the entities relate to each other and the behavior of the complex system as a whole.

In front of these two research areas, we aim at using network techniques to analyze mobility data to take advantage of the networks as a complementary view in order to provide an innovative perspective in analyzing mobility data. This chapter is

organized as follows. Section 1.1 presents the motivation, while Section 1.2 shows the contributions of this master dissertation. Next, Section 1.3 presents the publications originated from this master dissertation, and Section 1.4 presents the organization of the chapters remaining.

1.1 Motivation

The emerging of the network science and the progressive refinement of analysis techniques together with the high availability of complex mobility data have brought new opportunities to analyze mobility data under a perspective of complex networks. Modeling trajectory data as a network can offer another way to explain the interaction among the moving objects and the influence they have with each other. Furthermore, the science of networks also enables us to investigate the interaction among places (cities, regions, neighborhoods, points of interests) according to the displacements of the moving objects that visit them.

The challenging idea introduced in this master dissertation is to use complex network techniques to analyze mobility data. Mobility research area offers different methods, for example, to process trajectory data in order to discover mobility patterns, to measure similarity among trajectories, to name a few. On the other hand, the science of networks enables us to analyze the interaction among the entities, how they are connected to each other and how the system behaves as a whole. Therefore, each research area can contribute with complementary methods and techniques. Mobility area defines the entities (trajectories, places, etc) and relationships (similarities among trajectories, common visited places by the trajectories, etc). Science of networks in turn gives a global view of how those entities, based on a relationship, are organized and, consequently, which affects this structure may present. Indeed, the main motivation in network science comes from its capability to understand the relationships among entities and how these relationships may affect the system as whole, in our case, how the moving objects relate to each other or how places relate to each other according to the movements among them.

Therefore, two approaches are investigated in this master dissertation: in the first one, we consider the trajectories as the main entities in order to build a network constituted of trajectories representing the nodes, and some relationship between the trajectories to establish the edges between them; in the second approach, on the other hand, we consider the places visited by the trajectories as the principal entities to represent the nodes, and the movements of the trajectories between the places to correspond to the edges.

1.2 Contributions

This dissertation presents five contributions. The specific contributions are present in Chapter 4, 5 and 6 as follows.

- **Chapter 4** details two contributions. The first contribution is a method for devising a complex network from a trajectory dataset, called **trajectory network**. The aim of this method is to define specific steps for processing trajectory data in order to build and analyze the trajectory network. The second contribution is an algorithm for building a trajectory network given a trajectory dataset (set of spatio-temporal points);
- **Chapter 5** contributions are two fold. First, we propose a methodology for building a complex network combining Points of Interests (POIs) and traces of people movements, from which we build communities of POIs. Second, we apply this methodology in a real case study where trajectories are collected from private cars traveling in a city and Points of Interest are downloaded from the Web. We found different kinds of communities (e.g. *compact* where the movements are mainly inside the community, or *bridge* where the movements tend to connect two other communities);
- **Chapter 6** presents a software tool entitled MOBNET to analyze mobility data by complex network techniques according to the proposed methodologies present in Chapter 4 and 5.

1.3 Publications

We have published the following paper:

- [Brilhante et al., 2011] Igo Ramalho Brilhante, Jose Antonio Fernandes de Macedo, Chiara Renso, and Marco Antonio Casanova. 2011. **Trajectory data analysis using complex networks**. In Proceedings of the 15th Symposium on International Database Engineering & Applications (IDEAS '11). ACM, New York, NY, USA, 17-25,

and the following work was submitted and is under review:

- Igo Ramalho Brilhante, Roberto Trasarti, Michele Berlingerio, Chiara Renso, Jose Antonio Fernandes de Macedo and Marco Antonio Casanova. 2012. **ComeTogether: discovering communities of places in mobility data**. (Submitted to 13th International Conference on Mobile Data Management).

1.4 Organization

The remaining chapters are organized as follow.

Chapter 2 presents basic concepts in mobility analysis, such as trajectory definition, and algorithms for identifying stops and moves of trajectories; Chapter 3 introduces the concepts in complex network area, including basic concepts from graph theory, network properties, network models and community discovery. Chapter 4 and 5 present an introduction, the related works, the proposed methodology, experiments or a case study, and a conclusion. Chapter 4 presents the first approach where the trajectories correspond to the nodes in the networks and the edges are created between them with the help of a given function. Chapter 5, in turn, presents the second approach where the nodes are points of interest visited by users' trajectories, and the edges represent the movements of the trajectories between the places. Chapter 6 presents software tools in mobility and network analysis, and a developed software tool to support mobility analysis by complex network techniques. Finally, Chapter 7 summarizes the conclusions and future works.

CHAPTER 2

Mobility Analysis

The explosion of personal positioning devices, like GPS-enabled smartphones or vehicles tracking systems, have enabled the collection and storing of a huge amount of positioning data. People wearing these devices leave traces of their movements in the form of sequences of spatio-temporal positions, called *trajectories*. Trajectories are ubiquitous in the real world, i. e., trajectory are present almost everywhere, from people's movement to movement of animals or vehicles. In front of this context, the availability of trajectory data set has opened new perspectives for a large number of applications, ranging from transportation and logistics to ecology and anthropology, built on the knowledge of movements of objects [Spaccapietra et al., 2008].

Although the management of trajectory data dates back to the 1990s, when the first proposals for moving object databases came out, the challenging approaches towards the analysis and understanding of the movement complexity represented in the users' tracks is being faced only recently [González et al., 2008]. Even more challenging is the aspect of moving object interaction. How and how much do these moving objects interact? How do the *encounters* among moving entities globally characterize the movement of a moving community? Is there a specific law explaining the interactions of moving individuals? Is the movement of people in vehicles (e.g. cars in a road network) differs from people free movement and/or multi transportation trajectories? How do the individual movements of independent entities influence a crowd's movement pattern?

This chapter starts with preliminary definitions about mobility analysis in

Section 2.1. Section 2.2 presents two important features for understanding mobility data, i. e., *stops* and *moves*, and methods for identification of stops and moves.

2.1 Preliminaries

The study of mobility data started from the movements of entities or objects, which are called *moving objects*. Typical examples of moving objects under study include vehicles (cars, planes, ships), persons equipped with personal GPS devices, animals bearing a transmitter and hurricane tracking data from meteorological satellites. These movements are represented in form of spatio-temporal data, *trajectories*, where the spatial part identifies the position on earth, while the temporal part identify the instant when the moving object was at that position.

2.1.1 Trajectory

Trajectory is by definition a spatio-temporal concept. The strike difference among moving objects and non-moving objects refers to the fact that moving objects move to achieve a goal taking a finite amount of time and covering some distance in space. From users' viewpoint, the concept of trajectory is rooted in the evolving position of some object traveling in some space during a given time interval. But while moving may be seen as a characteristic of some objects that differentiates them from non-moving objects (e.g. buildings, roads), the concept of traveling object implies that its movement is intended to fulfill a meaningful goal that requires traveling from one place to another. Traveling for achieving a goal takes a finite amount of time (and covers some distance in space), therefore trajectories are inherently defined by a time interval. This time interval is delimited by the instant when the object starts a travel (t_{begin}) and the instant when the travel terminates (t_{end}). Identifying t_{begin} and t_{end} within the whole time-frame where the object is moving is an application decision, i.e. a user-driven specification. Therefore, [Spaccapietra et al., 2008] defined a trajectory as follows.

Definition 2.1 “A trajectory is the user defined record of the evolution of the position (perceived as a point) of an object that is moving in space during a given time interval in order to achieve a given goal.”

Trajectory: $[t_{begin}, t_{end}] \rightarrow space$.

The basic element of trajectories is a spatio-temporal observation consisting of a triple (ID, Location, Time), where ID is an unique identifier of the individual used throughout all recordings of that individual's movements, Location is a spatial descriptor (such as a coordinate pair, a polygon, a street address), and Time is the time stamp when the individual was at that particular location (such as a clock time in minutes or event time in years) [Spaccapietra et al., 2008]. A trajectory sample then can be regarded as a set \mathcal{T} of triples $\langle traj_{id}, x_i, y_i, t_i \rangle$, where $traj_{id}$ is the unique identifier, x_i, y_i are the spatial

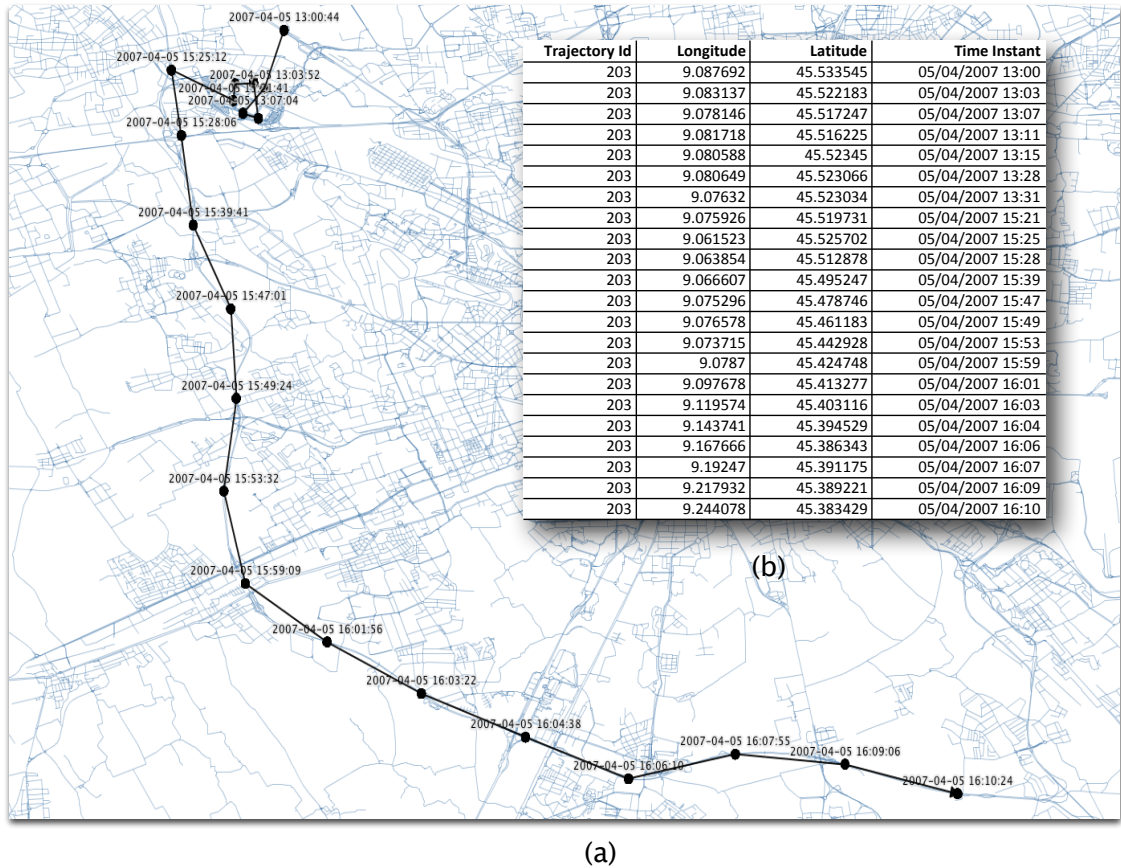


Figure 2.1: Example of a trajectory sample whose identifier is 223: (a) interpolation between the points; (b) “raw” data of the trajectory, the triples $\langle traj_{id}, x_i, y_i, t_i \rangle$

positions (points), such as latitude and longitude, and t_i represents the time instant. A trajectory sample is then defined as follows. Figure 2.1 depicts a trajectory sample whose identifier is 223. Hereinafter, we refer to trajectory sample as trajectory.

Definition 2.2 *Trajectory Sample:* A trajectory sample is a list of space-time points $\{p_0, p_1, \dots, p_n\}$, where $p_i = (x_i, y_i, t_i)$, $x_i, y_i \in \mathbf{R}$, $t_i \in \mathbf{R}^+$ for $i = 0, 1, \dots, n$, and $t_0 < t_1 < t_2 < \dots < t_n$.

2.1.2 Semantic Trajectory

Besides the trajectory as a set of time-stamped points, the trajectories can be semantically represented and they are referred to as *semantic trajectories*. The notion of semantic trajectory has been proposed by [Spaccapietra et al., 2008]. Semantic trajectories are enriched trajectories by geographic information (buildings, tourist places, restaurants, etc) or events (crimes, traffic accident, etc). Figure 2.2 shows trajectories (left) that apparently do not have meaning and the same trajectories enriched with geographic information (right), where we can infer the geographic location (Paris) and the intersection of trajectories with touristic place (e.g. Eiffel tower) and hotels [Alvares et al., 2007b]. Therefore, a trajectory is not represented anymore as a sequence of time stamped points, but as a

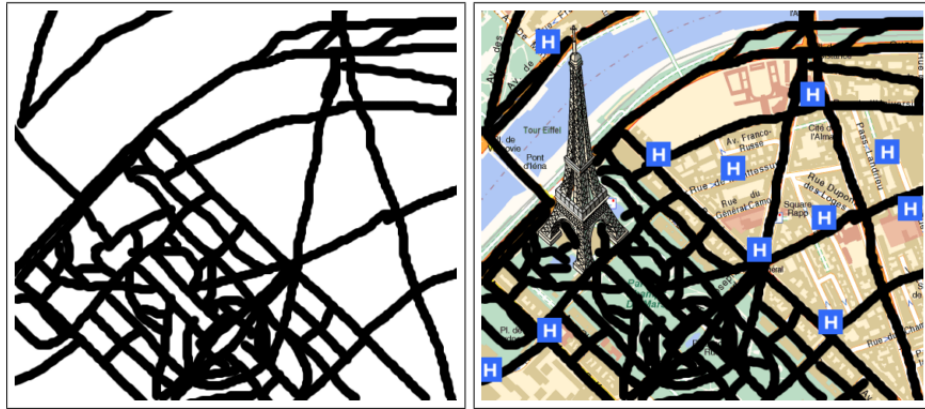


Figure 2.2: (left) trajectories and (right) trajectories with geographic information [Alvares et al., 2007b]

sequence of time stamped semantic location (restaurants, streets, etc). For example, a trajectory $tra_{jid} = 1$, which passed through a restaurant A, a square B and an open mall C, turns to be represented as a sequence $\langle (1, \text{restaurant A}, t_1), (1, \text{square B}, t_2), (1, \text{open mall C}, t_3) \rangle$ in Figure 2.3.

A trajectory passing through places, however, does not necessarily point out these places as important or interesting locations. Depending on the application the temporal aspect may be important, such as the temporal duration of the visit. For example, a stop of few second at the crossroad probably means the vehicle stopped at a traffic light, while a stop of one hour at the same place probably means there was a huge traffic jam, or the person parked the car to go shopping (Figure 2.3). This brings us two important concepts in mobility analysis: *stops* and *moves*; which are presented in the next section.

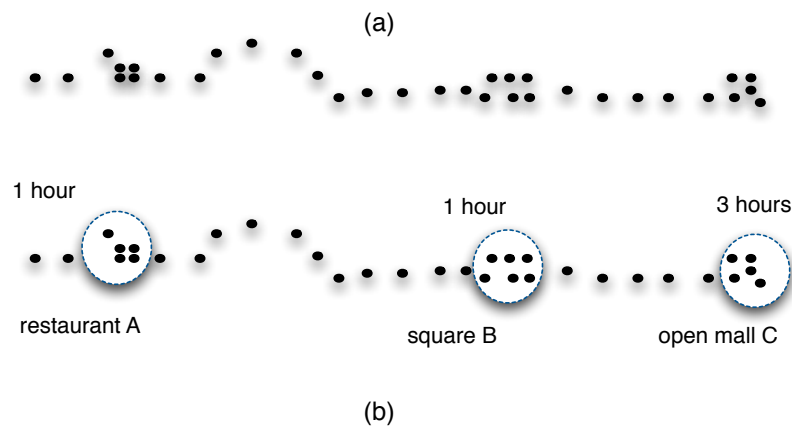


Figure 2.3: (a) a trajectory moving from left to right and (b) the same trajectory, but with semantic location associated with it. Besides of considering the passing through a location, the time spent on the place can be considered

2.2 Stops and Moves

People move every day throughout the city doing various activities like going to work, lunch, recreation, etc. What actually happens is that people, or moving objects in general, move for a while and then stop during a time span until move again. This notion demonstrates two fundamental concepts in mobility, i. e., the concept of *stops* and *moves* proposed by [Spaccapietra et al., 2008].

Intuitively, a *stop* represents a particular moment in a trajectory in which the moving object was kept in a fixed position, or at least with very little spatial displacement, during a time span. A *move* in turn represents the movement between two temporally consecutive stops, or it can be regarded as sub-trajectories where the t_{begin} is the first stop and t_{end} is the last stop: $[t_{begin}, t_{end}]$.

We can then see a trajectory as a sequence of moves connected by stops or a sequence of stops separating the moves. Take as example the salespersons on a business trip that stop at several locations where they planned to meet a customer, or the birds that depart for migration, stop somewhere for some time to feed, they fly again, then another stop to rest, and so on until they reach the final destination.

The identification of stops and moves plays an important role in the construction of semantic trajectories and they can be embedded into the semantic trajectory definition [Rocha et al., 2010, Yan et al., 2011] with aim of performing data mining algorithms to discover the most frequent/sequential patterns [Alvares et al., 2007a]. The raw points of a trajectory are replaced by the stops, or location associated with the found stops, forming a sequence of stops (locations) where there is a move between two temporally stops (locations). In addition, the identification of stops depends on the application. For instance, the stop of salespersons to drink a coffee may be irrelevant for the tracking application of the company, while the stops for meeting a customer are relevant.

2.3 Identifying Stops and Moves

Many algorithms for identification of stops and moves can be found in the literature [Xiao, 2005, Alvares et al., 2007a, Palma et al., 2008, Li et al., 2008, Yan et al., 2010]. We introduce three algorithms in this section: SMO_T (Stops and Moves of Trajectories) proposed by [Alvares et al., 2007a]; CB-SMO_T (Clustering-Based SMO_T) proposed by [Palma et al., 2008]; and Stay Point Detection by [Li et al., 2008]. The two firsts are dependent on an application to identify stops that intersect some geography. The last one identifies stops, called *stay points*, by getting the set of points in which the trajectory spent a duration of time. Those algorithms are presented as follows.

2.3.1 SMoT (Stops and Moves of Trajectories)

SMoT and CB-SMoT are both application-based methods, i. e., they depend on an *application* to find the stops through *candidate stops*. These notions were introduced by [Alvares et al., 2007a] and are presented as follows.

Definition 2.3 *A candidate stop C is a tuple (R_C, Δ_C) , where R_C is a (topologically closed) polygon in \mathbf{R}^2 and Δ_C is a strictly positive real number. The set R_C is called the geometry of the candidate stop and Δ_C is called its minimum time duration.*

Definition 2.4 *An application \mathcal{A} is a finite set $\{C_1, \dots, C_n\}$ of candidate stops with mutually non-overlapping geometries R_{C_1}, \dots, R_{C_n}*

Definition 2.5 *A stop of a trajectory T with respect to an application \mathcal{A} is defined as a tuple (R_{C_k}, t_i, t_{i+l}) such that $\langle (x_i, y_i, t_i), (x_{i+1}, y_{i+1}, t_{i+1}), \dots, (x_{i+l}, y_{i+l}, t_{i+l}) \rangle$ is a sub-trajectory of a trajectory T , there is a (R_{C_k}, Δ_{C_k}) in an application \mathcal{A} such that $\forall j \in [i, i+l] : (x_j, y_j) \in R_{C_k}, |t_{i+l} - t_i| \geq \Delta_{C_k}$ and this sub-trajectory is maximal (with respect to these two conditions).*

A move, in turn, is intuitively defined as follows.

Definition 2.6 *A move of a trajectory T with respect to an application \mathcal{A} is: (i) a maximal contiguous sub-trajectory of T in between two temporally consecutive stops of T ; OR (ii) a maximal contiguous sub-trajectory of T in between the starting point of T and the first stop of T ; OR (iii) a maximal contiguous sub-trajectory of T in between the last stop of T and the last point of T ; OR (iv) the trajectory T itself, if T has no stops.*

In other words, a stop is a polygon R_{C_k} such that part of the trajectory, a sub-trajectory, is within this polygon during a duration of time given by $|t_{i+l} - t_i|$, where t_i means the start of the stop, while t_{i+l} marks the end of the stop. Figure 2.4 shows an example with a trajectory and three candidate stops.

SMoT was proposed by [Alvares et al., 2007a] where stops are interesting spatial locations, also called spatial features, specified according to the application. For instance, traffic lights may be considered as stops in a transportation management application, but not in a tourism application. This algorithm is based on an application in order to verify parts of a trajectory that intersect those candidate stops of an application satisfying a duration of time. The algorithm verifies for each point of a trajectory T if it intersects the geometry of a candidate stop R_C to check then if the duration of the intersection is at least equal to a given threshold Δ_C . In the end, the algorithm returns a set of stops as well as a set of moves.

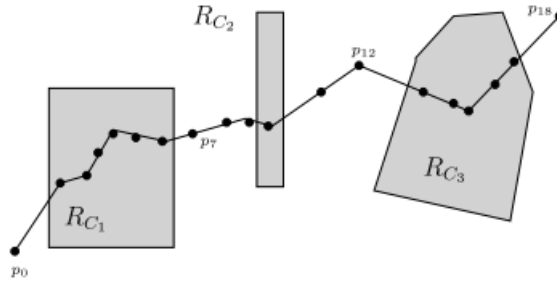


Figure 2.4: Example of application with three candidate stops ($R_{C_1}, R_{C_2}, R_{C_3}$). Imagine a trajectory sample running through from left to right and t_0, \dots, t_{18} are the time points of T . First, T is outside any candidate stop, so it starts a move. Then T enters R_{C_1} at time t_1 such that the duration is long enough, $t_6 - t_1 \geq \Delta_C$, then (R_{C_1}, t_1, t_6) is the first stop. When the trajectory enters R_{C_2} , it does not spend time enough inside that candidate, so it is not a stop. We then have a move until T enters R_{C_3} , which fulfills the requests to be a stop, and so (R_{C_1}, t_1, t_6) is the second stop of T . The trajectory ends with a move [Alvares et al., 2007a]

2.3.2 CB-SMoT (Clustering-Based SMoT)

CB-SMoT, proposed by [Palma et al., 2008], is based on the intuition that parts of a trajectory in which the speed is lower than in other parts of the same trajectory correspond to interesting places and, like SMoT, it is also dependent on an application. In a tourism application the tourists are visiting a new city and, therefore, they spend time visiting important monuments, a museum, going to their hotel and so on. Probably their trajectory has a lower speed around those places than they have in other parts, i. e., when they are moving from a place to another.

The proposed algorithm is two-step. In the first step slower parts of a trajectory, called *potential stops*, are identified by using a variation of the DBSCAN algorithm, well known density-based clustering algorithm [Ester et al., 1996], also proposed by them. This variation is related to the fact they are interested in finding clusters in a single trajectory and in considering time. They have changed some concepts of DBSCAN, where neighborhood should contain only points in the considered trajectory and the distance over the trajectory is taken into account instead of the direct distance between two points.

After applying the variation of DBSCAN to detect potential stops, CB-SMoT identifies in the second step where these potential stops found are located, considering the geography behind the trajectories. Each potential stop is tested with the candidate stops by both intersection and minimal stop duration. In case that a potential stop does not intersect any of the candidate stops, it can still be an interesting place. Then, in order to provide this information to the user, the algorithm labels such places as *unknown stops*. In the end, the algorithm returns a set of stops as well as a set of moves. Figure 2.5 illustrates these concepts [Palma et al., 2008].

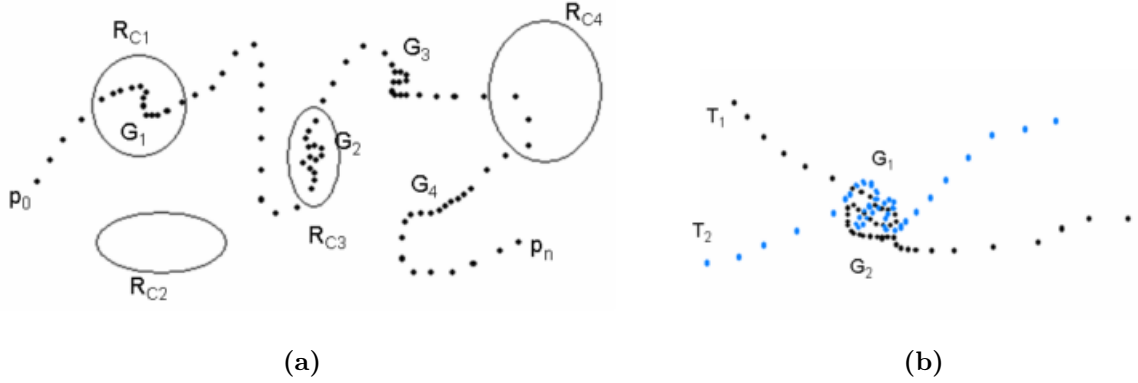


Figure 2.5: (a) a trajectory with four potential stops (G_1, G_2, G_3, G_4) and four candidate stops ($R_{C1}, R_{C2}, R_{C3}, R_{C4}$). G_1 and G_2 are stops intersecting R_{C1} and R_{C3} , respectively, while G_3 and G_4 are unknown stops since they do not intersect any candidate. (b) two trajectories with the same unknown stop. [Palma et al., 2008]

2.3.3 Stay Point Detection

Differently from the others presented so far, the algorithm proposed by [Li et al., 2008] refers to stops as *stay points*. Formally, a stay point s is characterized by a set of consecutive points $P = \langle p_m, p_{m+1}, \dots, p_n \rangle$, where $\forall m < i \leq n, Dist(p_m, p_i) \leq D_r$ (distance threshold), $Dist(p_m, p_{n+1}) > D_r$ and $Int(p_m, p_n) \geq T_r$ (time threshold). Therefore, $s = (x, y, t_a, t_l)$, where

$$x = \frac{\sum_{i=m}^n p_i * x}{|P|}, \quad (2.1)$$

$$y = \frac{\sum_{i=m}^n p_i * y}{|P|} \quad (2.2)$$

x and y are the average coordinates of the collection P , t_a is the user's arriving time on s and t_l is the user's leaving time.

The algorithm then detects temporally consecutive points whose distance is not greater than a given spatial threshold D_r and the duration of time measured by $Int(p_m, p_n)$ satisfies a given minimum time threshold T_r , where $p_m = t_a$ is the begin of the stop and $p_n = t_l$ is the end of the stop. When the set of points are detected, the algorithm computes the average coordinate of that set of points and sets it up as a tuple (p_m, p_n, t_a, t_l) .

The reason in which they detect stay point in such ways lies in two aspects. The first aspect is related to the fact that GPS devices lose satellite signal indoors, what hampers the finding of clusters, since the density of points recorded on such places will not fulfill the conditions to formulate a cluster. The second aspect matches with the fact that some regions, like road crossings, that a trajectory (user) iteratively passes do not carry semantic meanings, but they can be extracted [Li et al., 2008]. Furthermore, the computation of clustering will be extremely heavy as the number of GPS points is quite large compared to that of stay points.

2.4 Summary

This chapter presented basic concepts and definitions related to mobility analysis. First, a trajectory was defined as an evolution spatio-temporal of moving object in order to achieve a goal. Afterwards, it introduced the notion of semantic trajectories, which are important in the process of understanding mobility data once it enriches the raw trajectories semantically with geographic information.

The concepts of stops and moves were discussed as important features to analyze mobility data and to construct semantic trajectories as well. In addition, they can be embedded into semantic trajectories in order to perform data mining algorithms to capture the most frequent/sequential patterns. Finally, this chapter presented some algorithms that can be found in the literature for identifying stops and moves from trajectories.

CHAPTER 3

Complex Network

Research on complex network has been receiving considerable attention from the research community. Indeed, complex network is not a new research domain and preliminary works on this field came up with the birth of graph theory. Graph theory started with the mathematical Leonard Euler and the Königsberg problem. Königsberg is a city on the river Pregel in Prussia, now it corresponds to the city of Kaliningrad in Russia, formed by two island. The city is connected to the island by seven bridges, as showed in Figure 3.1(a). The people of Königsberg amused themselves with mind puzzles, one of which was: “*Can one walk across the seven bridges and never cross the same one twice?*”. In 1736, Euler proved that with the seven bridges such a path does not exist. He not only solved the Königsberg, but his proof originated the immense branch of mathematics known as *graph theory*.

Many consider the proof of Euler’s theorem as the first one in graph theory. Indeed, two centuries passed and graph theory became the basis of *complex networks*. *Complex networks*, or simply *networks*, are ensembles of elements represented by *nodes* (vertices, points) with some interaction between them, i.e., *edges* (links, ties). It is a multidisciplinary area and it has been applied in many different fields, such as Computer Science, Biology, Sociology and Physics. Complex network is the first approach to capture global properties of systems composed by a large number of highly interconnected dynamical units, such as chemical systems, neural systems, social interacting species, the Internet and the World Wide Web [Boccaletti et al., 2006]. On the one hand,

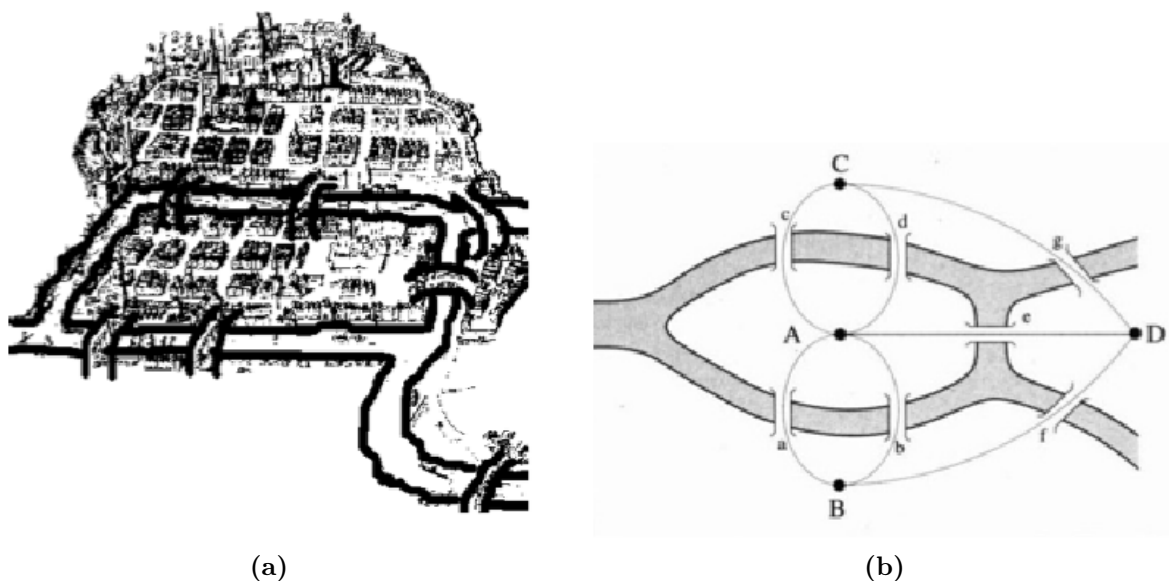


Figure 3.1: (a) Map of Königsberg (b) Map of Königsberg as a graph: nodes are pieces of lands (A, B, C, D), while edges are bridges (a, b, c, d, e, f, g) [Newman, 2006]

scientists have to cope with structural issues, such as characterizing the topology of a complex wiring architecture, revealing the unifying principles that are at the basis of real networks, and developing models to mimic the growth of a network and reproduce its structural properties. On the other hand, many relevant questions arise when studying complex networks' dynamics, such as learning how a large ensemble of dynamical systems that interact through a complex wiring topology can behave collectively.

Some categories of networks can be found in [Newman, 2003]: *social networks* represent groups of people with some interactions between them; *information networks* or “knowledge networks”. An example is the network of citations between academic papers; *technological networks* which are man-made networks designed typically for distribution of some resource, such as the electric power grid; and *biological networks* representing biological systems, such as metabolic pathways networks.

This chapter presents preliminary concepts in graph theory that are used in complex networks as well as some properties of complex network in Section 3.1. Next, Section 3.2 presents network models which aim at generating parenthetical networks in order to generate networks with specific properties. To conclude, Section 3.3 introduces some concepts and methods about community discovery, which is a branch of complex network research.

3.1 Preliminaries

3.1.1 Concepts and Basic Definitions

A network is formally defined as an undirected (directed) graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ where $\mathcal{V} = \{v_1, v_2, v_3, \dots, v_n\}$ is a set of nodes (vertices, points) and $\mathcal{E} = \{e_1, e_2, e_3, \dots, e_m\}$ is a set of edges (ties, links). In an undirected graph, each link is defined as unordered pair (v_i, v_j) of nodes v_i and v_j . Then, two nodes joined by an edge are referred to as *adjacent* or *neighbouring*. In a directed graph, the order of two nodes is taken into consideration: (v_i, v_j) is an edge from v_i to v_j and $(v_i, v_j) \neq (v_j, v_i)$. Furthermore, nodes and edges can carry out some properties, like weights: *weighted graphs*. More details about graph theory can be found in [Bollobás, 1998, Gilbert, 2011, West, 2001]. Figure 3.2 depicts 3 examples of graphs with 7 nodes and 14 links.

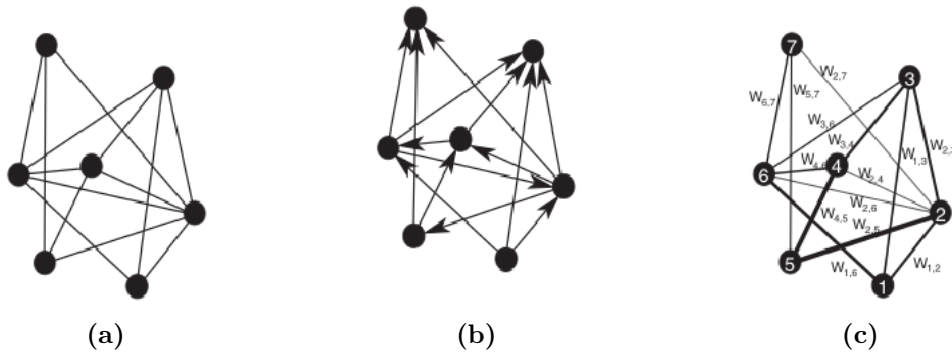


Figure 3.2: Undirected graph (a), directed graph (b) and weighted graph (c) in which the weight of an edge (i, j) is represented by $w_{i,j}$ and it is graphically represented by the link thickness [Boccaletti et al., 2006]

A central concept in graph theory is that of reachability of two different nodes in a graph [Boccaletti et al., 2006]. A *walk* from node i to node j is an alternating sequence of nodes and edges that begins with i and ends with j and its length is defined as the number of edges in the sequence. A *path*, in turn, is a walk in which no node is visited more than once. The walk of minimal length between two nodes is known as *shortest path* or *geodesic* and the longest shortest path defines the *diameter* of a graph. Yet, a graph can be *connected*, that is, there is a path from i to j for every pair of distinct nodes i and j , or *disconnected* when there is no such a path. From this, *component* is the largest subgraph such that is connected. The *degree* of a node i is the number of edges connected to i . It is not necessarily equal to the number of nodes adjacent to a node, since there may be more than one edge between any two nodes. In addition, a directed graph has both an *in-degree* and an *out-degree* for each node, which are the numbers of in-coming and out-going edges respectively.

3.1.2 Properties of Complex Networks

Typical issues addressed by network studies are *centrality*, representing the nodes that are best connected to others or have most influence; *connectivity* indicating how individuals are connected to one another through the network. Recent years however have witnessed a substantial new movement in network research, with the focus shifting away from the analysis of single small graphs and the properties of individual nodes or edges within such graphs to consideration of large-scale statistical properties of graphs [Newman, 2003].

This section presents some basic and useful network property related to the connectivity of a single node (local property) and the connectivity of the network as a whole (global property).

3.1.2.1 Degree and Degree Distribution

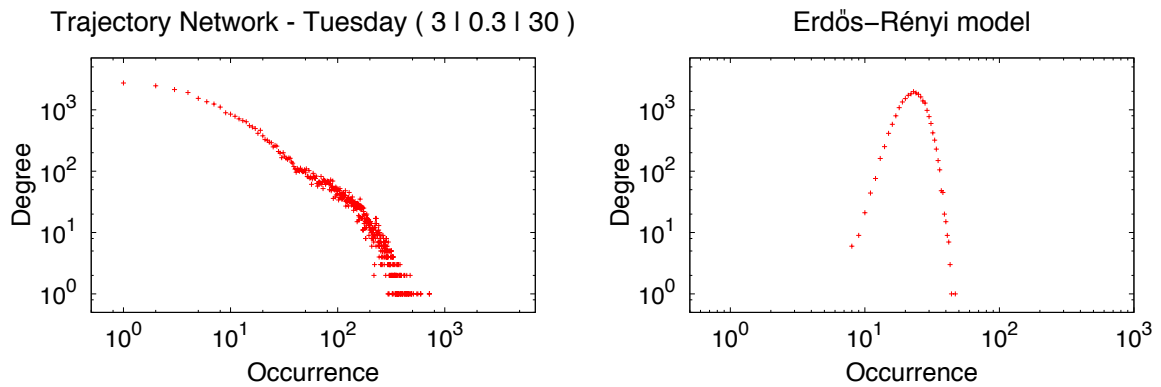
As discussed before, node degree is the number of edges that a node has and it can also consider the directness when we talk about directed networks. Degree is a measure of centrality in the network, where nodes more connected tend to be more central and have more importance when compared to lowly connected ones. Nodes with high degree are also considered “powerful” due to their connections. For instance, in a social network, these nodes correspond to people that know many others and, consequently, they are very important in the network.

Degree is a measure to identify individually important nodes by considering their edges. The degree distribution, however, is related to the network as a whole. It plays an important role when we want to characterize the connectivity of the nodes in the network. For instance, random graphs, graphs generated in a random way (Section 3.2.1), have a degree distribution of their nodes following a Poisson distribution, where the nodes tend to have the same degree: the average degree of the network. In real networks, on the other hand, the node distribution tends to follow a power law distribution, where the more connected nodes are more likely to receive new connections than the less connected nodes. Figure 3.3 shows two degree distribution, one following a power law distribution (Figure 3.3(a)) and another following a Poisson distribution (Figure 3.3(b)). Power law distribution is discussed further in Section 3.1.2.4.

To exemplify a degree distribution, let’s take the graph in Figure 3.2(c). First of all, the node degrees are calculated, Figure 3.4(a), and then the frequency of each found degree is also computed, Figure 3.4(b). Finally, a plot depicts the degree distribution, Figure 3.4(c).

3.1.2.2 Clustering Coefficient or Transitivity

In many real-world network it is found that if a node a is connected to node b and node b is connected to node c , then there is a high probability that node a will also be connected to node c . Intuitively it represents the idea that “a friend of my friend is also my friend”.



(a) Degree distribution following a power law from a network in Chapter 4
 (b) Degree distribution of a random graph generated from ER model in Chapter 4 following a Poisson distribution

Figure 3.3: Examples of degree distribution of two networks

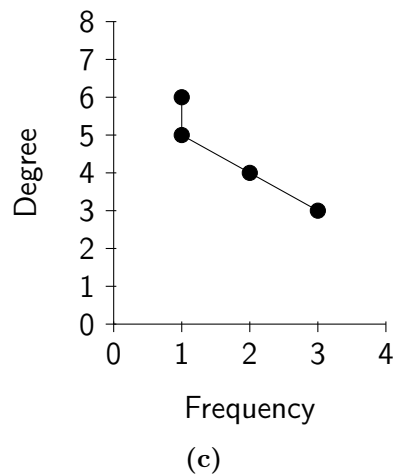
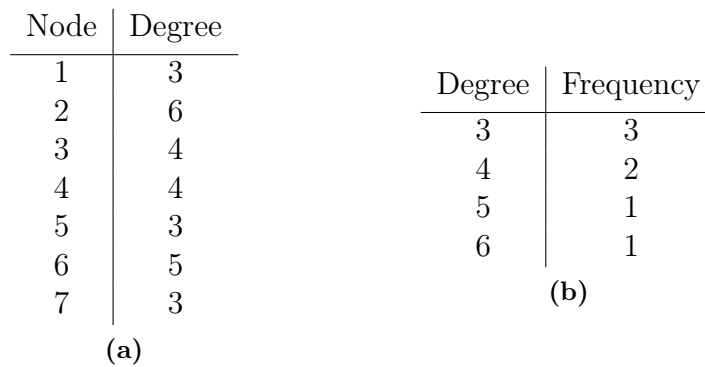


Figure 3.4: Example of degree distribution: (a) node degrees; (b) degree distribution; (c) degree distribution plot

This *transitivity*, also known as *clustering coefficient* means the presence of a number of triangles in the network, i.e., sets of three nodes each of which is connected to each of the others. There are different manners to compute clustering coefficient of a network: by finding triangles (global) or finding that coefficient for each node (local). The clustering coefficient is then defined by finding the triangles in the network as:

$$C = \frac{3 \times \text{number of triangles in the network}}{\text{number of connected triples of nodes}}, \quad (3.1)$$

where a “connected triple” means a single node with edges running to an unordered pair of others. C is the mean probability for two nodes that are neighbors of the same other node are also neighbors, hence C lies in the range $0 \leq C \leq 1$. It can also be written in the form

$$C = \frac{6 \times \text{number of triangles in the network}}{\text{number of paths of length two}}, \quad (3.2)$$

where a path of length two refers to a directed path starting from a specified node. Those definition of C are widely use in the sociology literature which is referred as the “fraction of transitive triples” [Newman, 2003]. On the other side, Watts and Strogatz [Watts & Strogatz, 1998] proposed an alternative definition by defining a local value:

$$C_i = \frac{\text{number of triangles connected to node } i}{\text{number of triples centered on node } i}. \quad (3.3)$$

Intuitively C_i of a node i represents the proportion of edges between its neighbors divided by the number of edges that there can be between them. For nodes with degree 0 or 1, their coefficient is defined as $C_i = 0$. Thus the clustering coefficient for the whole network is given by the average

$$C = \frac{1}{n} \sum_i C_i. \quad (3.4)$$

An interesting point about clustering coefficient refers to the fact that random graphs present low clustering coefficient when compared to real networks. So, heightened clustering coefficient and a large number of triangles are typical characteristics of real networks such as social network, biology networks and collaboration networks.

3.1.2.3 Shortest Path Length, Betweenness Centrality and Closeness Centrality

Nodes that are connected to others and these others are connected to many others and so on. Such connectivity offers an important role to the network in such a way that the nodes can change information or “reach” others by going through the edges. For instance, the routers that form a huge network changing package between them by the links. So, shortest paths are crucial for the nodes to reach other nodes besides being an important role in the characterization of the internal structure of a network. A measure of the typical

separation between two nodes in the network is given by the *average shortest path length*

$$L = \frac{1}{n(n-1)} \sum_{i \neq j} d_{i,j}, \quad (3.5)$$

where $n = |\mathcal{V}|$ and $d_{i,j}$ is the shortest path length between i and j . Random graphs and real networks are both characterized by small average shortest path lengths, i.e., the nodes tend to reach each others, on average, in a few steps. The reachability between two nodes i and j that are not neighbors depends on the nodes belonging to the paths connecting i and j . A node k that belongs to many shortest paths has an important position in the network: indeed, to transfer information from a node to another, this information passes through the node k that contributes to decrease the distances in the network and, consequently, to speed up the information propagation. This measure is known as *node betweenness*. Then, the betweenness of a node k is given by

$$b_k = \sum_{i,j,i \neq j} \frac{n_{ij}(k)}{n_{ij}}, \quad (3.6)$$

where n_{ij} is the number of shortest paths connecting i and j and $n_{ij}(k)$ is the number of shortest paths connecting i and j and passing through k . The concept of betweenness can also be used to edges, *edge betweenness*, which is defined as the number of shortest paths pairs of node that run through that edge. As degree, the betweenness is a measure of centrality in the network, since nodes with high betweenness play an important role in decreasing the average shortest path lengths. Other measure of centrality is the *closeness centrality*, which expresses the average distance of a node i to all others as

$$g_i = \frac{1}{\sum_{i \neq j} d_{ij}}. \quad (3.7)$$

Therefore, the nodes with shortest distances to the other nodes will be more central in the network.

3.1.2.4 Power Law Distribution

A distribution that follows a power law is a distribution in the form

$$p(x) = a * x^{-\omega}, \quad (3.8)$$

where $p(x)$ is the probability of x to occur, a is constant of proportionality and ω is the power law exponent [Newman, 2005]. Distributions that follow a power law are quite important for the understanding of natural and human phenomena. For instance, the population of cities and the intensity of earthquakes follow a power law distribution.

Other way to look at a power law distribution is to consider the popular saying “the richer get richer”, also known in Sociology as *Matthew effect* [Jackson, 1968]. Simon [Simon, 1955, Bornholdt & Ebel, 2001] showed that power laws arise when “the rich get richer”, when the amount you get goes up with the amount you already have.

Such phenomenon has also been found in real networks, i.e., degree distribution of the nodes follows a power law distribution. Such networks have many nodes with low degree, while a few nodes have high degree. Other works showed other distributions in network that also tend to follow a power law, such as the growth of the number of nodes and edges in evolving networks [Leskovec et al., 2006], and the number of triangles compared to the degree of nodes [Tsourakakis, 2008].

Although random networks and real networks present both a small average shortest path length, they are distinguished between them by their node distribution. As we have seen, degree distribution of real networks tends to follow a power law distribution. Random networks however tend to follow a Poisson distribution, i.e., the nodes tend to have the same degree. Networks with power-law degree distributions are sometimes referred to as *scale-free networks* [Barabási & Albert, 1999].

3.1.2.5 The Small-World Effect

The very famous experiment carried out by Stanley Milgram in the 1960s showed that letters passed from person to person were able to reach a designated target individual in only a small number of steps, sentence known as “*six degrees of separation*”. So, this result is one of the first direct demonstrations of the *small-world effect*, the fact that most pairs of nodes in most networks seem to be connected by a small short path through the network.

The small-world effect has implications for the dynamics of processes taking place on networks [Newman, 2003]. For instance, the spread of information across the network that occur quickly on most real networks, information like virus of computers, spam, diseases and even gossips. Networks with this behavior are characterized mainly by the two properties already discussed, i.e., they present a small average shortest path length and a high clustering coefficient when compared to random networks with the same size (number of nodes and edges) [Newman, 2003].

Recently, [Backstrom et al., 2011] discovered that this average number of acquaintances separating any two people in the United States was 4.37, and that the number separating any two people in the world was 4.74 by using data on the links among 721 million Facebook users.

3.2 Network Models

Network models aim at generating synthetic networks with peculiar characteristics. These models are useful tools in the comprehension of the growth and formation of the networks as well as in the studies of phenomena such as *small-world* and “*the rich get richer*”. In this section we present basic models starting from random graphs studied by Erdős and Rényi, passing through the small-world model proposed by Watts and Strogatz and ending up with the preferential attachment model by Barabási.

3.2.1 Random Graphs

The study of random graphs was initiated by Erdős and Rényi in 1959 with the original purpose of studying, by means of probabilistic methods, the properties of graphs as a function of the increasing number of random connections. Erdős and Rényi proposed a model to generate random graph with n nodes and m edges, which is known as ER random graphs [Erdős & Rényi, 1959, Erdős & Rényi, 1960]. There are two different ways to construct a network from this model:

- $\mathcal{G}_{n,p}$ is a random graph generated by taking some number n of nodes and connect each pair with probability p ;
- $\mathcal{G}_{n,m}$ is the ensemble of all graphs having n nodes and exactly m edges, each possible graph appearing with equal probability.

ER random graphs are the most studied among graph models, although they do not reproduce most of the properties of real networks, such as clustering coefficient. A small average shortest path length characterizes the random graphs, even though they do not present many triangles as we discussed early. In addition, random graphs present a Poisson distribution, while real networks tend to follow a power law distribution. Figure 3.5 depicts a random graph $\mathcal{G}_{n,p}$ with $n = 50$ and $p = 0.05$.

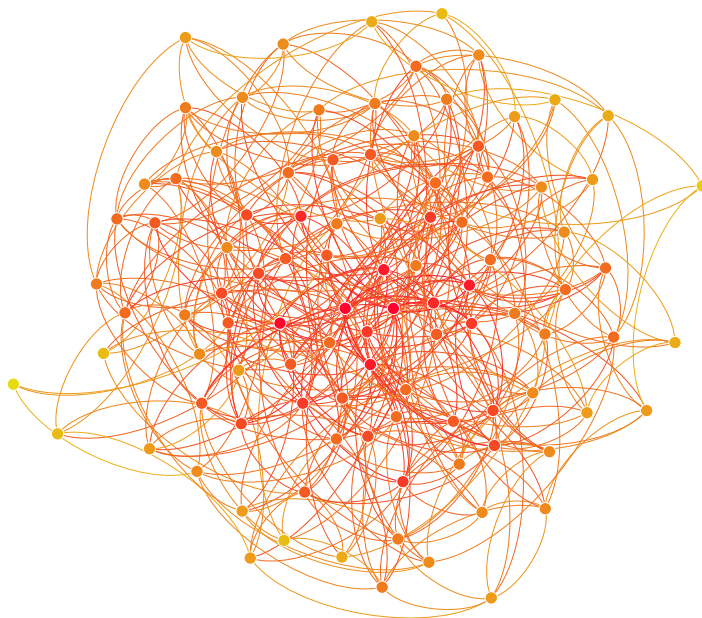


Figure 3.5: A random graph $\mathcal{G}_{n,p}$ generated with $n = 100$ and $p = 0.05$. Its average shortest path is 2.271 and its clustering coefficient is 0.099 (145 triangles)

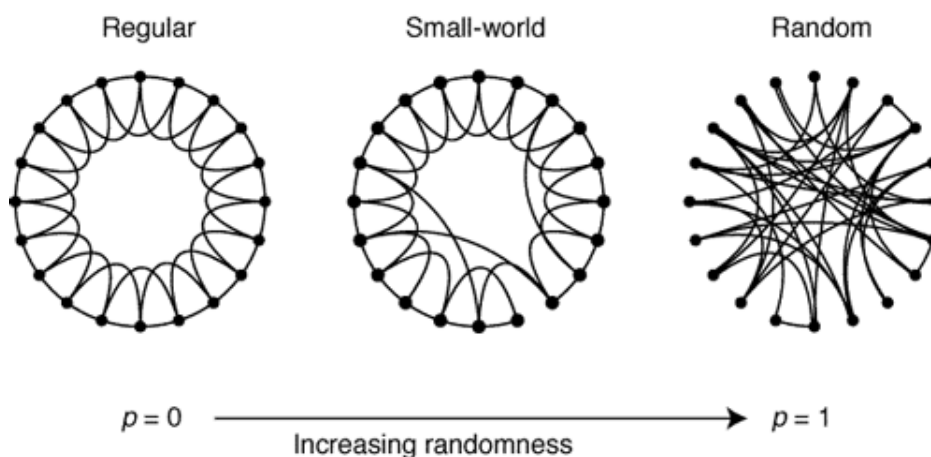


Figure 3.6: The small world model varying p . This figure illustrates well the network structure by varying the probability p : $p = 0$ generates a regular network, while $p = 1$ generates a random network [Watts & Strogatz, 1998]

3.2.2 Small-World Model

Watts and Strogatz proposed a model, known as *small-world* or *WS* model, to study and identify properties of *small-world effect*, i.e., a high transitivity and a small average shortest path length [Watts & Strogatz, 1998]. The small-world model starts with n nodes shaping a ring where each node is connected to its k nearest neighbors, that is, $k/2$ on its left side and $k/2$ on its right side. This first process generates a regular lattice. Forthwith, a process of “rewiring” is achieved, where each edge has its one end moved, with probability p , to a new location chosen uniformly at random from the lattice, except that no double edges or self-edges are ever created.

The rewiring process allows the small-world model to interpolate between a regular lattice and something similar to random graph. When $p = 0$, the generated network will be a regular lattice. However, the regular lattice does not show the small-world effect, since the path lengths tend to be large from a node to another. When $p = 1$, every edge is rewired to a new random location and the network looks like a random graph, with a small average shortest path, but with very low clustering coefficient. Therefore, p works as a balancer between regular lattice and something random as showed in Figure 3.6. A generated network from this small-world model is illustrated in Figure 3.7

Due to the simplicity of the original model proposed by Watts and Strogatz, where only one end of each chosen edge is rewired, no node is ever connected to itself and an edge is never added between node pairs where there is already one, many other small-world models were proposed [Monasson, 1999, Newman & Watts, 1999].

3.2.3 Models of Network Growth

The early models discussed so far take observed properties of the networks, such as degree or transitivity, to create networks that incorporate those properties without offering an

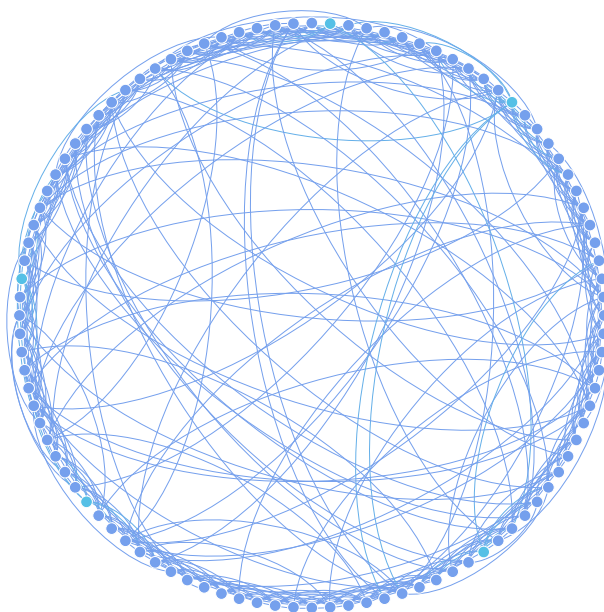


Figure 3.7: A WS network shaping a ring generated with $n = 100$ and $p = 0.2$ and $k = 10$. Its average shortest path is 2.548 and its clustering coefficient is 0.419 (624 triangles)

understanding how networks come to have such properties. In other words, they do not take into consideration the process of growth of networks and, hence, they are not models of network growth. From this, Barabási and Albert proposed the model known as *Barabási-Albert* (BA) or *Preferential Attachment* model [Barabási & Albert, 1999], which is based on two aspects: growth and preferential attachment.

The main idea of BA model goes towards the phenomenon “*the rich get richer*”. Speaking in network terms, the nodes with highest degrees (“the rich”) are likely to form new edges with other nodes (“get richer”), and those nodes are called *preferential attachment*. More precisely, an undirected graph $\mathcal{G}_{n,k}$ is constructed from BA model as follows. Starting with m_0 isolated nodes, at each time step $t = 1, 2, 3, \dots, N - m_0$ a new node j with $m \leq m_0$ links is added to the network. Then, the probability that a link will connect j to an existing node i is linearly proportional to the degree of i . Despite of being elegant and simple, BA model lacks some features that are present in the real World Wide Web, such as the directness of the edges. Figure 3.8 shows an example of $\mathcal{G}_{n,k}$.

Besides the model proposed by Barabási and Albert, there are other models related to the network growth, such as the Price’s model [Price, 1976, Price, 1965], which is very similar to BA model, and Dorogovtsev and Mendes [Dorogovtsev & Mendes, 2000].

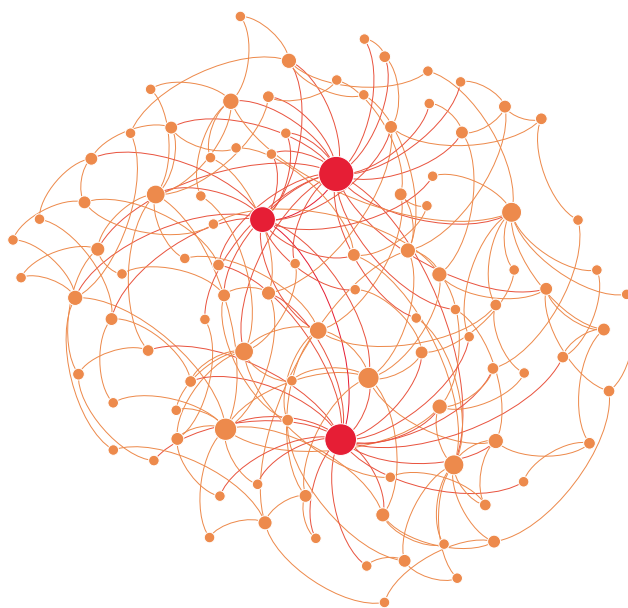


Figure 3.8: A BA network generated with $n = 100$ and $m_0 = 4$ and $k = 2$. The largest and red nodes represent the preferential attachments, highly connected nodes, which are more likely to establish new links with other nodes

3.3 Community Discovery

As we have discussed so far, many properties are computed over nodes and edges to catch global or local behaviors of the network. However, another important part of the network research is related to the network structure, that is, how the nodes connect to each other forming groups together called *communities*. Take as example people that form groups with other people in different contexts, such as our friends from work, university and even gym. Each context may correspond, somehow, to communities of a network of people.

According to [Newman, 2003], community discovery should not be confused with the technique of data clustering, which is a way of detecting groupings of data-points in high-dimensional data spaces. Community discovery and data clustering have some common features and algorithms for one can be adapted to the other, and vice-versa. For example, high-dimensional data can be converted into a network by placing edges between closely spaced data points, and then network clustering algorithms can be applied to the result. On balance, however, one normally finds that algorithms specially devised for data clustering work better than such borrowed methods, and the same is true in reverse.

Many algorithms have been developed to identify and extract communities from different types of network [Girvan & Newman, 2002, Radicchi et al., 2004]. In some cases the communities obey a recursive structure, where large communities can further be divided into smaller communities [Clauset et al., 2004, Guimera et al., 2007]. Some works proposed methods for community discovery in large networks based on heuristics [Blondel et al., 2008]. Despite of existing many algorithms to address this problem,

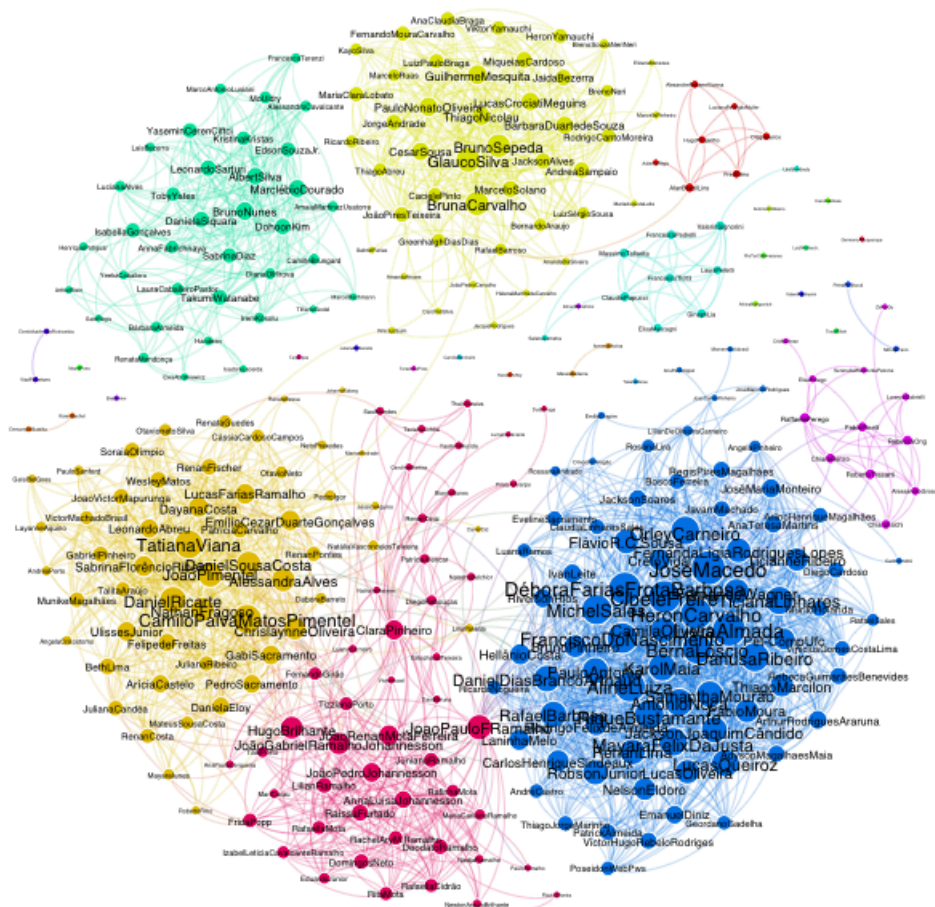


Figure 3.9: An example of community discovery by performing the method proposed in [Blondel et al., 2008]. This algorithm is available in Gephi [Bastian et al., 2009]. This network was built from a user’s profile in a social network and, thus, each community represented by a color shows a community of people. For instance, the blue community represents the friends from the university and the red one represents the family ties

community detection does not have a unique concept or definition. As consequence, a broad variety of methods have been proposed to discovery communities.

In front of these great variety of techniques, [Coscia et al., 2011] proposed a classification for community discovery methods in complex networks. Some methods take nodes as entities, that is, nodes are compared to each other to compute their similarity, while others consider edges as entities in order to group the edges and further the nodes. In addition, there are a number of interesting features of these communities that can be considered, such as hierarchical or overlapping configuration of the groups inside the network, the directness of the edges to give importance to this direction when considering the relations among entities and, yet, the dynamism of the networks, i.e., networks that evolve over time.

Since there are various different techniques, we only describe a community detection algorithm, the method proposed by [Ahn et al., 2010], in Section 5.2, since it is used in the case study in Section 5.4.

3.4 Summary

This chapter introduced basic concepts and definitions in complex networks, presenting some important definitions in graph theory, which is the basis of networks. In addition, this chapter presented some global and local properties of networks that are important not only for understanding of network topologies, but also for comprehension of network behaviors.

Some network models present in the literature were discussed, including the ER model proposed by [Erdős & Rényi, 1959, Erdős & Rényi, 1960], the WS model proposed by [Watts & Strogatz, 1998] to capture the behavior of networks with small-world effects, and the Preferential Attachment model proposed by [Barabási & Albert, 1999] in order to understand the growth of networks based on preferential attachment.

To conclude, this chapter presented an important branch of complex network research area that aims at discovering structure in the networks known as *communities*. Methods for community discovery have not been presented due the numerous available methods with different approaches. However, in Section 5.2 is presented the algorithm proposed by [Ahn et al., 2010] since it is used in the case study of Section 5.4.

CHAPTER 4

Trajectory Analysis using Complex Network

Although the management of trajectory data dates back to the 1990s, when the first proposals for moving object databases came out, the challenging approaches towards the analysis and understanding of the movement complexity represented in the users tracks is being faced only recently [Wang et al., 2009]. Even more challenging is the aspect of moving object interaction. How and how much do these moving objects interact? How do the *encounters* among moving entities globally characterize the movement of a moving community? Is there a specific law explaining the interactions of moving individuals? Is the movement of people in vehicles (e.g. cars in a road network) differs from people free movement and/or multi transportation trajectories? How do the individual movements of independent entities influence a crowd's movement pattern?

Inspired by these questions, this work poses a first step in experimenting the complex network analysis techniques applied to a trajectory dataset. The main aim here is to formalize interactions between moving objects as edges in a graph and study the behavior of this graph in terms of complex networks. The approach presented in this chapter can be placed between the discipline of mobility data analysis and complex networks, thus exploiting complex network properties to understand mobility of users. Also the challenging and innovative aspect of this experiment is that the network we computed is based on moving objects interactions, which is different from classical complex

networks experiments, which focuses on objects that are “static”, from the point of view of the spatial position.

The contributions present in this chapter are twofold. The first contribution is a method for devising a complex network from a trajectory dataset, hereinafter called *trajectory network*. The aim of this method is to define specific steps for processing trajectory data in order to build and analyze the trajectory network. The second contribution is an algorithm for building a trajectory network given a trajectory dataset (set of spatio-temporal points). Indeed, this is the first work on analyzing trajectory interactions through complex network techniques [Brilhante et al., 2011].

The proposed method has been evaluated using a real GPS dataset from vehicles moving in the City of Milan. All generated trajectory networks from this dataset presented the small world effect and the scale-free feature similar to the Internet and biological networks. However the interpretation of these features is an open issue, therefore we will discuss possible interpretations and exploitations of them.

This chapter is structured as follows. Section 4.1 reviews some basic definitions introduced in Chapter 2 and 3, and introduces related works. Section 4.2 presents the methods and algorithms used to build the trajectory complex network, whereas Section 4.3 reports experimental results carried on a complex network of vehicle trajectories. Section 4.4 draws conclusions and future work.

4.1 Basic Concepts and Related Work

4.1.1 Basic Definitions

As presented in Chapter 2, a *trajectory* can be defined as the spatio-temporal evolution of a moving object [Spaccapietra et al., 2008]. This evolution is typically represented as a sequence of sample points, representing the spatio-temporal positions detected by a tracking device, such as GPS tools or WIFI sensors. More formally, a *trajectory* T of an object O is represented as: $T_O = \{p_0, p_1, \dots, p_n\}$, where $p_i = (x_i, y_i, t_i)$, $x_i, y_i \in \mathbf{R}$ represent the spatial coordinates of the sample point, $t_i \in \mathbf{R}^+$ represents the timestamp for $i = 0, 1, \dots, n$, and $t_0 < t_1 < t_2 < \dots < t_n$.

In Chapter 3, a *complex network* is introduced as a network with thousands or millions of nodes whose structure is irregular, with non-trivial topology features [Boccaletti et al., 2006]. The following features typically characterize complex networks:

- Clustering coefficient: represents the density of triangles in the network. Sparse random graphs have smaller clustering coefficients, while real-world networks typically have larger coefficients;
- Average shortest path length: is the average node-to-node distance. Random graphs exhibit a small average shortest path length as well as real-world networks;

- Power law distribution: is a distribution that follows a power law function, $p(x) = a * x^{-\alpha}$, such that $p(x)$ is the probability of occurrence of x , a is a constant of proportionality and α is the power law exponent.

Complex networks can be characterized by the so called “small world” property when the average number of edges between any two vertices is very small and the clustering coefficient is large [Watts & Strogatz, 1998]. Intuitively, this represents a short path between two edges. This is also known as the “six degrees of separation”. Scale free networks are characterized by a degree distribution that follows a power law function. Intuitively, few nodes have many edges (the “hubs” or preferential attachment [Barabási & Albert, 1999]), many nodes have few edges.

4.1.2 Mobility Analysis

With the increasing availabilities of trajectory datasets collected from GSM or GPS equipped devices we have the possibility of studying people behavior from their movement traces. Several application areas would benefit from an extensive study on people trajectories such as traffic management, public transportation, commercial advertising, security and police, hazard evacuation management, location based services and so on.

The task of analyzing large trajectory datasets can be carried out in four different directions. First, basic statistics may be applied to trajectory data mainly to discover the distributions of people presence and origin-destination matrices [Calabrese et al., 2010]; other studies focus on trajectory data mining, that is, on the application of data mining techniques to trajectory data [Giannotti & Pedreschi, 2008]; other researches focus on representing and querying moving objects in database systems [Nguyen-Dinh et al., 2010, Güting et al., 2000]; finally, research originally coming from Physics studies mathematical models, such as complex networks, representing the general laws that describe human movement [González et al., 2008, Wang et al., 2009].

Trajectory mining aims at finding correlations in large datasets of trajectory data, collected by personal positioning devices. Techniques include: (1) clustering discovery - finding groups of objects moving together; (2) sequential pattern discovery - finding the most frequent sequences of places visited; (3) flock detection - extracting the convergence of people moving together for a certain amount of time [Dodge et al., 2008, Giannotti & Pedreschi, 2008].

Several works have investigated how to model and query movement data efficiently, in database literature a new class of database systems were created, called moving object database [Nguyen-Dinh et al., 2010, Güting et al., 2000]. However, these works were not focused in modeling or querying trajectory data as a first class object. In addition, they did not aim at exploring moving objects interactions.

Trying to model the basic laws governing the human motion is the aim of a broad research area coming from Physics [González et al., 2008]. Their objective is to

study the physical laws representing human movements. Social interactions is also in the scope of this research area. A typical example is the study of the spreading of cell phone viruses thru GSM phone calls [Wang et al., 2009].

4.1.3 Complex Network Analysis

As already presented in the previous section and Chapter 2, a network is a set of items, called vertices (or nodes), with connections among them, called edges (or links). The study of networks (in the form of mathematical graph theory) is one the fundamental pillars of discrete mathematics. Networks have also been extensively studied in different domains, such as Social sciences, Physics, etc. However, recent years have witnessed a substantial new movement in network research, focusing on developing methods and techniques to gather and analyze networks far larger than previously possible. Indeed, this new motivation is due to the inability of humans to draw a meaningful picture of a million vertices by direct eye analysis.

Network has been used as a mechanism of analyses of a huge amount of data with a set of objects which have a relationship or a interaction between them. For instance, the studies in psychology where a node represents a person and an edge represents friendship or that they work together or simply that they know each other or even they have sexual relationship; the studies in biology where the focus is on the species in an ecosystem and a interaction between them, that is, an edge (directed) from species A to species B indicates that A preys B [Pimm, 2002]. Therefore, networks offer a perspective of analyses basing on the relationships or interactions.

With respect to mobility analysis, [Guo et al., 2010] presents a graph-based approach to represent the trajectories by using representative points, a new set of points based on the original one, to generate a graph and find clusters of trajectories. However, they do not consider the properties of the complex network area such as clustering coefficient. On other hand, [Kaluza et al., 2010] analyzes global cargo ship movements by building a complex network whose nodes represent the ports and links represent the ship traffic between two ports. Differently from our approach, they do not represent the nodes as trajectories and, besides, the points of the trajectories are not taken into account, but only the ports that the cargo ships passed.

4.2 Complex Network and Trajectory

This section presents one approach about how to create a complex network from trajectory data. This approach constructs a simple graph where each node represents a trajectory and each edge represents a relationship among the nodes. A relationship between two nodes is established when there is an encounter between two trajectories in space and time with a minimum frequency of meetings.

In this approach, a set S of trajectories is represented as a network (N, E) with

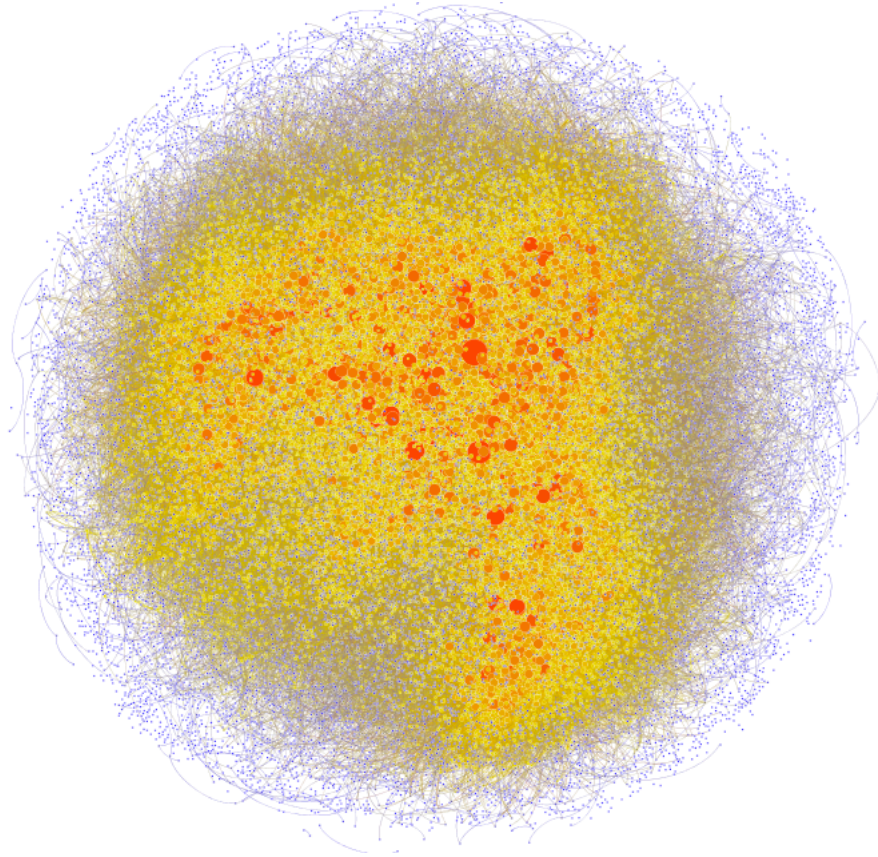


Figure 4.1: A plotted complex network composed by 36,824 nodes and 306,572 edges generated in our experiments

the help of a similarity function f between trajectories in S and a threshold constant c . This network is called *trajectory network*. The *trajectory network* (N, E) is constructed as follows: (1) each node in N represents a trajectory in S ; (2) there is an edge between two nodes n and m iff $f(m, n) \geq c$, that is, m and n represent trajectories whose similarity is above the given threshold.

In what follows, we will not distinguish between trajectories in S and nodes in N .

We therefore define a function f to capture the spatial and temporal proximity between two trajectories in order to establish an edge between them in the network. Let f be the similarity function for trajectories used to construct the trajectory network, and let s , t and k respectively be the spatial, temporal and frequency parameters of f . Given a trajectory T , the spatial and temporal parameters induce a buffer $B[s, t](T)$ around T . Given two trajectories T and U , we then define:

Definition 4.1 *meet (or collide):* T and U meet iff $B[s, t](T)$ and $B[s, t](U)$ overlap.

Definition 4.2 *encounter:* there is an encounter between T and U iff T and U meet more than k times.

Note that, as defined above, two trajectories meet independently of their direction. In addition, our definition of encounter does not constrain that two moving objects must be present in the same place at the same time instant.

Our methodology for analyzing trajectory data using this network approach encompasses 3 steps:

1. Build the trajectory network;
2. Analyze trajectory network features;
3. Identify relevant trajectories within the trajectory network.

In the rest of this section, we analyze networks constructed using the specified similarity function.

4.2.0.1 Step 1 - Build Trajectory Network

Intuitively, we define that two trajectories are similar iff they are within a certain distance of each other (spatial threshold) within a given time interval (temporal threshold) for a certain number of times (frequency parameter). The values of these parameters obviously depend on the application domain under study. For example, in the traffic management domain, we can establish that two vehicles meet with a minimum frequency of meetings (as few as 2-5 times, depending on the density of the dataset). This representation describes the interactions (given a spatio-temporal threshold and a minimum frequency) between vehicles regardless of the direction of their trajectories. But in other applications, where the study of the flows of vehicles is important, the direction should also be taken into account.

Algorithm 4.1 Compare two trajectories by their positions

Input: Two trajectories T_1 and T_2 , a temporal threshold Td to compute the temporal difference and a spatial threshold Sd to compute the spatial distance

Output: A frequency f

1. $f \leftarrow 0$
 2. **for each** position p in T_1 **do**
 3. **for each** position q in T_2 **do**
 4. $spatial \leftarrow spatialDifference(p, q)$
 5. $temporal \leftarrow temporalDifference(p, q)$
 6. **if** $spatial \leq Sd$ and $temporal \leq Td$ **then**
 7. // increment the meet between T_1 and T_2
 8. $f \leftarrow f + 1$
 9. **end if**
 10. **end for**
 11. **end for**
 12. **return** f
-

The main idea of the algorithm for building trajectory network (see Algorithm 4.2) is to compare each position (time, latitude, longitude) of a trajectory to all other trajectories. The comparison between two trajectories (see Algorithm 4.1) compares all points of these two trajectories (Algorithm 4.1 - line 2 to 10). Each comparison takes into account geographic position - latitude and longitude - and timestamp and the temporal and spatial distance thresholds (Algorithm 4.1 - line 4 and 5), respectively Td and Sd variables. When the comparison satisfies the thresholds Td and Sd (Algorithm 4.1 - line 6), then the frequency variable is increased by one unit (Algorithm 4.1 - line 7). After all comparisons (Algorithm 4.2 - line 9), we have computed the frequency of encounters of a trajectory with respect to another trajectory. If the frequency is greater than the value of the input parameter Frequency, denoted by variable c (Algorithm 4.2 - line 10), then an edge is created between two trajectories (Algorithm 4.2 - line 11).

Some improvements can be done by using spatial index structures or spatial database systems to take advantage of the index structure and the data management language as well. For instance, we could store the trajectory dataset on a spatial database system in order to index the points of the trajectories to perform more efficiently spatial comparisons and, besides, to use index structures on the timestamp of the points to compute the temporal difference among the points of the trajectories.

Algorithm 4.2 Trajectory Network Generator

Input: A trajectory dataset $traj$, a similarity function f to compute the similarity between two trajectories and a minimum threshold c to establish an edge between the trajectories

Output: A trajectory network TN

```

1.  $n \leftarrow |traj|$ 
2. create an undirected graph  $TN$ 
3. for each trajectory  $T$  in  $traj$  do
4.   //  $T$  represents a node in  $TN$ 
5.   create a node in  $TN$ 
6. end for
7. for  $i = 1$  to  $n$  do
8.    $similarity \leftarrow 0$ 
9.   for  $i = i + 1$  to  $n$  do
10.    // compute the similarity between  $traj[i]$  and  $traj[j]$  by using a given function
    //  $f$ : algorithm 4.1
11.     $similarity \leftarrow f(traj[i], traj[j])$ 
12.    if  $similarity \geq c$  then
13.      // there is an encounter between  $traj[i]$  and  $traj[j]$ 
14.      // create an edge between  $traj[i]$  and  $traj[j]$ 
15.      add edge  $(traj[i], traj[j])$  in  $TN$ 
16.    end if
17.  end for
18. end for
19. return  $TN$ 

```

4.2.0.2 Step 2 - Analyze Trajectory Network Features

In our approach we are interested in identifying the existence of two very important network features: the power-law distribution and the small-world effect. The analysis of the distribution degree of network vertices allows identifying whether such distribution is highly skewed, meaning that it has a power-law distribution profile. In this case, we conclude that few trajectories have many encounters, while most of the trajectories have very few encounters. The discovery of such property can be useful, for example, to identify trajectories having a high degree of encounters, which means that this trajectory has passed through paths with a high number of moving objects. Besides, small-world property may help to identify a set of trajectories that represent hubs in the trajectory network.

The small-world effect feature determines the mean shortest path length between pairs of trajectories as well as if the network has a high clustering coefficient. Through this measure we can quantify how well connected the trajectories in the network are. Besides a high clustering coefficient indicates the presence of a transitivity property among high connected nodes. This information can be very useful, for example, when analyzing bus trajectories and their encounters we can verify how buses lines are connected and how easy is it to move through the city using bus lines.

There are several free tools for computing network features, which can be used in this step without requiring devising an algorithm for that. One example of such a tool is the Network Workbench [Team, 2006] developed by Indiana University, Northeastern University, and University of Michigan. This tool provides several algorithms to calculate the properties of large complex networks.

4.2.0.3 Step 3 - Identify relevant trajectories within trajectory network

The third step in our approach aims at analyzing trajectories that have greater relevance within the network. The relevant trajectories are those that possess a high degree of connectivity. These trajectories are plotted on a map for visual analysis, allowing the user to give an interpretation of the relevance of these trajectories in the geographic context, Figure 4.5. This type of analysis will help reducing the number of trajectories to be analyzed. Furthermore, we can restore back the spatial information, which was lost during the creation of the network. Visualizing trajectories that are very connected, which we hereinafter call hub trajectories, is useful for understanding entities moving in the high dense paths with respect to the amount of moving objects.

4.3 Experiments

In this section, we presented the achieved experiments on a real dataset collected from vehicles in Milan, Italy. Firstly, some statistics of the dataset as well as the used parame-

ters for the Algorithm 4.2 are presented. Afterwards, we presented the computed features of the generated network described in Section 4.2.

4.3.1 Experiments on the vehicles' movements in Milan city

In this chapter, we propose a number of experiments using a mobility network that represents the trajectories of vehicles moving in the City of Milan, Italy (collected by GPS devices installed in the vehicles). We split the dataset into seven different files corresponding to the days of the week. The number of trajectories in each day of the week and their average length is depicted in Table 4.1. In this dataset, each trajectory corresponds to only one car.

Table 4.1: Information about trajectory dataset

Day	Number of trajectories	Average number of points per trajectory
Sunday	23535	8.290461
Monday	34812	8.927956
Tuesday	36824	9.206279
Wednesday	36023	9.467285
Thursday	35340	9.871647
Friday	33822	8.697179
Saturday	25576	7.746950

We can notice a high number of trajectories (i.e. vehicles) tracked each day; we also notice a decrease of moving vehicles during the weekend, as expected. The average number of sample points for each trajectory stays within 7 and 10. The low number of points per trajectory is due to a data cleaning process performed on the original data, which eliminated outliers and redundant points.

We start our experiments running Algorithm 4.2, Trajectory Network Generator, described in Section 4.2 with the following parameters:

- trajectory dataset: specifies the name of a trajectory dataset to be processed;
- minimum frequency of encounters: this parameter specifies a minimum number of meetings between two trajectories;
- a similarity function: Algorithm 4.1 with a spatio-temporal window for encounter. This window defines a temporal and spatial distance that is used to select trajectories that have a meeting in time and space.

We have generated 28 trajectory networks referring to the 7 days of the week with 4 different parameter configurations as presented in Table 4.2. We chose a spatial

distance threshold of 300 meters, and two different temporal intervals of 15 and 30 minutes. These values have been chosen starting from dataset statistics. For example, since the trajectory sample points are quite far from each other, we choose a quite broad notion of “meeting” setting this parameter to some hundreds meters. Obviously, denser sample points will correspond to smaller distance. Analogous is the consideration for the temporal threshold. The values 3 and AVG for minimum amount of encounters was defined taking into account the average number of encounters computed from our data set. The AVG value is related to the average number of points per trajectory in Table 4.1 for each day. For instance, on Sunday we get 8 as value to this parameter. Table 4.2 below summarizes the 4 parameters combinations that were used to setup our 4 experiments on the 7 daily trajectories datasets.

Table 4.2: Four different parameter combinations

Experiment	Minimum Frequency	Spatial Distance (km)	Temporal Distance (min)
1	3	0.3	30
2	3	0.3	15
3	AVG	0.3	30
4	AVG	0.3	15

4.3.2 Computed Trajectory Network Features

Tables 4.3, 4.4, 4.5, and 4.6 show the trajectory network features computed from our experiments. The computation of the trajectory network features was accomplished through the use of the Network Workbench Tool [Team, 2006]. Each table reports the number of nodes n (each node represents a trajectory), the edges m (an edge represents an encounter between two trajectories), the clustering coefficient C (represents the density of triangles in the network), the average shortest path length l and the diameter d .

We can notice that all 28 trajectory networks generated in our experiments are highly clustered networks. To arrive at this conclusion, we generated a Erdős and Rényi (ER) model [Erdős & Rényi, 1959, Erdős & Rényi, 1960] network, described in Section 3.2.1, with the same size, in nodes, as the networks generated in experiment 1 (Table 4.3), and with the number of edges proportional in order to have a similar network in size and structure. Table 4.7 shows the size and the computed features of the ER networks. They are generated randomly and characterized by their low average shortest path length as well as low clustering coefficient. Comparing Tables 4.3 and 4.7, one may verify that the level of clustering of the trajectory networks is much higher than that of the corresponding ER network. This means that we have a set of trajectories that have a high number of encounters. In combination, they present low values for average shortest path length (less than 6), and this has a clear interpretation as small world property

[Watts & Strogatz, 1998].

Also notice the presence of many trajectories that have few encounters and few trajectories that have a large number of encounters. This characterizes a phenomenon known as “the rich get richer”. This analysis is done by plotting a curve that shows the

Table 4.3: Experiment 1: frequency of 3, spatial threshold of 0.3 km and temporal threshold of 30 minutes

Day	n	m	C	l	d
Sunday	23535	229938	0.441	5.028	17
Monday	34812	515866	0.445	4.615	19
Tuesday	36824	590821	0.456	4.537	16
Wednesday	36023	587151	0.455	4.498	16
Thursday	35340	587399	0.457	4.393	16
Friday	33822	447244	0.437	4.776	19
Saturday	25576	210073	0.424	5.512	19

Table 4.4: Experiment 2: frequency of 3, spatial threshold of 0.3 km and temporal threshold of 15 minutes

Day	n	m	C	l	d
Sunday	23535	116671	0.418	5.691	22
Monday	34812	264897	0.417	5.196	22
Tuesday	36824	306572	0.432	5.099	19
Wednesday	36023	305338	0.432	5.058	20
Thursday	35340	307179	0.431	4.936	20
Friday	33822	229983	0.413	5.371	21
Saturday	25576	106424	0.398	6.363	22

Table 4.5: Experiment 3: frequency of AVG, spatial threshold of 0.3 km and temporal threshold of 15 minutes

Day	n	m	C	l	d
Sunday	23535	27789	0.431	6.574	22
Monday	34812	78204	0.449	6.014	22
Tuesday	36824	82795	0.487	6.052	17
Wednesday	36023	86811	0.488	5.995	20
Thursday	35340	92753	0.490	5.629	24
Friday	33822	63898	0.443	6.168	18
Saturday	25576	28633	0.399	8.065	25

Table 4.6: Experiment 4: frequency of AVG, spatial threshold of 0.3 km and temporal threshold of 30 minutes

Day	n	m	C	l	d
Sunday	23535	58901	0.464	5.906	21
Monday	34812	161632	0.480	5.328	18
Tuesday	36824	168823	0.511	5.319	20
Wednesday	36023	175310	0.510	5.267	17
Thursday	35340	185919	0.513	5.021	18
Friday	33822	132142	0.471	5.486	21
Saturday	25576	60438	0.431	6.614	23

correlation between the amount of nodes versus their respective degrees. Two graphs were created (Figures 4.2(a) and 4.2(b) showing the distribution of trajectories of Tuesday and Sunday, corresponding to the largest and the smallest networks generated by our experiments. In both graphs we have a power law curve.

By contrast, Figure 4.2(c) illustrates the degree distribution of the corresponding ER networks. Note that the node degrees of the ER networks do not follow a power law distribution, whereas the node degrees of the trajectory networks do. In fact, the trajectory networks have nodes that are hubs [Barabási & Albert, 1999]. This is an important result since all trajectory networks in our experiment present a small world and power law feature similar to the internet and biological networks.

Table 4.7: Random Graph - Erdős and Rényi

Day	n	m	C	l	d
Sunday	23535	278105	0.0010	3.539	5
Monday	34812	606707	0.0010	3.262	4
Tuesday	36824	679044	0.0009	3.221	4
Wednesday	36023	649475	0.0010	3.237	4
Thursday	35340	625219	0.0009	3.251	4
Friday	33822	572781	0.0009	3.283	4
Saturday	25576	328251	0.0009	3.483	5

Hereinafter, we analyze each trajectory network feature separately, namely clustering coefficient, average path length, and diameter. Figure 4.4(a) shows the clustering coefficient feature computed for all 28 trajectory networks. In this Figure, we can observe that the clustering coefficient increases on Tuesday, Wednesday and Thursday. This is consistent with the dataset statistics (see Figure 4.3) where we notice an increase of trajectories moving on Tuesday, Wednesday and Thursday compared to Friday (for that particular week the Friday was just before Easter holidays) and the weekend. Intuitively,

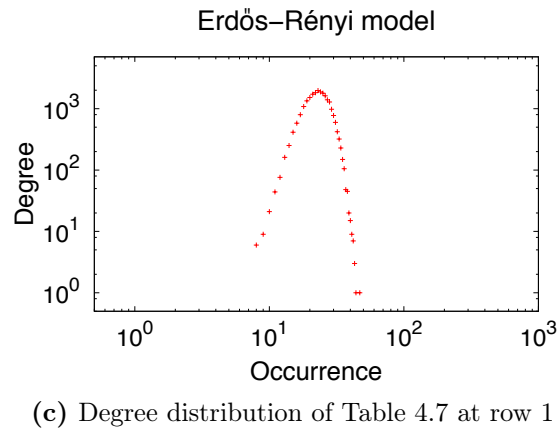
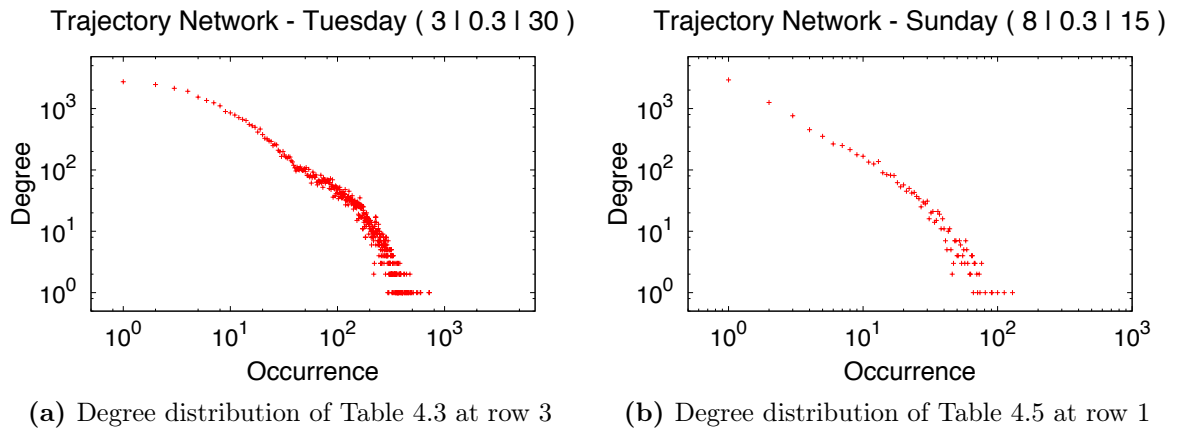


Figure 4.2: Degree distribution

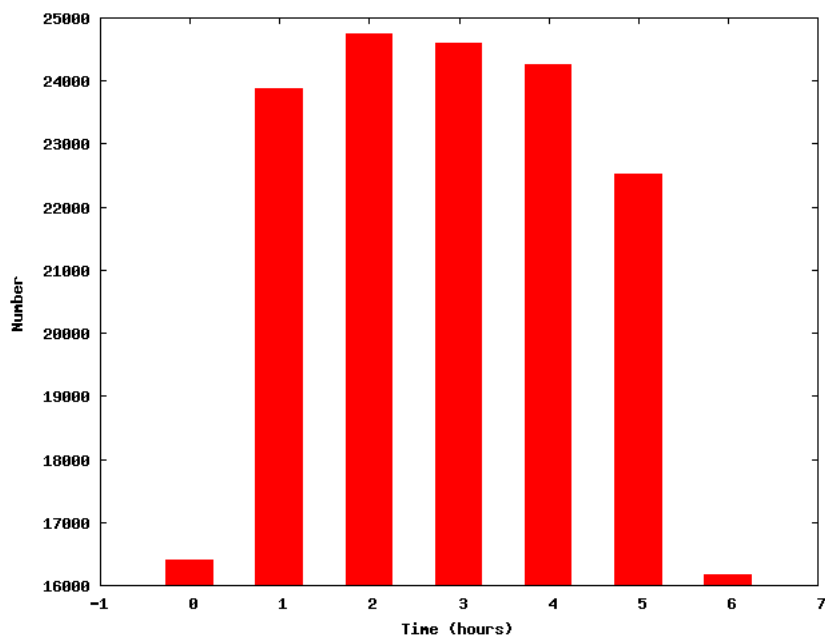


Figure 4.3: Plot showing the number of trajectories for each day of the week

this reflects the fact that we have more vehicle's encounters during these central days, that reduce when approaching to Friday and the weekend. From these trajectory networks, we deduce that Tuesday, Wednesday and Thursday are the days with more encounters than the other days of the week and this is consistent with dataset statistics.

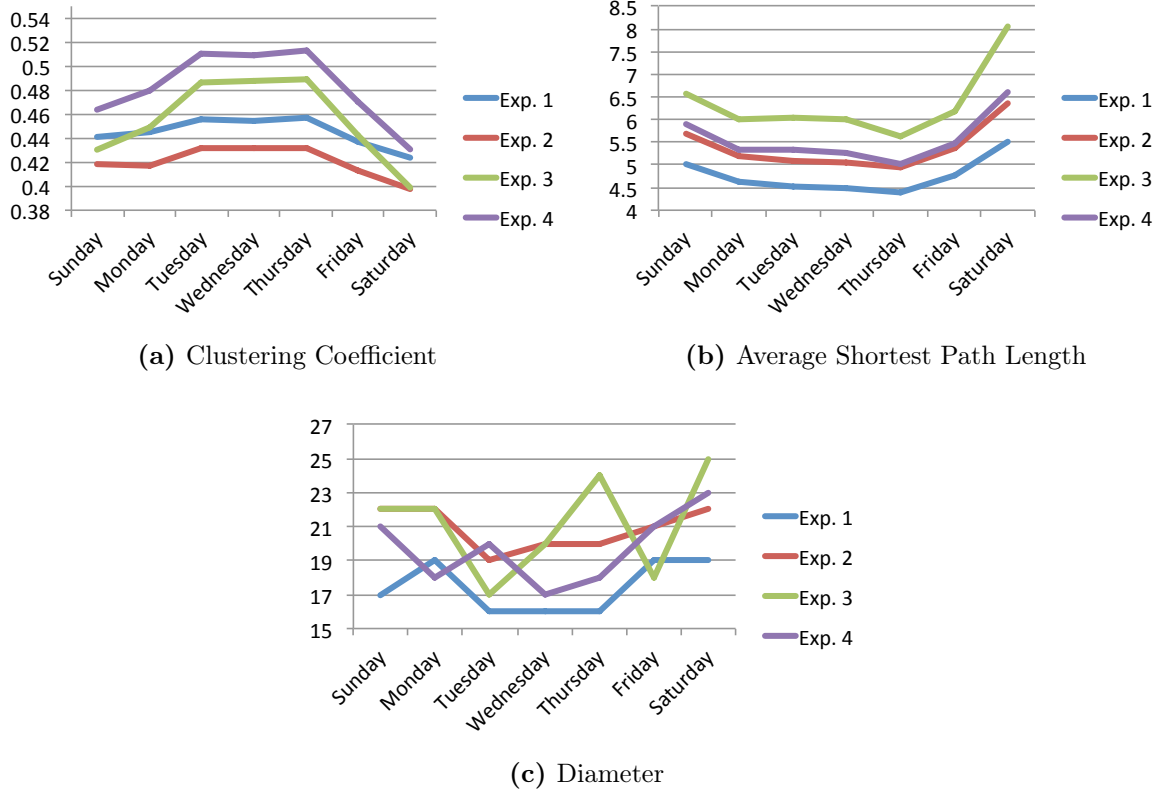


Figure 4.4: Network properties of tables 4.3, 4.4, 4.5 and 4.6

Figure 4.4(b) shows the average shortest path length feature for all 28 trajectory networks. We observe that in all trajectory networks the average path length increases during the weekend and reduces during the working days. Particularly, Thursday is the day with the lowest average shortest path length in all trajectory networks. Since the shortest path length characterizes the distance between two nodes, then when we have more encounters among vehicles we decrease the average path length. Thursday is the day with the largest number of encounters, which increases the probability of having traffic congestion. This result is coherent with the clustering coefficient feature graph (Figure 4.4(a)), where the highest clustering coefficient is also associated with Thursday.

The diameter feature (the largest distance between two nodes) is illustrated in Figure 4.4(c). In this graph we observe that there is no explicit correlation among diameter features from all the trajectory networks. This behavior is justified by the fact that the diameter feature is very sensitive to trajectory network topology. Besides, given the size of the trajectory networks, which are around 5 orders of magnitude, the variation in diameter along each day of the week (showed in Figure 4.4c) is insignificant.

Figures 4.5(a), 4.5(b), 4.5(c), 4.5(d), 4.5(e), 4.5(f) and 4.5(g) show the hub

trajectories for each day of the week. These trajectories are projected on GoogleMaps. Particularly, we note that a part of the road segment, the A51 highway, is presented in five of seven high connected trajectories (indicated in the figure by an arrow). This fact suggests that this road segment has a high concentration of cars, possibly because this is a ring highway that goes towards the Milano Linate Airport. In Figure 4.5b (Monday) we observe that the most connected trajectory represents a moving object that goes back and forth between the airport and Milan’s downtown, probably a shuttle service, which corroborates to increase the amount of encounters to this trajectory.

We can also remark that on Monday, Tuesday, Thursday, Friday and Saturday (respectively Figures 4.5(b), 4.5(c), 4.5(e), 4.5(f) and 4.5(g)) the corresponding hub trajectory has some movement within the airport. This observation reveals that the airport is a hot spot (with respect to encounters of moving entities) in Milan’s city. Beyond this, we can observe that ring highways appear on six of the seven hub trajectories, which reinforces that this is a road that is very much used by vehicles. Indeed, several analyses can be done by using hub trajectories, such as comparing several hub trajectories of the same trajectory network, or analyze relevant parts of hub trajectories. Due to time and space limitations we left this investigation for future work.

4.4 Conclusion

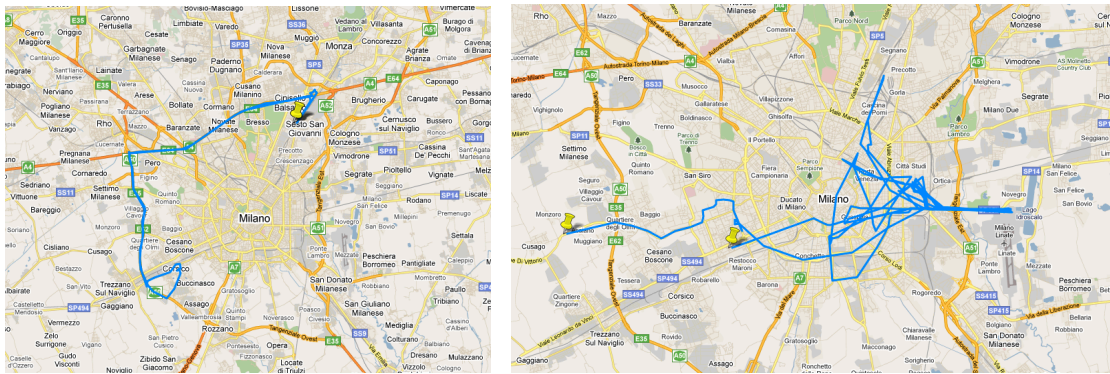
In this chapter we have presented a theoretical analysis and experimental results for moving object trajectories on complex networks. Our motivation comes from the desire to understand the influence of moving objects trajectories interaction on the traffic dynamics. We have defined a method for devising trajectory network from a dataset of moving object trajectories. In addition, we have built 28 trajectory networks from a real dataset of trajectories of vehicles. We have computed three network features (i.e. clustering coefficient, average shortest path length and diameter), for each trajectory networks and compare them.

Our analysis reveals that all trajectory networks are scale free network, presenting small world and power law features. Our results have practical implications for investigating moving objects interactions from complex network perspective. Although we have provided basic methods for building trajectory networks and analyzing their features, future investigation is needed in order to define how to interpret such features taking into account the application domain knowledge.

Comparing to existing data mining and statistical methods, our proposed approach provides another method for analyzing trajectories from the potential *interaction* perspective. Besides, complex network technique is adequate to analyze relationships among a large set of entities by computing topological features of the graph. Although building a trajectory network is time consuming, computing its properties is not. Thus, we believe that this technique can open new opportunities in mining the network structure of interactions between a large number of moving object trajectories.

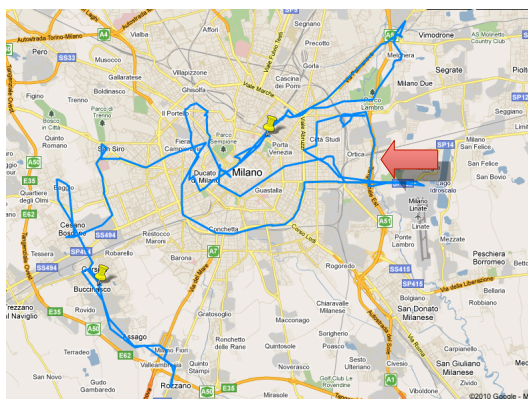
Our analysis can be performed on other applications as well. Study of football players, for instance, to identify the players that move near their opponents to block their game. Other applications go towards hospital environment, where doctors and patients could wear GPS-enabled devices or chips to collect their mobility and, then, to analysis the risk of contagious of the network built by their encounters, that is, if a doctor, that could be in touch with patients with contagious disease, encounters many others, doctors and patients, and, consequently, might spread these diseases.

The future research focus is on further analyzing the interactions between trajectories and space (i.e. landmarks, point of interest,etc), or between trajectories and time (i.e. hush hours, weekend, etc), or between trajectories and events (i.e. soccer match, festival, etc), to name a few.

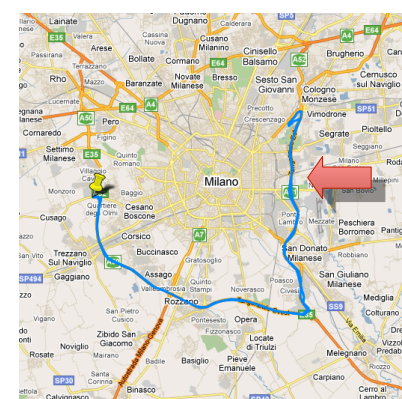


(a) Sunday

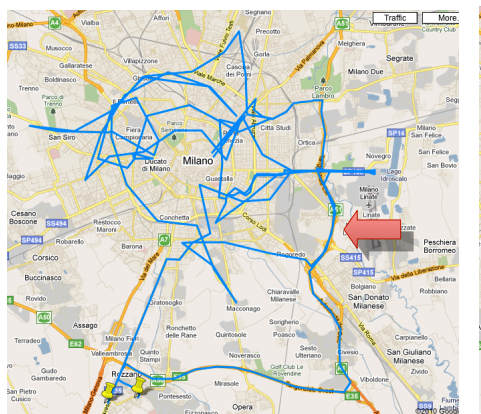
(b) Monday



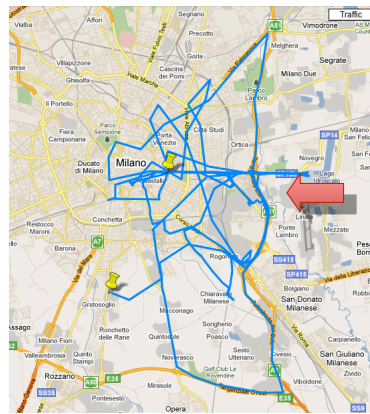
(c) Tuesday



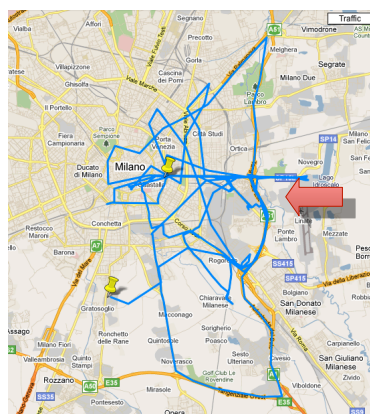
(d) Wednesday



(e) Thursday



(f) Friday



(g) Saturday

Figure 4.5: Trajectory plot of the days of the week

CHAPTER 5

COMETOGETHER: discovering communities of places in mobility data

People live in an environment where they move from one place to another. Therefore “places” are not only “static geographical objects” but they are also part of people lives. There is a two-way relationship between how the movements of people are affected by the location of places of interest, and how the places themselves are characterized and connected by the mobility of people. The way people move towards these places and the way they visit these places affects the overall movements, or mobility, of the environment. But the other way around is also true: city places like the Points of Interests (or POIs such as shops, restaurants, banks, hospitals and any locations that can be of interest for individuals) can be featured based on how people globally access them.

In this chapter is presented a new perspective on observing how places are connected based on the mobility among them in a urban context. We believe that just counting, for example, the number of visits of a given place, although certainly giving a measure of the attractiveness of that place, is not enough to get a deep understanding on how that particular place is “lived by” people and how this place “relates” to other places. Are there “communities” of places characterized by common mobility? In other words, which is the relationship between two or more places in terms of the mobility that connects them? Do people tend to visit places in the same communities? Are these communities related to mobility issues like traffic congestion or public transportation optimization?

These are only a few examples of questions that inspired our research. To answer these questions we need to analyze the city places from a “mobility” point of view. In other words, we need to build connections between places based on the trajectories of people accessing them. We believe that this perspective could be complementary to state-of-the-art mobility analysis techniques and furthermore it could improve the understanding of how places in a city are connected. Indeed, not necessarily spatially close POIs belong to the same community, as we will see in Section 5.4. Several applications can take advantage of this analysis ranging from traffic management to advertising, but also to municipality administration issues or human behavior studies.

In this chapter we face this intricate problem of relating places with the mobility at a global scale proposing a complex network framework to analyze the POIs in relation to the mobility of people accessing them. This paradigm gives a vision of the interrelation of places with the trajectories visiting them that is not explicitly faced by "standard" spatio-temporal analysis methods, as discussed in Section 5.1. Here, we concentrated on the *community* aspect, a well known analysis method in complex networks: POIs are grouped together based on the common trajectories that visited them.

The contributions present in this chapter are twofolds: on one hand, we propose a methodology for building a complex network combining Points of Interests and traces of people movements, from which we build communities of POIs. On other hand, we also experimented this methodology in a real case study where trajectories are collected from private cars traveling in a city and Points of Interest are downloaded from the Web. We found different kinds of communities (e.g. *compact* where the movements are mainly inside the community, or *bridge* where the movements tend to connect two other communities). We discuss the possible exploitation of these results in the mobility and advertisement application fields.

This chapter is organized as follows. Section 5.1 shows the novelty of the approach comparing the present approach to some related works in the field of mobility and complex network and in mobility data mining. Section 5.2 introduces some concepts used in the methodology introduced in Section 5.3. Section 5.4 reports on the experimental results using a real dataset. Section 5.5 contains the conclusions and describes future works.

5.1 Related Work

In this chapter we offer a new perspective in understanding human mobility in terms of finding the communities of places based on the user movements they share. To this end, we build a complex network from POIs connected by trajectory data. In Chapter 2, we have introduced a trajectory as the spatio-temporal evolution of a moving object by [Spaccapietra et al., 2008]. This evolution is typically represented as a sequence of positional observations represented by x and y coordinates of time-stamped sample points as collected by a tracking device, such as GPS tools or WIFI sensors. Trajectories

representing the movement evolution of individuals has witnessed an increasing interest in the last decade, especially due to the increasing availability of personal tracking device, ranging from GSM phone to the more sophisticated GPS-enabled smartphones. Mobility analysis has become a hot research topic since several methods on data mining and statistical techniques, tailored to trajectory data, have been proposed in the literature, [Giannotti & Pedreschi, 2008, Giannotti et al., 2011, Zheng & Xie, 2010] to mention a few.

The task of analyzing large trajectory datasets can be carried out towards three different directions. First, basic statistics may be applied to trajectory data mainly to discover the distributions of people presence and origin-destination matrices [Calabrese et al., 2010]; other studies focus on trajectory data mining aiming at finding correlations in large datasets of positioning data [Giannotti & Pedreschi, 2008]. Techniques to extract movement patterns include: (1) clustering discovery - finding groups of objects moving together; (2) sequential pattern discovery - finding the most frequent sequences of places visited; (3) flock detection - extracting the convergence of people moving together for a certain amount of time [Dodge et al., 2008, Giannotti & Pedreschi, 2008, Wachowicz et al., 2011]. These techniques are based on the geometric properties of trajectories thus trying to extract similarities or common behavior from the spatio-temporal dimension of the data. The connection to the places that people access during their movements is not explicitly taken into account during the mining task. The concept of *semantic trajectory*, Section 2.1.2, [Spaccapietra et al., 2008] as a sequence of stops (locations associated to the absence of movement) and moves (where the object is actually moving) is a first step in including places of interest visited by the user into the trajectory definition. The POIs are associated to stops and thus they are embedded into the semantic trajectory definition [Rocha et al., 2010, Yan et al., 2011]. Later, data mining algorithms are applied to discover the most frequent/sequential patterns [Alvares et al., 2007a]. In these approaches the POIs are linked to the stops of a single trajectory, but there are no explicit connections between the POIs and the trajectories at a global scale. Therefore, what is missing in these lines of approaches is a global perspective of the connection between the POIs based on the mobility of people accessing them.

As we have seen, the specific aspect of understanding how the objects interact at a global scale is usually associated to the paradigm of complex networks. The study of networks, or Network Science, is broadly interdisciplinary and important developments have occurred in many fields, including mathematics, physics, computer and information sciences, biology, and the social sciences [Newman, 2010] and have been receiving increasing attention by the scientific community, Inspired by real-world scenarios such as social networks [Aiello et al., 2000, Castro & Grossman, 1999], technology networks [Adamic et al., 2001], the World Wide Web [Leskovec et al., 2010, Donato, 2010], biological networks [Jeong et al., 2001, Jeong et al., 2000], and human movement [González et al., 2008, Wang et al., 2009] the last few years have seen a wide, multidisciplinary, and extensive research devoted to the extraction of non trivial knowledge from such networks. Finding social interactions at a global scale is also in the scope of this research area. A typical example is the study of the spreading of cell phone viruses thru GSM phone calls

[Wang et al., 2009, Barabási & Albert, 1999]. However, to the best of our knowledge, the approach introduced in this chapter in community discovery from complex network of POIs based on the trajectories visiting them, has not been faced in previous work.

5.2 Background

As presented in 2, a network $G = (V, E)$ is an object in which entities (the nodes in V) are linked by ties (the edges in E), representing any sort of connection, similarity or interaction. Since networks are usually modeled by graphs, network analytics has focused to the characterization and measurement of local and global properties of such graphs, such as diameter, degree distribution, centrality, connectedness - up to more sophisticated discoveries based on graph mining, aimed at finding frequent subgraph patterns and analyzing the temporal evolution of a network.

As introduced in Section 3.3, a branch of complex network research has been focusing on the discovery of structures called *communities*. Communities are groups of nodes highly interactive, densely connected, or, more in general, highly similar, for a given definition of similarity between any two individuals.

Several approaches have been proposed so far to perform community discovery [Coscia et al., 2011]: from divisive graph partition algorithms, to random walk based approaches, from label propagation based methods, to clique percolation techniques. However, the literature is still missing a unique definition of the concept of community, and the diverse techniques lead all to different results, sometimes hard to compare to each other. Although a few measures of the quality of the results have been proposed so far (among which, the modularity), their definitions are still questionable (the modularity, for example, has a well known problem of resolution, and approaches that try to maximize it tend to create very large communities).

Some of the existing approaches for community detection focus on finding groups of nodes, while others put the links among entities at the center of the investigation. Since we are interested in analyzing movements between places visited by trajectories and in grouping places according to trajectories visiting them, we consider the edges as the main entities to be grouped. In addition, we also want to consider the possible overlap between different communities. Different places can, in fact, take part into more than a community, due to their role of spatial “bridges” between them.

The authors of [Ahn et al., 2010] proposed an algorithm detecting communities from the links and that considers the node overlapping. In this approach, the authors start from the assumption that whereas nodes belong to multiple groups (e.g. individuals have families, co-workers and friends), links often exist for one dominant reason (two people are in the same family, work together or have common interests). They define a similarity between two edges based on the Jaccard coefficient. Firstly, the inclusive neighbors of a node i is defined as:

$$n_+(i) = \{x | d(i, x) \leq 1\}, \quad (5.1)$$

where $d(i, x)$ is the length of the shortest path between nodes i and x . The set simply contains the node itself and its neighbors. This measure computes the ratio of nodes shared by two edges, or, in formula:

$$S(e_{ik}, e_{jk}) = \frac{|n_+(i) \cap n_+(j)|}{|n_+(i) \cup n_+(j)|}, \quad (5.2)$$

where e_{ik} and e_{jk} are two links sharing the node k . With this similarity function, a hierarchical clustering algorithm is performed to build a dendrogram where each leaf is a link from the original network and branches represent link communities. In addition, in this dendrogram, links occupy unique positions whereas nodes occupy multiple positions, owing to their links. After obtaining this dendrogram, it is necessary to find the best way to cut it. The best height is found thanks to the usage of a natural objective function, the partition density, D , based on link density inside communities.

For a network with M links, $P = \{P_1, \dots, P_c\}$ is a partition of the links into C subsets. P_c has $m_c = |P_c|$ links and $n_c = |\bigcup_{e_{ij} \in P_c} \{i, j\}|$ nodes. Therefore, the density of a partition P_c is

$$D_c = \frac{m_c - (n_c - 1)}{n_c(n_c - 1)/2 - (n_c - 1)}, \quad (5.3)$$

and the partition density, D , is the average of D_c , weighted by the fraction of present links:

$$D = \frac{2}{M} \sum_c m_c \frac{m_c - (n_c - 1)}{(n_c - 2)(n_c - 1)} \quad (5.4)$$

The equation 5.2, however, does not consider the weights of the links as metric of similarity. Thus, we use the generalization of Jaccard coefficient, i. e. the Tanimoto coefficient. Let $\mathbf{a}_i = (\tilde{A}_{i1}, \dots, \tilde{A}_{iN})$ with

$$\tilde{A}_{iN} = \frac{1}{k_i} \sum_{i' \in n(i)} w_{ii'} \delta_{ij} + w_{ij}, \quad (5.5)$$

where w_{ij} is the weight on edge e_{ij} , $n(i) = \{j | w_{ij} > 0\}$ is the set of all neighbors of node i , $k_i = |n(i)|$, and $\delta_{ij} = 1$ if $i = j$ and zero otherwise. The similarity between edges e_{ik} and e_{jk} is:

$$S(e_{ik}, e_{jk}) = \frac{\mathbf{a}_i \cdot \mathbf{a}_j}{|\mathbf{a}_i|^2 + |\mathbf{a}_j|^2 - \mathbf{a}_i \cdot \mathbf{a}_j} \quad (5.6)$$

Therefore, we are able to compute the link similarity by considering their weights, i. e., the number of trajectories or users carried by the links between two places.

5.3 Problem Definition and Methodology

In this section we introduce the problem definition and the methodology to build the complex network of POIs. However, before going into the details of the complex network

construction algorithm, let us introduce the concept of Point of Interest.

Definition 5.1 *A Point of Interest (POI) is a geographical object that is interesting for a specific application, usually associated to a human activity. Formally, we define a POI as a triple $POI = (c, r, l)$ where c is the representative spatial point, r is the spatial area representing the extent of the object and l is the label of the form $cat:n$ where cat is the category of the POI and n is the POI name.*

An example of POI is the Eiffel Tour: the representative spatial point c is the center of the tower, the extent is the area covered by the base of the tower and the label is the category (which can be, for example, “tourist attraction” or “monument” or “tower”, depending on the application) and the name “Eiffel Tour”

The starting point of our process is the set of user position observations. Therefore, we define the mobility history of a single user as:

Definition 5.2 (User Mobility History) *Given a set of user’s observations D_u , the user’s history is defined as an ordered sequence of spatio-temporal points $H_u = \langle p_1 \dots p_n \rangle$ where $p_i \in D_u$, $p_i = (x_i, y_i, t_i)$, x_i, y_i are spatial coordinates, t_i is an absolute timepoint and $\forall(i, j)t_i \leq t_j$ holds.*

Problem Definition Given m traced moving users, a set of Points of Interest (POIs) V and the dataset collecting the users’ histories: $D = \{D_1, \dots, D_m\}$, we want to group the POIs in V into groups (or communities) connected by the common mobility of the users.

To solve this problem, we must overcome the limitations of standard methods of grouping locations like spatial clustering, which is based only on the geographical aspect, to move towards a communities perspective where POIs are grouped by the mobility of the users.

The proposed methodology combines different aspects of mobility and graph analysis and it is composed of two main steps: the first step *builds a network* where each link represents the *relations* between two POIs in terms of mobility; the second step *extracts the communities* that identify groups of POIs which share a common mobility context. Furthermore, we define some measures to evaluate and compare the discovered communities. These steps are illustrated below.

5.3.1 Building the Network

The network is composed of a set of nodes which correspond to the set of POIs where the moving users stopped to perform some activity. In order to find these POIs we need to first distinguish the single *trajectory* as the part of the user history representing the

movement associated to a specific activity, such as *going to work*, *shopping* etc. In order to distinguish between the different trajectories in a user history, we need to detect when a user stops for a long time so that this stop can be considered the end of that particular trajectory and the beginning of the next one. Section 2.3 introduces some methods present in the literature.

However, for computational efficiency reasons here we propose a different method as a trade off between precision and efficiency. We search the points that change only in time. i.e. points that stays in the same spatial position for a certain amount of time quantified by the temporal threshold *MinStopTime*. A spatial threshold *MaxStopArea* is used to remove both the noise introduced by the imprecision of the device and the small movements that are of no interest for a particular analysis. These thresholds are used for detecting the *candidate stops* as defined below, where *area()* is a function computing the size of the minimal convex region including a set of points and \preceq is the operator of sequential inclusion without gaps.

Definition 5.3 (User’s candidates stops) *Given the user history H_u , we define the sequence of candidate stops $S_u = \langle s_1 \dots s_m \rangle : s_k = (a, t, d)$, $a = \langle p_i, \dots p_j \rangle \preceq H_u$, $area(a) \leq MaxStopArea$, $t = p_i.t$, $d = p_j.t - p_i.t \geq MinStopTime$.*

From this set of candidates we want to build the set of user trajectories by removing the cases of slow movements or long stops in a place. Examples of this long stops may be the home and the work places since usually users spend the night at home and the day at work. For this reason we use a threshold called *MaxMoveTime* to break the user history into distinct trajectories. This trajectory partitioning step is presented in [Zheng & Xie, 2010]. Therefore, we define the user trajectories set as follows, where *contains* is a spatial inclusion predicate between two spatial regions:

Definition 5.4 (User’s trajectories set) *Given a set of Points of Interest V and given the sequence of candidates stops S_u for the user u , we define the user’s trajectories set as $T_u = \{t_1, \dots, t_h\}$ where each trajectory is the maximal sequence $t = \langle v_1.l, \dots, v_k.l \rangle : \forall_{(i,j)}, 1 \leq i < j \leq k, \exists_{(w,q)}(s_w, s_q) \preceq S_u$, $contains(v_i.r, area(s_w.a))$, $contains(v_j.r, area(s_q.a))$ and $s_w.t - s_q.t \leq MaxMoveTime$.*

Having all the trajectories of all the users $T = \bigcup_{1 \dots m}^u T_u$, we compute the POIs network as:

Definition 5.5 (Points of Interest (POI) network) *Given a set of POIs V and a set of users trajectories T , we build the points of interest network $P_{oi}^N = (V, E, W)$ where $E = \{e_{i,j} : \exists t \in T, \langle v_i, v_j \rangle \preceq t\}$ and $W = \{w_{i,j} : w_{i,j} = |\{t_1, \dots, t_m\}|, \langle v_i, v_j \rangle \preceq t\}$.*

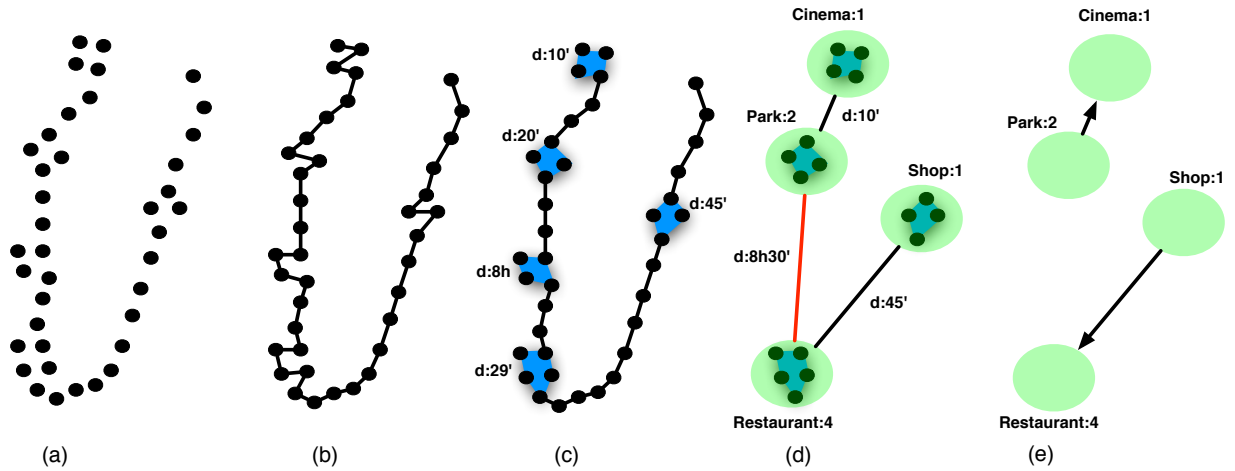


Figure 5.1: The building process of places network from one user history: From positional observations in (a) to the user history in (b), the candidates stops in (c). The trajectories set in shown in (d) where a move of duration of 8h30' (thus exceeding 4 hrs) splits the user history into two trajectories. The POI network is depicted in (e)

In other words, the POI network is a directed and weighted graph which summarizes all the trajectories of the users and each edge is weighted by the number of trajectories which share the movement between the same pair of POIs.

An example of the stop computation process is presented in Figure 5.1. In Figure 5.1(a) the set of positional observation of a user are shown. The process starts building the user history as a continuous sequence of points ordered by time Fig.5.1(b). In Fig.5.1(c) the stops are identified considering $MaxStopArea = 50m^2$ and $MinStopTime = 30$ minutes. Then the stops are spatially intersected with the set of POIs V as shown in in Fig.5.1(d): the red edges between the two stops has a duration which is grater than the $MaxMoveTime$ (e.g. 4 hours) therefore it is removed cutting the user history into two trajectories. Finally, the two trajectories will contribute to the edges shown in Fig.5.1(e) where w and w' are the number of trajectories which share the same path respectively $Shop : 1 \rightarrow Restaurant : 4$ and $Park : 2 \rightarrow Cinema : 1$.

The network building process is summarized by the pseudo-code of the algorithm Points of Interest Network Builder: Algorithm 5.1.

5.3.2 Communities of Points of Interests

Having the POIs network P_{oi}^N we can identify communities of POIs that are grouped based on the movements between them. This can be done using the state-of-the-art algorithm [Ahn et al., 2010] presented in the section 5.2, thus obtaining a set of communities $C = \{C_1, \dots, C_n\}$ where each community is a subgraph of P_{oi}^N . Moreover, in order to evaluate the quality of discovered communities, we introduce three measures: the *Nodes similarity* measuring how similar communities are based on the nodes shared by them; the *Trajectories similarity* giving a measure of how the communities are similar from the point of view of the trajectories which pass through their edges; and the *Compactness*

Algorithm 5.1 Points of Interest Network Builder

Input: A set of positional observations D , a set of POIs V , a temporal threshold $MinStopTime$ as $S_{minTime}$, spatial threshold $MaxStopArea$ as $S_{maxArea}$ for stop detection and a temporal threshold $MaxMoveTime$ as $M_{maxTime}$ for creating users' trajectories

Output: A points of interest network $P_{oi}^N = (V, E, W)$

1. $P_{oi}^N.V \leftarrow V$
 2. $P_{oi}^N.E \leftarrow \emptyset$
 3. $P_{oi}^N.W \leftarrow \emptyset$
 4. **for each** $D_u \in D$ **do**
 5. // create users' history
 6. $H_u \leftarrow \text{userHistory}(D_u)$
 7. // identify candidate stops
 8. $S_u \leftarrow \text{candidateStop}(H_u, S_{minTime}, S_{maxArea})$
 9. // create users' trajectories
 10. $T_u \leftarrow \text{userTrajectorySet}(S_u, M_{maxTime})$
 11. **for each** $t \in T_u$ **do**
 12. **for each** $\langle v_i, v_j \rangle \preceq t$ **do**
 13. // create edge e_{ij}
 14. $e_{ij} = \langle v_i, v_j \rangle$
 15. $P_{oi}^N.E \leftarrow P_{oi}^N.E \cup e_{ij}$
 16. // update the weight w_{ij} of the edge e_{ij}
 17. update w_{ij} in $P_{oi}^N.W$
 18. **end for**
 19. **end for**
 20. **end for**
 21. **return** P_{oi}^N
-

measuring how much the trajectories creating a community move inside the community itself. Formally:

$$Similarity_{Node}(C_i, C_j) = \frac{|V_i \cap V_j|}{|V_i|}. \quad (5.7)$$

$$Similarity_{Traj}(C_i, C_j) = \frac{|T(C_i) \cap T(C_j)|}{|T(C_i)|}. \quad (5.8)$$

$$Compactness(C_i) = \frac{|E_i|}{|distinct(T(C_i))|}. \quad (5.9)$$

where $T(C_k)$ is the set of trajectories traversing a community C_k and $distinct(T)$ set of edges traversed by the set of trajectories T .

5.4 Case Study

In this section we present the experiments carried out using a real trajectory dataset and a set of POI existing in the geographical area of the movements. Furthermore, we analyze and evaluate the generated network according to the analysis and measurements presented in Section 5.3. In our experiments we use a set of positional observations collected by an Italian insurance company which offers a discount to the users who have an embedded GPS device in their car. The set of collected observations in one week in Milan (Italy) is composed by 1,806,293 points for 17,087 users in the Milan area. The POIs dataset of Milano has been downloaded from the web (OpenStreetMap [OpenStreetMap, 2011]) obtaining a set of 2501 locations corresponding to commonly used POIs semantic categories such as banks, restaurants, cinemas, theaters, museums, etc.

According to the methodology defined in Section 5.3, from the set of positional observations we computed the set of candidate stops for each users. The parameters used are 20 minutes as *MinStopTime* and 150m² as *MaxStopArea* (i.e. a car with speed less than 0.5 km/h) thus obtaining 216,523 candidate stops. Due to the fact that (1) It is possible to retrieve the representative point of the POIs, but not the precise extent and (2) the observations refer to the position of the car and not the user himself, we use an approximated area around the POIs of 150m. However a single stop matches since the POIs are very close (e.g. an open mall or the city center). To solve this problem we propose to perform a preprocessing step to group together the POIs that match a stop thus defining a *composite POI* represented by the union of all the extent of the close POIs. This has been done with spatial clustering (i.e. T-Optics [Andrienko et al., 2009]) and each cluster will be handled as a single POI for the purpose of the network construction. The number of POIs after this clustering process is 347 for the single POI and 77 composite POI.

Having the candidates stops computed as described in Section 5.3 with the

specified thresholds and the set of POIs, we can build the trajectory sets and then the POIs network. For this steps we used a *MaxMoveTime* of 4 hours obtaining a network with 77 nodes and 677 links corresponding to the movements of the trajectories between the POIs. Figure 5.2 illustrates the generated POI network in City of Milan.

5.4.1 POI Network Characteristics



Figure 5.2: The plot of the POIs network generated from our experiments with 77 nodes and 677 edges: nodes represent the composite POIs; and edges represent the movement of users' trajectories between the nodes

Figure 5.3 shows the distribution of the edge weights, which represents the number of trajectories. In Figure 5.3 we observe that the distribution follows a power law, i.e. there are few edges with a large number of trajectories while there is a large number of edges with a small number of trajectories. Intuitively we can conclude that few composite POIs are very popular, having many movements through them..

In order to understand the characteristics of the generated POI network, we computed the following network measures, widely used in standard complex network analysis: clustering coefficient, average shortest path and diameter. The clustering coefficient of P_{oi}^N is 0.329, the average shortest path length is 2.584, the diameter is 7. These results are similar to the ones found in many real world networks, such as biological networks, social networks, and citation networks [Newman, 2003]. They highlights a small world phenomenon where there is a high clustering coefficient and a small average shortest path.

Another interesting analysis is to understand how composite POIs are related, which may help in answering interesting questions such as: do the highly connected composite POIs preferentially connect other high-degree composite POIs? Or do they

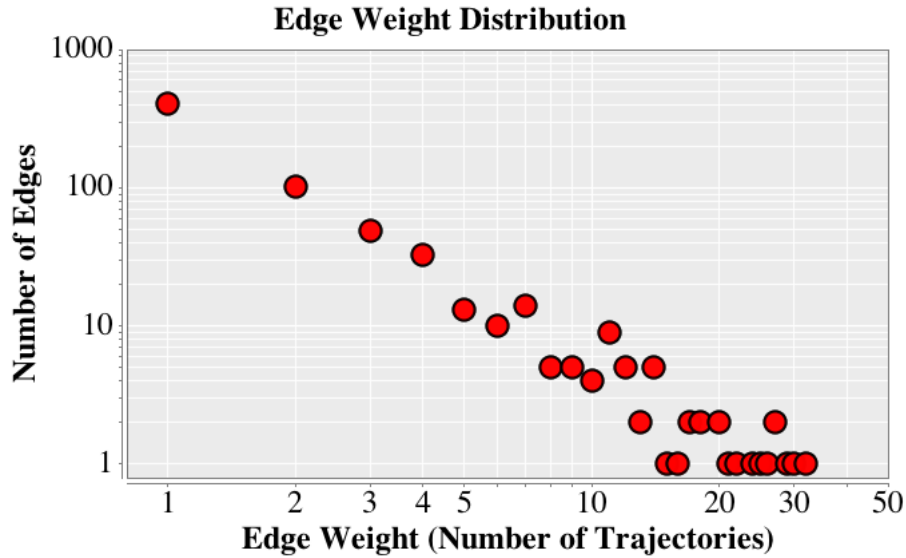


Figure 5.3: Edge weight distribution of the network P_{oi}^N

prefer to connect to low-degree ones? In other words, are we facing degree correlation? Degree correlation is a special case of *assortative mixing* of node degree [Newman, 2002, Newman, 2003]. The correlation r is calculated by means of Pearson correlation between of the nodes at either end of the edges. Table 5.1 shows the correlation r . As we see, in Table 5.1(a) nodes with high in-degree tend to connect to nodes with high in-degree; in Table 5.1(b) high in-degree nodes tend to connect high out-degree nodes; in Table 5.1(c) we see high out-degree nodes connecting to high in-degree nodes; Table 5.1(d) shows nodes with high out-degree connecting to nodes with high out-degree too. Summing up, Table 5.1 shows that the network is generally assortative.

When we compare this result to the literature, we can see that social networks tend also to be assortative, whereas other networks seem to be disassortative [Newman, 2002]. Together with the above, this means that the obtained network has typical characteristics of a social network. Indeed, a point of interest network is generated from the users' trajectories which may represent social aspects of the users, such as the places where the users perform some activities. To exemplify the concept of assortativity, we can take as an example the node 135 from the obtained network. This node represents a composite points of interest, including a coffee bar, a fast-food restaurant and a parking space. Its in-degree is 11 and its out-degree is 10. It connects to nodes whose average in-degree is 15.4 and out-degree is 17.2, and receives connection from nodes whose average in-degree is 14.8 and out-degree is 17.2. This example reflects the behavior of people in the real world where people tend to move from a popular place to another popular one.

5.4.2 Communities Analysis

Discovering communities structures within a complex network is the key for finding tightly connected groups of nodes. In our scenario, communities represent places that are at-

Table 5.1: Degree correlation of P_{oi}^N

	source	target	r
(a)	in-degree	in-degree	0.3715
(b)	in-degree	out-degree	0.3533
(c)	out-degree	in-degree	0.3375
(d)	out-degree	out-degree	0.3126

tended together, in a social network communities represent individuals belonging to social communities, while communities in genetic networks may be associated to functional modules. Thus, community discovery is a powerful tool for understanding the functioning of the network [Boccaletti et al., 2006].

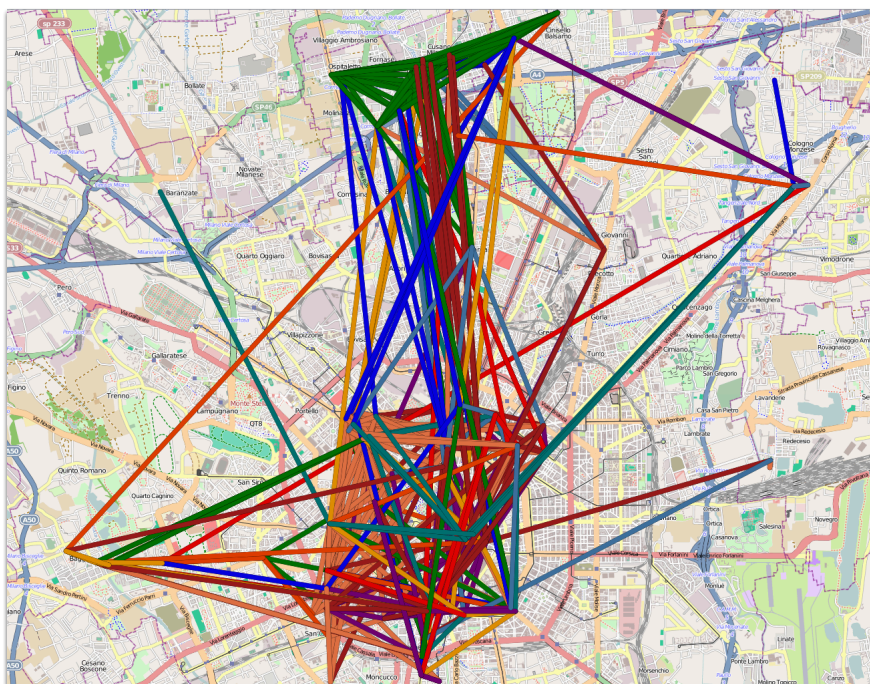


Figure 5.4: The 109 communities discovered from the POIs network. The edge color identify the different communities

After executing the community detection algorithm described in Section 5.2, we obtained 109 communities from the generated POI network. Figure 5.4 illustrates P_{oi}^N in which the colors of the edges identify each discovered community. Furthermore, Figure 5.5 shows the community size distribution considering the number of edges. It is worth noting that a number of communities have only one edge while few communities have a large number of edges. Indeed, small communities can be formed by a single movement between two composite POIs while large communities require a large number of distinct composite POIs and movement among them. Moreover, a set of composite POIs that participate in large communities may form small community among them.

Since communities are defined from trajectories, it is important to understand

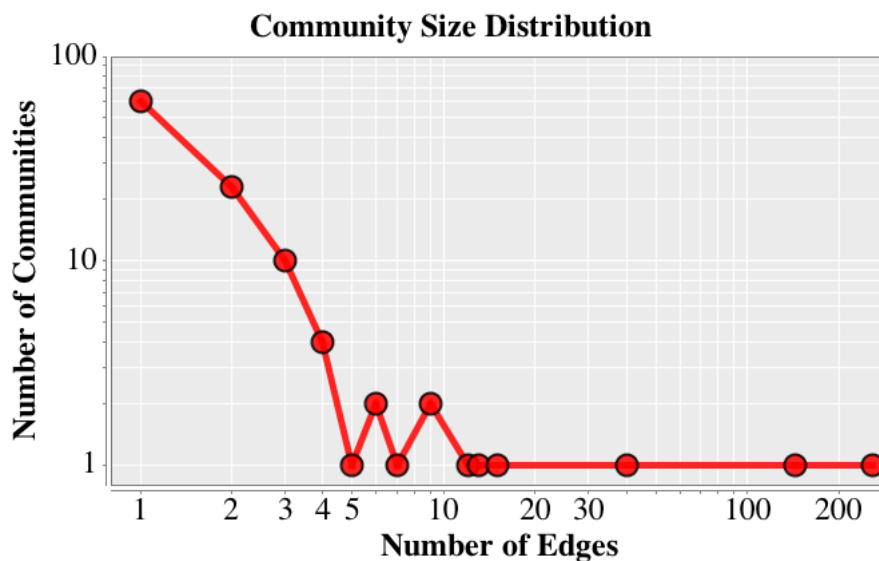


Figure 5.5: Community size distribution considering number of edges. Many communities are formed by a few edges, whereas a few communities are composed by a higher number of edges

the relationship between them. Do trajectories that define a community tend to have their moves on the edges of that community or they also have some moves in another community? From this question we analyse the community by means of *Compactness* measure defined in Section 5.3.

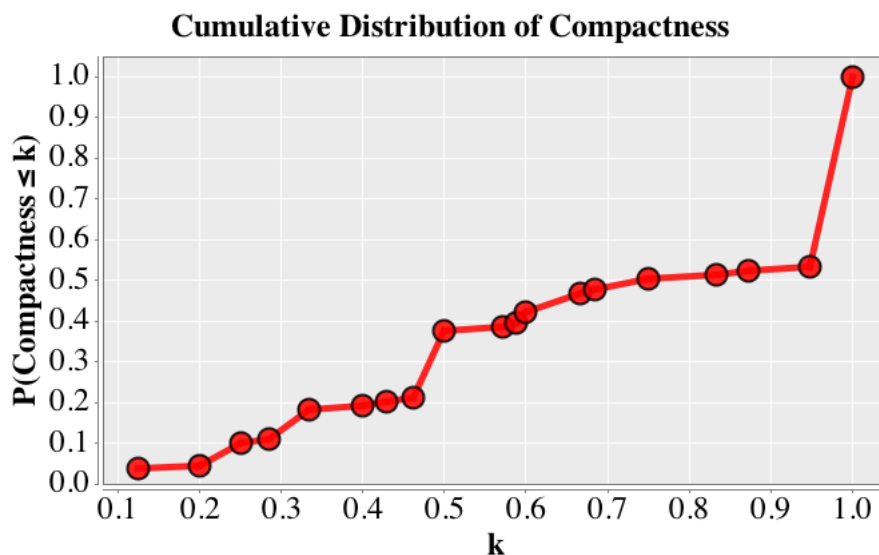
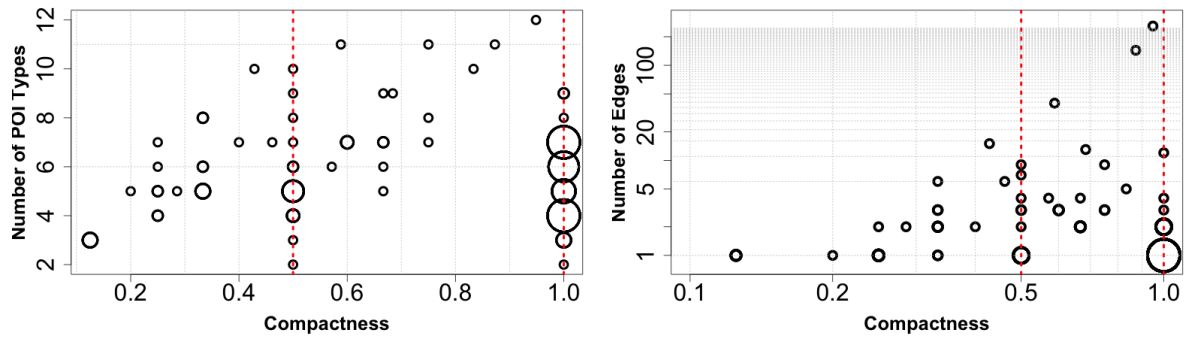


Figure 5.6: Cumulative Distribution of *Compactness*. $P(\text{Compactness} \leq k)$ indicates the probability that *Compactness* takes on a value less than or equal to k

Figure 5.6 shows the *Compactness* cumulative distribution. This plot illustrates these measures in the discovered communities. As we can notice, there is an evident gap at 0.5. It suggests two different behaviors of *Compactness*: one for the interval $(0, 0.5)$ corresponding to less compact communities and other for the interval $(0.5, 1.0)$ related to the communities more compact. Looking into these two groups, we analyze what may

influence the *Compactness* such as number of edges and number of types of places of interest involved into the community. Then, we analyze the correlation between some community properties and the *Compactness* value. The analysis are presented in the following and depicted in Figure 5.7.

For the interval $(0, 0.5)$ we can notice that small communities seem to be less compact, since the trajectories tend to go towards outside the community. In fact, the average community size in this interval is 2.4782 edges. Nonetheless, what may influence the *Compactness*? We have found that the highest correlation of *Compactness* was with the number of POI types of 0.7874 (Figure 5.7(a)). This means that, for this interval, communities tend not to be so large and that communities with more POI types tend to be more compact. In other words, we could interpret this as the trajectories that tend to keep themselves inside the community since this community present different POI type.



(a) Correlation between *Compactness* and the number of type of points of interest. For the interval $(0, 0.5)$ the correlation is 0.7874, and for $(0.5, 1)$ the correlation is 0.6371

(b) Correlation between *Compactness* and the number of edges (logarithmic scale). For the interval $(0, 0.5)$ the correlation is 0.5741, and for $(0.5, 1)$ the correlation is 0.7257

Figure 5.7: Correlation

The situation is changed for the interval $(0.5, 1.0)$ since, in this case, the communities tend to be larger with an average of 29.4705 edges. There is still a correlation between compactness and the POIs categories since the values is 0.6371. However, the highest correlation found for this interval was between *Compactness* and the number of edges, corresponding to 0.7257 (Figure 5.7(b)). Therefore, in this case the edges contributes for larger values in the compactness of the community. This means that the communities tend to be larger in number of POIs and, consequently, the trajectories tend to remain inside the community.

As consequence of the previous analysis we discovered two interesting types of communities: (i) the big communities with a large number of edges which cover multiple types of POIs, thus becoming compact, and (ii) the communities which are not large but since they cover several POIs categories they tend to form a compact structure. In the following we focus our attention on these two types of communities.

5.4.3 Large Communities

In this section, we focus on the top three larger communities with respect to the number of edges (72, 20 and 25). In Figure 5.8 we can see how they are distributed in space and how the communities 20 and 25 are interconnected since they share some nodes of the network while community 72 is very well separated. In fact, these three communities highlight how the center of the city is essentially divided into two major communities of POIs. This is confirmed by the Table 5.2a and 5.2b where the $Similarity_{Node}$ is 50% and 28% but the $Similarity_{Traj}$ is only 4% and 1% highlighting that only few users use both the communities. The community 72 is in a peripheral area of the city and describes a new gravitational point for the activities of the people. From Tables 5.2(a) and 5.2(b) we can see that it is completely separated from the others considering both nodes and trajectories.

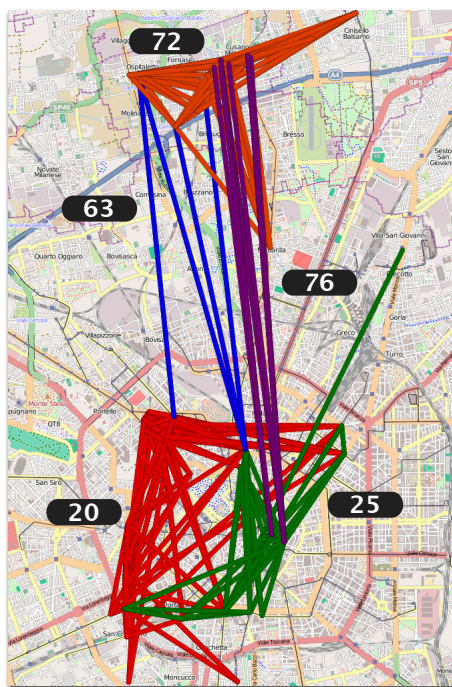


Figure 5.8: The selected communities: the three largest communities by the number of edges are 72, 20, 25; and the two communities 76 and 63 act like a "bridge" between them characterizing the movement between two regions of the city

Looking at the other communities discovered in Figure 5.4 we note that there exists *bridge* communities, which connect this three large communities. Intuitively, a *bridge* community C_b is a community that shares nodes with two other communities C_i and C_j such that C_i and C_j do not share nodes with each other. Including them in the analysis (communities 63 and 76 in Figure 5.8) we can understand how they connect the center of the city with the peripheral area. We computed the similarity measure and we discovered that they share a large percentage of nodes and trajectories among them.

The results of such analysis could very valuable for a broad range of urban actors such as a mobility agency or an advertising company, which can understand the

Table 5.2: Similarities between the communities in Figure 5.8

	(a) $Similarity_{Node}$					(b) $Similarity_{Traj}$				
	20	25	72	63	76	20	25	72	63	76
20	–	0.28	0.00	0.08	0.00	20	–	0.01	0.00	0.00
25	0.50	–	0.00	0.07	0.14	25	0.04	–	0.00	0.02
72	0.00	0.00	–	0.15	0.35	72	0.00	0.00	–	0.007
63	0.40	0.20	0.60	–	0.00	63	0.00	0.25	0.50	–
76	0.00	0.22	0.77	0.00	–	76	0.00	0.00	0.66	0.00

dynamics and the interconnection of the city and be more accurate in their actions. For example, an advertising company could use this information to understand where to locate their advertising posters to optimize the spreading of the information to all the three major communities exploiting, for example, of the *bridges* communities. Hence, considering the $Similarity_{Traj}$ and $Similarity_{Node}$ between the large communities and the bridges, the better places to put the advertise posters are the shared nodes between community 20 and 25 and the shared nodes between these two communities and the *bridges*. Indeed, although communities 76 and 63 do not influence the central area, they share a large number of trajectories with community 72.

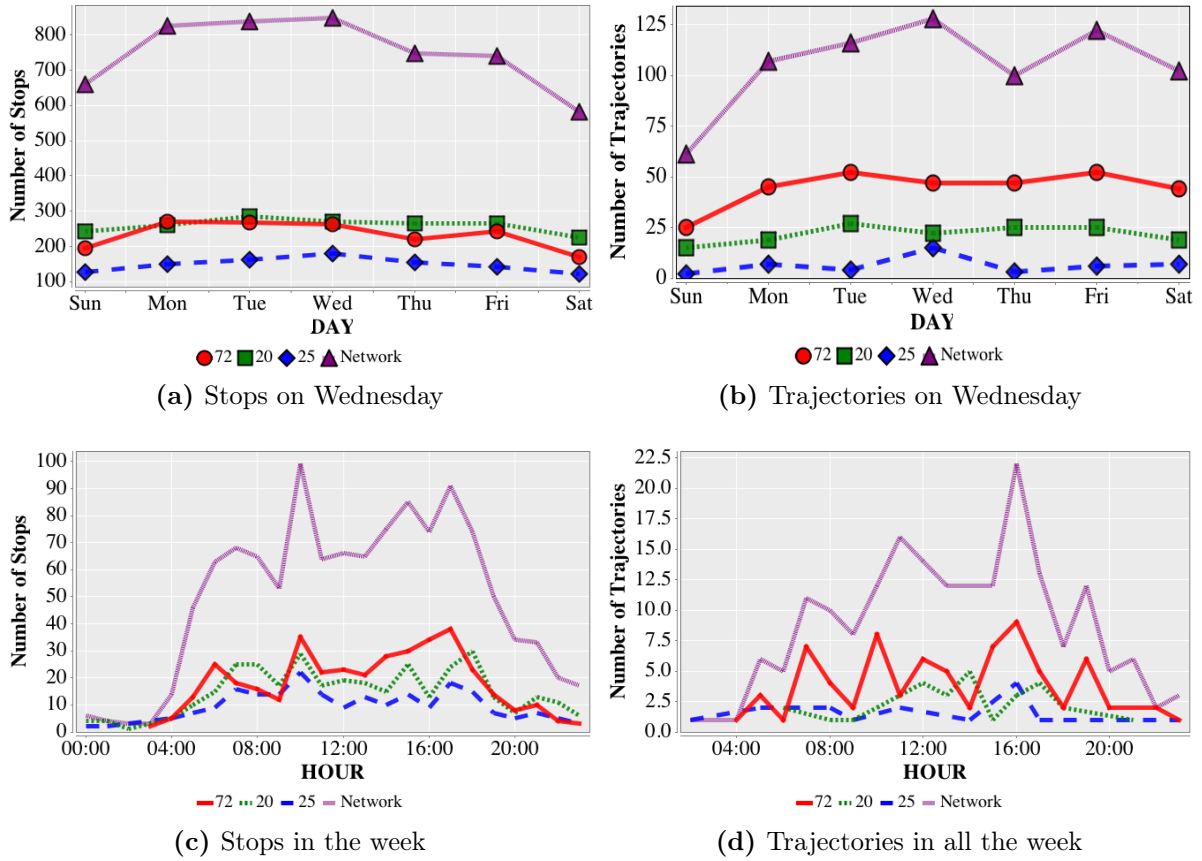


Figure 5.9: Temporal analysis of the network P_{oi}^N and communities 72, 20 and 25 showed in Figure 5.8

In the case of a mobility (or traffic) agency further information can be obtained by analyzing the temporal aspect of the communities. In fact, the communities change over time as shown in Figure 5.9. The communities are analyzed and compared to the entire network usage in terms of number of users who stop in a POI and the number of users who move between POIs is reported comparing them. The temporal analysis is performed using two different granularities, *days* and *hours*. We can observe how each community has its own distribution which follows the general behavior of the network. However, there are some specific periods where they clearly diverge. This result can be used by the mobility agency to better organize public security. For example, considering Figure 5.9(d), traffic agents could be allocated in community 72 at 16:00h (high movement) in order to guard movement among community POIs. It's important to notice how a simple spatial clustering is not sufficient to obtain this result since these kinds of algorithms tend to partition the space (and therefore the groups of POIs) not considering the mobility information.

5.4.4 Compact Communities

In this section we focus on *compact* communities, which are characterized by containing trajectories that tend to remain inside the community. As we have shown and discussed in Figure 5.7(a), there is an high correlation between the number of POI types and *compactness*. Here we focus on six compact communities selected from the intervals of compactness that have been discussed in Figures 5.7 and 5.6. We discover again the communities number 20 and 72 are among the most compact ones, but not the 25 (see Table 5.3).

Focusing on the other communities in Figure 5.10 (i.e. 13, 6, 43 and 86), we can notice that some of them seem to be similar to the larger communities: community 86 is high related to the community 25, sharing a large percentage of nodes and trajectories (respectively 77% and 40%). Moreover, if we consider the topology of the community, this suggests the presence of a central core connecting almost all the POIs of the large community. A different relation exists between the communities 6 and 20 where the percentage of shared nodes is 90% but the shared trajectories is 0%: this means that it represents a different community which uses the same POIs. This observation highlights the complexity of the mobility in a city and the method discussed in this paper is a further step in trying to get an understanding of the phenomena.

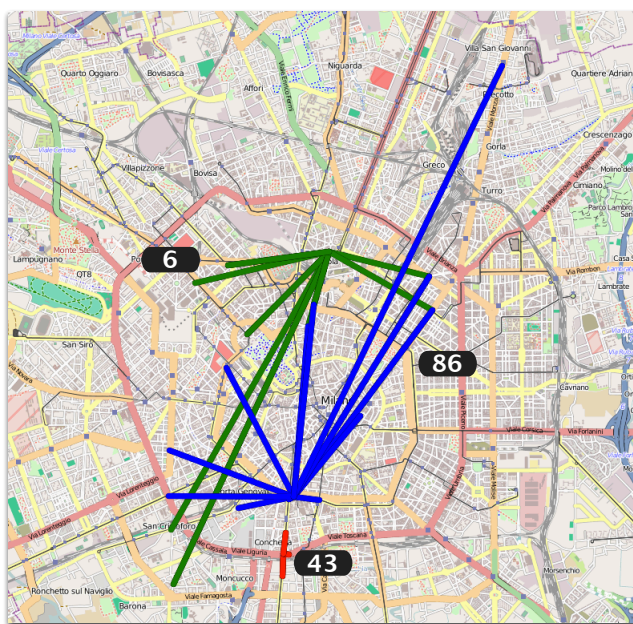


Figure 5.10: Communities to illustrate the measure *Compactness* considering different degrees of *compactness*: communities 104, 86 and 43 are less compact communities; 6 and 13 are more compact communities

Table 5.3: Compactness of the some communities of Figures 5.8 and 5.10

Community C_i	<i>Compactness</i> (C_j)
86	0.42
43	0.46
25	0.58
6	0.68
20	0.87
72	0.94

5.5 Conclusion

Unveiling the complexity of people movement is a challenging task, as witnessed by the recent increasing interest in the literature. Relating people movement with the places of interests visited by the travelling individual at a global scale is even more complicated since it combines the spatio-temporal dimension of the movements with the geographical and semantic aspect of the urban locations. In this paper, we proposed an explorative study on the relation between people mobility and Points of Interest at the global scale, based on the complex network paradigm. We presented an algorithm to build a complex network that combines locations that people visit with the mobility of users represented as trajectories.

From this network we computed the communities as the subgroups of Points of Interest related by the common users trajectories visiting them. An explorative analysis has been conducted in a real case study where a complex network has been built combining Point Of Interests with traces of moving cars in Milan, Italy, and communities of places grouped by common mobility are extracted. We defined some interesting features of these communities such as the compactness or the presence of "bridge" communities. We observed these measures discussing the possible interpretations in terms of applications such as traffic management or advertising.

Future works follow several directions. First of all, alternative ways of computing the stops and associating the POIs may be applied in order to better represent the actual activity of the user. Furthermore, we plan to extend this methodology to other real datasets to further validate the results. Naturally, we intend to investigate more in deep the possible applications that can be benefit from this analysis, for example going to the direction of POIs recommendation systems.

CHAPTER 6

MOBNET: a software tool to analyze mobility through complex network

The development of new methodologies to analyze data has provided the creation of new software tools to support specialist and non-specialist users in several different areas. Therefore, from the proposed methodologies in Chapter 4 and 5, we have developed a software tool to analyze mobility data using complex network techniques. This application is named MOBNET.

MOBNET aims at analyzing mobility data by means of complex network techniques, such as the graph representation, local and global properties and the methodologies proposed in Chapter 4 and 5. Furthermore, it also intends to present the results to the users to make decisions and interpretation, by visualizations the networks or communities on the map, for instance.

This chapter is organized as follows. Firstly, we present some software tools in mobility analysis and network analysis in Section 6.1 and 6.2, respectively, that can be useful for users that intend to understand these type of data. Afterwards, Section 6.3 introduces the software MOBNET, presenting an overview of this tool and concluding with the main capabilities present in MOBNET. Finally, Section 6.4 draws the conclusions.

6.1 Tools in Mobility Analysis

Due to growth of techniques and the availability of mobility data, tools become more and more necessary. These tools are important to help user at analyzing mobility data by applying several developed techniques, such as stops and moves detection, creation of trajectories and semantic trajectories. In this section we present two tools for mobility analysis: *Weka-STPM* and *M-Atlas*.

6.1.1 Weka-STPM

Weka, [Holmes et al., 1994], is a free and open source non-spatial data mining toolkit developed in Java. It has a non-spatial data preprocessing module named *weka.Explorer*, where data can be obtained from a database, a web site, or an arff file (specific format for Weka).

A great advantage in Weka is the possibility to create *modules* and aggregate them to the existing system. Hence, [Bogorny, 2011, Alvares et al., 2010] proposed a module for Weka called *STPM*. The module *STPM* is fully integrated into Weka in order to automatically access the database and add semantics to trajectory data. In addition, *Weka-STPM* is an extension of Weka for spatio-temporal data and it is interoperable with all databases constructed under Open GIS Consortium (OGC) specifications [OGC, 2008]. As a module of Weka, it allows the user to directly apply the several mining algorithms available in Weka to mine semantic trajectories.

The module *STPM* extends the Weka database connection interface. So, the database schema is provided by the user and *STMP* loads all geographic database tables to the boxes *Trajectory* and *Relevant Features*. This allows the user to choose the target trajectory table and the spatial feature types of interest. This spatial feature types of interest is related to the candidate stops used by the methods *SMoT* and *CB-SMoT* discussed in Section 2.3. Consequently, *STMP* offers both methods as tools to generate semantic trajectories, which receives as inputs a minimum time threshold, to consider a stop, and a spatial threshold that represents the size of a buffer that in turn is the zone around relevant features, represented by points or lines, to overcome spatial uncertainty. Figure 6.1 shows the interface of *STMP*.

Therefore, *Weka-STPM* is a developed module for Weka that works with Open GIS Consortium specification in order to generate semantic trajectories by identifying stops and moves provided by the methods *SMoT* and *CB-SMoT* and some relevant features. With the set of semantic trajectories, the user is able to achieve several data mining techniques available in Weka over the semantic trajectories, such as techniques of association rules and sequential patterns.

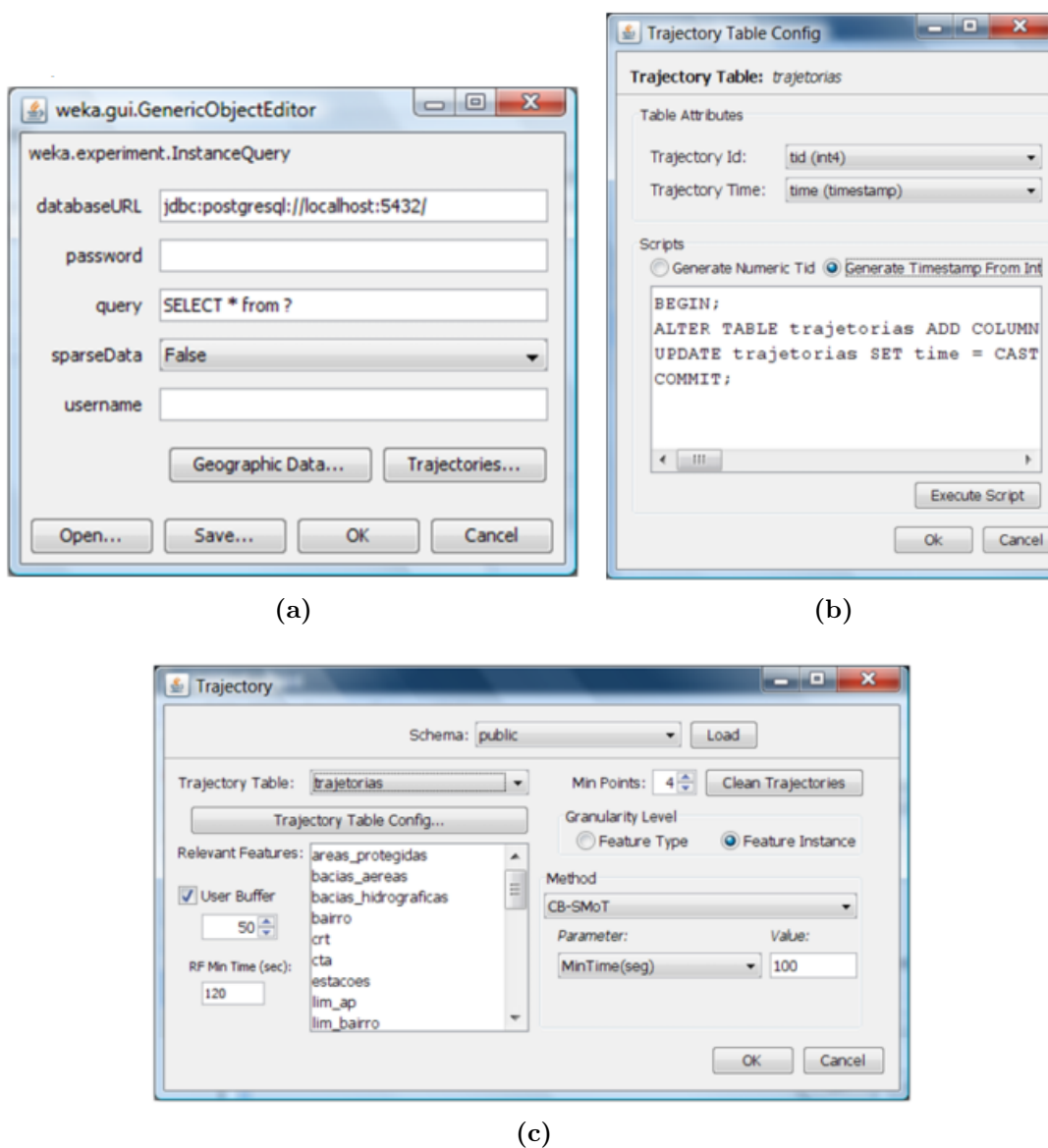


Figure 6.1: (a) STPM module, (b) *Trajectory Table Config* to perform other transformation or generation over the attributes *trajectory id* and *trajectory time*, (c) database schema provided by the user [Alvares et al., 2010]

6.1.2 M-Atlas

[Trasarti et al., 2010] argues the idea that a system able to master the complexity of the knowledge discovery process over mobility data needs to support at least four aspects:

- i. trajectory data need to be created, stored and queried through spatio-temporal primitives;
- ii. trajectory models and patterns representing collective behavior have to be extracted using trajectory mining algorithms;
- iii. such patterns and models representing have to be represented and stored in order to be re-used or combined;
- iv. and new mining algorithms may be added.

[Trasarti et al., 2010] proposed a system called M-Atlas that combines those four aspects through a Data Mining Query Language (DMQL) [Giannotti et al., 2011]. This language in turn can be used to express the knowledge discovery process as a sequence of queries to be submitted to the system.

For mobility understanding, M-Atlas supports several statistical analysis on a dataset: *movement distribution analysis* to estimate the active movements in each hour of the week; *cumulative lengths distribution* to represent the cumulative number of trajectories having the same length; *density of length over speed* to analyze the variance of each speed value where lower densities are represented by cold colors, while higher densities are represented by warm colors. Figure 6.2 illustrates these tasks. M-Atlas also integrates a set of data mining tools in order to discovery mobility behaviors. It supports the construction of Origins-Destinations Matrix, the construction of georeferenced density maps according to different measures, extraction of mobility patterns, such as T-Patterns [Giannotti et al., 2007], T-Clustering [Andrienko et al., 2009], T-Itineraries [Benkert et al., 2008] and T-Prediction [Monreale et al., 2009].

M-Atlas therefore allows the user to combine tools (statistical methods, data mining algorithms) in order to build his own discovery knowledge process in an iterative and interactive way. Furthermore, this system also support the construction, storage and retrieval of trajectories.

6.2 Tools in Complex Networks

Complex network area has received many attention and, consequently, many methods has been developed to support analysis in networks in several areas, such as Sociology, Biology, Computer Science and so on. From this, software tools are important to aid analysis on network data in different areas, offering manners to compute properties and investigate the network structure.

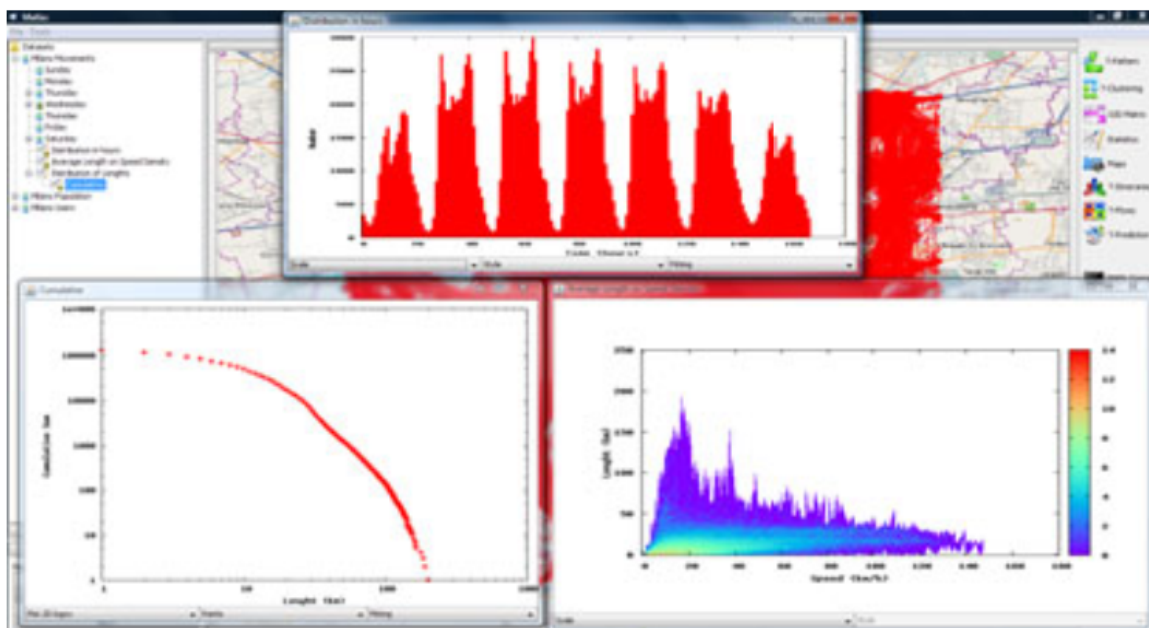


Figure 6.2: Statistical analysis performed by M-Atlas: movement distribution (top), cumulative lengths distribution (left) and density of length over speed (right) [Trasarti et al., 2010]

This section presents three tools to support users in network analysis: *Cytoscape*, *Network Workbench* and *Gephi*.

6.2.1 Cytoscape

Motivated by the explosion in experimental technologies for characterizing molecular interactions and states, researchers have turned to a variety of software tools to process and analyze the resulting large-scale data. However, those software tools are not able to integrate both molecular interactions and state measurements together in a common framework, and to then bridge these data with a wide assortment of model parameters and other biological attributes. [Shannon et al., 2003] proposed a general-purpose and open-source software environment for the large scale integration of molecular interaction network data called Cytoscape.

Cytoscape Core software component provides basic functionality for integrating arbitrary data on the graph, a visual representation of the graph and integrated data, selection and filtering tools, and an interface to external methods implemented as plugins. Cytoscape integrates data with the graph model using attributes, where an attribute is a single predicate of a node or edge. In addition, it provides a representation of a hierarchical classification, ontology, by using annotations in order to structure more specific descriptions of groups of nodes or edges. These annotations typically correspond to an existing repository of knowledge that is large, complex, and relatively static, such as an ontology database.

Visualization is another important functionality provided by Cytoscape, which supports a variety of automated network layout algorithms as well as it provides an attribute-to-visual mapping to control the appearance of nodes and edges, such as node color, shape, size and so on. In addition, it also provides a filtering mechanism to select nodes and edges according to a wide variety of criteria, including selection by name, by a list of names, or on the basis of attribute (e.g. node degree). The possibility of extending the Core with plug-in modules is a powerful means of implementing new algorithms and additional network analyses. Figure 6.3 illustrates some feature of Cytoscape.

Cytoscape is therefore a powerful software of analyzing large network data whose initial proposal was to face manners of analyzing molecular interactions and state measurements through networks and their properties. Nonetheless, many new algorithms have been developed for supporting different types of analysis in several areas, such as social network, bioinformatics and semantic web [Shannon et al., 2003].

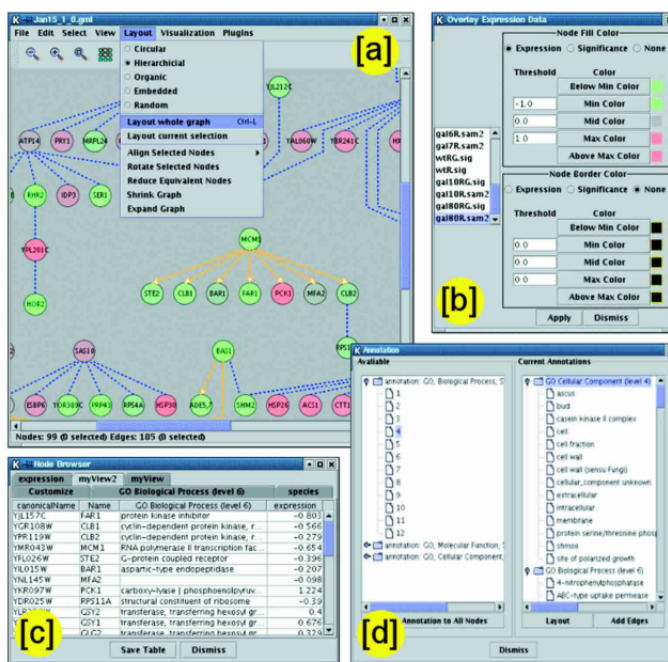


Figure 6.3: Some functionality of Cytoscape. (a) Network layout algorithms. (b) Data attribute-to-visual mapping to control the appearance of their associated nodes and edges and data types as well. (c) Attribute of the selected nodes and edges. (d) Annotations are transferred to node and edge attributes by choosing the desired ontology and hierarchical level from a list of those available [Shannon et al., 2003]

6.2.2 Network Workbench

Network Workbench (NWB), proposed by [Team, 2006], aims at analyzing large-scale networks as well as providing a toolkit for modeling and visualization for biomedical, social science and physics research. NWB performs network analysis with the most known algorithms. In addition, it is able to generate, run and validate models to advance their understanding of the structure and dynamics of particular networks.

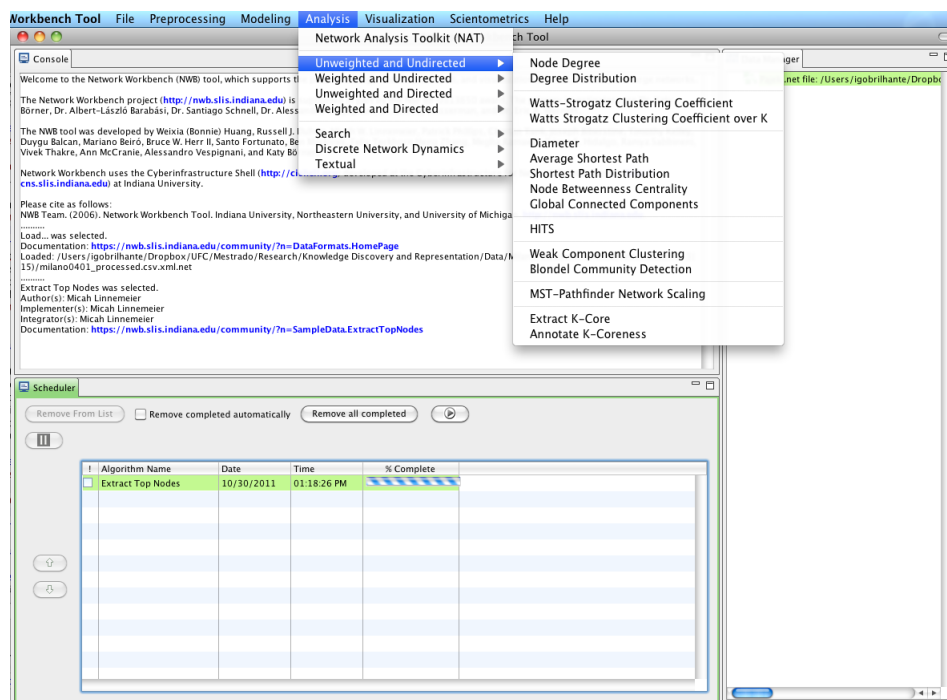


Figure 6.4: NWB interface with the menu to compute network analysis on different types of networks, such as node degree, degree distribution, clustering coefficient and community detection

NWB performs several analysis on different types of networks, i.e, weighted, unweighted, directed, undirected networks. Methods of preprocessing can be executed to remove isolated nodes, highly connected nodes or node in a random manner. In addition, network models can be generate as focus of study to understand the characteristics present in each model, and the results are store in plain text files that can be plotted using an external tool. Moreover, additional algorithms and data formats can be integrated into the NWB using wizard driven templates.

6.2.3 Gephi

In order to developed a network exploration tool with high quality layout algorithms, data filtering, clustering, statistics and annotation, [Bastian et al., 2009] proposed the Gephi project, focusing on analysis clarity and on modern user interface, to both experts and uninitiated audience (Figure 6.5).

Gephi is an open source network exploration and manipulation software inspired by WYSIWYG editors like Adobe Photoshop - “*Like PhotoshopTM for graphs*” - [Bastian et al., 2009]. Gephi can import, visualize, spatialize, filter, manipulate and export all types of networks and it can deal with large network (i. e. over 20,000 nodes) and, because it is built on a multi-task model, it takes advantage of multi-core processors. Node design can be personalized with a shape, a panel or a photo. The highly configurable algorithms can be run in real-time. Labels can be shown on the visualization window from any data attribute associate to nodes and, besides, a special algorithm named *Label Adjust*

avoids label overlapping.

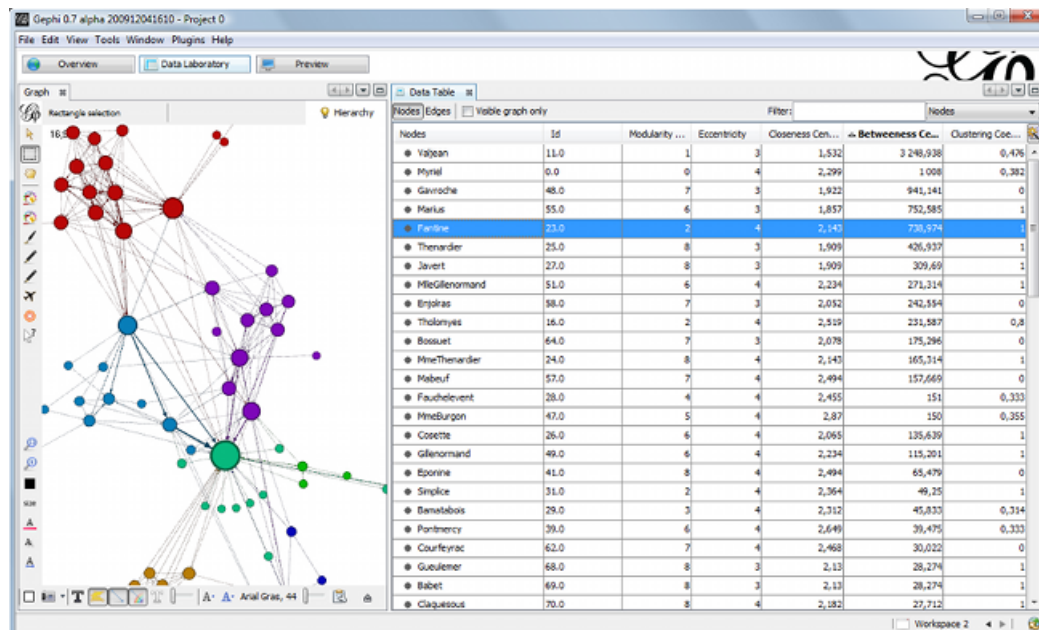


Figure 6.5: Interface of Gephi showing the *Data Laboratory* with *Data Table*, nodes and edges with their properties, and a graph visualization window [Bastian et al., 2009]

As well as Cytoscape, Gephi supports node attributes and edge attributes and they can be used to control the appearance of nodes and edges and as a mechanism of filtering as well. Besides importing different network formats (e.g. adjacent list), it can load networks from database systems by querying tables. The architecture is interoperable and data source can be created to communicate with existing software, third parties databases or web-services. Gephi also supports network model generation, hierarchical networks and dynamic networks. Furthermore, all the generated visualization of the networks can be exported as SVG file.

6.3 MobNet

In this section we introduce the developed software tool MOBNET that is fruit of the achieved work in this dissertation. Firstly, we present an overview of the tool, including the its layers and interface. Afterwards, we show the main capabilities that are based on the contributions in Chapter 4 and 5.

6.3.1 Overview

MOBNET is a multi-platform software that supports users to analyze mobility data through complex network techniques. This software tool receives some mobility data as input to create a network structure of the data according to a proposed methodology presented in this dissertation. MOBNET is composed by three main layers:

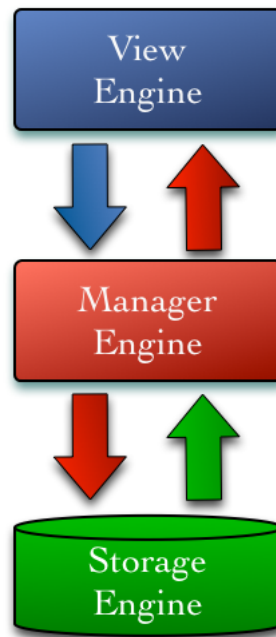


Figure 6.6: MOBNET architecture consisting of three main layers: view engine, manager engine and storage

- **View Engine** aims at interacting with the user. It provides ways to collect the inputs and present the results to the user, including the nodes properties, edges properties and communities properties. The built networks can be plotted on the map to ease the understanding of the results. With the built *trajectory networks* or *poi networks*, the View Engine can perform a visualization to node as trajectories as well as nodes as points of interests and edges as movements of trajectories among the points;
- **Manager Engine** is responsible for managing the process of retrieving, storing and analyzing of the networks. It is the core of the application, where is defined the model to store and retrieve the networks and their properties in the storage layer. Besides, Manager Engine also provides the methods to compute the properties of the networks, such as clustering coefficient, average shortest path length, community detection, etc.
- **Storage Engine** corresponds to a database system to storage and retrieve the networks and their properties, such as clustering coefficient, communities, etc. In addition, it offers the indexing structure and query language to Manager Engine in order to take advantage of the database system to perform some activities, such as the computation of some properties.

This architecture is illustrated in Figure 6.6. Figure 6.7 shows MOBNET and some of its components.

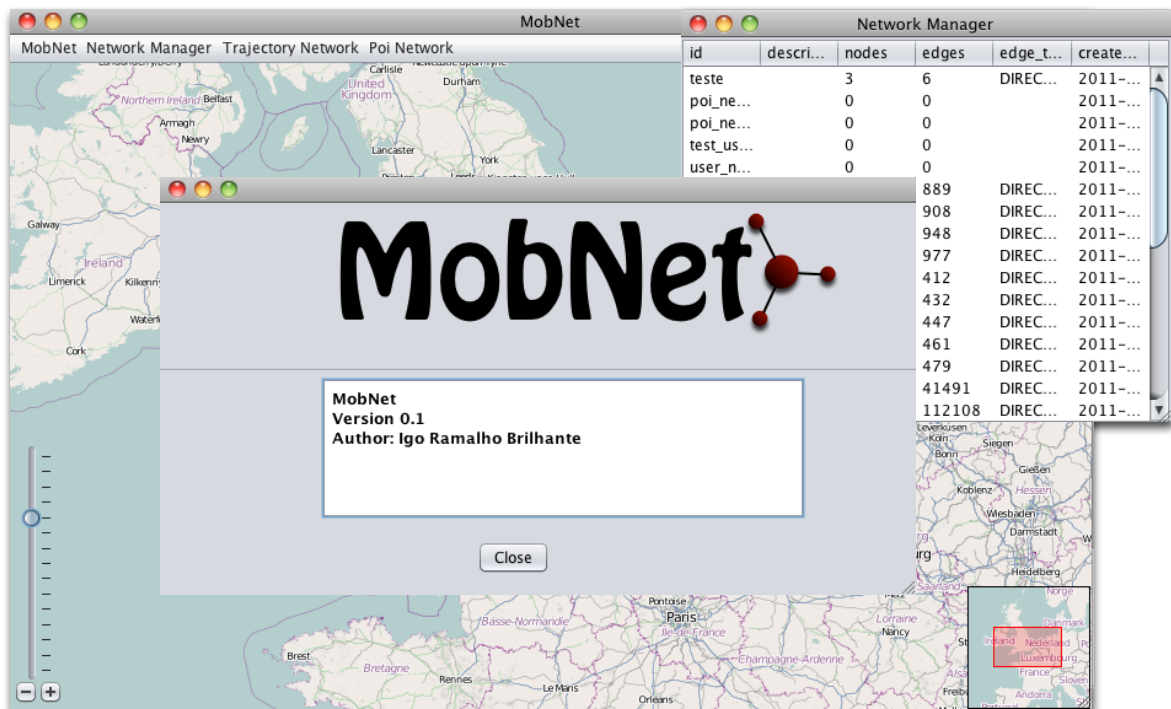


Figure 6.7: MOBNET interface: in the menu we can choose between *trajectory network* and *poi network* to build and visualize networks; *Network Manager* performs activities on the built networks

In the next two sections, we present the main functionalities of MOBNET according to the proposed methodologies, starting with *trajectory network* and concluding with *poi network*.

6.3.2 Trajectory Network

In Chapter 4 we have presented a methodology to build a network from a trajectory dataset, where the nodes represent the trajectories and the edges are formed according to a given similarity function. Thereby, MOBNET provides a interface to build a *trajectory network* from a trajectory dataset (Figure 6.8) and a mechanism of visualizing the nodes (trajectories) of such network (Figure 6.9).

The network is built by the *trajectory network builder* and, after that, the properties are computed and stored for further analysis (Figure 6.8). With a built trajectory network, the user can investigate the network, node and edges properties and, then, visualize all the nodes, the trajectories, or a set of nodes selected by the user (Figure 6.9).

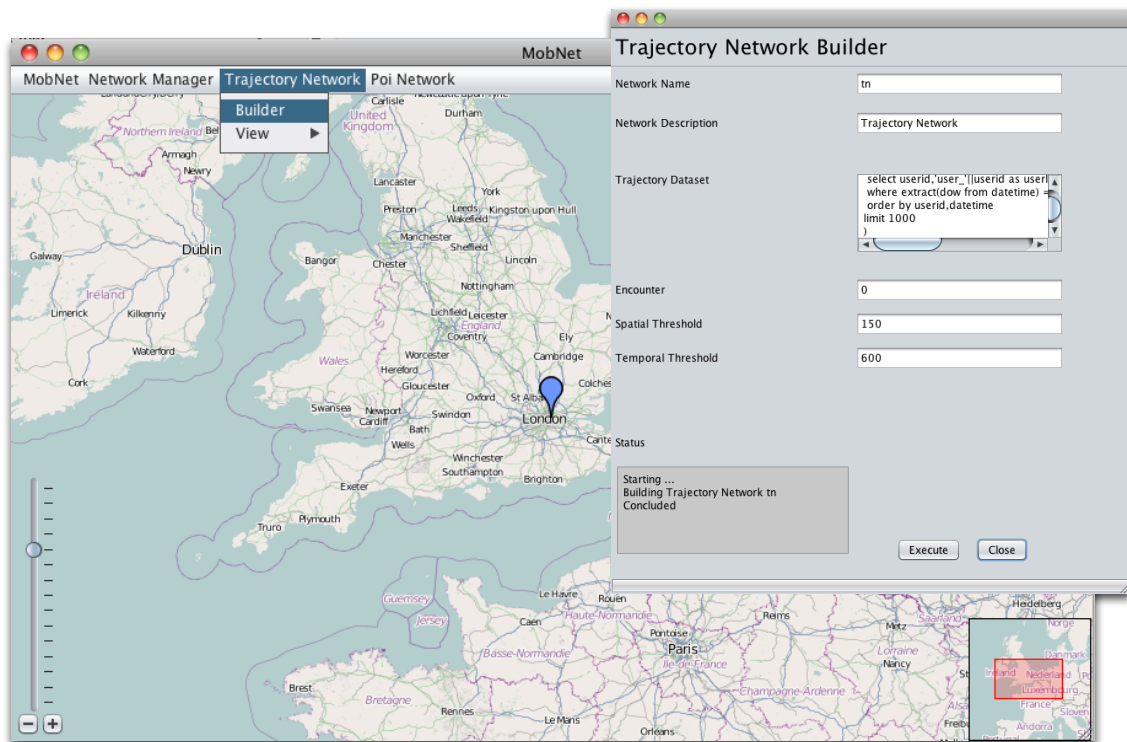


Figure 6.8: Building a *trajectory network* where the nodes are the trajectories from the dataset

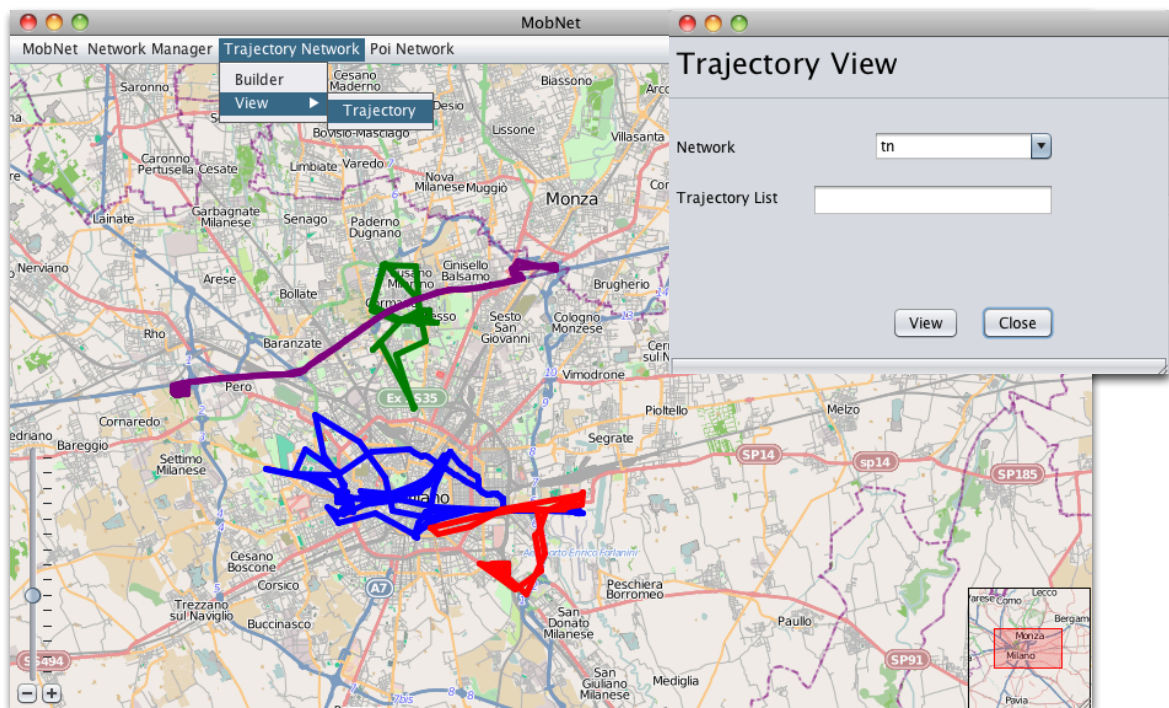


Figure 6.9: Visualizing the nodes (trajectories) of a built *trajectory network*: this is a trajectory network with four nodes representing the four trajectories depicted on the map

6.3.3 Points of Interest Network

In Chapter 5 we have presented a methodology to build a *poi network* from a trajectory dataset and a poi dataset. Thereby, MOBNET also provides mechanisms to build a *poi network*, to detect communities and their properties, and to visualize the networks and the communities.

In this case, MOBNET builds a poi network according to the proposed methodology in which the user is able to analyze the network properties and draw the results on the map. In addition, the user can perform a community discovery process on the network, investigate the community properties and see the discovered communities on the map as well. Figure 6.10 shows the *poi network* builder. Figure 6.11 illustrates the process of visualization of communities: select all the communities or some of them.

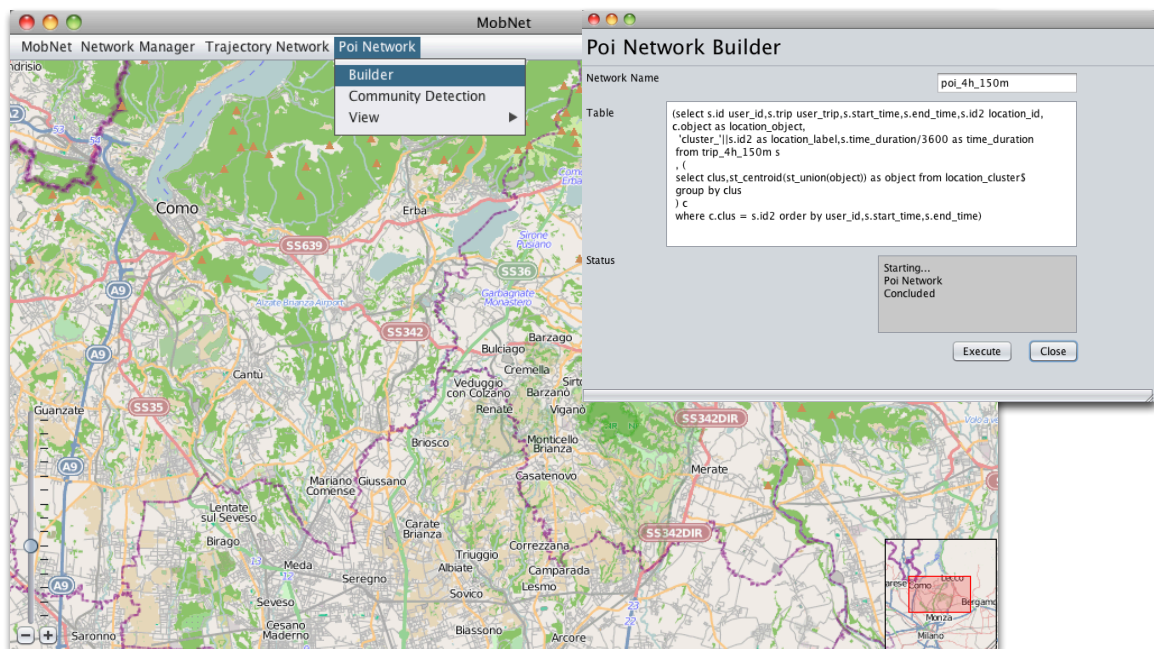


Figure 6.10: Building a *poi network* where the nodes represent the points of interest, and the edges correspond to the movement of the trajectories between the points

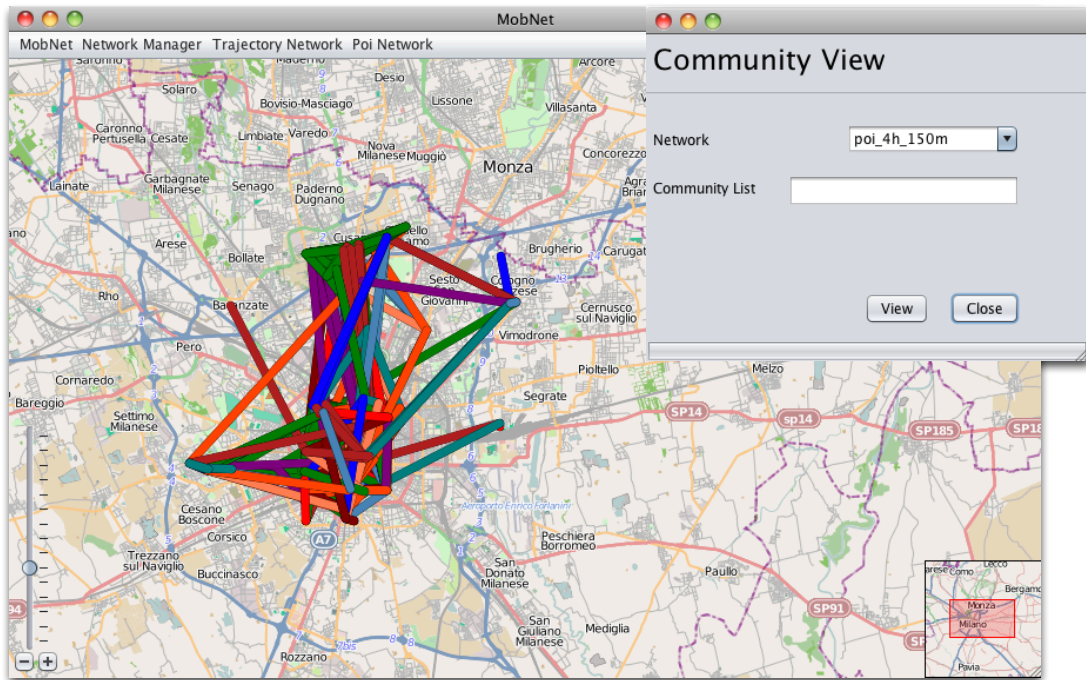


Figure 6.11: Visualizing the discovered communities of a built *poi network*

6.4 Conclusion

In this chapter we presented some software tools in mobility and network analysis, and a proposed software tool fruit of the main idea of this dissertation: mobility analysis under a perspective of complex networks. Firstly, we presented some software tools whose goal is to analyze mobility data, including the building of trajectories, semantic trajectories, the identification of stops and moves, etc. Next, we introduced some software tools to support users in network analysis, including network properties, network models, etc.

In the end, we introduced the proposed software tool called MOBNET whose the main goal is to analyze mobility data using complex network techniques. The two main capabilities are based on the methodologies proposed in Chapter 4 and 5, i. e., it is based on the building of trajectory and poi networks and the analysis on them.

CHAPTER 7

Conclusions

In this master dissertation we have presented a multidisciplinary study combining mobility and complex network areas. Firstly, we introduced the basic concepts used throughout this dissertation in Chapter 2 and 3. Afterwards, we presented the contributions of this dissertation in Chapter 4 and 5, showing different approaches to analyze trajectory data using complex network techniques, and Chapter 6 introduction the developed software tool to analyze mobility data using the proposed methodologies.

7.1 Conclusion

In Chapter 2 we introduced basic definitions in mobility research field, including the concepts of *trajectory* and *semantic trajectory*. Moreover, we presented the notions of *stops* and *moves* as well as some methods to identify them from a trajectory dataset. In Chapter 3, on the other hand, we presented the basic concepts in complex network areas, covering some global and local properties of the networks, network models and community detection. Finally, we have presented some tools to support users in analyzing network data.

In Chapter 4, we presented the first approach, in which the trajectories correspond to the nodes in a *trajectory network*, and the relationship between the nodes, trajectories, is achieved with a given similarity function between two trajectories. Thereby,

we introduced a similarity function called *encounter* in order to capture the spatial and temporal proximity between two trajectories to establish an edge between them if their similarity satisfies a given minimum threshold, and we proposed an algorithm to generate a *trajectory network* based on the *encounters* of the trajectories. Besides the theoretical analysis, we presented experimental results for moving object trajectories on complex networks. Our analysis reveals that all trajectory networks are scale free network, presenting small-world and power law features. In addition, the results have practical implications for investigating moving objects interactions from complex network perspective. This approach provides another method for analyzing trajectories from the potential *interaction* perspective, when it is compared to existing data mining and statistical methods. This approach can be performed on other applications: in hospital environment, where doctors and patients could wear GPS-enabled devices or chips to collect their mobility and, then, to analysis the risk of contagious of the network built by their encounters, that is, if a doctor, that could be in touch with patients with contagious disease, encounters many others, doctors and patients, and, consequently, might spread these diseases; Study of football players, for instance, to identify the players that move near their opponents to block their game.

In Chapter 5, we proposed an explorative study on the relation between people mobility and points of interest, POIs, at the global scale, based on the complex network paradigm. In this contribution, we presented an algorithm to build a complex network, named *poi network*, that combines locations that people visit with the mobility of users represented as trajectories. From this network we computed the communities as the subgroups of points of interest related by the common users trajectories visiting them. An explorative analysis was conducted in a real case study where a complex network was built combining points of interest with traces of moving cars and communities of places grouped by common mobility were extracted. We defined some interesting features to characterize these communities such as the compactness or the presence of “bridge” communities. We observed these measures discussing the possible interpretations in terms of applications such as traffic management or advertising.

We introduced in Chapter 6 the developed software tool, named MOBNET, to encapsulate the proposed methodologies to analyze mobility data using complex networks techniques. Yet, we presented some software tools to support users in mobility analysis and network analysis as well.

7.2 Future Works

Future works follow several directions. First of all, alternative ways of computing the stops and associating the POIs may be applied in order to better represent the actual activity of the user. Furthermore, we plan to extend this methodology to other real datasets to further validate the results. Naturally, we intend to investigate more in deep the possible applications that can be benefit from this analysis, for example going to the direction of

POIs recommendation systems.

Other direction will focus on creating a framework to analyze mobility data under a complex network viewpoint, whose goal is not only at exploiting complex network methods, but also integrating mobility techniques found in the literature (statistics or data mining) with network techniques. This is an open and challenging issue. Which is the potential synergic mobility knowledge we can get from the cooperation of data mining and complex networks techniques?

This new topic offers several research opportunities. For example, since complex networks are strongly related to social media (social networks, etc.), future works also intend to take advantage of the huge amount of information available about Internet users. For instance, map services offer comments and rates of the places written by people that visited them; associate the traffic flow with what people say on social media (Facebook, Twitter, etc.).

These future works aim at targeting people that are not only specialists, but non-specialist as well. Certainly the methods to support traffic specialists or urban designers are quite important. However, enable citizens to understand the traffic is quite relevant, since they can make choices in order to use the urban space for the benefits of all.

References

- [Adamic et al., 2001] Adamic, L., Lukose, R., Puniyani, A., & Huberman, B. (2001). Search in power-law networks. *Physical review E*, 64(4), 046135.
- [Ahn et al., 2010] Ahn, Y.-Y., Bagrow, J. P., & Lehmann, S. (2010). Link communities reveal multiscale complexity in networks. *Nature*, 466(7307), 761–4.
- [Aiello et al., 2000] Aiello, W., Chung, F., & Lu, L. (2000). A random graph model for massive graphs. In *Proceedings of the thirty-second annual ACM symposium on Theory of computing* (pp. 171–180).: Acm.
- [Alvares et al., 2007a] Alvares, L. O., Bogorny, V., Kuijpers, B., de Macedo, J., Moelans, B., & Vaisman, A. (2007a). A model for enriching trajectories with semantic geographical information. In *Proceedings of the 15th annual ACM international symposium on Advances in geographic information systems* (pp.22). New York, New York, USA: ACM.
- [Alvares et al., 2007b] Alvares, L. O., Bogorny, V., Kuijpers, B., Moelans, B., Antonio, J., & Palma, A. T. (2007b). Towards Semantic Trajectory Knowledge Discovery.
- [Alvares et al., 2010] Alvares, L. O., Palma, A., Oliveira, G., & Bogorny, V. (2010). Weka-STPM: from trajectory samples to semantic trajectories. In *Proceedings of the Workshop on Open Source Code*, volume 1 (pp. 1–6).
- [Andrienko et al., 2009] Andrienko, G., Andrienko, N., Rinzivillo, S., Nanni, M., & Pedreschi, D. (2009). A Visual Analytics Toolkit for Cluster-Based Classification of Mobility Data. *SSTD 09 Proceedings of the 11th International Symposium Advances in Spatial and Temporal Databases*, 5644, 432–435.
- [Backstrom et al., 2011] Backstrom, L., Boldi, P., Rosa, M., Ugander, J., & Vigna, S. (2011). Four Degrees of Separation. *ArXiv e-prints*.
- [Barabási, 2002] Barabási, A.-L. (2002). *Linked: The New Science of Networks*, volume 71. Perseus Publishing.

- [Barabási & Albert, 1999] Barabási, A.-L. & Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439), 11.
- [Bastian et al., 2009] Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi : An Open Source Software for Exploring and Manipulating Networks. *American Journal of Sociology*, (pp. 361–362).
- [Benkert et al., 2008] Benkert, M., Gudmundsson, J., Hubner, F., & Wolle, T. (2008). Reporting flock patterns. *Computational Geometry*, 41(3), 111–125.
- [Blondel et al., 2008] Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10), P10008.
- [Boccaletti et al., 2006] Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., & Hwang, D. (2006). Complex networks: Structure and dynamics. *Physics Reports*, 424(4-5), 175–308.
- [Bogorny, 2011] Bogorny, V. (2011). Weka-STPM: a Software Architecture and Prototype for Semantic Trajectory Data Mining and Visualization. *Transactions in GIS*, 15(2), 227–248.
- [Bollobás, 1998] Bollobás, B. (1998). *Modern Graph Theory*, volume 184 of *Graduate Texts in Mathematics*. Springer.
- [Bornholdt & Ebel, 2001] Bornholdt, S. & Ebel, H. (2001). World Wide Web scaling exponent from Simon’s 1955 model. *Physical Review E - Statistical, Nonlinear and Soft Matter Physics*, 64(3 Pt 2), 035104.
- [Brilhante et al., 2011] Brilhante, I. R., de Macedo, J. A. F., Renso, C., & Casanova, M. A. (2011). Trajectory data analysis using complex networks. In *Proceedings of the 15th Symposium on International Database Engineering & Applications, IDEAS ’11* (pp. 17–25). New York, NY, USA: ACM.
- [Calabrese et al., 2010] Calabrese, F., Colonna, M., Lovisolo, P., Parata, D., & Ratti, C. (2010). Real-Time Urban Monitoring Using Cellular Phones: a Case-Study in Rome. *IEEE Transactions on Intelligent Transportation Systems*.
- [Castro & Grossman, 1999] Castro, R. D. & Grossman, J. W. (1999). Famous trails to Paul Erdős. *Mathematical Intelligencer*, 21, 51–63.
- [Clauset et al., 2004] Clauset, A., Newman, M. E. J., & Moore, C. (2004). Finding community structure in very large networks. *Physical Review E*, 70(6).
- [Coscia et al., 2011] Coscia, M., Giannotti, F., & Pedreschi, D. (2011). A classification for community discovery methods in complex networks. *Statistical Analysis and Data Mining*, 4(5), 512–546.
- [Dodge et al., 2008] Dodge, S., Weibel, R., & Lautenschütz, A.-K. (2008). Towards a taxonomy of movement patterns. *Information Visualization*, 7(3), 240–252.
- [Donato, 2010] Donato, D. (2010). Graph Structures and Algorithms for Query-Log Analysis. In *CiE* (pp. 126–131).

- [Dorogovtsev & Mendes, 2000] Dorogovtsev, S. N. & Mendes, J. F. F. (2000). Scaling Behaviour of Developing and Decaying Networks. *Europhysics Letters*, 52(1), 7.
- [Erdős & Rényi, 1959] Erdős, P. & Rényi, A. (1959). On random graphs. *Publ. Math. Debrecen*, 6(290-297), 290–297.
- [Erdős & Rényi, 1960] Erdős, P. & Rényi, A. (1960). *On the evolution of random graphs*, volume 5. Akad. Kiadó.
- [Ester et al., 1996] Ester, M., Kriegel, H.-p., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Computer*, 1996(6), 226–231.
- [Giannotti et al., 2011] Giannotti, F., Nanni, M., Pedreschi, D., Pinelli, F., Renso, C., Rinzivillo, S., & Trasarti, R. (2011). Unveiling the complexity of human mobility by querying and mining massive trajectory data. *Very Large Database*, 20(5).
- [Giannotti et al., 2007] Giannotti, F., Nanni, M., Pinelli, F., & Pedreschi, D. (2007). Trajectory pattern mining. *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '07*, (pp. 330).
- [Giannotti & Pedreschi, 2008] Giannotti, F. & Pedreschi, D., Eds. (2008). *Mobility, Data Mining and Privacy - Geographic Knowledge Discovery*. Springer.
- [Gilbert, 2011] Gilbert, J. D. (2011). Graph Theory. *The Mathematical Gazette*, 85(502), 176.
- [Girvan & Newman, 2002] Girvan, M. & Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(12), 7821–7826.
- [González et al., 2008] González, M. C., Hidalgo, C. a., & Barabási, A.-L. (2008). Understanding individual human mobility patterns. *Nature*, 453(7196), 779–82.
- [Guimera et al., 2007] Guimera, R., Sales-Pardo, M., & Amaral, L. A. N. (2007). Module identification in bipartite and directed networks. *Physical Review E*, 76(3), 1–8.
- [Guo et al., 2010] Guo, D., Liu, S., & Jin, H. (2010). A graph-based approach to vehicle trajectory analysis. *J. Locat. Based Serv.*, 4(3-4), 183–199.
- [Gütting et al., 2000] Gütting, R. H., Böhlen, M. H., Erwig, M., Jensen, C. S., Lorentzos, N. a., Schneider, M., & Vazirgiannis, M. (2000). A foundation for representing and querying moving objects. *ACM Transactions on Database Systems*, 25(1), 1–42.
- [Holmes et al., 1994] Holmes, G., Donkin, A., & Witten, I. H. (1994). WEKA: a machine learning workbench. *Proceedings of ANZIIS 94 Australian New Zealand Intelligent Information Systems Conference*, 24(3), 357–361.
- [Jackson, 1968] Jackson, R. (1968). The Matthew Effect in science. *International Journal of Dermatology*, 27(1), 16.
- [Jeong et al., 2001] Jeong, H., Mason, S. P., Barabasi, A. L., & Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature*, 411(6833), 41–42.

- [Jeong et al., 2000] Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., & Barabási, A. L. (2000). The large-scale organization of metabolic networks. *Nature*, 407(6804), 651–654.
- [Kaluza et al., 2010] Kaluza, P., Kölzsch, A., Gastner, M. T., & Blasius, B. (2010). The complex network of global cargo ship movements. *Journal of the Royal Society, Interface / the Royal Society*, 7(48), 1093–103.
- [Leskovec et al., 2010] Leskovec, J., Huttenlocher, D., & Kleinberg, J. (2010). Predicting positive and negative links in online social networks. In *Proceedings of the 19th international conference on World wide web* (pp. 641–650).: ACM.
- [Leskovec et al., 2006] Leskovec, J., Kleinberg, J. M., & Faloutsos, C. (2006). Graph Evolution: Densification and Shrinking Diameters. *ACM Transactions on Knowledge Discovery from Data*, 1(1), 2–es.
- [Li et al., 2008] Li, Q., Zheng, Y., Xie, X., Chen, Y., Liu, W., & Ma, W.-Y. (2008). Mining user similarity based on location history. *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems - GIS '08*, (c), 1.
- [Monasson, 1999] Monasson, R. (1999). Diffusion, localization and dispersion relations on “small-world” lattices. *European Physical Journal B*, 12(4), 555–567.
- [Monreale et al., 2009] Monreale, A., Pinelli, F., Trasarti, R., & Giannotti, F. (2009). WhereNext: a location predictor on trajectory pattern mining. *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, (pp. 637–645).
- [Newman, 2002] Newman, M. E. J. (2002). Assortative Mixing in Networks. *Phys. Rev. Lett.*, 89(20), 208701.
- [Newman, 2003] Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM review*, 45(2), 167–256.
- [Newman, 2005] Newman, M. E. J. (2005). Power laws Pareto distributions and Zipf’s laws. *Contemporary Physics*, 46, 323–351.
- [Newman, 2006] Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 103(23), 8577–8582.
- [Newman, 2010] Newman, M. E. J. (2010). *Networks: an introduction*. Oxford Univ Pr.
- [Newman & Watts, 1999] Newman, M. E. J. & Watts, D. J. (1999). Renormalization group analysis of the small-world network model. *Physics Letters A*, 263(4-6), 4.
- [Nguyen-Dinh et al., 2010] Nguyen-Dinh, L. V., Aref, W. G., & Mokbel, M. F. (2010). SECONDO: A Platform for Moving Objects Database Research and for Publishing and Integrating Research Implementations. *Data Engineering*, 33(2), 46–55.
- [OGC, 2008] OGC (2008). Opengis standards and specification.

- [OpenStreetMap, 2011] OpenStreetMap (2011). OpenStreetMap. <http://www.openstreetmap.org/>.
- [Palma et al., 2008] Palma, A. T., Bogorny, V., Kuijpers, B., & Alvares, L. O. (2008). A clustering-based approach for discovering interesting places in trajectories. *Proceedings of the 2008 ACM symposium on Applied computing SAC 08*, (December), 863.
- [Pimm, 2002] Pimm, S. L. (2002). *Food Webs*. University of Chicago Press, second edition.
- [Price, 1965] Price, D. D. S. (1965). Networks of Scientific Papers. *Science*, 149(3683), 510–515.
- [Price, 1976] Price, D. D. S. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, 27(5), 292–306.
- [Radicchi et al., 2004] Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., & Parisi, D. (2004). Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(9), 2658–2663.
- [Rocha et al., 2010] Rocha, J. A. M. R., Times, V. C., Oliveira, G., Alvares, L. O., & Bogorny, V. (2010). DB-SMoT: A direction-based spatio-temporal clustering method. In *IEEE Conf. of Intelligent Systems*.
- [Shannon et al., 2003] Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11), 2498–504.
- [Simon, 1955] Simon, H. A. (1955). On a class of skew distribution functions. *Biometrika*, 42(3-4), 425–440.
- [Spaccapietra et al., 2008] Spaccapietra, S., Parent, C., Damiani, M. L., Demacedo, J., Porto, F., Vangenot, C., & de Macêdo, J. A. F. (2008). A conceptual view on trajectories. *Data & Knowledge Engineering*, 65(1), 126–146.
- [Team, 2006] Team, N. (2006). Network Workbench Tool. Indiana University, Northeastern University, and University of Michigan. <http://nwb.slis.indiana.edu>.
- [Trasarti et al., 2010] Trasarti, R., Rinzivillo, S., Pinelli, F., Nanni, M., Monreale, A., Renso, C., Pedreschi, D., & Giannotti, F. (2010). Exploring Real Mobility Data with M-Atlas. *Matrix*, (pp. 624–627).
- [Tsourakakis, 2008] Tsourakakis, C. E. (2008). Fast Counting of Triangles in Large Real Networks: Algorithms and Laws. *Machine Learning*, 1401(2008), 608–617.
- [Wachowicz et al., 2011] Wachowicz, M., Ong, R., Renso, C., & Nanni, M. (2011). Finding moving flock patterns among pedestrians through collective coherence. *IJGIS*, 25(11).
- [Wang et al., 2009] Wang, P., González, M. C., Hidalgo, C. A., & A.-L. Barabási (2009). Understanding the spreading patterns of mobile phones viruses. *Science*, (324), 1071–1076.

- [Watts & Strogatz, 1998] Watts, D. J. & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(6684), 440–2.
- [West, 2001] West, D. B. (2001). Introduction to Graph Theory, 2nd ed. *Upper Saddle River New Jersey Prentice Hall*.
- [Xiao, 2005] Xiao, J. (2005). Clustering Spatial Data for Join Operations Using Match-based Partition. In *CIMCA/IAWTIC*.
- [Yan et al., 2011] Yan, Z., Chakraborty, D., Parent, C., Spaccapietra, S., & Aberer, K. (2011). SeMiTri: a framework for semantic annotation of heterogeneous trajectories. In *International Conference on Extending database Technology - EDBT*.
- [Yan et al., 2010] Yan, Z., Parent, C., Spaccapietra, S., & Chakraborty, D. (2010). A hybrid model and computing platform for spatio-semantic trajectories. *The Semantic Web: Research and Applications*, 6088(978-3-642-13485-2), 60–75.
- [Zheng & Xie, 2010] Zheng, Y. & Xie, X. (2010). Learning Location Correlation from GPS Trajectories. *2010 Eleventh International Conference on Mobile Data Management*, (49), 27–32.