DIGITAL TRACES OF HUMAN MOBILITY AND INTERACTION: MODELS AND APPLICATIONS

by

ANTONIO LIMA

A thesis submitted to
The University of Birmingham
for the degree of
DOCTOR OF PHILOSOPHY

School of Computer Science
College of Engineering and Physical Sciences
The University of Birmingham
June 2016

Abstract

In the last decade digital devices and services have become increasingly popular and have permeated many aspects of everyday life. They generate massive amounts of data that provide insightful information about how people move across geographic areas and how they interact with others. The availability of this kind of detailed information about a large group of people now makes it easier to investigate several aspects of human behaviour and in particular their mobility and interactions. Therefore, the central thesis of this dissertation is that *the analysis of mobility and interaction traces generated by digital devices and services, at different timescales and spatial granularity, can be used to gain a better understanding of human behaviour, build new applications and improve existing services.* In order to substantiate this statement I develop analytical models and applications supported by three main sources of mobility and interaction data: online social networks, mobile phone networks and GPS traces. First, I present three applications related to data gathered from online social networks, namely the analysis of a global rumour spreading in Twitter, the definition of spatial dissemination measures in a social graph and the analysis of collaboration between developers in GitHub. Then I describe two applications of the analysis of country-wide data of cellular phone networks: the modelling of epidemic containment strategies, with the goal of assessing their efficacy in curbing infectious diseases; the definition of a measure of individual risk, associated to the mobility profiles of individuals, which can be used to identify who needs targeted treatment. Finally, I present two applications based on GPS traces: the estimation of trajectories from spatially-coarse temporally-sparse location traces and the analysis of routing behaviour in urban settings. These studies demonstrate that digital traces coming from various data sources, with different granularities and limitations can be used advantageously in several fields, making it possible both to implement new services and to improve existing systems.

Ai miei genitori e a tutti i miei insegnanti,
per avermi trasfuso l'importanza
dell'istruzione, la passione per la conoscenza
e la curiosità verso il nuovo.

ACKNOWLEDGEMENTS

First of all, I would like to thank my supervisor Mirco Musolesi for advising me throughout my PhD course. His guidance during these years has invaluably helped me to navigate the multi-faceted world of academia and to develop my research skills. I thank Giovanni Mangioni, who caused my interest in research to spark in the first place, back when I was a MSc Student in Catania. I thank Marta C. González, who was advising me at M.I.T., for making me feel "at home" while overseas, in a Civil Engineering department, and for fuelling my enthusiasm.

I thank all my collaborators and co-authors, who contributed to make me a better researcher. In particular Manlio De Domenico, who has been both a valuable collaborator and a dear friend, and the collaborators at Telefonica Research, especially Ilias Leontiadis, Rade Stanojevic and Dina Papagiannaki, who made my internship enjoyable and fruitful.

My greatest gratitude goes to my parents, who have loved me, supported me in pursuing good education and encouraged me to follow my interests. Thanks to my sister Violetta, who has been close to me even when geographically far. Thanks to my whole family, Isotta, Diletta, Salvo, M. Grazia, Pina, because of all the memorable moments I have spent with them. And thanks to my *nonni*, whose memories I cherish dearly.

I thank all the friends I have met in these years, in particular Andrea, Shoshana, Veljko, Tylor, Luca, Olle, Thiago, Jeremy, Suma, Serdar, Roberta, Morgan, Serafina.

Finally, a big thank you to my dearest Claire. She has followed me across the world, she has been reading my drafts (and "fixing" my Americanisms) and she has encouraged me in these last years, sustaining me during the highs and lows of the PhD journey. These memories will be everlasting.

CONTENTS

CHAPTER 1

INTRODUCTION

In recent years we have witnessed a dramatic rise in the popularity of digital devices and services. People use them to interact with each other, to access information, to generate information and to disseminate it. Such devices, increasingly portable, always-connected, and equipped with several sensors are quickly permeating many aspects of everyday life for millions of people. The deluge of digital data continuously produced by these devices contains a great deal of information about human behaviour. Data related to a single smartphone, for example, is so detailed that it can provide specific information about the device owner, what kind of person they are, their neighbourhoods, their friends, their environment, i.e., all aspects that are intimately related to the identity of the owner. More importantly, data related to several distinct individuals is much more than just a collection of identities, since it embodies information about general patterns of human behaviour. In other words, the scientific investigation of large scale datasets does not have as an objective the single identities, but instead generic aspects of human behaviour. For this reason, while the possibility of analysing such sensitive data about a large number of individuals warrants reasonable privacy concerns, it also represents an unprecedented opportunity to analyse human behaviour at a large scale.

In the last decades, scholars have started exploring this unique opportunity. Studies related to several fields, ranging from sociology to geography, from psychology to urban planning, that were previously conducted by scientists at small scales [Jac61; Mil67;

Rav85], through ad-hoc experiments involving surveys and real tasks, can now be performed at a large-scale, either passively analysing digital traces, or actively delivering experiments to individuals through digital devices. The famous small-world experiment designed by Stanley Milgram and implemented using 296 letters [Mil67] has recently been reproduced using online social networks composed of millions of individuals [BBR+12]. The expensive surveys and census data used to predict street network flows and generate origin-destination matrices can now be replaced with the analysis of mobile phone networks [HD14; JFY+13].

This important data revolution, focussed on sociality and locality, in my opinion has been mainly driven by three fundamental trends:

- the availability of precise positioning services such as Global Positioning System (GPS) [HLC13] and Wi-Fi-based localisation [Ben07].
- the rise of popularity of mobile phones and, more recently, of smartphones equipped with location sensors [Rai10];
- the spread of Internet connection, both in terms of increased provision and reasonable cost [Per15; Rai10].

Some scholars had envisioned that such technological advantages would have produced the so-called "Death of Distance" [Cai01; Gra98]. This prediction was logically motivated by the advent of the communication age, which makes interaction between people easier, regardless of how distant they are and whether they know each other already. Instead, although interaction is indeed easier today, a few phenomena suggest an opposite effect from what was predicted, pointing out that geographic and social distance are now more important than ever. The urban population has steadily risen in the last decades [Coh03] and surpassed half of the world population in 2009, a sign that people are brought closer by the prospect of short-distance interaction. A large number of services available to individuals are aware of social and location contexts, using the information about the whereabouts and the connections of individuals for a variety of applications, for example

to tailor recommendations [MLR03], to optimise their infrastructure [BSM10; SMM+11] and to share under-utilised resources [CMF+14]. Diffusion processes are dominated by the activity of super-spreaders and structural holes [PMA+14], indicating that geographic and social distance still matters for these important processes [LLN+95; SMM+10].

A large body of novel research studies has recently unveiled interesting results about human behaviour by investigating digital traces coming from very diverse sources. The most common data sources that have been used for this purpose have been: online social networks [BBR+12], mobile phone networks [GHB08] and GPS traces [ZL15]. While this list is not exhaustive, it includes data that is quite diverse, in terms of spatial granularity, accuracy and time sparsity. Such a diversity can be potentially representative of similar data sources not analysed here. For example, studies based on credit card data are quite rare because of privacy concerns [dMRS+15], but the spatio-temporal properties of the data points can be expected to share some similarities with check-ins in retail shops collected from location-based social networks.

1.1 Thesis and contributions

As I have discussed in the previous paragraphs, digital traces constitute a novel kind of data about human mobility and interactions, which, depending on the source they are collected from, can greatly vary in terms of geographic granularity, temporal scale and accuracy. Despite this degree of variability, these sources make it possible to study human behaviour at a large scale and open up important opportunities both in the definition of theoretical models and in the development of practical applications that make use of them.

As a consequence, the central **thesis of this dissertation** is that *the analysis of mobility and interaction traces generated by digital devices and services, at different timescales and spatial granularity, can be used to gain a better understanding of human behaviour, build new applications and improve existing services.* In order to support this statement, I will focus on two classes of human behaviour, human mobility and interactions between

individuals. I will analyse digital traces collected from sources of different nature, such as online social networks, mobile networks, and GPS devices, and I will detail how these analyses can be used for modelling and for practical applications in real world scenarios. My research makes two main contributions. First, the analysis of data sources that, while being similar, as they describe the movements and interactions between individuals, can vary greatly in their time sparsity and space granularity. For example, GPS traces are spatially fine-grained (metres) and collected at a high temporal rate (seconds to minutes), while mobile network logs describe the position of an individual when a network event occurs (minutes to hours), with a coarse spatial resolution (the range of the antenna, typically a few kilometres). Second, the construction of context-specific models that make use of this data for specific purposes, and allow for it to be used in practical applications, such as counteracting epidemic outbreaks with novel techniques and estimating movement trajectory from coarse and sparse location traces.

In order to substantiate this statement, I first discuss the characteristics of these digital traces, from what sources they are collected, what information they contain and, most importantly, what their limits are. Second, I present, for each data source, how these traces can be used to improve existing services and design new applications based on their analysis.

1.2 Thesis outline

This dissertation is organised as follows:

- In Chapter 2, I introduce the main sources of digital traces that are available for analysis of mobility and interaction. I discuss how they are collected, what their limits are in terms of spatial granularity and time sparsity and what opportunities they give for analysis and applications. I finally review existing studies in the area and how my work builds upon this.

- In Chapter 3, I focus on digital traces collected from online social networks. I first analyse a global information spreading event on Twitter and model its dynamics over the social links between its users [DLM+13]. Motivated by the strong influence of geography on similar processes, I then show how popular node measures typically used for social networks, such as degree centrality, can be extended to include a spatial component and then be used in the context of location-based social networks [LM12]. These measures capture the effects of social links on the spreading of information in a given area; I evaluate the effectiveness of the measures by means of two practical applications. I conclude the chapter by analysing a geo-social network of different nature: the network of open-source collaboration on GitHub [LRM14]. I find that, for this kind of interaction, geography has a strong influence, as people tend to collaborate with people who are at close range.

- In Chapter 4, I turn my attention to the analysis of mobile network data, a data source that describes the movements and communication patterns of a large fraction of the population of a country. I investigate how mobile network traces collected at country level can be used to counteract epidemic outbreaks considering a real scenario. I study this problem both at macro/population scale and micro/individual level. At macro scale I predict the evolution of disease spreading and evaluate different types of countermeasures [LDP+15]. In particular, I build a model that simultaneously accounts for epidemic spreading and dissemination of information that can slow down the disease and then I use aggregated traces to assess whether it is possible to use the social network to disseminate information against the disease and to what extent the strategies based on this model can be used to curb the diffusion. At micro scale, I use individual location traces to evaluate the risk related to an individual carrying the disease, when moving between regions [LPR+15].

- In Chapter 5, I focus on individual mobility trajectories, collected from mobile network data (CDRs) and from GPS traces. Fine-grained GPS traces capture in-

dividual mobility pattern with a great level of detail that is not available with other data sources. I first outline a method that converts raw GPS traces collected from cars into structured data about routine trips and route choices, while being agnostic of the underlying street network [LSP+16]. Using this approach, I find that drivers often follow routes that are not optimal. I then characterise how far from the ideal optimal route they are willing to go while on their way, finding that routes are typically contained by an elliptic area of high eccentricity. While this first problem consists of finding structure from dense mobility data, I then move to a related inverse problem of making sparse data denser, estimating the mobility trajectories followed by mobile network subscribers from time-sparse and spatially coarse-grained data about their association to mobile towers [LLK+14]. In particular, I analyse and model the antenna towers' coverage in urban areas and I use this model to inform routing along the location waypoints reached by the user. While this method can estimate trajectories at block-level less accurately than GPS, the former comes at a fraction of its cost and energy requirements. It also allows network operators to mine information about how people move between places and provide novel services based upon it.

- In Chapter 6, I conclude by reviewing the contributions of this thesis and outlining how future research can build on it.

The various data sources used throughout the dissertation all tell a story about human mobility and interaction, with each one giving a particular angle, coming with specific opportunities, limitations and biases. For this reason, I find throughout my investigation that a data source might be particularly tailored to certain types of analysis and unsuited to others. The combined perspective of different data sources might be able to overcome such limitations; the methodology described in Section 5.2 is an example of this integrated approach.

Each model was built around the specific properties of the data source it uses and around a specific domain context. Such models might be reused in other instances of the same domain, for example on datasets with similar spatio-temporal granularity.

1.3 List of publications

During my PhD I have authored the papers detailed below. Papers related to this dissertation:

Chapter 3

- [LM12] Antonio Lima, Mirco Musolesi. Spatial Dissemination Metrics for Location-Based Social Networks. In *Proceedings for the 4th ACM International Workshop on Location-Based Social Networks (LBSN 2012)*. Colocated with ACM UbiComp 2012. Pittsburgh, Pennsylvania, USA. September 2012.

- [DLM+13] Manlio De Domenico, Antonio Lima, Paul Mougel, Mirco Musolesi. The Anatomy of a Scientific Rumor. In *Scientific Reports* 3:2980. Nature Publishing Group. October 2013.
  Awarded best publication of the month (October 2013) in the College of Enginnering, University of Birmingham.

- [LRM14] Antonio Lima, Luca Rossi, Mirco Musolesi. Coding Together at Scale: GitHub as a Collaborative Social Network. In *Proceedings of the 8th AAAI International Conference on Weblogs and Social Media (ICWSM 2014)*. Ann Arbor, Michigan, USA. May 2014.

Chapter 4

- [LDP+15] Antonio Lima, Manlio De Domenico, Veljko Pejovic, Mirco Musolesi. Disease Containment Strategies based on Mobility and Information Dissemination. In *Scientific Reports* 5:10650. Nature Publishing Group. June 2015.

Awarded First prize (best overall) at the Orange Data for Development Challenge 2013.

- [LPR+15] Antonio Lima, Veljko Pejovic, Luca Rossi, Mirco Musolesi, Marta C. González. Progmosis: Evaluating Risky Individual Behavior During Epidemics Using Mobile Network Data. Submitted to the Orange Data for Developement Challenge (D4D) 2014.

Chapter 5

- [LLK+14] Ilias Leontiadis, Antonio Lima, Haewoon Kwak, Rade Stanojevic, David Wetherall, Konstantina Papagiannaki. From Cells to Streets: Estimating Mobile Paths with Cellular-Side Data. In *Proceedings of the 10th ACM International Conference on Emerging Networking Experiments and Technologies (CoNEXT 2014).* Sydney, Australia.

- [LSP+16] Antonio Lima, Rade Stanojevic, Dina Papagiannaki, Pablo Rodriguez, Marta C. González. Understanding Routing Behaviour. In *Journal of Royal Society Interface* 13:116 (Mar. 2016).

Papers not related to this dissertation:

- [LM14] Antonio Lima, Mirco Musolesi. The Rebirth of Locality: Information, People and Places in a Connected World. In *Proceedings of the GeoHCI 2013 Workshop.* Colocated with CHI 2013. Paris, France. April 2013.

- [DLM13] Manlio De Domenico, Antonio Lima, Mirco Musolesi. Interdependence and Predictability of Human Mobility and Social Interactions. *Pervasive and Mobile Computing.* 9:6 (Dec. 2013)
  Awarded First prize at the Nokia Mobile Data Challenge (open track).

- [ÇLG16] Serdar Çolak, Antonio Lima, Marta González. Understanding congested travel in urban areas. *Nature Communications.* 7:10793.

Some of the publications related to this dissertation are the result of fruitful collaborations. In [DLM+13] I collected the data and developed the model, Manlio De Domenico provided support in developing the model and implemented it, all the authors participated in the experimental design and wrote the paper. In [LLK+14] I developed the methodology and carried out code implementation, Ilias Leontiadis collected the experimental dataset, provided support with the development and carried out code implementation, while all the authors provided support with the experimental design and wrote the paper. In [LDP+15] I designed the model and carried out the code implementation, Manlio De Domenico provided support in developing the model and also carried out the code implementation, all the authors participated in the experimental design and wrote the paper. In [LPR+15] I designed the model, designed the experiments and carried out the code implementation, while all the authors participated in the experimental design and wrote the paper.

CHAPTER 2

DIGITAL TRACES OF HUMAN BEHAVIOUR: AN OVERVIEW

In this chapter I will introduce the content of the thesis by describing the main sources of digital data that have enabled the study of human behaviour, particularly with regards to human mobility and interactions. In the first part of the chapter I will focus on the different classes of data sources that I will use, discussing for each of them advantages, pitfalls and relevant literature. Then, I will present the main results related to the use of these data sources for investigation of human mobility and interactions and, in particular, how my dissertation relates to the previous body of work in related areas.

2.1 Sources of digital traces

With the term *digital trace* I refer to records of human activity that are collected through and stored by means of digital devices [LPA+09]. The term is very generic and includes a large number of forms of data coming from devices of various nature and related to several aspects of human behaviour. However, the term does not include pieces of data that are stored digitally but do not relate to human activity (e.g., the temperatures recorded at a certain location). Similarly, the term does not refer to data that is about people but has been collected manually and possibly stored later in devices (e.g., census information). A complete classification and study of digital traces would be prohibitive to make. Instead,

	Online social networks	Call data records	GPS traces
Trace collection	Manual (on explicit user action, e.g., check-in, profile update)	Automatic (periodically and on user action, e.g., call, SMS, movement)	Automatic
Time sparsity	Sparse, minutes to hours	Sparse, minutes to hours	Dense, seconds to minutes
Spatial granularity	Fine to very coarse, metres to country level	Typically coarse (a few kilometres)	Very fine, metres
Number of individuals	Thousands to millions	Thousands to millions	Thousands

Table 2.1: Properties of digital traces in relation to the data sources they are collected from.

the goal of this thesis is to focus on two salient aspects of human behaviour, namely human mobility and interaction between individuals. I will also restrict my investigation to digital traces generated by three classes of systems: 1) **online social networks** (OSNs); 2) **mobile phone networks**; 3) **positioning devices** (e.g., GPS systems), such as those used by smartphones and car navigation systems. On one hand, these three classes of systems are similar, since each of them can be used to investigate both the location and the interactions of an individual, although in different ways. Online social networks typically allow users to specify their location and have interactions with other users (messages, follow/friend links, etc). Mobile network carriers keep logs of network activity, which can be used to identify the regions and the contacts of an individual and possible co-locations with other individuals. Finally, positioning systems like GPS can be used to extract a very precise trajectory of user movements and can also reveal when two or more people are co-located, i.e., they are at a short distance from each other. While all these data sources allow researchers and practitioners to analyse location and interaction, they are also very different in various aspects, as summarised below:

- **Trace collection.** A first distinction can be made depending on how the traces are created, i.e., either *passively*, through sensors or automatic systems (e.g., GPS

positions, IP address logs, phone call logs) or *actively*, through reporting (e.g., surveys, OSNs profile information, check-ins, user prompts). The collection method has a strong influence on the validity of data: passive traces are vulnerable to system errors and sensing noise; active traces can be misreported by mistake and more easily falsified.

- **Temporal sparsity.** Data can be recorded at various rates, depending on the underlying nature of the data and the system in use. The temporal sparsity can go from mostly static (e.g., for nationality and city reported on a OSN profile, which rarely changes in time), to frequent updates every second (e.g., GPS trajectory traces). This property often depends on whether the data is recorded in relation to a specific event or is sampled automatically at defined time-intervals. For example call data records, like those analysed in Chapter 4, are typically recorded when there is network activity (either a text message, a phone call or data transmission); GPS traces, instead, might be sampled regularly, as in the case of the study described in Section 5.1.

- **Spatial granularity.** The location information contained in a digital trace can refer to areas of different size, from large areas, such as a country or a city, to very specific geographic locations, such as those defined by geographic coordinates or a postal address. Common cellular towers, for example, can be associated to coverage areas with a radius of a few kilometres, while GPS traces typically have an accuracy of a few metres [HLC12].

- **Representativity.** Depending on how many people are using the system that collects the data and how people belonging to different demographic classes are likely to use it, the data might suffer from sampling biases and might not correctly represent the general population. For example, social media are typically very biased towards younger and educated people [Dug15].

While this thesis will not deal with each of these aspects specifically, it is important to keep them in mind as inherent limitations of the data that is being analysed and for that reason I will refer back to these properties throughout the dissertation. In Section 5.2 I will offer an example of how it is possible to overcome some of the limitations by combining multiple data sources. Additional data sources that do not strictly fall in the three categories previously described (OSNs, cellular data, GPS) might share similar features to one category or another. For example credit card traces are similar to social-networks check-ins in time sparsity and space granularity [dMRS+15].

2.1.1 Online social networks

Broadly speaking, an online social network (OSN) is a platform built to support social interactions between its members [MMG+07]. A social network can be represented as a graph $\mathcal{G} = (V, E)$ with N nodes and K links, where nodes are users and links are the social connections between them. While social networks can be abstracted by similar models (i.e., directed/undirected monopartite/bipartite graphs) and they all support social interactions of some kind, it is important to keep in mind that each of them is used in a different context and links between users have a different meaning. Each social network is typically built to primarily support a specific type of interaction, for example friendship ties in real life in Facebook, shared interests in Twitter and Pinterest, work connections in LinkedIn, open-source software collaboration in GitHub, places recommendation in FourSquare. Despite this diversity, some common traits, shared by most online social networks, have been observed in the past, most importantly: heavy-tailed distributions, often following a power-law distribution, indicating that few nodes have a high number of links, while the majority has only a small number [MMG+07; New03a]; high clustering coefficient and community structure, indicating that subsets of nodes tend to be very densely connected between each other [GN02]; and a deep connection between the geographic and social features of the nodes in the network [LK07; LNK+05; SNM11].

In order to study the two aspects of human behaviour this thesis is focussing on (i.e., human mobility and interactions) from OSNs data, it is first necessary to understand how information related to them is available in such services. For this reason, in the remainder of this section I will briefly discuss how this information is generated and can be retrieved from these services.

There are two main sources of location information: individual profiles and geotags. Individuals can customise their personal profile with information about themselves. The profile information disclosed by a user can include their main location, typically specified at country or regional level. Since this information is usually specified as a free text field, an additional step, called *geocoding*, is required to convert the textual name into a location [RK10]. Geocoding is not trivial to perform accurately and has limitations when the specified location is inaccurate, ambiguous or simply incorrect. The name "Cambridge" does not make it clear whether it is the city in England or in Massachusetts. Similarly, geocoding fails when a string identifies several places (e.g., "Rome and NYC") or fictitious locations (e.g., "Everywhere" or "in outer space"). Some services, such as GitHub, check that the location provided by the user is valid; in these cases the geocoding process will work well compared to when the validity check is not performed. As I will show later in Subsection 3.2.2, most users on Twitter prefer to share this information at town/city level, whereas a minority discloses coarser or finer location identifiers. Because of its coarseness, changes in the text field are typically rare and this information is mostly static.

Some social networks, for instance FourSquare, Twitter and Facebook, also allow individuals to share several locations, in terms of absolute geographic coordinates or as well defined points of interest (a shop, a building, a museum, a train station, ...), and related to well defined date and time. This spatio-temporal information can be attached to other content and it is referred to as a *geotag*. Several issues can compromise the validity of this data source. First of all, the demographics of the user base might be strongly biased, as has been found in Twitter [MLA+11]. Secondly, the action of sharing

a location is not automatic, but manual and self-reported, hence people might forget or deliberately avoid to check-in at times. It is also explicit and related to a very precise location, with defined semantic meaning; depending on what people want to share with others, what they want to keep secret and how they want to appear to the outside world, they will more eagerly check-in in some places (perhaps too eagerly, without having really been there) than in others. As an example, people might not check-in every time at their workplace, thinking it might be uninteresting and repetitive; they might not check-in as often in bars and pubs, if they are afraid to appear unprofessional to their boss and work contacts. Thirdly, the selection of the check-in venue is usually aided by GPS and Wi-Fi based location sensors, making it difficult to report incorrect locations; however, it is ultimately the user who chooses the place to be shared, hence he could select one of the neighbouring places. Finally, a few services encourage check-ins by giving out awards to prolific sharers, in the attempt to increase user activity, and encourage people to remember to check-in in each place they go to; on the other hand, this might also encourage false reporting. These concerns have been studied in [ZZZ+13]. While the concerns do not automatically make investigations that make use of this data invalid, it is sensible to take them into consideration before drawing general conclusions from results based on these data sources. It is also worth noting that this study has been criticised for selection bias, over-representing "power users" who are strongly motivated to over-report check-ins.

While for the analysis of location information from OSNs the aspects discussed above need to be taken into consideration, the analysis of social interaction is relatively more straightforward. As previously described, social networks allow users to *follow*, to *add as a friend*, or otherwise add another user as a contact (the exact terminology of the "connection" varies from service to service). However, OSNs typically allow for several types of interactions and each of these can be represented using a different network. Just as a matter of example, Twitter allows users to *follow* other users, but also to *retweet* messages posted by others, to *reply* and to *mention* other users. Each of these action has

a different meaning [CHB+10], so has the network it defines [ODA15]. In Subsection 3.1.2 I will analyse the mention and reply networks on Twitter.

In some cases the interaction between one or more individuals may be mediated by an intermediate entity. For example, Facebook users might be members of the same *group* and GitHub users can collaborate on the same project (as I will explain in more detail in Section 3.3). This kind of interaction is best represented through a bipartite network [HLE+03], where nodes are partitioned into two categories and links are only allowed between nodes of different categories. In the GitHub case, discussed in Section 3.3 users do not collaborate directly with each other, but collaborate on projects; the network representation of such a system is a bipartite network connecting user nodes to project nodes. In order to make the analysis simpler, it is possible to perform a one-mode projection of the bipartite network, transforming it into one of two possible monopartite networks. In the GitHub example, for instance, the collaboration network can be projected on project nodes, linking projects that have at least one collaborator in common; it can also be projected on user nodes, linking users who have collaborated on at least one project.

2.1.2 Mobile network traces

Mobile network traces are logs maintained by cellular networks operators, which detail the activity of customers on their infrastructure. They include CDRs ("call detail records" or "charging data records"), which only report billable actions performed by their customers, such as SMSs, calls and data usage. While similar logs are also kept for landline subscriptions, mobile networks traces are particularly interesting as they can reveal both mobility and interaction patterns of individuals. Therefore, this data source has a great research potential and can be used to unveil human behaviour patterns previously not known [BDK15; GHB08; SKW+10]. It also raises important privacy concerns related to how this data is obtained, analysed and shared, to ensure correctness and replicability of the research results, while preventing misuse [Pen09]. Whereas the specifics of mo-

bile network traces vary depending on the network carrier and the equipment used, they commonly include the following information:

- the phone number of the originating customer;
- the phone number of the receiving line;
- the action type (call, SMS, transmission of data, ...);
- a timestamp;
- details about the action, such as the call duration;
- information about the base transceiver station (BTS) that received the communication event;
- additional technical information, used to investigate faults in the network.

Telephone operators initially collected this data for billing and technical purposes, in order to be able to run the network. Over time, it became gradually clear those data also had commercial value and research potential, once appropriate privacy-preserving measures were ensured (for example, one-way hashing of phone numbers). Because of that, fairly recently some researchers were finally able to analyse this data and find interesting results about human mobility patterns [GHB08; PMJ+14; SGM+12] and their predictability [SQB+10].

The main advantage of mobile network data, over other kinds of mobility and interaction data, is the number of users it relates to. Data of this kind can be available at country level and, because of the much higher penetration rate of mobile phones than online social networks, it suffers less from selection bias. The spatio-temporal accuracy of the mobility inferred from this kind of data is limited: each location data point is typically kept when there is some network activity (a call, a text message, a data transfer, a handover); each data point also specifies the location in terms of the tower sector the device is connected to, which typically involves an area of a few km^2. On the other hand,

network data capture social interaction between individuals in great detail, allowing for instance to understand not only who the social contacts of a person are, but also how many times they typically communicate via call or text.

While data used for these initial studies is not available to the whole research community, in recent years some initiatives, promoted by telecommunication companies themselves, have tried to move to new sharing paradigms that strike a balance between restricted and open data access. The aim is to make the data available to a larger group of researchers, while preventing misuse. Some examples of these initiatives are the Nokia Mobile Data Challenge[1], the Orange Data for Development Challenges[2], and the Telecom Italia/TIM Big Data Challenges[3]. In all cases the data made available is heavily redacted and anonymised, through various kinds of hashing, geographic fuzzying and spatio-temporal aggregation. Researchers who want to obtain access to the data have to clearly state beforehand what the aims of their research projects are, while agreeing to use it only for those purposes.

2.1.3 Positioning services

Positioning services provide location information anywhere on Earth. While the most widely known positioning service is probably the Global Positioning System (GPS), there are several other services, based on satellite technology, like GPS, or other technologies, such as Wi-Fi and GSM signal. Often these technologies are combined, to provide higher precision and compensate for limitations of one system with other systems' strengths. For example, the typical position of a smartphone mainly relies on Wi-Fi, cell network and GPS at the same time. While in this dissertation I will mainly use GPS data generated from either GPS-only navigation systems or GPS-enhanced data sensed by mobile smartphones, the results obtained might also be valid for positioning systems with similar accuracy properties, which might be available in the future.

[1] http://www.idiap.ch/project/mdc/
[2] http://www.d4d.orange.com/en/Accueil
[3] http://www.telecomitalia.com/tit/en/bigdatachallenge.html

The typical accuracy of GPS traces is a few metres [HLC12]. Since this system is based on satellite communication, the accuracy can greatly degrade indoors and in urban areas with "street canyons", because of multipath and shadowing effects caused by skyscrapers and tall buildings [WGZ13]. Because of the high accuracy of GPS, its use for navigation purposes is widespread among users. The analysis of GPS traces has been successfully used in several studies where accuracy is important, such as those related to routing behaviour [LZ13] and to the identifiability of trajectories [RWM15].

A typical entry from a GPS track includes the following information:

- a timestamp;
- the geographic coordinates (latitude, longitude, altitude);
- the current speed;
- the current bearing;
- an expected current accuracy.

Continuous logging of user position is a power intensive operation in most recent smartphones. Some methods which aim to mitigate this problem by using other sensors, like accelerometers and Wi-Fi, have recently been developed and are used by some services like Moves.app[1] and Google Location History. Because of reasonable privacy concerns, accessing data about a high number of individuals for research purposes is extremely difficult.

2.1.4 Comparison

Which data source is the best? Which data source is the most informative? There is no single answer, as the "best" data source depends on the kind of analysis needed and the application domain. Indeed, throughout the thesis I will use different data sources for

[1] https://www.moves-app.com

different kinds of analysis and applications, although all related to human mobility and interaction.

Moreover, great potential is hidden behind the combined analysis of these data sources. I will give an example of this in Chapter 5, by describing a model that uses insights gathered from a GPS dataset to enhance the spatial and temporal resolution of CDR logs. The result of this process is mobile traces that share the advantages of the CDRs (high penetration, low battery consumption, ...) with the advantages of GPS (high temporal and spatial resolution).

2.2 Applications of the analysis of digital traces

2.2.1 Human mobility

One of the first important studies about human mobility based on large-scale digital data was performed in 2008 by González et al. and it was based on mobile phone data [GHB08]. Before that study, mobility had mainly been studied through census and survey-based data, from the earliest attempts made to characterise migration [Rav85] until the more modern efforts to understand the urban space and improve its transportation [SG07]. One of the main findings of the pioneering study by González was that human movement is characterised by significant regularity, because individuals are returning frequently to a few key locations, such as home or work. Another study performed two years earlier [BHG06] used the displacement of dollar bills as a proxy of human movement. While this study had found some properties later confirmed in the mobile phone study, it could not identify individual regularity: properties related to personal mobility patterns were not possible to investigate through items that frequently change owner, such as banknotes.

The high regularity of human movement was later found to be also paired with slow exploring patterns [SKW+10], making individuals' unique behaviour highly predictable [dMHV+13; dMRS+15; SQB+10]. In general, the analysis of mobile phone data

has fuelled a vast number of studies related to a deeper understanding of the city and its transportation network [AJM+15; ÇAA+15; LLP+15; SBC+13; TÇS+15]. As discussed in the previous section, while this data source describes the mobility of a significant fraction of the population, it suffers problems of being spatially coarse and temporally sparse. I will describe a methodology that tries to address this problem in Section 5.2.

In recent years other studies have used data sources that are typically more precise, such as GPS [AS03], and traces collected from online social networks, which typically carry additional semantic information in the form of metadata, e.g., FourSquare [NSL+12a; SNL+11]. Such detailed data has made it possible to investigate theories about human mobility to a greater extent, for example to try to understand how different place features lead to the formation of friendship links [SNM11]. GPS data, in particular, have been used to investigate problems that require a high level of spatial precision, such as routing behaviour [LZ13; ZL15].

In the area of human mobility, the contribution of this thesis is threefold. Firstly, while most recent advances about human mobility are about individual's visit pattern, i.e., what venues are visited and how often, there is relatively little research on how individual actually perform such movement, in terms of choosing and following a route. Most results involve small-scale studies from transportation research [ASW+03; BS10; MAC15] and psychology [BSU00; BAM+12; GSP+85] about the several *factors* that influence route choice [Gol95]. Despite evidence that many reasons influence route choice, most route assignment models accept Wardrop's principles [War52] and assume every driver aims to minimise travel time. In this thesis, in Chapter 5 I try to *measure* how many routes are used by individuals, how often and how dispersed in space they typically are.

Secondly, there is much research about estimating travel paths from noisy location data. When the location data is sampled frequently ($\approx$ 10s) and has low spatial noise ($\approx$10m) approaches referred to as map-matching are used [LZZ+09]; other approaches that use more noisy radio sources, like Wi-Fi and GSM/UMTS rely on signal attenuation models, trilateration and fingerprinting [MR07] and need access to the signal strength.

No prior studies have attempted to estimate fine-grained trajectories from coarse cellular tower association logs, without information about the signal strength (e.g., RSSI). In Chapter 5 I will describe a solution that allows this type of estimation, using spatial probability density distributions to bias route-selection on the street network.

Thirdly, mobility has been widely studied as the principal driver of disease spreading [BCG+09; CBB+06; PTC12]. However, there have been no efforts aimed at understanding the interaction between social networks and mobility in epidemic spreading, apart from a study [GGA13] which appeared after ours [LDP+13]. In this thesis I try to fill that gap in Chapter 4 and I model to what degree countermeasures spreading through social contacts can be effective in delaying the contagion. Some studies have explored the effects of self-initiated changes in the mobility of individuals during an epidemic outbreak [MPA+11] and other studies have demonstrated the importance of super-spreaders during the epidemic [BL12; KGH+10; Ste11]. This thesis presents a novel methodology to rank important spreaders based both on their mobility and the state of the regions involved in their movement, and to act on high-risk spreaders in order to appreciably delay the diffusion of the disease.

2.2.2 Social interactions

Even long before the advent of large datasets of digital traces, scientists have been trying to understand the fundamental mechanisms underlying the interaction between individuals, for example in terms of friendship links [Gra73], diffusion of diseases [Sno55], information [DW04], innovation [Rog62] and collaboration [New04a]. The vast body of research in this field was started by sociologists [WF94] and economists, sparked by the necessity to describe the flow of resources, either material or virtual, between individuals and representing it as *relational ties* between *actors* in a social network [WF94].

More recently, the availability of large-scale datasets of digital traces has sparked the interest of researchers from other areas, such as physics and computer science. These studies have validated older results on much larger datasets [BBR+12; Mil67] and dis-

covered new insights, for instance related to the structure of interaction networks, their dynamics and the processes taking place in them. The digital traces, which such studies are based on, mainly come from online social networks and personal devices.

In particular, the possibility of tracking face-to-face short-distance interactions through personal devices, for example through the analysis of Bluetooth signals, made it possible for scientists to run several ground-breaking studies in this area [CP03; MCM+12; SVB+11]. Some of the sensed interactions [SCH+06] were used to develop mobile network protocols. It is worth noting that, while the data used in these studies was purposely collected to track interactions via specific hardware and software, in my thesis I will also use data that was collected for other reasons (i.e. mobile network traces as opposed to data collected through specialised smartphone "apps"). On one hand this may give less flexibility during the data collection and it may limit the scope of the research investigation; on the other hand it makes it possible to analyse the behaviour of much larger groups of individuals, in a way that would be too expensive or infeasible for tailored experiments.

The structure of social networks was frequently found to include a few recurring aspects: heavy-tailed distributions, sometimes consistent with power-law distribution [MMG+07; New03a], high clustering coefficient and community structure [GN02] and a strong relationship between the geographic and social dimensions of the networks [LK07; LNK+05; SNM11].

With regard to the processes happening on social networks, spreading processes have received great attention. Historically, the first models of spreading appeared in epidemiology, with the goal of understanding the spread of disease. Although the first mathematical model of disease spreading was described by Bernoulli in 1760, it was only in the 20th century that further models were developed, culminating in the first compartmental model formulated by Kermack-McEndrick in 1927 [KM27]. In compartmental models the population is divided into compartments, depending on their disease state. Individuals pass from one to another according to rules defined analytically.

In their simplest version, the SIR model, there are three compartments: susceptible, infected and resistant (also sometimes referred to as removed or recovered). Each person must belong to one of these compartments at a time. Assuming a fixed population of $N(t) = N$ individuals, then we have:

$$S(t) + I(t) + R(t) = N(t) \tag{2.1}$$

where $S(t)$, $I(t)$ and $R(t)$ represent the temporal evolution of the number of individuals belonging to susceptible, infected and resistant compartments, respectively.

Several variations of this model are present. The SIS model suppresses the resistant state to reflect spreading of diseases for which immunity is short-lived (for example, cold and influenza). The SEIR includes an *exposed* state that represents individuals who are infected but not infectious, to model diseases that have a latent period.

Compartmental models have also been used to model spreading phenomena other than epidemic outbreaks. The *Bass Model* [Bas69] was formulated to describe the adoption of new products in a population and it is an extension of the SIS model. Apart from the different vocabulary and context, its novelty is the added possibility of infection from an external force by *innovation*, rather than only from infection of other individuals by *imitation*.

A related view of how collective behaviour unfolds is given by threshold models [Gra78; Sch78]. These models build on the assumption that each person has a personal threshold; when the fraction of people who adopt a certain behaviour is above the threshold, the person will adopt the same behaviour. Granovetter suggested [Gra78] that these models could be successfully applied to various situations, including the diffusion of innovations, rumours, diseases, strikes, voting, riots, migrations, etc. A support to the generality of such models comes from the fact that they have inspired recent models of online spreading phenomena [AA05; CMG09; KKT03; SMM+11].

Among the assumptions that make compartmental models unrealistic, especially when considering large communities, such as countries, is homogeneous-mixing. Under this assumption, each individual interacts in the same way with the others and no heterogeneity or class-class correlations are considered. This limitation has been overcome through the use of networks, mainly in two ways [BGM07], through contact network models [KE05] and network population models [CV08].

Contact network models [KE05] represent each individual as a vertex in a complex network and interactions between individuals as links. While this allows for very realistic modelling of the spreading phenomena, such an approach is often impractical, for several reasons. Firstly, collecting reliable and complete data about interactions within a real population is a resource-consuming task and prone to errors. Secondly, different networks would need to be built differently for different diseases, depending on the level of interaction required for a contagion-prone contact to occur. Nevertheless, these networks are interesting from a theoretical standpoint because they make it possible to investigate how the spreading of the disease is influenced by the network topology.

Network population models, also referred to as metapopulation models, instead, try to overcome the homogeneous mixing limitation with an hybrid approach, by allowing several populations to have loosely coupled epidemic evolution [CV08]. Within each population the homogeneous-mixing assumption still holds and a distinct epidemic process for each of them evolves simultaneously. However, individuals can jump from one population to another population, according to some mobility rules, for example a Markov matrix that establishes the fixed rate of individuals that move between population i and j during a time unit. These models are effective in describing systems of spatially structured populations, like cities and geographic regions, hence I will use them instead of the others in Chapter 4. Mobility flows between metapopulations are typically inferred from census statistics, air-traffic [BCG+09; CBB+06] and cell phone devices [TBD+14], depending on the scale considered in the study.

My contribution with regard to the field of social interactions analysis is threefold. Firstly, I try to analyse and model spreading of a global scientific rumour. Some other studies have dealt with information spreading online. Adar and Adamic [AA05] have investigated how URLs propagate across weblogs. Vallina-Rodriguez et al. [VSH+12] analysed tweets to understand the spatial extent of the "los indignados" movement and to assess whether it served more as a medium for mass communication or for grassroots communication.

Secondly, I quantify the potential influence that individuals have over a geographic region by extending measures of centrality taken from networks analysis [New10]. A previous study [KKT03] has looked into finding the most important individuals to target in order to maximise information spreading in a social network. However, this study focusses on information spreading, assuming spreading by threshold or cascade; this study also does not take into account the spatial component.

Thirdly, I analyse the collaboration networks emerging from open source projects on GitHub. First studies on collaboration networks are mostly centred around coauthorship networks of scientific literature [DWS+14; New04a; New04b]. More recently, Wikipedia's encyclopedic entries [BKL+09] and open-source projects [HGH08; VS07] have been analysed. However, these studies focus on the analysis of interactions between contributors over specific projects, rather than analysing collaboration of multiple users on several projects. With respect to this body of work, to the best of my knowledge, my thesis presents the first systematic quantitative analysis of the interactions in GitHub, discovering global patterns of interactions that cannot be obtained by means of small-scale and interview-based studies.

2.3 Summary

In this chapter I have reviewed the main sources of digital traces currently available for analysis of human mobility and interactions, i.e., online social networks, cellular data and

positioning services. While all these sources have something in common, as they all allow researchers and practitioners to analyse aspects of human movements and interactions, they significantly differ in several aspects. These differences are very important as they can make one class of digital traces better suited for a specific analysis.

Then I have outlined the main areas of applications that make use of digital traces. As discussed in Section 2.2, the study of digital traces has advanced scientific knowledge of general aspects of human behaviour, such as mobility, information dissemination and social interactions. Several studies have focussed on more practical aspects related to a myriad of different areas and applications, such as optimisation of transportation networks [ÇLG16; TÇS+15], place recommendation [NSL+12b], epidemic outbreak simulation and content delivery in computer networks [SMM+11].

In Chapter 3 I will use data collected from online social networks to investigate human interaction in the form of information dissemination and collaboration and how geography affects it. In Chapter 4 I will use cellular network data to design and evaluate mitigation strategies against epidemic outbreaks. In Chapter 5, I will use GPS data to learn behavioural patterns of drivers, which can be used in traffic modelling and simulation, which can be an alternative to optimisation-based rules. I will show how the analysis of both cellular data and GPS traces, when concurrently collected, makes it possible to design a methodology that estimates the position of many mobile network subscribers with increased accuracy. Future work might further investigate general methods that combine knowledge gathered from multiple data sources, with different degrees of granularity, accuracy and precision, with the goal of mitigating data noise, incorrect entries and errors.

CHAPTER 3

INFORMATION PROCESSES IN ONLINE SOCIAL NETWORKS

The widespread availability of Internet broadband connections today allows people who are distant to communicate with one another [Rai10]. Online social networks have recently gained popularity as one of the main media of online communication [Per15]. The large number of individuals who regularly use online social networks has made them very suitable to study human behaviour at an unprecedented scale, as demonstrated in several studies [BSM10; CML11; KGA08; KLP+10].

In this chapter I will focus on the use of data collected from online social networks to uncover patterns of human interactions. The aim of this chapter is to demonstrate that data collected from online social networks can be used to analyse human behaviour and these insights can then be beneficial for practical applications. The first part of the chapter will be devoted to the modelling of an information spreading process and the analysis of its spatio-temporal patterns. I will then move to the related problem of quantifying potential influence on social networks in which nodes have location information attached to them. Finally, I will analyse patterns of open-source coding collaboration and will show how geography affects it. This chapter is structured as follows:

- Firstly, I analyse the spreading of information on the Twitter Social Network[2] during an event of global interest. I observe the spatio-temporal patterns generated by

[2] `http://twitter.com`

online discussions concerning the event and I design models of information spreading that reproduce the observed data. The observed spatio-temporal patterns suggest that processes like information spreading over social ties are highly influenced by the location of people who are involved in it.

- Motivated by this, I extend traditional measures of influence to geo-social networks to include the geographic domain. I define these measures and suggest how they could be used in common real-world scenarios using data collected from two location-based online social networks, namely Twitter and FourSquare[1].

- Finally, I conclude the chapter by examining collaboration around open-source projects, in order to investigate how collaboration-driven platforms compare to other online social networks. To this end I conduct an extensive analysis of activity in GitHub[2]. I find that this kind of interaction, compared to those happening in other social networks, is less common and happens at smaller geographic distances.

3.1 Global information spreading on Twitter

I begin the analysis of online social interactions by focussing on the interactions observed on Twitter during an event of global importance. The salient aspects of Twitter and its usage are described in Chapter 2.

The event analysed in this study is the announcement of the discovery of a particle compatible with the features of the so-called "Higgs boson", as observed on Twitter. This discovery will be remembered as one of the milestones of the scientific endeavour of the 21st century. The association of the Higgs boson to the idea of a deeper understanding of our Universe and the possibility of the Grand Unified Theory [ADF91; CMP+79; EGH+89; Lan81] is likely to be responsible for the huge popularity of this discovery both in academic

[1] `http://foursquare.com`
[2] `http://github.com`

and non-academic circles. Indeed, the interest from both specialised and mass media increased after the "God particle" nickname was assigned to the Higgs boson [Led93].

The announcement of this discovery was the first of this kind in the era of global online social media: the entire world followed and discussed the news and updates through them, commenting and providing personal views about the event. All this information is publicly generated online and represents an extremely interesting source of data for analysing the global dynamics of this scientific rumour around the world.

On the 2nd July 2012, initial results were presented by the Tevatron team, but they were not sufficient to claim a scientific discovery. The statistical significance of all the combined analyses was 2.9 sigma, equivalent to a 1-in-550 chance that the signal was due to a statistical fluctuation [Ttt+12]. Although of remarkable importance for the scientific community, such an announcement had a weak impact on the general public. Following this, there was a strong expectation, accompanied by rumours, for the corresponding results from the CERN teams. An unofficial video was even leaked during those days [Sam12]. The spreading of these rumours about a possible discovery attracted the interest of media, also outside the academic community, until the official day of the announcement on 4th July during the International Conference on High-Energy Physics 2012 in Melbourne, Australia.

The events before and after the discovery of the boson, can be separated into 4 different periods:

- **Period I:** Before the announcement on the 2nd July, there were some rumours about the discovery of a Higgs-like boson at Tevatron;

- **Period II:** On the 2nd July at 1 PM GMT, scientists from CDF and D0 experiments, based at Tevatron, presented results indicating that the Higgs particle should have a mass between 115 and 135 GeV/c^2 (corresponding to about 123-144 times the mass of the proton) [Ttt+12];

- **Period III:** After the announcement and before 4th July there were many rumours about the Higgs boson discovery at LHC [Sam12];

- **Period IV:** The main event was the announcement on 4th July at 8 AM GMT by the scientists from the ATLAS and CMS experiments, based at CERN, presenting results indicating the existence of a new particle, compatible with the Higgs boson, with mass around 125 GeV/c^2 [AAA+12; CKS+12]. After this, mass media around the world covered the event.

Here I present a characterisation of the spreading of this scientific rumour by analysing the related Twitter user activity before, during and after the announcement. More specifically, this study considers the messages posted in Twitter about this discovery between 1st and 7th July 2012. My contributions are twofold. First, I present an in-depth spatio-temporal characterisation of the information diffusion process extracted from the dataset. I report evidence for non-trivial spatio-temporal patterns in user activities at individual and global levels, as tweeting, re-tweeting or replying to existing tweets. Well-defined trends can be associated to different periods of time. Abrupt changes can be linked to the key events around the announcement. Second, I propose a model that describes the dynamics of the information spreading process over the Twitter network. I then analyse the activity patterns of the individuals that tweeted about this discovery over the period considered. I propose a model for the information spreading over the Twitter network, assuming memoryless individuals where the activation process is driven by social reinforcement at neighbourhood level. Finally, I show that these models are able to reproduce the global behaviour of more than 500,000 individuals with remarkable accuracy.

3.1.1 Description of the dataset

The dataset I analyse in this section consists of messages posted on the Twitter social network, crawled by means of the Application Programming Interface (API) made available by the service itself. It is composed of tweets sent between 00:00 AM, 1st July 2012

and 11:59 PM, 7th July 2012 containing at least one of the following keywords: `lhc`, `cern`, `boson`, `higgs`. The query that was used matched messages containing the selected words, and their hashtag version (e.g., prepended with hash symbol) and the search was case-insensitive.

The query was executed using the Twitter Streaming API. By accessing this service, it is possible to retrieve a stream of tweets in real time as they are posted. During this data collection the host that was used for the data collection did not experience any network error or disconnection. Sometimes, however, when the rate of messages is too high, Twitter removes a certain number of tweets from the stream. When this happens, it reports information about how many tweets are missing. During the data collection phase of this study, Twitter removed 102 tweets out of 985,692 total tweets satisfying the query, accounting for a negligible fraction of them. For this reason, to the best of my knowledge, this dataset is composed of about 99.99% of all the tweets satisfying the search query. The original list of relevant keywords was initially larger, including terms such as `alice` and `cms`. However, the amount of tweets retrieved using these keywords but not related to the Higgs boson was not negligible and, for this reason, tweets that were containing only these ambiguous keywords were discarded.

While the Streaming API provides access to messages and some user profile information, it does not provide specific information of the social links (who each user *follows*), which has to be collected through the Twitter REST API. The social graph was obtained by retrieving the *following* list for each user participating in the process. Because of rate-limiting, it took 25 days in total to retrieve the social graph. It is safe to assume that changes of the network that took place in that period are negligible. In order to support this assumption, it is possible to estimate the amount of missed links, starting from a previous longitudinal study of "follow" events presented in [MKS+11]. Accounting for the different network size in terms of number of nodes and edges, in the worst case 1% of newly created links were missed by the data collection process. This is a quite conser-

vative upper bound in my opinion, considering that the cited work [MKS+11] relates to accounts of celebrities, which are probably more active than common users analysed here.

Recent studies make use of the retweet network (i.e., who retweets whom) and the mentions network (i.e., who mentions whom) to investigate, for instance, the patterns of sentiment expression [BKH+12]. However, it is worth remarking that the main source of information consumption on the Twitter website and clients is the home *timeline*, which contains messages from all social contacts, regardless of the number of reciprocal interactions. Therefore, it is safe to assume the timeline also stimulates users' activity depending on their preferences: people read their contacts' messages and contribute to topics of discussion they find interesting. This study uses the follower network to investigate the temporal dynamics of the fraction of people involved in the process of spreading the information about the Higgs boson discovery as a function of time.

The final number of collected tweets was $985,590$. The corresponding social network of the authors of the tweets is composed of 456,631 nodes and 14,855,875 directed edges. Nodes correspond to the authors of the tweets and edges represent the "user follows" relationships between them. 70,838 users were discarded from the original dataset containing 527,469 users because of the non accessibility of the list of their followers and followed users due to privacy settings. Twitter users can specify their location by filling the *Location* field of their profile, on an optional basis and at different levels of granularity (e.g., United States, New York, Chelsea, etc.). When available, this information is used to assign a geographic position to each tweet: the resulting number of geo-located tweets is 632,027). In order to convert the textual geographic information to geographic coordinates, the Google Geocoder API was used.

Fig. 3.1 shows the distributions of the in-degree, out-degree and total degree of the users that tweeted about the Higgs boson. Intriguingly, the underlying topology is not trivial. While all three distributions are heavy tailed, the out-degree exhibits a lighter tail than the others. Users who follow more than a thousand users are very rare in the dataset, while people who *are followed* by more than a thousand users are quite

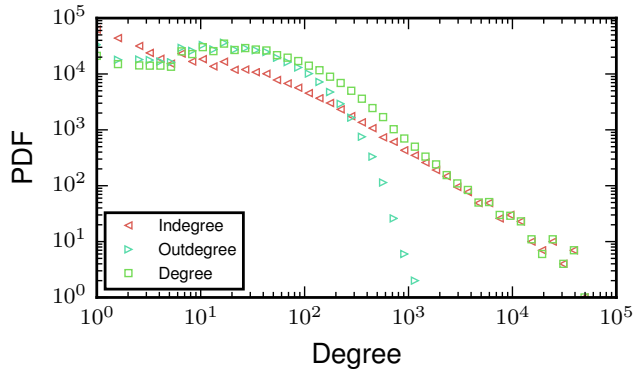


Figure 3.1: Probability density of in-degree, out-degree and total degree of the nodes that tweeted about the Higgs boson.

common. A standard method to uncover the presence of correlations in the network is to investigate the assortative mixing of its nodes [New02a; New03b]. In fact, the nodes in the network with a large number of links may tend to be connected to other nodes with many connections (assortative mixing with positive assortative index) or to other nodes with a few connections (disassortative mixing with negative assortative index). In both cases, the network shows degree correlations resulting in an assortative index different from zero, at variance with an uncorrelated network where this index is close to zero. This study finds a value of about -0.14, indicating the presence of correlations in the network, with disassortative mixing of users. A possible explanation for this consists in the network analysed here actually being a subgraph of the social network. This subgraph, composed by users who mentioned one of the keywords taken into consideration at least once, might exhibit more disassortative links than the original full network, where no topic-restriction is made. This might suggest that, at least for this specific topic, information exchange between high-degree nodes (information hubs) and low-degree nodes (information consumers) prevails.

3.1.2 Spatio-temporal analysis

I first investigate both spatial and temporal features of user activity, as observed on Twitter. More specifically, user behaviour is studied using two different analyses: the

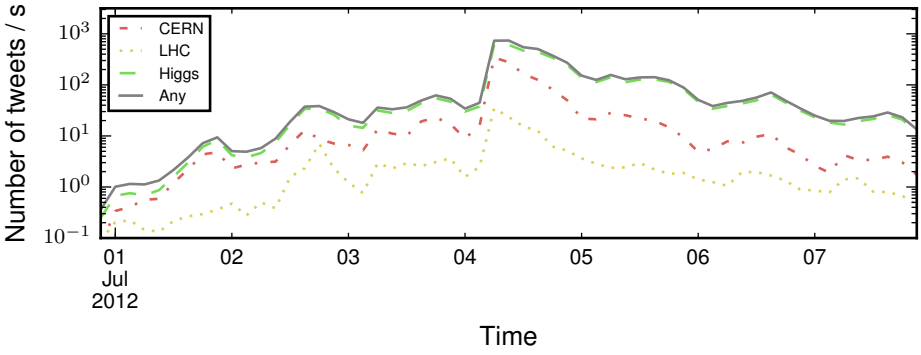


Figure 3.2: Number of tweets per second as a function of time during the period of data collection. The curves correspond to tweets containing only the `CERN`, `Higgs`, `LHC` keywords (as the result of a non case-sensitive search) and at least one of them, respectively.

first one is performed at a global (macroscopic) level ignoring social ties, while the second one is performed at an individual (microscopic) level and based on social ties.

Macroscopic level

Fig. 3.2 shows the evolution of the rate of tweets containing the `CERN`, `Higgs`, and `LHC` keywords, in logarithmic scale. The rate shows a rapidly increasing trend up to the day of the announcement of the CERN teams, after which it slowly decreases. The rate of tweets for all the keywords considered increases from approximately 36 tweets/hour at the beginning of Period I up to about 36,000 tweets/hour at the beginning of Period IV. The rumours anticipating the presentation of results at Tevatron caused the initial spreading of tweets about the Higgs boson. This was further sustained by the subsequent comments to these initial postings and the rumours about the results to be presented by the scientists belonging to the ATLAS and CMS experiments. During a few hours after the announcement of the discovery, the rate increased by more than one order of magnitude, while it slowly decreased in the following days.

The top panels of Fig. 3.3 show the density of tweets before (left panel), during (middle panel) and after (right panel) the main event, on 4^{th} July 2012. The bottom panels in the same figure show the corresponding networks of users built from re-tweets. The impact of the announcement on the 4^{th} was truly global. Instead, before and after this main event

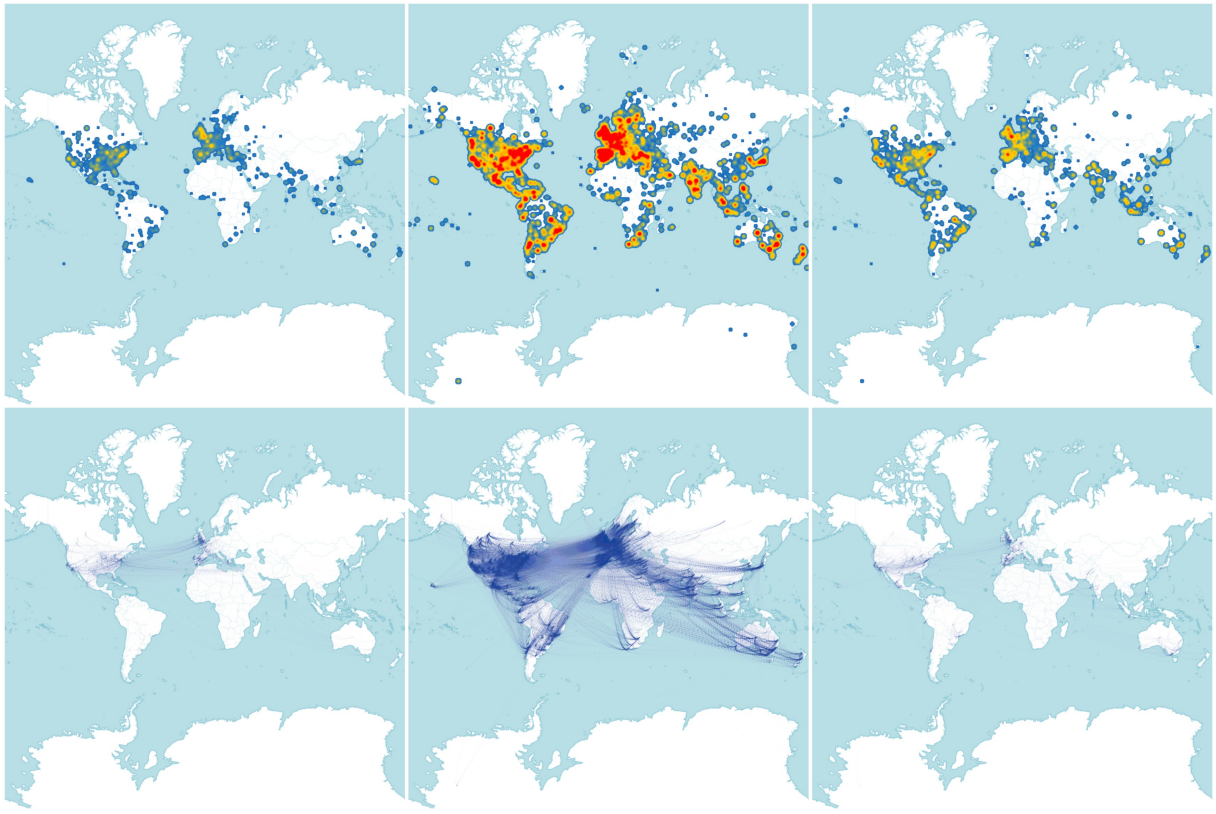


Figure 3.3: Top: heatmap for the density of tweets before (left panel), during (middle panel) and after (right panel) the main event on 4th July 2012. Bottom: corresponding networks of re-tweets between users. During the announcement, the Twitter activity is truly global, whereas before and after the announcement, the most active countries were European and American, due to the large presence of scientists in these geographic areas. Map data: © OpenStreetMap contributors, available under the Open Database License.

the countries with a significant number of tweets were European, probably due to the fact that CERN is in Switzerland and the largest number of scientists working there are from Europe. A large number of tweets were also observed from the United States, which hosts a very large community of scientists. Also, it is possible to notice that while before and after the event retweets are mainly within and between Europe and United States, during the key event those two regions are much more connected to the rest of the world.

I consider the entire set of individuals as a large-scale complex system of interacting entities, and I analyse the dynamics at a macroscopic level of such a system by inspecting spatio-temporal patterns of consecutive tweets. The first goal of this study is to gain insights into the spatial and temporal patterns of this complex geographic social network: in order to do so, I study the time interval and the space distance between consecu-

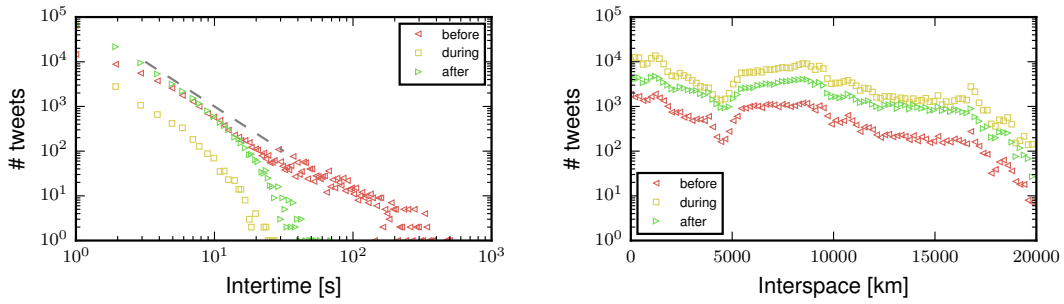


Figure 3.4: Global spatio-temporal activities of *any* user in the social network. Distribution of intertimes (left panel) and interspaces (right panel) between consecutive tweets sent by any user, before, during and after the main event on 4$^{\text{th}}$ July. The dashed line indicates a power law $\sim \tau^{-2}$ and is for guidance only.

tive tweets posted in the network, calling them *intertimes* and *interspaces*, respectively. Fig. 3.4 shows the distributions of these two quantities, grouped by the time interval when they occurred, either before, during or after the main event. While the distribution of interspaces is roughly the same regardless of the time window taken into consideration, the distribution of intertimes in the three windows is very different. From a global point of view, this Twitter activity exhibits long tails before the main event, with a large number of tweets sent within a few seconds, and a small number sent within a few minutes. On the other hand, the dynamics of the process change dramatically during the main event, when the intertimes between consecutive tweets is likely to be less than two seconds and no more than six seconds, indicating a frenetic user activity. After the main event, the activity decreases its intensity, but an exponential cut-off limits the highest values of intertimes to one minute.

Microscopic Level. The user dynamics is now analysed at a microscopic level (i.e., treating individuals separately) by inspecting intertimes of user activities, such as tweeting, replying and re-tweeting. In the following, the intertime for user u is defined by $\tau_u(i) = t_u(i+1) - t_u(i)$, where $t_u(i)$ and $t_u(i+1)$ indicate the times when user u sent the i-th and the $i+1$-th tweets, respectively.

Fig. 3.5 shows the distribution of user activity intertimes τ (i.e., between consecutive tweets by the *same* user) during, before and after the main event. Intriguingly, before

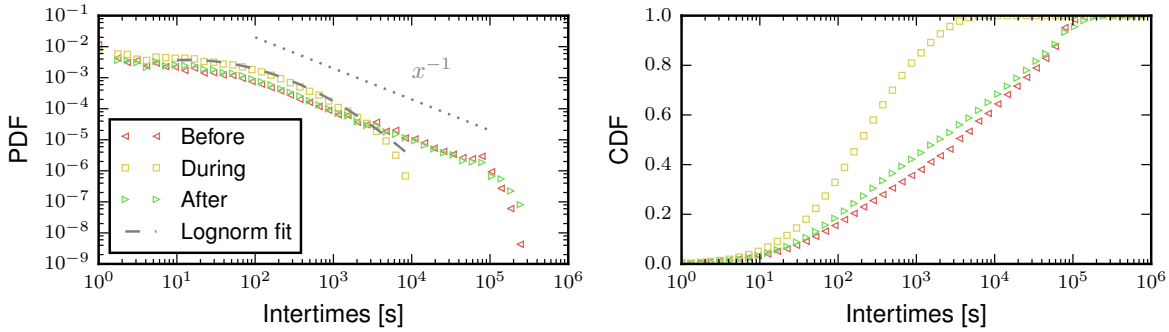


Figure 3.5: Probability and cumulative density distributions of intertimes, before and after the main event on 4$^{\text{th}}$ July. Power scaling behaviour is visible for certain ranges of values, dashed line for guidance only.

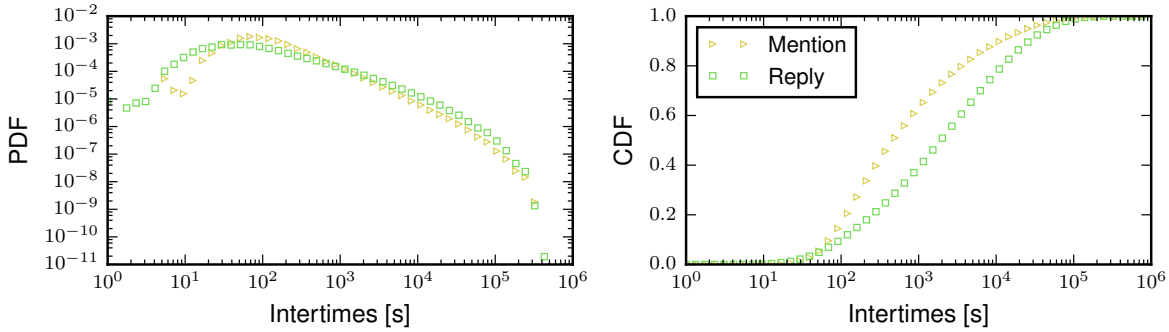


Figure 3.6: Probability and cumulative density distributions of intertimes between consecutive replies and retweets by the same user.

and after the main event, such distributions show power-law scaling of type $P(\tau) \propto \tau^{-\alpha}$, with $\alpha \approx 1$, over three decades of intertimes, from the scale of one minute to the scale of one day.

Timing of user activities is usually modelled using a Poisson distribution. However, there is evidence that intertimes between subsequent user actions follow a non-Poisson statistics, characterised by bursts of rapidly occurring events separated by long periods of inactivity [DWF03; KKB+12; MMW03]. The bursty nature of user behaviour has been recently attributed [Bar05] to decision-based queuing processes [Cob54], where individuals tend to act in response to some perceived priority. According to this model, the timing of tasks to be executed is heavy-tailed, with rapid responses in the majority of cases and a few responses with very long waiting times [Bar05]. Moreover, it has been shown that

bursty user activity patterns might have a remarkable impact on the spreading dynamics over complex networks: this dynamics might be related to the waiting time distribution but it is not sensitive to the network topology [MGV11].

A similar dynamics is observed in this study: the distribution of intertimes shown in Fig. 3.5 reflects the bursty nature of user activities in online social networks, where individuals are more likely to send several tweets in quick succession within a few minutes, followed by long periods of no or reduced activity, up to one day. Inter-tweet times distribution during the main event shows a very different behaviour, more compatible with a log-normal law instead of a power-law scaling relationship. In this case, if τ is the random variable representing the time to the next tweet, the random variable $\log \tau$ is normally distributed with mean μ and standard deviation σ. If a user starts to tweet because he or she is triggered by tweets of users in his or her social neighbourhood, the total number of tweets can pass a threshold value above which a cascading effect may occur in the network. If this is the case, intertimes are distributed following a log-normal distributions if the number of vertices involved in the cascade is large enough [Mit04]. The log-normal law with $\mu = 5.627 \pm 0.008$ and $\sigma = 1.742 \pm 0.006$ describes the observed activities during the main event with remarkable accuracy.

Fig. 3.6 shows the distributions of intertimes for replies and re-tweets during the entire data collection period. Intriguingly, user activities are still characterised by bursty behaviour. Intertimes for replies follow a power-law $P(\tau) \propto \tau^{-1.2}$ from a few minutes up to one day: such a scaling can be explained with the existence of bursty behaviour in the timing of user actions, previously discussed in the case of intertimes before and after the main event. It is worth noting that the scaling exponent is larger for intervals between the original tweets and replies than for re-tweets. For temporal scales larger than one day, the power-law scaling is not present; an exponential cut-off does not model the observed decay.

The case of re-tweets deserves particular attention. From the time scale of a few minutes up to the time scale of a few hours, there is a power-law scaling relationship

$P(\tau) \propto \tau^{-0.8}$, with a cut-off on the time scale of one day. This behaviour can be modelled as a power-law with an exponential cut-off $P(\tau) \propto \tau^{-0.8} \exp(-\tau/\tau_0)$, with cut-off scale $\tau_0 \approx 11$ hours. It is worth remarking that power-law scaling relationships with exponent $\alpha \leq 1$ cannot be normalised and do not occur in nature unless the scaling deviates from power law after some threshold value, the cut-off scale, above which the distribution rapidly falls to zero. Even in such cases, phenomena exhibiting scaling exponents smaller than unity are very rare [New05].

3.1.3 Rumour spreading

In this section, I will model the dynamics of information spreading that occurred among Twitter users who tweeted about the Higgs boson. Despite the fact that information spreading shares some general dynamical features with the spreading of diseases, their nature is deeply different. For instance, disease epidemics depend on the physical contacts between individuals and the different biological characteristics of both the infectious agent and the carrier, as well as many other factors [AM91], whereas information can also be spread through non-physical contacts making use of communication infrastructures such as telephone, television and Internet [MZL12]. Information is very volatile and it is not subject to incubation periods: it is only worth spreading or not and this decision is made by individuals, unlike the case of disease spreading. In Chapter 4 I will show how this crucial difference can be potentially used to mitigate epidemics.

In the last decade, the study of contagion dynamics, involving either information or disease transmission, has greatly benefited from key results in complex networks modelling [KE05; LKG+07; MPV02; New02b; NFB02; PV01a; PV01b; RMK11]: in fact, the structure of social relationships plays a fundamental role for any type of spreading dynamics [BLM+06; DGM08; New03a]. If the underlying topology of the network is homogeneous, the dynamics can be studied by adopting a mean-field approximation and the spreading occurs only if the rate of transmission of information exceeds an epidemic threshold. Conversely, heterogeneous structures like scale-free networks require heterogeneous

mean-field approximation [PV01a; PV01b], involving the single-site equation governing the time evolution of the relative density of "infected" vertices with given connectivity k, i.e., the probability that a vertex with degree k is infected. Moreover, such networks have the peculiar property of facilitating the spreading of infections: in fact, if the corresponding degree distribution shows diverging second moment, then the epidemic threshold is zero independently of the degree correlations [BPV03]. Although mean-field approximations are fundamental tools to capture the main features of the spreading dynamics, particularly in the early stage, the models are less efficient when the finite size of the population becomes a significant factor. More recent approaches focus on the probability of transmission of individual vertices [GAB+10] and non-perturbative formulation of the heterogeneous mean-field approach [GGM+11]. This analysis will distinguish between two different states for users in the social network: "active" and "non-active" vertices. I will indicate with "tweeting activation" or "rumour spreading" the user-to-user interaction process for transmitting information related to a particular topic. In the following, I will indicate with $A(t)$ and $D(t)$ the number of active and non-active users at time t, respectively, with $A(t) + D(t) = N$, where N is total number of users considered in the social network.

The observed social network of active users is shown in Fig. 3.7, where a visualisation based on k-core decomposition and component analysis is presented [ADB+05; BAB08]. The k-core of a graph is defined as the maximal connected subgraph in which all vertices have a degree of at least k. In practice, a k-core is obtained by recursively removing all vertices with a degree less than k, until the degree of all remaining vertices is greater than or equal to k. The k-coreness of a vertex is the index of the highest k-core containing that vertex. Vertices with the highest k-coreness act as the most influential spreader of information in the network. In fact, it has been recently shown that in some plausible circumstances the best spreaders are not the most highly connected or the most central people but those with higher k-coreness [KGH+10], and there is evidence of a positive correlation between k-coreness and the size of cascades of messages, suggesting that users

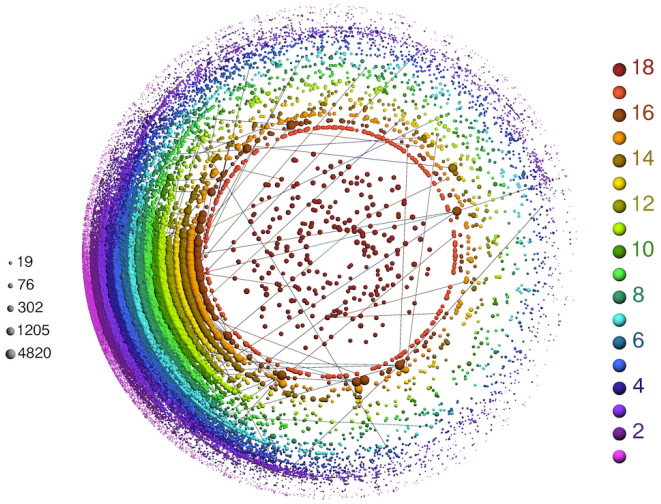


Figure 3.7: Visualisation of the social network of active users, based on k-core decomposition and components analysis. The size of each vertex is proportional to its degree, whereas colour codes the k-coreness. A sample of 10% of the whole network has been used for this visualisation.

at the core of the network are more likely to be the seeds of global chains of information diffusion [GBR+11].

The k-core decomposition allows some salient features to be identified in the observed social network of active users, uncovering structural properties due to its specific topology. In Fig. 3.7, the presence of an inhomogeneous distribution of vertices in the shells is a signature of non-trivial correlations. Moreover, the presence of vertices with a high degree in any k-shell, i.e., a very low correlation between degree and shell-index, indicates that hubs are likely to be found also in external shells, a behaviour typical of networks without an apparent global hierarchical structure like the World Wide Web [ADB+05; BAB08].

Modelling the dynamics of user activation without social ties

As the first step, I do not consider the influence of the structure of the network on the process. I define a user as active (in the sense that they contribute to the spreading of information) at time t if he or she has tweeted at least once about the Higgs boson before the instant of time t. In the following, I indicate with $A^\star(t)$ and $a^\star(t)$ the number and the

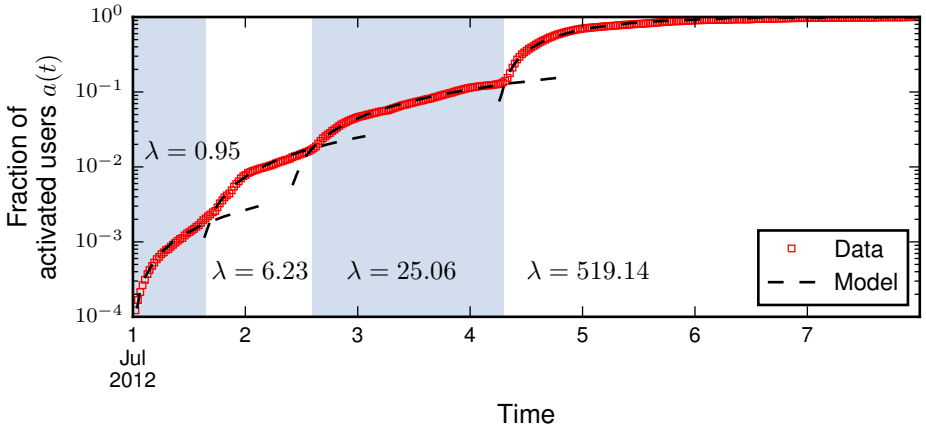


Figure 3.8: Points indicate the fraction of users who are active at least once (see the text for more detail) with respect to the total number of users in the dataset at the end of the period taken into consideration, i.e., $A^\star(t = 8 \text{ July } 2012)$, as a function of time. Lines indicate the fitting results obtained separately for each temporal range by adopting the model given by Eq. (3.3). The rate of activation $\lambda^\star$ for each period is reported at the bottom of the figure.

fraction of active users at time t, respectively. Hence, the number $A^\star(t)$ of active users is expected to be a monotonic increasing function of time.

I divide the whole period of data into four temporal ranges of interest, corresponding to periods I, II, III and IV, described above. In Fig. 3.8 I show for each period the observed evolution of the fraction $a^\star(t) = A^\star(t)/N$ of infected users versus time, where N is the total number of users in the dataset in the data collection period.

In order to model the evolution of the active users over time, I firstly tried to exploit classic susceptible-infected (SI) models in an unstructured population [KR08], but this led to a very poor fit of the data. For this reason, I developed a new model starting from the observation of the specific characteristics of the dataset. I make the simplifying assumption that a user will not tweet again after the tweeting activation. This assumption is consistent with the observed bursty behaviour and a typical scenario where users post messages in rapid succession and then "sleep" for a longer time. In this case, the number of newly active vertices at time t is proportional to the number of users who have not been active before:

$$A^\star(t+\Delta t) = A^\star(t) + \lambda^\star[N - A^\star(t)]\Delta t, \tag{3.1}$$

where $\lambda^\star$ is a constant activation rate. In the limit of small Δt, it is possible to obtain the following ordinary differential equation:

$$\frac{da^\star(t)}{dt} = \lambda^\star[1 - a^\star(t)], \tag{3.2}$$

corresponding to the model for the fraction of users tweeting at least once about the Higgs boson. The evolution over time of $a^\star(t)$ is the solution of Eq. (3.2), given by

$$a^\star(t) = 1 - [1 - a^\star(t_k)]e^{-\lambda^\star(t-t_k)}, \tag{3.3}$$

where $k =$ I, II, III and IV indicates the period of interest, t_k is the starting date of period k and $a^\star(t_k)$ is the corresponding initial fraction of users active at least once.

The evolution function given by Eq. (3.3) can be fitted to the observed data for each period of interest: the resulting model for each case is shown in Fig. (3.8), demonstrating the agreement with the data. The activation rate increases during the four intervals of time taken into consideration, from about one user per minute on 1$^{\text{st}}$ July 2012, up to about 519 users per minute in the last period.

Modelling the dynamics of user activation considering social ties

The goal of this subsection is to understand how likely it is for a person to posts a message with the specified keywords, based on the number of the users they follow who have used the same keywords. Here I also drop the assumption that a user is always considered active in the spreading dynamics. A user is considered non-active in a given time window Δt if he or she has not tweeted in that time interval. In other words, in this refined model, an active user can become non-active again (de-activated) if he or she does not keep tweeting about the Higgs boson. In the time interval between t and $t + \Delta t$ active

users can become non-active after a certain amount of time for any reason: I indicate with $\beta(t)$ the probability per unit of time for the transition from active to non-active state. Hence, the number of users that become non-active in the interval Δt is given by $\beta(t)A(t)\Delta t$. By introducing de-activation, I also account for the limited visibility of tweets on the timelines of Twitter clients, i.e., newer tweets replace older ones. Moreover, I observe that the number of non-active users at time t that will become active at time $t + \Delta t$ is a function of both their in-going degree and the out-going degree of active users at time t.

A non-active user connected to more than one active user at the same time is more likely to become active with respect to non-active users connected to only one active user. Let us indicate with j_A the number of active users connected to a non-active user. If $\lambda(t)$ indicates the activation probability per unit of time per link, for a non-active user with degree k^{in} the probability per unit of time of changing from non-active to active state is given by $p_\lambda(t; j_A) = 1-[1-\lambda(t)]^{j_A}$. In general, the probability that such a non-active user is connected to j_A active users at the same time depends on the out-going degree of active users, i.e., on network vertex-vertex correlations. More specifically, such a probability depends on the conditional probability of observing a vertex with out-going degree k^{out} connected to a vertex with in-going degree k^{in}.

It has been shown that a pure scale-free degree distribution with exponent between 2 and 3 is a sufficient condition for the absence of an epidemic threshold in unstructured networks with arbitrary two-point degree correlation function [BPV03], i.e., correlations at neighbourhood level do not affect the spreading dynamics. I use this result as a simplifying assumption for modelling the spreading in the network, exhibiting a scale-free degree distribution with exponent 2.5 for $k > 200$. Therefore, it is possible to neglect correlations and to estimate the probability that a non-active user, with in-going degree k^{in}, is connected to j_A active users, with *any* out-going degree, by

$$\tilde{p}(t; j_A, k^{in}) = \frac{\binom{A(t)}{j_A}\binom{N-A(t)-1}{k^{in}-j_A}}{\binom{N-1}{k^{in}}}, \tag{3.4}$$

accounting for all the possible ways to arrange $A(t)$ activations within j_A users from the total number of possible combinations of the remaining $N-1$ users within k^{in} users.

Hence, the probability that a non-active user with in-going degree k^{in} is activated by at least one active user in its neighbourhood is given by

$$P_{\lambda,k^{in}}(D \longrightarrow A) = \sum_{j_A=1}^{k^{in}} \tilde{p}(t; j_A, k^{in}) p_\lambda(t; j_A). \tag{3.5}$$

It follows that the total probability that non-active users will become active during a time unit is given by

$$\Theta_\lambda(t) = \sum_{k^{in}} \mathcal{P}(k^{in}) P_{\lambda,k^{in}}(D \longrightarrow A), \tag{3.6}$$

where $\mathcal{P}(k^{in})$ is the probability density of the in-going degree. It follows that $(N - A(t))\mathcal{P}(k^{in})$ indicates the number of non-active users with in-going degree k^{in} at time t. I model the dynamics of the number of active users in the time interval Δt by

$$A(t+\Delta t) = A(t) + [-\beta(t)A(t) + (N - A(t))\Theta_{\lambda(t)}(t)]\Delta t.$$

Therefore, by choosing $\Delta t = 1$, i.e., equal to the time unit of observation, I obtain the general discrete model

$$A(t+1) = (1 - \tilde{\beta}(t))A(t) + (N - A(t))\Theta_{\tilde{\lambda}(t)}(t), \tag{3.7}$$

valid for the general case of activation and de-activation rates that change over time.

In Eq. (3.7) the parameters $\tilde{\beta} = \beta\Delta t$ and $\tilde{\lambda} = \lambda\Delta t$ indicate probability instead of probability rates. However, in the particular case of $\Delta t = 1$ it is possible to mix rates and probabilities because both will have the same values, even though their units are different [GGM+11]: for the sake of simplicity, in the following I use the notation $\beta = \tilde{\beta}$ and $\lambda = \tilde{\lambda}$. Eq. (3.7) represents the balance equation indicating that the number of active

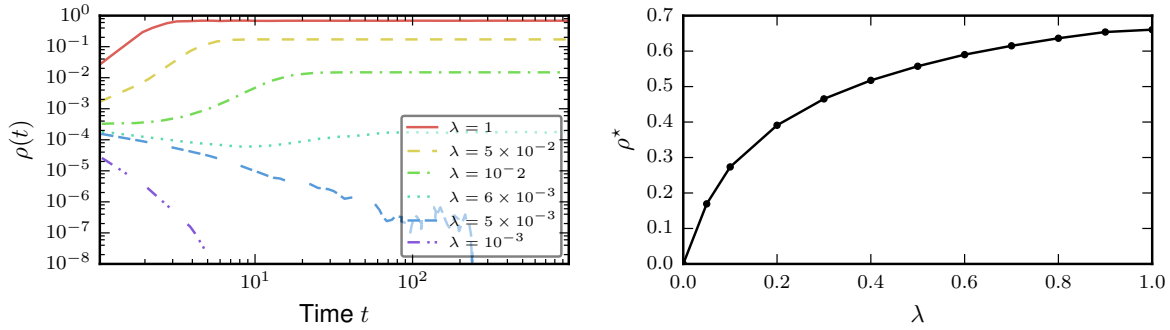


Figure 3.9: Evolution of the density of active users versus time obtained from simulations of spreading dynamics (left panel). The de-activation rate is $\beta = 1$ and the different curves correspond to different values of the activation rate λ. Each curve corresponds to the ensemble average of 200 random independent realisations. Average value of the density of active users in the stationary state, as a function of λ (right panel).

users at a certain instant is given by the number of vertices that at the previous instant did not change from active to non-active state plus the number of newly active users. In the following I will consider the density of active users defined by $\rho(t) = A(t)/N$, leading to the evolution equation

$$\rho(t+1) = (1 - \beta(t))\rho(t) + (1 - \rho(t))\Theta_{\lambda(t)}(t). \tag{3.8}$$

In general, the solution of Eq. (3.7) and Eq. (3.8) cannot be obtained analytically because of the complexity of $\Theta_{\lambda(t)}(t)$: therefore, some simplifying assumptions or numerical methods should be adopted instead.

Let us focus only on Period IV, i.e., during and after the main event, from 03:00 AM on 4$^{\text{th}}$ July to the end of the data collection period. The initial fraction of active users is approximately $\rho(0) = 0.1\%$ of the total number of total users in the dataset.

In order to assess the validity of the analytical model, I perform large-scale Monte Carlo simulations of the spreading dynamics through the network of observed connections among users. More specifically, I consider the case where activation and de-activation rates do not change over time: I vary their values from 0 to 1, independently; for each possible configuration corresponding to the pair (β, λ) I perform 200 random independent

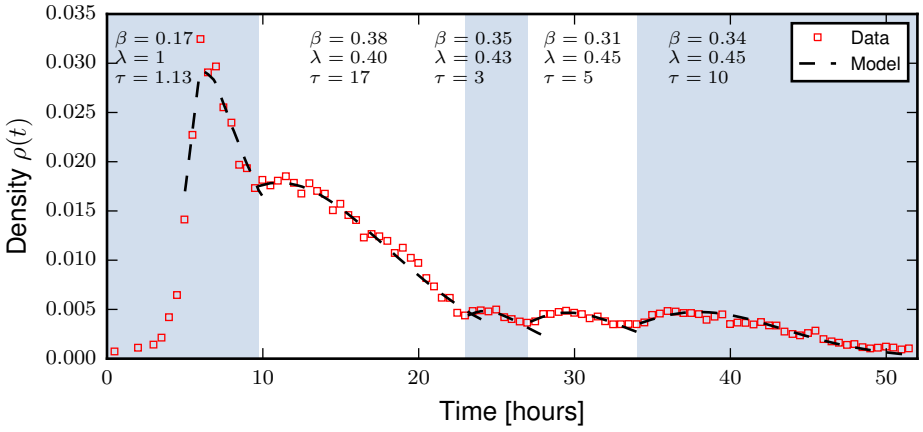


Figure 3.10: Observed evolution of the density of active users versus time (points) in Period IV, i.e., during and after the main event, from 03:00 AM, 4$^{\text{th}}$ July. Curves indicate the predictions obtained from the model defined by Eq. 3.8 coupled to Eq. 3.9, where the values of the corresponding parameters are reported in the figure for different sub-periods. The reported λ refers to the initial value of the activation rate.

realisations of rumour spreading and I calculate the ensemble average at each time step t to obtain an estimation of the expected value of the density $\rho(t)$. The results for the case with $\beta = 1$ and several different values of λ are shown in Fig. 3.9. In the first panel, I show the evolution of $\rho(t)$ versus time: for $\lambda < 6 \times 10^{-3}$ the density $\rho(t)$ tends to decrease to zero for increasing time, while for $\lambda \geq 6 \times 10^{-3}$ the density $\rho(t)$ tends to reach a stationary state, indicating that the spreading becomes endemic. I indicate with $\rho^{\star}$ the stationary value reached by $\rho(t)$ after the transient time. In the second panel, I show the value of $\rho^{\star}$ versus the activation rate: the endemic state, where $\rho^{\star} > 0$, is quickly reached for small values of λ. This result is qualitatively confirmed by the analytical model (see Eq. (3.8)) and it is in agreement with the result reported in [BPV03], stating that the epidemic threshold of an endemic state tends to zero for increasing network size with a scale-free topology. However, such results do not reproduce the observed spreading dynamics, whose density $\rho(t)$ is shown in Fig. 3.10. The data show a quickly increasing number of users within a few hours, with a maximum value reached at the beginning of the International Conference on High-Energy Physics. Such a fast increasing behaviour can be explained by tweets related to the excitement for a possible announcement of the discovery of the

Higgs boson. In fact, the number of active users in the following hour rapidly decreases by about 40%, staying stable for the subsequent 2 hours and then decreasing again.

In case of epidemics with constant activation rate in scale-free networks with a large number of nodes we would expect the appearance of an endemic state. However, this is not the case. For this reason, I modify the model by introducing a variable activation rate λ, accounting for the decreasing interest on a tweet over time, according to recent studies suggesting the existence of a natural time scale over which attention fades [WH07]. I model the evolution of λ as follows:

$$\lambda(t+1) \;=\; (1-\xi)\lambda(t), \qquad 0<\xi<1, \tag{3.9}$$

which is the discrete counterpart of the continuous equation whose solution is the exponential decay $\lambda(t) = \lambda(t_0)e^{-\xi(t-t_0)}$. Here, ξ can be interpreted as the inverse of a characteristic scale τ regulating the decay dynamics. I use the coupled equations (3.8) and (3.9) to model the observed spreading dynamics.

During the whole Period IV, I identify five sub-periods, each one characterised by an increasing number of active users followed by a decreasing one. I then vary the parameters β, λ and τ in order to try to reproduce the data in each sub-period. The dashed curves in Fig. 3.10 correspond to the model (Eq. (3.8) and Eq. (3.9)) using the set of parameters minimising χ^2. The rapid increase of active users in the first sub-period of Period IV is followed by a fast decrease, with time scale $\tau \approx 1.13$ hours, initial activation rate $\lambda_0 = 1$ and $\beta = 0.17$. Such a fast decreasing trend is slowed after about 9 hours, approximately at the time when the rumour has reached the other side of the world in the early morning: from this time instant up to the end of the observation, the values of de-activation probability and the initial value of the activation probability are almost constant (ranging from 0.31 to 0.38, and from 0.40 to 0.45, respectively). In the following sub-periods only the decay time scale τ significantly varies from 3 to 17 hours.

3.2 Quantifying influence in geo-social networks

In the previous section I modelled information dissemination in a network of half a million people using the structure of the social network on which the process occurred. The presence of few high-connected nodes and high k-core nodes shows that large diffusion processes like this often revolve around a small fraction of nodes who are the most influential in the network. The problem of identifying the key influential actors in a social network is well-known in sociology [WF94] and network science [New10] and has recently been applied also to online social networks [CHB+10]. Identifying these key individuals can be important for several applications. For example, it could be used for predicting the popularity of items in a social network by monitoring only a few individuals used as sensors [GMC+14], or it could suggest how to control the dissemination acting on the most influential nodes [KKT03].

Classic measures of centrality have been developed to capture potential influence [New10; WF94]. However, these measures have neglected the geographical aspects of networks. To make a simple example, in a small region a local celebrity might be much more influential than a global celebrity, who might happen to be unpopular in that specific region.

In this section, I propose centrality measures that capture and quantify geographic importance and centrality of users in geo-social networks. I evaluate these measures by associating users to one or more locations, using datasets extracted from Twitter and FourSquare. The measures focus on the structural properties of the geo-social networks and not on the processes happening over them, such as information cascading and retweeting. Moreover, by separating structure and dynamics, they can be used as quantitative generic tools for evaluating the *potential* role of each node in disseminating information in the geographic space.

The need for modelling spatial social networks and finding measures for quantifying geographic centrality and influence comes not only from the ambition to study the complex interactions between the social and spatial dimensions more comprehensively, but also

from a variety of potential practical applications which could benefit from this analysis. These include:

Targeted information spreading. Being able to measure geographic centrality allows us to rank users according to the number of contacts they have in a certain area. Consequently, they can be used to select individuals to be targeted for spreading information. Applications include not only support for advertisement campaigns of certain products or promotions restricted to given areas, but also the design of systems for dissemination of emergency alerts in natural or man-made disaster situations, where information should be disseminated in a spatially-limited area (for example in the case of security alerts in parts of a city or for weather alerts in a certain region).

Models of cultural influence. OSNs are an invaluable source of data for studies in social sciences that were simply not possible in the past [Kle08; LPA+09]. In particular, estimating social and political influence can be very important and relevant for analysing and interpreting several cultural phenomena. For example, a person tweeting in London might have influence also outside it, for example in his or her hometowns, and in case of recent immigrants, in his or her country of origin. Other possible fields include health studies [CF07] and economics [Sor03]: until now research in these fields has focussed mainly on the structure of the social networks without considering geographic aspects.

3.2.1 Spatial information dissemination measures

A social network can be represented as a graph $\mathcal{G} = (V, E)$ with N nodes and K links, where nodes are users and links are the social connections between them[1]. I define a *spatial social network* as a social network where each user i is assigned a set of n_i geographical points $\mathcal{P}_i = \{p_0^{(i)}, p_1^{(i)}, \dots p_{n_i}^{(i)}\}$ including locations that are important to him/her (e.g.,

[1]This representation can be considered as a snapshot of the graph at a given time t. A treatment considering the time-varying nature of the social graphs is outside the scope of this work.

hometown, workplace, favourite restaurant, etc.)[1]. I will firstly introduce a set of accessory definitions that will be used in the remainder of this section.

- As far as the social graph is concerned, I define the neighbours (or connections) of node i as the set of nodes j that are reachable from i through the out-link e_{ij} (content can flow from i to j). The *social neighbourhood* $\mathcal{N}_i$ of a node i is the set of all the k_i neighbouring nodes of i (e.g., all the followers of user i in Twitter); k_i is often referred to as the degree of node i. This set is defined only by social ties and does not take into consideration any geographic information.

- As far as the spatial dimension is concerned, I use the notation $d_G(p_1, p_2)$ to indicate the geodesic distance between two points on Earth p_1 and p_2. I then define the *spatial neighbourhood* $\mathcal{S}$ as an arbitrarily shaped part of the geographic surface; this is a continuous set of geographic points. For simplicity, in this work I will often consider circular regions specified by their centre and radius but the definitions presented here can be applied to regions of any shape.

- Given a node j and a geographic region $\mathcal{S}$, the intersection $\mathcal{P}_j \cap \mathcal{S}$ contains all the significant points of j falling inside the region. I define the *socio-spatial neighbourhood* $\mathcal{N}_{i,\mathcal{S}}$ of the node i with respect to $\mathcal{S}$ as the set of neighbours j who have at least one significant point inside $\mathcal{S}$:

$$\mathcal{N}_{i,\mathcal{S}} = \{j \in \mathcal{N}_i : \mathcal{P}_j \cap \mathcal{S} \neq \emptyset\}. \tag{3.10}$$

  With $k_{i,\mathcal{S}}$ I denote the number of users in this set. An example is presented in Fig. 3.11.

[1]In the simpler case each user can be assigned a single significant location. In the evaluation section I will present two examples covering both cases.

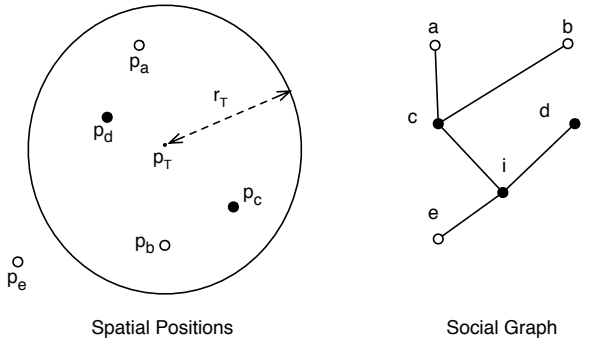


Figure 3.11: Example of social graph (on the right) and spatial dimension (on the left). In this example the social neighbourhood $\mathcal{N}_i$ of i is composed by the nodes c, d, and e; the points of interest of a, b, c and d are inside the spatial neighbourhood $\mathcal{S}$, which is the circle of centre p_T and radius r_T. As a consequence, the socio-spatial neighbourhood $\mathcal{N}_{i,\mathcal{S}}$, indicated with full black dots, does not include node e because it falls outside the spatial neighbourhood and also excludes nodes a and b because their positions are located outside the social neighbourhood.

Spatial degree centrality

In general, in a social graph degree centrality is used to rank users according to the number of ties they have within the network [New10]; its value is a simple indicator of *influence* and prestige [WF94]. Methods based on degree centrality are generally used to select the best nodes for spreading information [KKT03]. I extend the concept of degree centrality to spatial social networks with respect to a given spatial neighbourhood $\mathcal{S}$ by introducing the concept of *spatial degree centrality*:

$$C_{i,\mathcal{S}} = \sum_{j \in \mathcal{N}_i} |\mathcal{P}_j \cap \mathcal{S}| \,. \tag{3.11}$$

This value indicates how many significant points the social neighbourhood of user i has got inside the considered spatial neighbourhood $\mathcal{S}$. If every user is associated with only one significant point, this value indicates the size of the audience of user i in the region. In the general case of many significant points for each user, this also takes into account the strength of the potential audience in the region (i.e., social connections with many significant places inside the region give a larger contribution than those with fewer).

The size of the considered region $\mathcal{S}$ affects the calculation of the values of the measures. For this reason, the size should be set according to the characteristics of the dataset (measurement granularity and precision) and the goal of the analysis itself (for example, researchers might be interested in an analysis at city level). Since the degree of each node also affects this value, a normalisation of this measure might also be necessary. The normalisation is particularly convenient when comparing users who have a number of followers that differs by orders of magnitude. This might be the case that happens when comparing accounts of news agencies and celebrities, often followed by hundreds of thousands of users, with users who have dozens or hundreds of followers. I call the normalised version *spatial degree ratio*, formally defined as:

$$\rho_{i,\mathcal{S}} = \frac{1}{\sum_{j \in \mathcal{N}_i} n_j} C_{i,\mathcal{S}} \tag{3.12}$$

where n_i is the number of significant places of the user i; this is equivalent, for the one-place case, to:

$$\rho_{i,\mathcal{S}} = \frac{1}{k_i} C_{i,\mathcal{S}}. \tag{3.13}$$

This measure has values in the range $[0, 1]$. It represents the ratio of connections of i that are inside the area $\mathcal{S}$, therefore it allows for the comparison of nodes that have different degrees in the graph.

These centralities might be considered as simple measures of spatial influence, which can be used in the selection of a user for spreading information to a certain geographic region. However, as they are based on the concepts of geographic membership and social membership, they might not be entirely sufficient to describe the geographic distribution of the neighbours of users. For this reason, in the next subsection I will introduce measures that also take into account geographic distances.

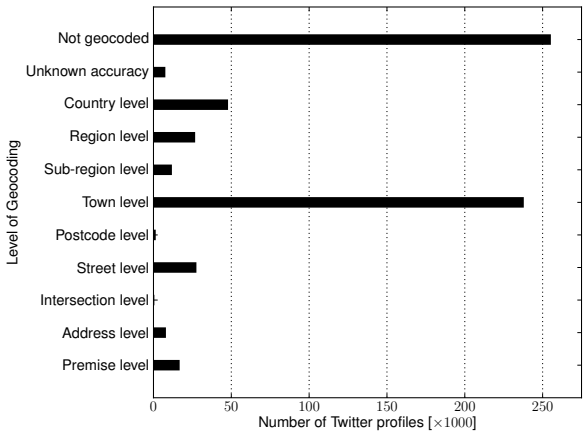


Figure 3.12: Precision of geocoding for the Twitter dataset.

Spatial closeness centrality

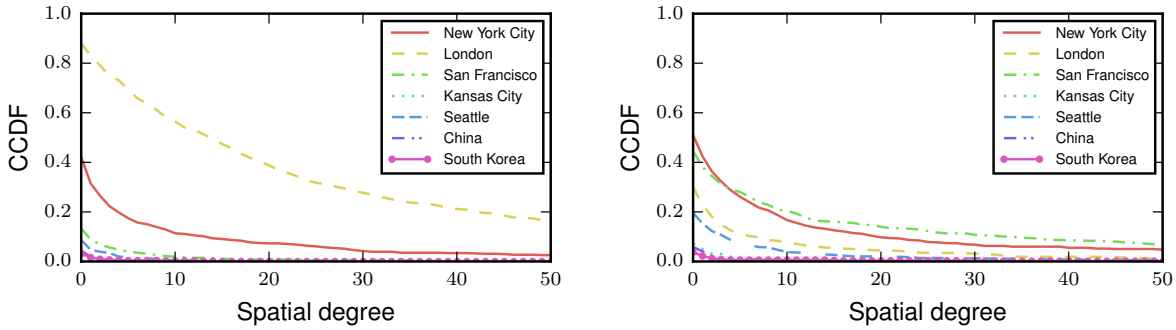
I have defined spatial degree centrality relating to a region. Now I will define a measure of centrality concerning a *punctual* location. Given a target point $p^\star$ on Earth, I define the *spatial closeness centrality* for a user i towards this point as its average geographic distance from all the significant places of his or her connections, formally:

$$C^C_{i,p^\star} = \frac{1}{\sum_{j\in\mathcal{N}_i} n_i} \sum_{j\in\mathcal{N}_i} d_G(p_j, p^\star). \tag{3.14}$$

This definition is an indicator of how the influenced audience of a user is geographically close to the target point. It can be considered as the spatial counterpart of closeness centrality, which for complex networks is defined as the average distance of the shortest path from the node to all the other nodes [New10]. It is used as a heuristic when selecting nodes in information diffusion processes [KKT03]. However, this measure might have some drawbacks in specific scenarios given the fact it is calculated as an average of all the distances. This measure can be generalised to the case of multiple locations.

Spatial efficiency

In order to deal with the problem of very large distances which might skew the value of spatial closeness centrality, I define *spatial efficiency* of user i with respect to a point $p^\star$



(a) Spatial degree centrality from London. (b) Spatial degree centrality from San Francisco.

Figure 3.13: Spatial degree centrality of Twitter users in London and San Francisco.

as follows:

$$C^{E}_{i,p^\star} = \frac{1}{k_i} \sum_{j \in \mathcal{N}_i} \frac{1}{d_G(p_j, p^\star)}. \tag{3.15}$$

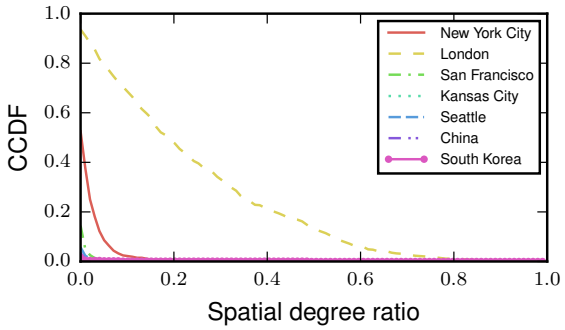
This measure can be thought of as a spatial version of efficiency of traditional graphs [LM01]. However, this definition has also a potential drawback: if the neighbour location p_j coincides with $p^\star$ this formula is not defined. For this reason, I modify the above formula by introducing a smoothing decay term as follows:

$$C^{E}_{i}(p) = \frac{1}{k_i} \sum_{j \in \mathcal{N}_i} e^{-d_G(p_j, p^\star)/\gamma} \tag{3.16}$$

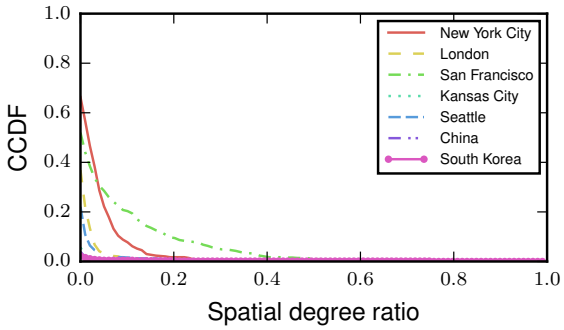
where γ is a scaling factor that can be used to give different weights to the distance $d_G(p_j, p^\star)$. In this formula, the contribution for every neighbour j is at most 1. It is equal to 1 if the neighbour location p_j coincides with $p^\star$, whereas it is negligible if the point is very distant (asymptotically zero if the distance is infinite). This definition can be generalised to multiple locations in a similar way to the formulae presented above.

3.2.2 Datasets

In order to evaluate the measures presented above, I analyse two popular real-world OSNs, Twitter and FourSquare. In general, datasets were acquired using 2-hop snowball sampling, seeded with random users chosen in some urban geographic areas. Due to

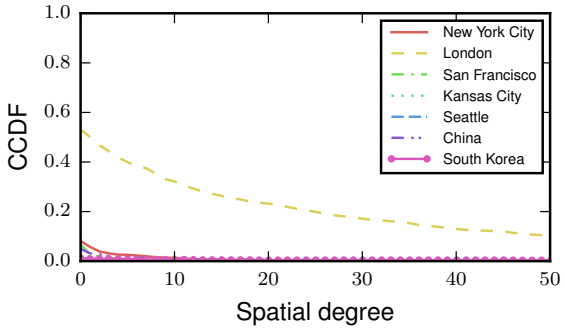


(a) Spatial degree ratio from London.

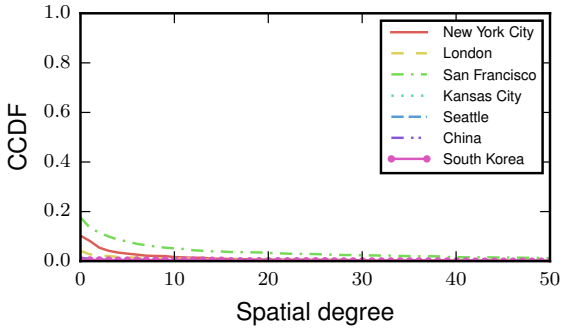


(b) Spatial degree ratio from San Francisco.

Figure 3.14: Spatial degree ratio of Twitter users in London and San Francisco.



(a) Spatial degree centrality from London.



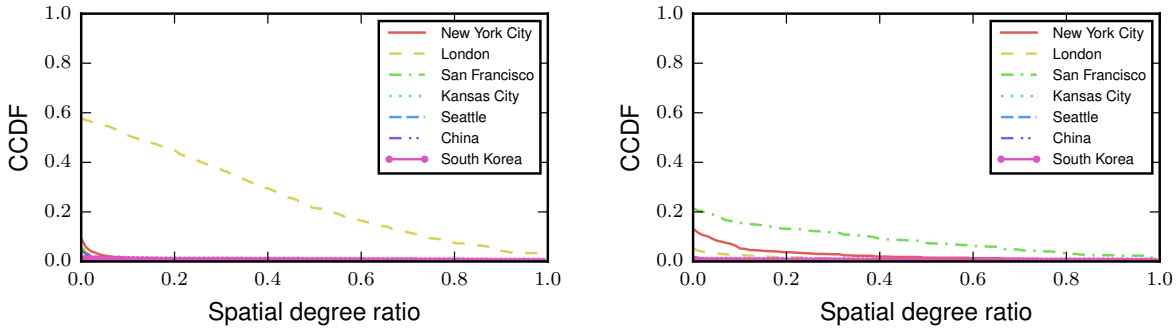
(b) Spatial degree centrality from San Francisco.

Figure 3.15: Spatial degree centrality of FourSquare users in London and San Francisco.

different properties of the two social networking services taken into consideration, the two datasets were obtained following different methodologies, as explained below.

With respect to Twitter, I crawled a dataset containing information about 657,777 users, starting from two evenly distributed sets of 1375 seed users. These were chosen randomly among users that were tweeting from two urban areas, London, UK and San Francisco, California[1]. This location bias was necessary given the nature of the investigation, which requires to have a statistically significant sample of users in the area. It is also worth noting that this can be considered as a practical way of retrieving these users for a potential deployment of the algorithms in a real-world system. I assigned a single significant place to each user, by fetching the information in the "location" field of their

[1]The locations for the "seeds" were retrieved from geotags, i.e., spatial tags which are associated to tweets either by automatic geographic sensors as GPS or manually by the user.



(a) Spatial degree ratio from London.

(b) Spatial degree ratio from San Francisco.

Figure 3.16: Spatial degree ratio of FourSquare users in London and San Francisco.

personal profile, and converting it to geographic coordinates through the Google Geocoding API. The geocoder was able to identify locations for 378,829 users, with different levels of precision; the majority of the identified locations were at town level, according to the distribution shown in Fig. 3.12, similar to that shown in [HHS+11]. I am indeed aware of the fact that locations are not precise and the data are noisy.

In FourSquare, a location-based online social network, users "check-in" at venues to let their friends know about their whereabouts, to keep track of their habits and to explore places related to the interests they have in common with other people. The user with the highest number of check-ins over the last 60 days is called the "mayor" of the venue in FourSquare jargon. For this reason, mayorship provides information of potentially strong spatial significance of a certain place for that user. This is also fine-grained information, as venues are commonly specified at premises level. For this reason, I used the collection of mayorships' locations to build the set of significant places. I crawled a dataset of 177,809 users. Since the number of connections in FourSquare is typically smaller than the number of followers in Twitter and the former tend to link with spatially close people [SMM+10], the sampling strategy used here followed a different approach in order to avoid geographically sparse data. I selected a group of interesting urban areas and I crawled venues in the area using the FourSquare API. It is worth noting that these considerations are of great importance for a practical implementation of systems for

calculating these measures in (quasi) real-time, also considering the crawling limitations of the APIs[1].

Finally, I make the simplifying assumption that the rate of change of the network topology is negligible with respect to the information dissemination process taking place over it. This assumption seems reasonable in networks such as Twitter or FourSquare, where the rate of change of links is usually very low at the scale of 1 day, for example. In fact, the number of new added and removed followers and friends is quite low for a given user after an initial period where a large number of users is added.

3.2.3 Evaluation

In this section I will present a selection of measurements for each measure. More specifically, I choose to compare areas that are heterogeneous from a cultural point of view and different in size.

I also consider two practical case studies. The first is related to the London riots that took place in August 2011: I measure the centrality of Londoners on Croydon, which was one of the theatres of the most violent acts in the British capital. This scenario is an example of usage of this technique in the case of emergency. In other words, this methodology can be used to answer the following question: *what is the best set of people to target in order to have localised influence through social media in the case of natural and man-made emergency and disasters?*

The second consists of quantifying the centrality of San Franciscans with respect to people living in Chinatown, and of people living in Chinatown with respect to people living in China. It can be seen as an application to the area of geo-demographics [AL05], aimed at quantifying the potential cultural influence of the inhabitants of certain areas of the city over other areas.

[1]The FourSquare API returns at most 50 venues per call and does not allow to paginate over all venues in a given large area. Therefore, I queried for venues in categories in small-radius areas (i.e., with a 50 m radius) randomly selected inside the larger areas considered.

Spatial degree centrality and ratio. Fig. 3.13 reports the complementary cumulative distribution Function (CCDF) of spatial degree centrality of the users located in London and in San Francisco towards four cities (New York, Kansas City, London and Seattle) and two countries (China and South Korea) using the Twitter dataset. I selected the countries by considering the presence of a non-negligible percentage of their population belonging to these ethnic groups. It is possible to observe that both cities have a high degree centrality with respect to themselves, as expected. It is surprising, though, that the degree centrality of Londoners on themselves is very high; in comparison, San Franciscans are not significantly central with respect to their fellow-citizens, and have a self-centrality similar to the centrality shown towards New Yorkers. In my opinion, a possible cause might be that many people who spend most of the day in San Francisco (for example, because their workplace is based there), actually live in the neighbouring areas and commute everyday. While 9 users out of 10 in London have at least 1 follower from their own city, only 1 user out of 2 in San Francisco has at least a fellow-citizen reading his content. San Franciscans have some limited potential influence on Londoners, though not as much as on New Yorkers. Users from London and Seattle are also potentially influenced in a substantial way, though not as much as New Yorkers. China and South Korea score very low centrality measures in both cases, and their curves overlap with those related to Kansas City, the city on which both Londoners and San Franciscans influence the least.

It is worth noting that these results could be influenced by a culture-related tendency to include location information: users from some locations might be keener to include the real personal location, compared to users from other places, due to a different sensibility about privacy issues [ABL15]. Unfortunately, I do not have hard evidence about this fact.

Similar observations can be made for the CCDF for Spatial Degree Ratio in Fig. 3.14. The high degree centrality of London with respect to itself is actually connected to a low spatial heterogeneity of followers: nearly one Londoner out of two has *at least* 20 followers living in the same city, while in San Francisco only one out of ten satisfies this property. This peculiar characteristic might be explained both with the tendency of Londoners

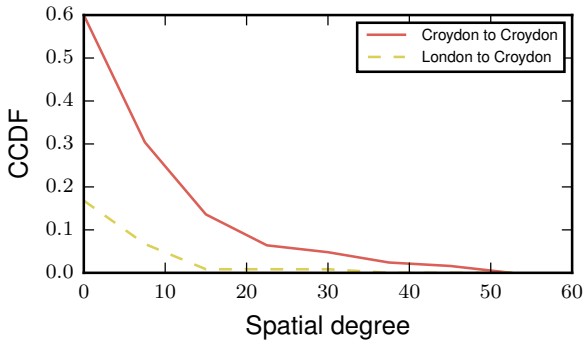
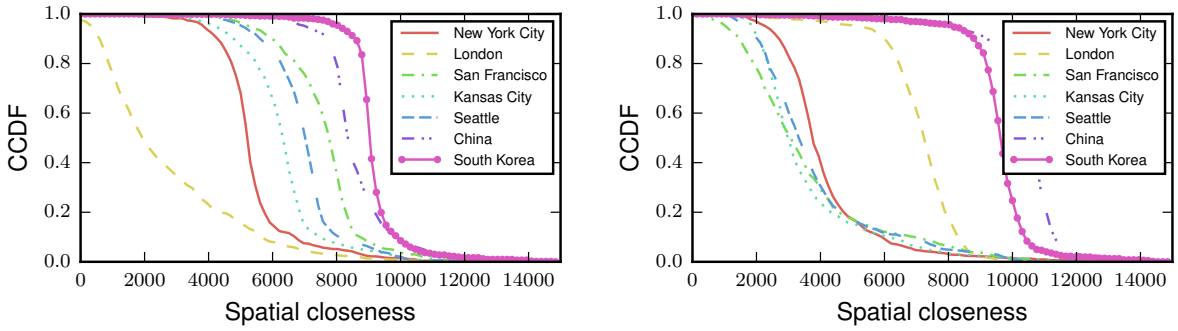


Figure 3.17: Spatial degree centrality of FourSquare users in Croydon and London towards Croydon.



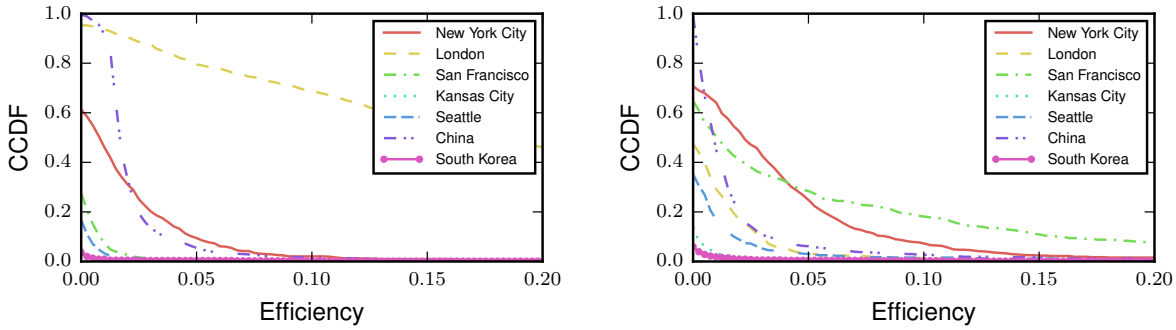
(a) Spatial closeness centrality from London. (b) Spatial closeness centrality from San Francisco.

Figure 3.18: Spatial closeness centrality of Twitter users in London and San Francisco.

to follow people from London and with a low interest shown by non-Londoners for the content shared by Londoners. The two highest curves of ratio show a more linear progress, compared to their spatial degree centrality counterparts.

I also perform a similar analysis using the FourSquare dataset. The lower penetration of FourSquare leads to a lower average degree (i.e., on average users in FourSquare have a smaller number of connections than in Twitter) and consequently to smaller centrality values, which include many zeros. However, results shown in Fig. 3.15 and Fig. 3.16 are still in accordance with those observed for Twitter. Considering the characteristics of the users in the city, London is again a place of high centrality with respect to itself.

While this city-level analysis can be carried out on the Twitter dataset, I cannot use it for a meaningful analysis at a finer scale, given the nature and quality of the data.



(a) Spatial efficiency centrality from London. (b) Spatial efficiency centrality from San Francisco.

Figure 3.19: Spatial efficiency centrality of Twitter users in London and San Francisco.

Therefore, I use instead the FourSquare dataset, in particular to study the potential influence of Chinatown towards San Francisco and China. The measure is able to identify 8% of users that have a non-null centrality on China and to *rank* them according to their centrality (which quantifies their potential influence over China). When analysing the average values of the measure, it is interesting to note that the centrality of San Francisco towards Chinatown and the centrality of Chinatown towards itself are comparable (3.2 vs 3.06). This might support the hypothesis that the district is considerably influenced by people living in other parts of the city and that choosing to deliver information to people in Chinatown instead of San Francisco might not have a significant impact on how the information is spread in Chinatown itself. Moreover, given its history and ethnic composition it is not surprising to discover that the average centrality of Chinatown towards China is almost 3 times the average centrality of the city of San Francisco on China (32.24 vs 11.87).

For the Croydon scenario, Fig. 3.17 reports the centrality of Croydon and London on the Croydon area itself. Users in Croydon appear to have substantially higher values of degree centrality compared to all the users in London. This suggests that when disseminating information, targeting people in the area of Croydon, instead of the whole of London, might give an advantage in reaching the area of Croydon itself.

Spatial closeness centrality. Fig. 3.18 shows the probability distribution function for the seven distributions of spatial closeness centrality. For each curve, a dashed vertical line represents the median. We can firstly notice that for both cities taken into consideration, London and San Francisco, the closeness centrality curves are more spread out compared to the other curves, which are generally narrower and characterised by a series of peaks. London shows this behaviour with stronger emphasis; this can be another evidence of the high locality of London followers. By definition, geographic constraints have a strong impact on this measure; therefore, we would expect that the peak and the median are very close to the physical distance between the considered points. Indeed, this is the case for all the pairs reported in the figure.

Spatial efficiency. In order to characterise spatial efficiency, I set the value of γ equal to the maximum radius of the geographic area taken into consideration. Fig. 3.19 shows the CCDF for the Twitter users in London and San Francisco with respect to the areas considered for the other measures. As this measure emphasises the role of neighbours which are close to the location taken into consideration, we can see how the efficiency of London with respect to itself stands out from all the other curves.

3.3 Patterns of online collaboration

So far I have analysed interactions on online networks that are mainly driven by social activities, such as visiting places, meeting friends (FourSquare) and discussing various topics (Twitter). An interesting question that might arise is whether the nature of the social networks has an effect on its properties and to what extent. To explore this aspect, I turn the focus of the investigation to a quite different online social network, namely GitHub, a collaborative-platform for open source projects. While GitHub has several social features that make it possible for people to follow each other, it is still a work-related network, where interactions are mainly motivated by the goal of advancing open-source projects.

GitHub is indeed the most popular repository for open source code [Fin11]. It has more than 3.5 million users, as the company declared in April 2013, and more than 10 million repositories, as of December 2013. It has a publicly accessible API and, since March 2012, it also publishes a stream of all the events occurring on public projects. Interactions among GitHub users are of a complex nature and take place in different forms. Developers create and fork repositories, push code, approve code pushed by others, bookmark their favourite projects and follow other developers to keep track of their activities.

3.3.1 Dataset

The full list of public events that have happened on GitHub is available on the GitHub Archive website[1]. Here I analyse events that happened on GitHub over a period of 18 months, between the 11$^{\text{th}}$ March 2012 and the 11$^{\text{th}}$ September 2013, retrieved from that archive. This dataset includes various types of events performed by users on public repositories or following events between users (i.e., when a user starts following another user). The total number of retrieved events is 183,540,210 and events fall into 18 categories[2]. Each event, regardless of its kind, usually includes some metadata about the entities involved (e.g., the profile information of a user, his or her number of followers, the language of a repository, etc.). Fig. 3.20 shows how events are distributed among the various categories. The user `Try-Git`, which shows an uncommonly high number of collaborations, is a learning tool that pushes code automatically to other users' repositories. I discarded this outlier from the dataset.

In order to explore the geographic features of users, I investigate the location information that can be found in the user profiles. In this dataset, 345,625 users have a non-empty location field. As the field is optional, there is little incentive to fill it with fake information. Therefore, it is reasonable to assume that most of the non-empty entries are correct. In order to convert the text field to an unambiguous location, I use the MapQuest Open

[1] `http://www.githubarchive.org`
[2] `http://developer.github.com/v3/activity/events/types/`

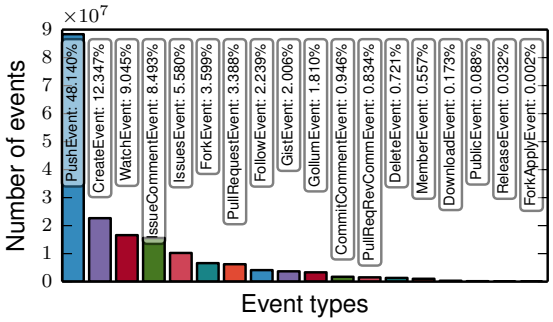


Figure 3.20: Number of events detected in the GitHub stream.

Geocoding API[1]. I evaluate the validity of the geocoder by considering a sample of 1,000 users in the population of users with non-empty location field and assessing the fraction of correctly geocoded elements by manually labelling them. I find that 106 elements are incorrectly geocoded. From the analysis of this sample, therefore, I can say that the geocoder fails to correctly convert to coordinates in $10.6 \pm 1.91\%$ cases of the original population, with 95% confidence level. Incorrectly geocoded entries in the sample fail mostly for the following reasons: because they describe multiple locations (for example `"London and Nottingham"`), because they have no geographic meaning (e.g., `"localhost", "emacs"`) because they are ambiguous (e.g., `"San Jose"`, rather than `"San Jose, CA"`).

Here I present a characterisation of GitHub, as both a social network and a collaborative platform. To the best of my knowledge, this is the first quantitative study about the interactions happening on GitHub. I analyse the logs from the service over 18 months (between March 11, 2012 and September 11, 2013), describing 183.54 million events and I obtain information about 2.19 million users and 5.68 million repositories, both growing linearly in time. I show that the distributions of the number of contributors per project, watchers per project and followers per user show a power-law-like shape. I analyse social ties and repository-mediated collaboration patterns, and I observe a remarkably low level of reciprocity of the social connections. I also measure the activity of each user in terms of authored events and I observe that very active users do not necessarily have a large

[1]Data: © OpenStreetMap contributors, available under the Open Database License. Geocoding: courtesy of MapQuest (`http://www.mapquest.com`).

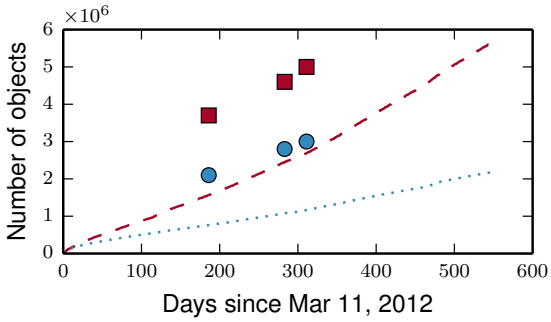


Figure 3.21: Number of unique repositories and unique users detected from the stream since March 11, 2012. The dashed and dotted blue lines show the number of repositories and the number of users detected from the event stream, respectively. The three squares and circles indicate the number of repositories and users in three specific dates as advertised by GitHub itself on its website.

number of followers. Finally, I provide a geographic characterisation of the centres of activity and I investigate how distance influences collaboration.

It is important to be aware that this data source suffers a time bias, since the archive does not include events that happened before March 2011. Fig. 3.21 shows the number of unique users and public repositories seen in the event stream since the 11$^{\text{th}}$ March 2012. As previously discussed, I am able to retrieve metadata when entities are involved in an *event*. In other words, I do not have information about dormant entities that were created before the 11$^{\text{th}}$ March 2012 and no longer generate *any* event during the subsequent 18 months (e.g., an inactive user, an abandoned repository). I am also not able to extract pre-existing following relations from the stream. After a short transitory period, which is present because of the temporal bias of the data collection process based on events, both curves show linear growth with different coefficients, with the ratio describing the number of repositories over the number of users reaching a steady value of approximately 2.59. The figure also reports the number of repositories and users (indicated using squares and circles, respectively) publicly declared by GitHub. This dataset contains a lower number of users and repositories for two reasons. Firstly, the official numbers include *all* the users and repositories created since the launch of the website in 2008, whereas this dataset contains only the *active* users and repositories in the period taken into consideration.

Secondly, the official statistics probably include private repositories, which do not appear on the public events timeline. For these reasons, we can conclude that a large number of users do not actively use the website (i.e., do not generate events) or they act exclusively on private repositories. These figures also suggest that a large number of repositories are either abandoned or private.

3.3.2 Structural analysis

In this section I define, extract and analyse several networks generated from the event stream, which describe interactions between users and repositories.

- I represent users' following relations by means of a directed graph G_F, which I call *followers graph*. I am able to reconstruct this network by looking at *follow events* in the stream.

- I represent the collaborations of users on repositories as a bipartite graph G_C, the *collaborators graph*, where repository nodes are connected to their collaborators nodes. I am able to infer this network by extracting from *push events* information about who uses write permission and on which repositories. I refer to $G_C^{\perp}$, the *projected collaborators graph*, as the graph obtained by projecting the collaborators graph onto the set of users. In this projected graph users who collaborate in at least one repository are connected to each other.

- I represent users assigning a star to a repository as a bipartite graph G_S, the *stargazers graph*. This network can be generated using the information found in *watch events*.

- Finally, I build the *contributors graph* G_N by analysing the content of every *push event*, which includes authorship information of the pushed commits.

For the static analyses I consider these networks as they appear on the final day of the time window taken into consideration.

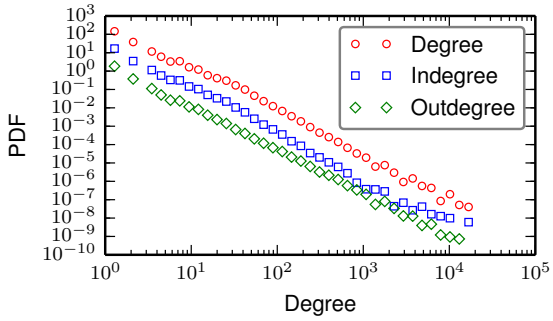


Figure 3.22: Distribution of degree, in-degree and out-degree of the social graph. The distributions were shifted along the y-axis to put in evidence their structure. The three distributions exhibit a power-law scaling behaviour, with different exponents, for values in the range from 20 to 1000.

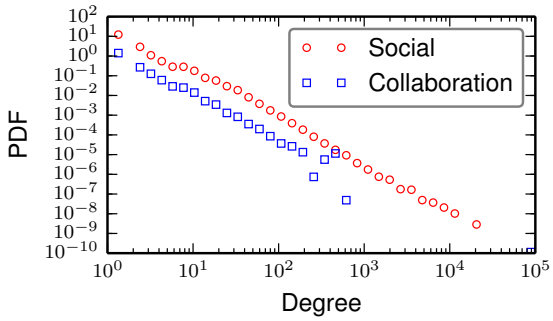
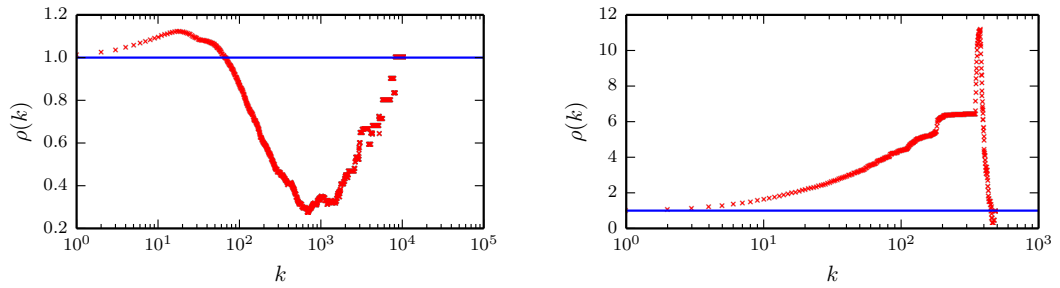


Figure 3.23: Distribution of the number of followers per user (circles) and the number of total collaborators per user (squares), which corresponds to the degree distribution of the users projection of the collaborators bipartite graph.

Followers and collaborators networks

As previously explained, a user follows other users in order to be regularly updated about events regarding them (e.g., forks, created repositories, starred repositories, and so on). The followers graph G_F has a total of 671,751 nodes and 2,027,564 edges, with a resulting graph density of 4.4932e-06 and an average degree of 3.019. The low graph density and average degree indicate that on GitHub the follow action is associated with a high cost, as following many developers results in receiving many notifications from them. This result also reflects the fact that following links in GitHub do not play the same important role they have in other social networks, such as Facebook or Twitter.



(a) Followers

(b) Collaborators

Figure 3.24: The normalised rich-club coefficient $\rho(k)$ as a function of the node degree. The blue horizontal line represents the value of $\rho(k)$ on a maximally random graph. Values of $\rho(k)$ above (below) 1 correspond to the presence (absence) of the rich-club phenomenon with respect to the random case. The two networks show remarkably different rich-club behaviours, due to their distinct nature.

Fig. 3.22 shows the distributions of the in-degree, out-degree and total degree of the users in G_F. All the three distributions show a power-law scaling behaviour, characterised by different regimes. For degrees smaller than $k \approx 20$, in all the three cases the scaling relation is not satisfied. Interestingly, the degree distributions of G_F and of $G_C^{\perp}$ follow the same power-law regime, as shown in Fig. 3.23. However, the node degree in the followers graph grows considerably larger than in $G_C^{\perp}$.

The followers network is also characterised by low reciprocity: only 9.6% of the pairs of users have a reciprocal relation between them, while the remaining 90.4% are one-way. Other studies on social networks have reported considerably higher levels of reciprocity, such as 22.1% for Twitter [KLP+10], 68% for Flickr [CMG09] and 84% for Yahoo! 360 [KNT10]. The consistently lower reciprocity in GitHub is partially motivated by the presence of a few popular programmers, the so-called "rockstar programmers", who exhibit high in-degrees and low out-degrees. However, the profoundly different nature of GitHub, compared to other social networks, might also play a role in this. In fact, social networks are mostly used for leisure and they thrive on distractions coming from noisy timelines; on the contrary, the productivity of GitHub developers might be critically disrupted by non-relevant notifications, which are hence kept to a minimum. In other words,

establishing links has a high cost in GitHub, as people do not "follow-back" unless they are professionally interested in the activity of their followers.

In order to uncover the presence of node degree correlations, I first measure the degree assortativity. A network shows an assortative mixing [New02a] if nodes with a large number of links tend to share edges with high degree nodes. Conversely, if nodes with a small number of links tend to share edges with low degree nodes the network shows a disassortative mixing. This network has a degree assortativity of -0.0386, which suggests a tendency to a disassortative mixing of users. The rich-club coefficient ϕ [ZM04] measures the tendency of high degree nodes to form tightly interconnected communities. Although apparently similar to the concept of assortative mixing, the rich-club phenomenon is not necessarily associated with the latter, as one can define a disassortative network that still shows evidence of a rich-club phenomenon. Let E_k denote the number of edges among the N_k nodes having a degree higher than k. The rich-club coefficient $\phi(k)$ is defined as follows:

$$\phi(k) = \frac{2E_k}{N_k(N_k - 1)} \tag{3.17}$$

It represents the fraction of edges connecting nodes in N_k out of the maximum possible amount they can share, i.e., $\frac{N_k(N_k-1)}{2}$. More specifically, here I use the normalised rich-club coefficient proposed by [CFS+06], where the normalisation is introduced to account for the fact that high degree nodes have a higher probability of sharing edges than low degree ones. Fig. 3.24a shows the rich-club index of the followers graph, for increasing degree k. I use the definition of the rich-club index for G_F considering it as an undirected graph. Interestingly, low degree nodes show a less accentuated rich-club phenomenon, while high degree nodes do not. In other words, the plot indicates that hubs, i.e., popular developers, tend to share links with lower degree nodes rather than being tightly interconnected among them.

Compared to the followers graph, the collaborators graph $G_C^{\perp}$ also shows disassortative mixing of the nodes, with a value of -0.0518. However, the characteristics of the rich-club phenomenon are remarkably different. Fig. 3.24b shows the rich-club index of $G_C^{\perp}$,

for increasing values k of the degree. The plot shows that up to $k \approx 30$ the nodes show a strong rich-club phenomenon, with a pronounced increase followed by a sudden drop around $k \approx 40$. This effect is amplified by the projection operation itself, as each group of collaborators forms a clique in $G_C^{\perp}$.

I also measure the clustering coefficient [WS98] of $G_C^{\perp}$ and I compare it with that of the followers graph. Again, we expect the average clustering coefficient of the network to be high due to the way in which $G_C^{\perp}$ is constructed. Indeed, I find a value of 0.395 for $G_C^{\perp}$ and of 0.012 for G_F. Note, however, that this implies that users contributing to the same repositories do not necessarily follow each other, as in that case we would expect the average clustering coefficients of the two networks to be similar. Once again, this underlines the fact that the social interactions captured by the two structures are rather different.

I now investigate the relation between the number of followers of a user and his or her contributions to GitHub. We would expect popular users in terms of contributions to be followed by a higher number of people. In order to evaluate this, I measure the Spearman correlation coefficient [Leh06] between the number of followers and the number of contributions per user, and I find a value of 0.2568, with p-value < 0.01, indicating the lack of a clear correlation between the two dimensions. This result is unexpected, as it would seem reasonable to assume *active* users, i.e., users that contribute to a large number of repositories, to be more popular in terms of followers.

Interactions on repositories

Despite the large number of repositories hosted at GitHub, developers work only on a consistently small fraction of them. Only 62.90% of the total number of repositories available in the dataset experience at least one code commit during the 18 months taken into consideration. Only 74.22% of these repositories have at least two contributors, meaning that one active repository out of four is exclusively authored by a single individual. This might happen for a variety of reasons: the project might not look promising to other users or the

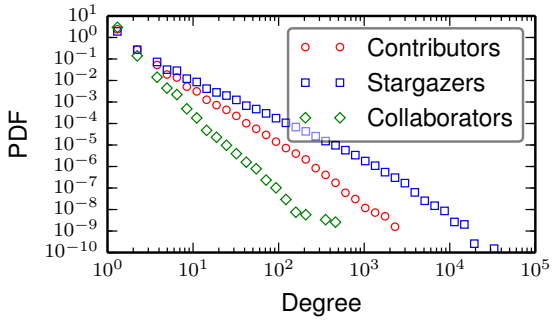


Figure 3.25: Distribution of the number of contributors, collaborators and stargazers per repository. The contributors and stargazers distributions are best fitted by a power law distribution with exponential cut-off and exponent α equal to 2.34 and 1.77, respectively. The collaborators distribution is consistent with a power law with $\alpha = 3.39$. The distributions parameters are estimated using the approach presented in [CSN09].

owners of the repository might reject contributions. This fraction includes activity both from one-time and habitual collaborators. Commonly, long-term contributors are turned into collaborators, so that they can help developing big projects. However, this kind of collaboration is quite rare, as only 9.61% of the repository has at least 2 of them. This is not surprising: collaborators need to be trusted individuals who have full understanding of the project goals and structure, as they have write access on the repository and they determine which contributions should be accepted. Fig. 3.25 reports the distribution of number of contributors, stargazers and collaborators per repository.

Forking and repository tree structure

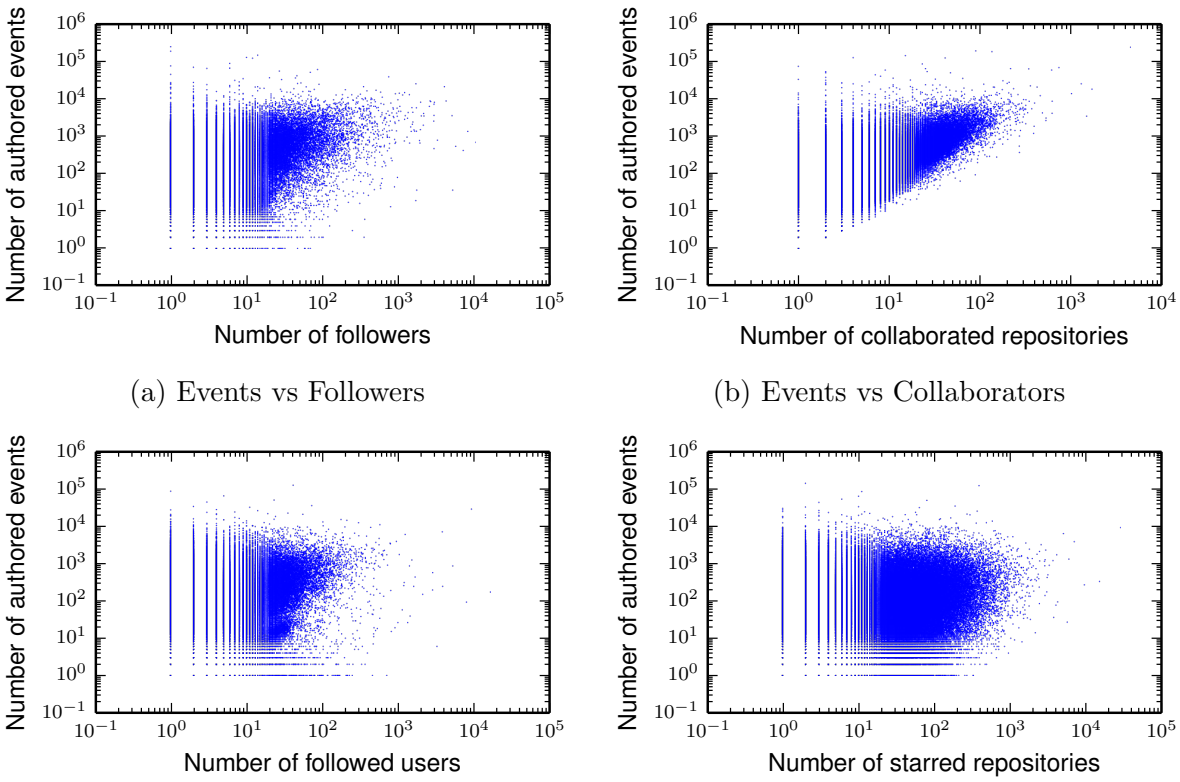
The fork operation is intended to let users actively contribute to a project. This action produces a copy of the parent repository and essentially generates a simple tree structure. Further forks on the leaves of the tree increase its depth, while forking an internal node results in an increased width of the set of its children. The repository tree can be thought as a directed acyclic graph, where the fork operation generates a directed edge from the parent repository to its child. In the following I refer to the *depth* of the tree as the longest path from the root to its leaves, and to its *width* as the maximum number of children over the internal nodes or 1 if the root has no children.

For a few repositories the maximum depth goes up to 12. However, these few structures are hardly the result of collaboration, in my opinion. In fact, user accounts involved in their creation do not exist any more. For this reason, I suppose these accounts have been removed due to abnormal or suspicious activity. I also find that the average depth is 3.0695, but the mode is 0, indicating that the majority of repositories has a low number of contributions. The width, on the other hand, goes up to 10,256, which is normal considering that many people fork to contribute to popular packages, such as `mxcl/homebrew`. Top repositories include `heroku/node-js-sample`, `YOU-LOST/THE-GAME` (apparently, a playful non-software repository) and `facebook-tornado`. The overall average width is very low (1.0653), showing that just a few popular repositories get forked, while the vast majority of them (93.91%) have a width of just 1. This, together with the observation that the majority of the repositories has depth equal to 0 and width equal to 1, seems to suggest that forks on GitHub happen on a limited number of key projects.

3.3.3 Activity, social presence and indirect rewards

Human activities are commonly driven by reward mechanisms of some kind: people work to earn money and achieve a social status, they play games because they have fun, they travel because they enjoy seeing new places. A recent study has found that areas of the brain connected to rewards are activated during the use of social network websites [MMH13]. One of the aspects that drives activity in GitHub, among others, is self-promotion [DST+12]. I hypothesise that for a hybrid service like GitHub, both a social network and a collaboration network, some kind of indirect reward mechanism might potentially underpin user activity. Even if it is not possible to provide definitive evidence about that, in the following I will show some interesting correlations between the activity of a user and some indirect rewards in terms of "social prestige" in GitHub.

In social networks, a common measure of user popularity and influence is given by the in-degree [WF94]. Therefore, it is reasonable to consider new connections as a reward for those users receiving them, as they increase their popularity. In order to investigate this



(a) Events vs Followers

(b) Events vs Collaborators

(c) Events vs Followed users

(d) Events vs Starred repositories

Figure 3.26: Number of actions executed by a user against (a) user followers, (b) number of repositories with write privileges, (c) followed users, (d) starred repositories.

aspect in GitHub I will search for correlation between user activity and degrees in the several graphs I have introduced. Fig. 3.26a shows the number of authored events (i.e., for which the user *actively* executes an action) for each user against the number of followers. We firstly note that people with a higher number of followers are commonly more active and people with lower levels of activity generally have fewer followers. However, we also observe that many users with a very high number of events have a very low number of followers: a higher level of activity does not directly translate into a larger number of followers. A similar phenomenon is also visible in Fig. 3.26b, where I plot the number of authored events against the number of repositories for which a user is a collaborator or the repository owner. Being the collaborator can also be seen as a kind of indirect reward, as it is more important and prestigious than being a contributor. Collaborators

Rank	Country	%	City	%
1	USA	30.14	San Francisco, US	3.84
2	UK	6.43	London, GB	3.33
3	Germany	5.28	New York City, US	2.93
4	China	5.11	Beijing, CN	1.98
5	India	4.05	Paris, FR	1.80
6	France	3.87	Tokyo, JP	1.69
7	Canada	3.69	Seattle, US	1.59
8	Brazil	3.60	Berlin, DE	1.49
9	Russia	3.14	Chicago, US	1.39
10	Japan	2.83	Shanghai, CN	1.34
11	Australia	2.00	Bangalore, IN	1.32
12	Spain	1.92	Toronto, CA	1.23
13	Netherlands	1.84	Moscow, RU	1.17
14	Sweden	1.51	Austin, US	1.12
15	Ukraine	1.37	Boston, US	1.07
16	Italy	1.32	Los Angeles, US	1.01
17	Poland	1.02	Sydney, AU	0.94
18	Switzerland	0.86	Portland, US	0.88
19	Belgium	0.75	Melbourne, AU	0.85
20	Mexico	0.74	Stockholm, SE	0.81

Table 3.1: Top 20 countries and cities, ranked by absolute number of users.

receive permissions to modify the repository, whereas contributors only contribute their code through pull requests.

I am also interested to see whether a higher out-degree on the social graph is an indicator of a higher activity. However, in Fig. 3.26c it is possible to note that a much weaker correlation between these two quantities is present. A similar behaviour can be observed in Fig. 3.26d, which shows activity versus the number of starred (i.e., bookmarked) repositories. In other words, users who follow many other users or bookmark many repositories are not much more active than those who do not.

3.3.4 The geography of collaboration

Fig. 3.27 shows the geographic distribution of GitHub users in the world. The majority of users are located in Europe and North America, while other geographic regions have a consistently smaller number of users. Tab. 3.1, listing the 20 most common countries and cities indicated in GitHub user profiles, confirms this observation. The popularity of GitHub among developers living in the USA is really prominent, as 3 users out of 10 are based there.

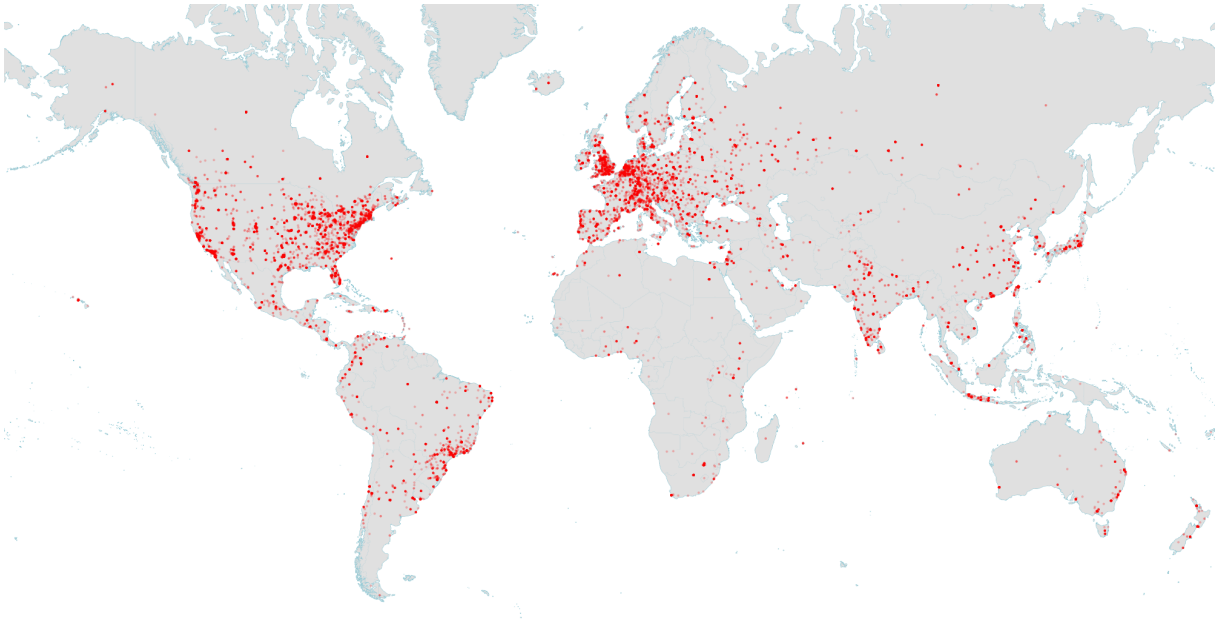
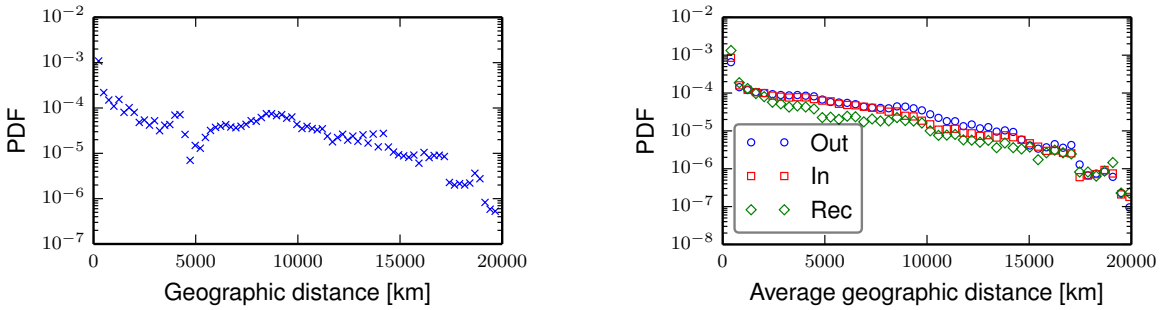


Figure 3.27: Distribution of GitHub users in the world. For each user, a partially transparent point is drawn on the map. The majority of users are located in North America and in Europe. The leading countries are the United States and the United Kingdom. A 15% random sample of the original distribution was used to make this figure.

Impact of geographic proximity

I also analyse the impact of physical proximity on the patterns of collaboration between different users. Are people more likely to follow people who are closer to them? Fig. 3.28a shows the distribution of the distance covered by each pair of users connected by a directed link in the social network. The first part of the distribution shows that links decrease with distance, until $x = 5000\ km$: these are intra-continental links. The sudden drop at $x = 5000\ km$ is due to the ocean separating North America and Europe, that are the two regions where GitHub is mostly popular. For larger distances, the distribution increases again, showing a big presence of intercontinental links. This analysis, however, considers all the links, without discriminating them on a per-user basis.

I now want to assess how *local* or *global* is the neighbourhood of a user, depending on how far his connections are located. In order to do that, I calculate for each user the average distance of their followers, their followed users and reciprocated links (i.e., users that are both followers and followed). Fig. 3.28b I show the probability density function



(a) (b)

Figure 3.28: (a) Distribution of the inter-user distance covered by each follow link. The distribution has a maximum at the lowest distance and gradually decreases for high distances. (b) Distribution of the average geographic distance of a user's outgoing, incoming and reciprocated links, respectively represented by blue circles, red squares and green diamonds.

of the values of this measure. As expected, the distribution of these values decreases as the distance increases, indicating that users tend to interact with people that are close. We also notice that in the majority of the cases the average distances of the reciprocated connections of a user, usually considered as evidence of friendship or mutual acquaintance or collaboration, tend to be smaller compared to the other two types of links.

Globality and distant collaboration

I now investigate if geographic proximity has an impact on the collaboration between users. In this case, it is not possible to compute the geographic distance between collaborators of a certain repository and the repository itself, as a repository does not have specific geographic coordinates. Project collaborators might be sparse around the globe or concentrated in a single city. In order to quantify how sparse they are, I define the *globality* of a set of users $\mathcal{S}$ as follows:

$$G = \frac{1}{N d_{max}} \sum_{i,j \in \mathcal{S}} d_{ij} \tag{3.18}$$

where d_{max} is the maximum distance between two points on Earth where two generic users are localised and N is the number of users taken into consideration. This measure

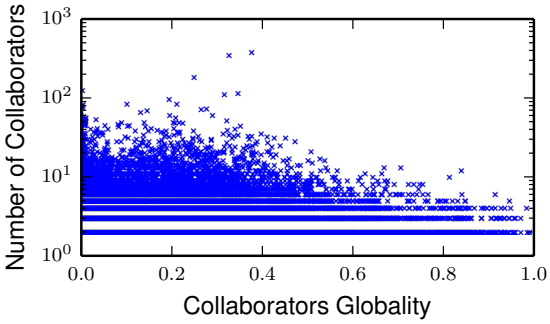
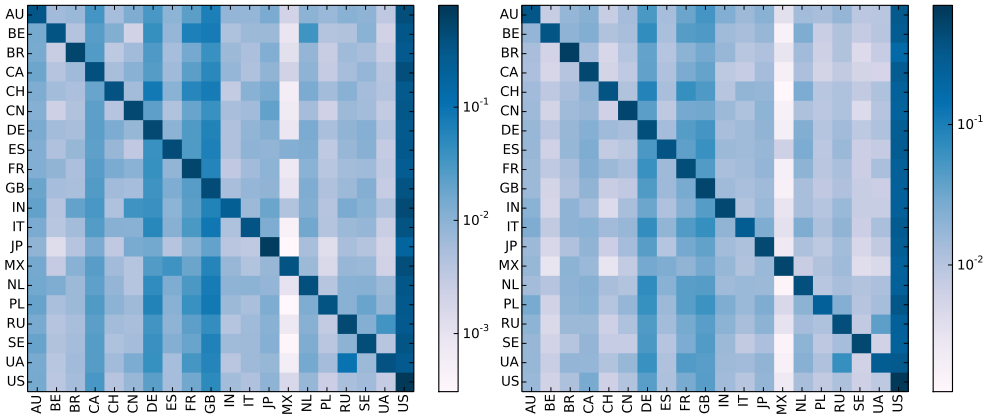


Figure 3.29: Scatter plot showing, for each repository, the number collaborators and the globality calculated over their geographic points. Intriguingly, repositories with a high number of collaborators exhibit smaller values of globality.

is the normalised average of distances between all the points in the set. When all points coincide the measure is 0, whereas when the points are evenly distributed at the antipodes the measure is 1.

Fig. 3.29 shows the value of globality against the number of collaborators for all the repositories that have at least two collaborators with location information in their profile. For a repository with a low number of collaborators, globality reaches values close to its maximum. As the number of collaborators goes up, the value of globality is found to be



(a) Followers (b) Contributors

Figure 3.30: The ties among users of the top 20 countries in terms of number of users. The rows of the matrix are normalised to sum to unity. Note that the most followed users are in the United States, while the least followed country is Mexico. As expected, the matrix also shows a strong level of interaction between users from the same country. The left matrix is built from the followers graph, while the right matrix is built from the contributors graph.

lower. This suggests that repositories with a low number of collaborators tend to have them concentrated around one or more key locations rather than scattered around the globe.

I now investigate how social and collaboration links are distributed among countries. In order to do that, I build two square matrices M_{G_F} and $M_{G_C^\perp}$ describing the number of links between countries. An element m_{ij} of the matrices indicates the number of links from people in country i to people in country j. The rows are normalised to sum to unity. This matrix has a precise meaning: each row shows how links coming from the people in a given country are distributed geographically. For clarity, in Fig. 3.30 I show the normalised matrices only for the top 20 countries, although the following measures are calculated on the full matrices. I first note that both matrices have a strong diagonal component (on average 0.245 and 0.346 for followers and collaborators, respectively), in accordance with the fact that links are more likely to be directed to the same country of origin. The two matrices are also significantly similar, as confirmed by the low average cosine distance, amounting to 0.277.

3.4 Summary

In this chapter I have explored patterns of human interaction happening on online social networks. First, I have analysed a large-scale process of information dissemination on Twitter. I have modelled its dynamics and analysed the spatio-temporal patterns of its activity. Then, I have presented a set of measures that extend classical network centrality measures to include the geographic domain and I have showed how they can be applied to real scenarios on data collected from Twitter and FourSquare. Finally, I discussed an analysis of patterns of collaborations on open-source projects hosted on GitHub. Although GitHub makes it possible to collaborate at long distances, I found that most social links and collaborations stay within national boundaries.

Although the OSNs I have analysed are built around different domains and although the connections and interactions that take place in them have different semantic meaning, in all of them the effects of geography have a prominent importance in the way users are connected to each other. In particular, in Twitter, where interaction (tweet, reply, retweet) has a relatively low cost, I found that the long-distance interaction can be quite common, for example during an event of global interest. In GitHub, instead, where interaction has a higher cost and involves a longer and more intense intellectual task (write code, commit, review, push), we see that interactions usually happen at small distances, while long-range interactions are rare. Future work might investigate whether the cost associated to a long-distance online interaction has indeed an effect on how likely it is to happen.

CHAPTER 4

DESIGNING EPIDEMIC CONTAINMENT STRATEGIES WITH MOBILE NETWORK DATA

Mobile web searches have recently surpassed searches performed from desktop computers, and for many people in the world the first computing device is a smartphone. These two facts illustrate well how access to services is increasingly shifting towards mobile devices.

While smartphones are prevalent in developed countries, they are still markedly less common in developing countries. These countries show a strong preference towards more affordable feature-phones [Pew15] (also referred to as dumb-phones), which, for example, are used as part of payment systems, letting people exchange money in form of airtime credit. Apart from this difference, for the vast majority of countries the penetration of mobile phones (regardless of it being a smartphone or not), is very high, above 70% [Pew14]. Because of this, mobile network data can describe the behaviour of countries' populations, both in terms of mobility and communication. The quality of the data is so high that methods based on mobile network data have been proposed as a replacement for surveys and for traditional mapping, because they are cheaper and quicker to complete [DLM+14].

In this chapter I will focus on the use of data collected from mobile networks to model and simulate country-wide epidemic spreading processes. In particular, I will describe the design of novel epidemic containment strategies, both at global and individual level, and I will evaluate their effectiveness through computer-based simulations. This chapter is structured as follows:

- Firstly, I describe the nature of the mobile network datasets that I will be using throughout the chapter.

- Then, I present an epidemic model that describes two concurrent processes: the spread of the disease in a country, fuelled by people moving between regions, and the spread of information to prevent the diffusion of the epidemic itself in the country (e.g., information about vaccination campaigns), travelling through the social ties of the individuals living there. I will then evaluate to what extent this second process can be effective in delaying the disease spreading, and how it compares to more drastic measures, such as the restriction of mobility between regions.

- Finally, I propose a methodology to quantify how each individual contributes to the disease spreading with their movement, as a risk measure. The risk measure can be used to prioritise treatment and design containment strategies of the disease. I show with computer-based simulations how targeting individuals according to this risk measure leads to a decrease the incidence of the disease.

4.1 Description of the datasets

A call detail record (CDR) contains information about each phone call, and includes, among other fields, calling and called party phone numbers, time and duration of the calls, and a coarse-grained location of the communicating parties. Throughout this chapter, I will use the datasets provided for the Orange Data for Development D4D challenges: a) CDRs collected from the Orange mobile network in Ivory Coast for a 5-month period (from 1st December, 2011 to 28th April, 2012) [BEC+12], and (b) CDRs from the Sonatel network in Senegal for 12 months (from 1st January, 2014 to 31st December, 2014) [dMST+14]. While the numbers of users considered and sampled in each dataset vary, both datasets are similar for the type of information they contain, which is:

- **Antenna-to-Antenna traffic.** The total number of calls and the total call duration, specified for each pair of antennas, aggregated by hour.

- **Fine-grained short-term mobility.** The trajectories at antenna level for a random sample of 50,000 (a) / 300,000 (b) users. The user sample is updated at random every two weeks for the whole window of observation.

- **Coarse-grained long-term mobility.** The trajectories at sub-administrative level (also called sous-préfecture or arondissement) for a sample of 50,000 (a) / 146,352 (b) randomly selected users, during the whole window of observation.

The mobility events in the dataset are specified with minute-precision. As explained in Subsection 2.1.2, mobile traces generate a location entry only in the event of a phone call, text message or data transmission. Because of this, movements happening between consecutive events of this type will not be not captured; for the same reason, people who generate few such events will exhibit lower mobility compared to those who originate many of them.

Another limitation to be aware of is related to the accuracy of the location reported by mobile network data. As I will later detail in Section 5.2, mobile towers are typically covering an area that can be a few kilometres away from their position. Moreover, contrary to what is commonly believed, often a mobile phone is not connected to the closest mobile antenna: several factors come instead into play when deciding which antenna a device is connected to. Since in this chapter I will be using the coarse-grained dataset, both these limitations are expected to have little impact on the validity of the analysis.

4.2 Extracting regional patterns of mobility and communication

I now describe how these datasets can be used to capture and represent realistic mobility and communication flows happening in the country under consideration. Such flows will

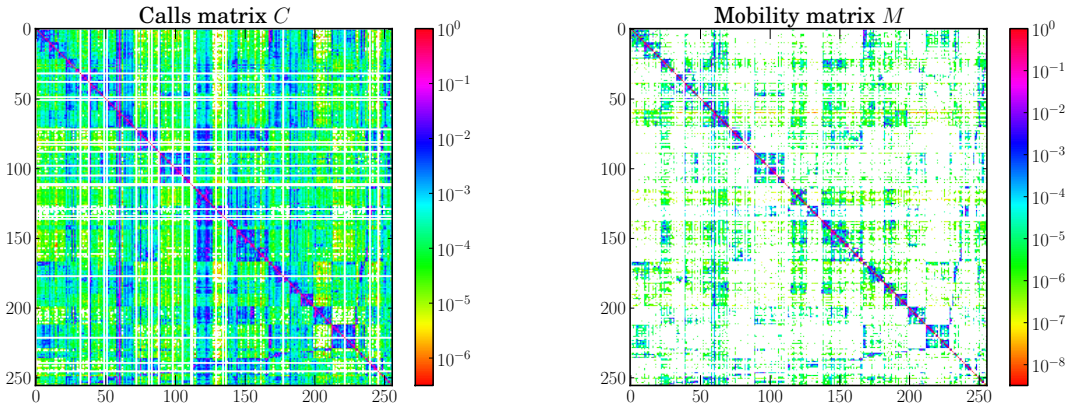


Figure 4.1: Logarithmic representation of the call matrix (a) and the mobility matrix (b). Null values are indicated using the colour white. For both matrices highest values are mostly concentrated along the diagonal, showing that communication and movement between sub-prefectures is highly uncommon. However, the calls matrix is visibly denser than the mobility matrix, confirming that phone contacts between different sub-prefectures are more usual than movement.

then be used in combination with the epidemic model that will be detailed in the next section to run computer-based simulations. The observations drawn from simulations make it possible to evaluate the effectiveness of epidemic-containment policies.

I extract patterns of individual mobility and communication from the available datasets, obtaining two matrices: the *mobility matrix* M and the *communication matrix* C, which are shown in Fig. 4.1. They represent the mobility flow and the communication flow between geographic areas. In Fig. 4.2 the geographic networks of calls (left panel) and mobility (right panel) are shown, respectively, where nodes are positioned using the geographic locations of the sub-prefecture they represent.

In particular, I use the coarse-grained long-term mobility dataset to estimate the probability that an individual moves from the sub-prefecture i to the sub-prefecture j:

$$m_{ij} = \frac{\sum_u \mathcal{M}_{ij}^u}{\sum_k \sum_u \mathcal{M}_{ik}^u}, \tag{4.1}$$

where $\mathcal{M}_{ij}^u$ is the number of times user u moves from the sub-prefecture i to j, during the entire period of observation. The numerator counts how many times users who are in i move to j; the denominator normalises this number by the total number of transitions

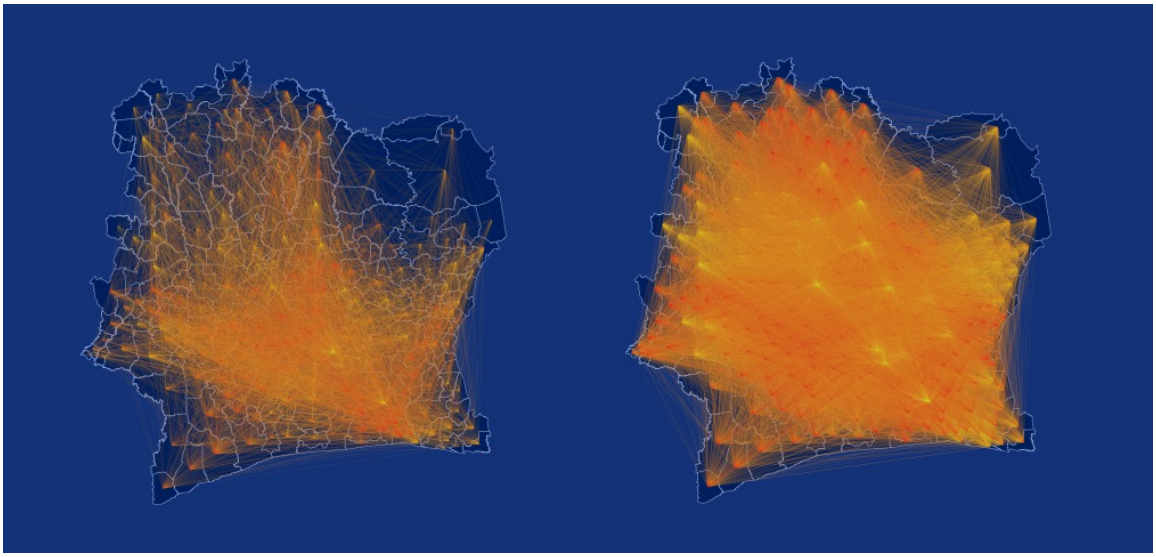


Figure 4.2: Geographic network obtained from mobility traces (a) and call logs (b), where nodes represent sub-prefectures. This map was generated by a custom d3 script. Map data: © OpenStreetMap contributors, available under the Open Database License.

from i to any sub-prefecture k. When $i = j$, m_{ij} is defined as the number of people who remain within their own metapopulation. A *mobility matrix* M is built using these values. It is worth remarking that this matrix is not built on a particular time window – it does not represent the number of calls per hour or per day. Instead, each row in the matrix shows in which proportion the calls originating from a sub-prefecture are directed to other sub-prefectures. By using this matrix, human mobility in the country can be modelled as a Markov process [Nor98]. The matrix is quite sparse and the highest values are concentrated along the diagonal. As the representation is in logarithmic scale, this demonstrates that the movement between sub-prefectures is present, but rather uncommon. The information found in this dataset was also used to allocate the population to different geographic areas, depending on the average number of users present.

Similarly, the antenna-to-antenna traffic dataset is used to observe macroscopic information about communication in the country. Each cell phone tower is associated with the sub-prefecture it is located in, by using the supplied geographic position. Then, the

probability of a call being established between sub-prefectures i and j is estimated as:

$$c_{ij} = \frac{\mathcal{C}_{ij}}{\sum_k \mathcal{C}_{ik}}, \tag{4.2}$$

where $\mathcal{C}_{ij}$ is the number of phone calls initiated from the sub-prefecture i and directed to the sub-prefecture j, during the entire period of observation. The term at denominator indicates the total communication flux out of i and it is used to normalise the probability. Using these values a *communication matrix* C is built. This matrix also shows high values along the diagonal, but it is distinctly denser, showing that calls between sub-prefectures are more common than movement. The vertical line at $x = 60$ identifies calls directed to the sub-prefecture that contains the capital.

Finally, it is worth mentioning that, although I do not use it for these analyses, the high-resolution short-term dataset could be used to estimate the number of potential connections that an individual might have in a certain area, served by a cell phone tower.

4.3 Disease spreading in the presence of mobility and information

In the following I will present two models: a model of disease spreading as a function of the mobility patterns of individuals and a model for information spreading among the same population, considering the underlying social structure. The multiplex structure is used to model the interactions between the mobility layer, which represents movement of users between regions, and the communication layer, which captures calls between regions and, consequently, information spreading across the country. Each model will be evaluated using the data inferred from the CDRs.

4.3.1 Epidemic spreading with mobility

I now present a model that represents the evolution of an epidemic taking place on a network of metapopulations. The aim of the model is to describe how the system evolves under the action of two processes: contagion and mobility. The population is distributed in n different metapopulations, where the i-th metapopulation has $N_i[t]$ individuals at time t. I make the simplifying assumption that there are no deaths and births in the considered time window, i.e., at each time $t = 1, 2, \dots T$ the total population is constant $\sum_{i=1}^{n} N_i[t] = N$.

I assume that contagion happens inside each metapopulation following a standard SIS model [KR08]. I indicate the number of infected and susceptible individuals at time t in a metapopulation i with $I_i[t]$ and $S_i[t]$, respectively. At each time t a person is either infected or susceptible, therefore $N_i[t] = I_i[t] + S_i[t]$.

Simultaneously, individuals move through the metapopulation network according to the *mobility matrix* M of dimension $n \times n$. The generic element m_{ij} of the matrix represents the probability that a person moves from the metapopulation i to j. Hence, the state variables $N_i[t]$ evolve over time as follows: $N_i[t+1] = \sum_{j=1}^{n} m_{ji} N_i[t]$. Under the assumption that individuals inside the classes I and S move consistently, the last relation can also be written for the state variables $I_i[t]$ and $S_i[t]$. This assumption can be relaxed if data about the behaviour of different classes of individuals is available, i.e., when a different matrix M can be defined for each class – I and S. The contagion-mobility combined system can then be described by the following set of equations:

$$
\begin{aligned}
I_i[t+1] &= \sum_{j=1}^{n} m_{ji} \left[I_j[t] + \lambda \frac{S_j[t]}{N_j[t]} I_j[t] - \gamma I_j[t] \right] \\
S_i[t+1] &= \sum_{j=1}^{n} m_{ji} \left[S_j[t] - \lambda \frac{S_j[t]}{N_j[t]} I_j[t] + \gamma I_j[t] \right],
\end{aligned} \tag{4.3}
$$

for each metapopulation $i = 1, 2, \dots, n$, with λ being the product of the contact rate and the contagion probability and γ being the recovery rate. The formulae inside the square

brackets describe the evolution of n SIS models, one for each metapopulation. They are multiplied for the elements of the mobility matrix, which accounts for individuals moving between metapopulations.

This analytical model describes the expected outcome of a stochastic model where the following actions occur at each time step: first, each infected person in the metapopulation j causes the infection of new $\lambda\frac{S_j}{N_j}$ individuals inside j (this step is repeated for each metapopulation); successively, a new position i is assigned to each individual in the metapopulation j according to the probability density function $[m_{j1}, m_{j2}, \ldots, m_{jn}]$ (this step is repeated for each metapopulation).

The model I have presented can also be described in a more compact matrix form. Let us indicate with

$$\mathbf{N}[t] \equiv \begin{bmatrix} N_1[t] \\ N_2[t] \\ \vdots \\ N_n[t] \end{bmatrix}$$

the state vector of the population at time t. Analogously, the state vectors $\mathbf{S}[t]$ and $\mathbf{I}[t]$ can be defined for the corresponding number of susceptible and infected people, respectively. Hence, the equation describing human mobility is given by

$$\mathbf{N}[t+1] = M^{\mathrm{T}}\mathbf{N}[t] \tag{4.4}$$

where T denotes the transpose operator. Assuming the equation also holds for $\mathbf{A}[t]$ and $\mathbf{S}[\mathbf{t}]$, we have

$$\mathbf{I}[t+1] = M^{\mathrm{T}}\left(\mathbf{I}[t] + \frac{\lambda}{\mathbf{N}[t]} \circ \mathbf{S}[t] \circ \mathbf{I}[t] - \gamma\mathbf{I}[t]\right) \tag{4.5}$$

$$\mathbf{S}[t+1] = M^{\mathrm{T}}\left(\mathbf{S}[t] - \frac{\lambda}{\mathbf{N}[t]} \circ \mathbf{S}[t] \circ \mathbf{I}[t] + \gamma\mathbf{I}[t]\right) \tag{4.6}$$

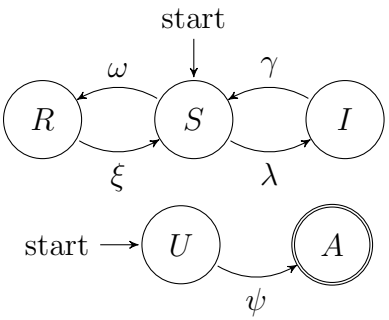


Figure 4.3: State machines describing the state transitions of a person with respect to the disease contagion (R=Resistant, S=Susceptible and I=Infected) and with respect to the information spreading (U=Unaware, A=Aware), respectively. A person starts in the susceptible and unaware states. It is assumed that aware individuals spread the information and cannot go back to the unaware state.

where the operator $\circ$ is the Hadamard product, otherwise called entrywise product. It is straightforward to verify that, for $n = 1$, it reduces to the classical SIS model for a population with homogeneous mixing.

4.3.2 Information spreading model

The model I presented in the last section describes the spreading of a disease in a population where individuals change locations over time. The aim of this work is to analyse some scenarios and study the effectiveness of possible containment techniques. In particular, as anticipated, the final goal of this study is to investigate whether a collaborative effort of the population is able, in theory, to reduce the spread of the disease and to what extent. The population can disseminate, using pre-existing personal social ties, *immunising* information (e.g., information about prevention techniques, hygiene practices, advertisement of nearby vaccination campaigns, and in general any information that can lead to a reduction of the number of contagion events).

In order to take into consideration these aspects, I now use a SIR model for each metapopulation in the mobility layer of the multiplex network, so that each person either belongs to the susceptible (S), infected (I) or resistant (R) category. At the same time, another simultaneous epidemic happens on the network of metapopulations, disseminating

information that can make individuals resistant to the disease, actually working against the disease epidemic. I model this network by means of a communication layer, where nodes are the same metapopulations existing in the mobility layer.

In fact, a person also belongs to the category of unaware (U) or aware (A) individuals, with respect to the immunising information. More formally, for every metapopulation i we have that:

$$N_i[t] = I_i[t] + S_i[t] + R_i[t] = A_i[t] + U_i[t] \tag{4.7}$$

It is worth noting that this "immunising epidemic" goes *beyond* the boundaries of metapopulations: in other words, it is a *distance contagion*. It is also important to remark that the states "aware" and "resistant" are substantially different. An unaware person that receives the information (i.e., has an "information contact") becomes aware with rate ψ; since the person is aware, he or she will start spreading the information as well. An infected person that receives the information becomes immune with rate ω. Additionally, individuals who have acquired immunity through information can lose it with rate ξ (for example, because they forget or abandon disease prevention practices). The transition rates between states are summarised in Fig. 4.3. The model can be described by the following set of equations, specifying how state vectors evolve over time:

$$\begin{aligned}
I_i[t+1] &= \sum_{j=1}^{n} m_{ji}\left[I_j[t] + \lambda\frac{S_j[t]}{N_j[t]}I_j[t] - \gamma I_j[t]\right] \\
S_i[t+1] &= \sum_{j=1}^{n} m_{ji}\left[S_j[t] - \lambda\frac{S_j[t]}{N_j[t]}I_j[t] + \gamma I_j[t] + \xi R_j[t] - \omega S_j[t]p_j[t]\right] \\
R_i[t+1] &= \sum_{j=1}^{n} m_{ji}\left[R_j[t] - \xi R_j[t] + \omega S_j[t]p_j[t]\right] \\
A_i[t+1] &= \sum_{j=1}^{n} m_{ji}\left[A_j[t] + \psi U_j[t]p_j[t]\right] \\
U_i[t+1] &= \sum_{j=1}^{n} m_{ji}\left[U_j[t] - \psi U_j[t]p_j[t]\right]
\end{aligned} \tag{4.8}$$

for every $i = 1, 2, \ldots, n$, where

$$p_j[t] = \frac{\sum_{k=1}^{n} c_{kj} A_k[t]}{\sum_{k=1}^{n} c_{kj} N_k[t]} \tag{4.9}$$

represents the probability that a call from an aware person occurs in the metapopulation j. This term is responsible for the interaction between the two layers of the multiplex and it models the distance-contagion. It is possible to verify that if the matrix is identical (absence of contacts between populations) it reduces to $A_k[t]/N_k[t]$, falling back to a model where contagion occurs only inside metapopulations.

This analytical model describes the expected value of a stochastic model where the following actions occur at each time step t: i) each infected person in the metapopulation j causes $\lambda \frac{S_j}{N_j}$ new individuals to get infected inside j; ii) each unaware person in the metapopulation j becomes aware with probability $\psi p_j[t]$; iii) each person in the metapopulation j who is susceptible, becomes resistant with probability $\omega p_j[t]$; iv) a new position i is assigned to each person in the metapopulation j according to the probability density function $[m_{j1}, m_{j2}, \ldots, m_{jn}]$. It is worth mentioning that each step is repeated for each metapopulation.

4.3.3 Simulations

I study the evolution of the epidemic outbreak on multiple scenarios by using the analytical model, considering a large range of values for the key parameters, and conducting a series of Monte-Carlo simulations for multiple sets of parameters. Each simulation is run on a distinct set of parameters and is deterministic. The estimated population size of Ivory Coast for July 2012 was 21,952,093 [Cen12]. Ivory Coast is organised in 393 sub-prefectures (*sous-préfectures*), which will be used to define metapopulations. Hence, I initialise each scenario by allocating 22 million individuals to sub-prefectures, according to information found in CDRs. In each scenario I bootstrap the spreading process by

infecting a fraction of the population (0.1%) distributed across metapopulations according to different criteria:

- **Uniform distribution**: every sub-prefecture gets a number of infected proportional to their population, i.e., every sub-prefecture has the same fraction of infected population.
- **Random**: a single sub-prefecture, chosen randomly, is the origin of the infection.
- **Centrality based**: the sub-prefectures are chosen according to their centrality rank, in particular focusing on the first 1, 5 or 10 highest ranked. Although eigenvector centrality is used, it is worth remarking that the top ranked sub-prefectures are very similar for other types of centralities, as reported in the Tab. 4.1.

Betweeness	Closeness	Degree	Eigenvector
60	60	60	60
39	58	58	58
89	39	39	39
58	69	69	69
75	138	138	250
144	250	64	138
138	64	144	64
165	144	250	144
212	182	122	122
168	122	182	182

Table 4.1: IDs of the highest ranked sub-prefectures, according to different definitions of centrality, calculated on the mobility network. The 10 top ranked sub-prefectures ordered by centrality are very similar.

In the following, I present the results obtained from these simulations.

Unconstrained epidemic spread

First, I explore the evolution of the epidemic in the case where no countermeasures are taken. In order to analyse the evolution of the system in more detail, two measures are considered: the fraction of infected population i^{∞} at the stationary state and the time

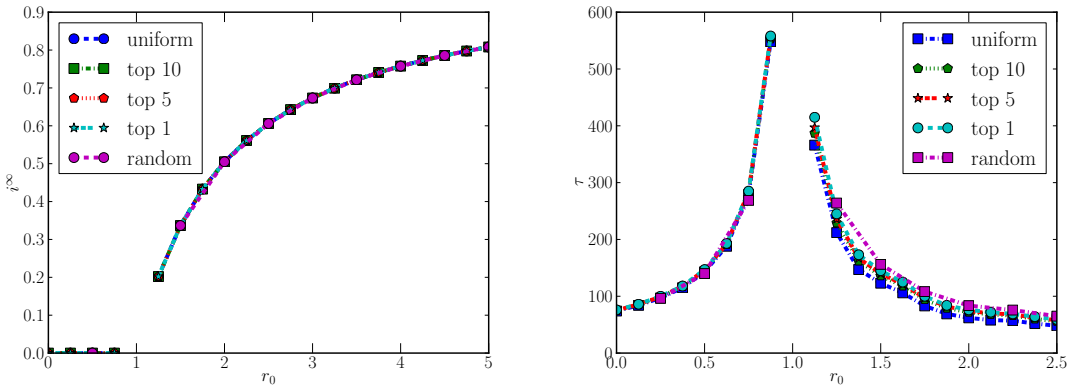


Figure 4.4: Fraction of infected population at the stationary state (left panel) and time required to reach the stationary state (right panel), for different values of r_0 and for different initial conditions. Missing values in the curves mean that, for the corresponding values, no stationary state is reached during the period of observation.

required to reach the stationary state τ. Fig. 4.4 shows their values versus $r_0 = \frac{\lambda}{\gamma}$, which is the basic reproductive ratio of a classic SIS model [KR08]. For $r_0 = \frac{\lambda}{\gamma} < 1$ there is no endemic state (i.e., the final fraction of infected population is zero), whereas for $r_0 > 1$ a non-null fraction of population is infected. Values for $r_0 = 1$ are missing since no stationary state is reached within the observation window. In other words, for this particular scenario, experimental results show that the basic reproductive ratio of the model is very close to r_0; this is a consequence of the low inter-subprefectures mobility. We can also notice that the initial conditions do not affect i^∞ at all. Before the critical point (i.e., $r_0 = 1$) the choice of the initial conditions also has no impact on the delay time, whereas for $r_0 > 1$ it slightly affects the delay: epidemics that initially involve more sub-prefectures are slightly faster than the others.

Geographic quarantine

I now analyse the effects of curbing the mobility between sub-prefectures, i.e., forbidding all the incoming and outgoing movement for a group of sub-prefectures. To this aim, I calculate the centrality values of each sub-prefecture in the mobility network. I present the results for eigenvalues centrality, because the ranking based on other centralities is very similar. Then, for the quarantine operations, I select those with the highest centrality

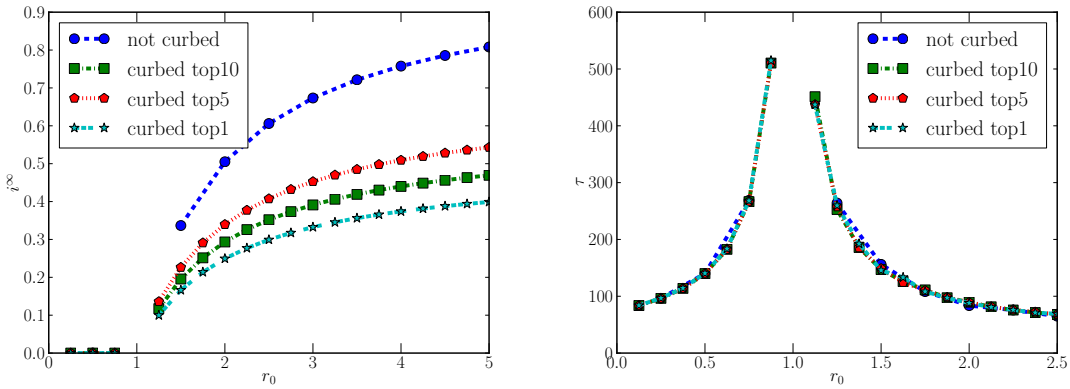


Figure 4.5: Fraction of infected population at the stationary state (left panel) and time required to reach the stationary state (right panel), for different values of r_0 when the epidemic starts from a random sub-prefecture, and different levels of geographic quarantine are applied. Missing values in the curves mean that, for the corresponding values, no stationary state is reached during the period of observation.

values. From a practical point of view, this is achieved by simply changing the $i-$th row and column in the mobility matrix, so that all the elements m_{ij} and m_{ji} are null, except for the elements $m_{ii} = 1$. It is important to note that in this process I recompute m_{ji}s in order to conserve the probability, i.e., so that each row of M sums up to one. For these scenarios, I randomly choose a single sub-prefecture where the initial individuals are infected, and then I average i^∞ and τ over all the Monte Carlo simulations. As shown in Fig. 4.5, the fraction of the infected population is sensibly affected by this measure, as the population inside the quarantined areas is protected from contagion. However, contrary to the intuition, the delay is not affected by the quarantine, even when the countermeasures involve 10 sub-prefectures, which account for almost half of the population. This suggests that such an invasive, expensive and hard to enforce measure considerably reduces the endemic size, but does not slow down the disease spreading in the rest of the country. For this reason, I now investigate a radically different approach to protect the population.

Information campaign (social immunisation)

I now show how a collaborative information campaign could help in contrasting the spread of the disease, following the model I presented in the last section. The scenario is initialised

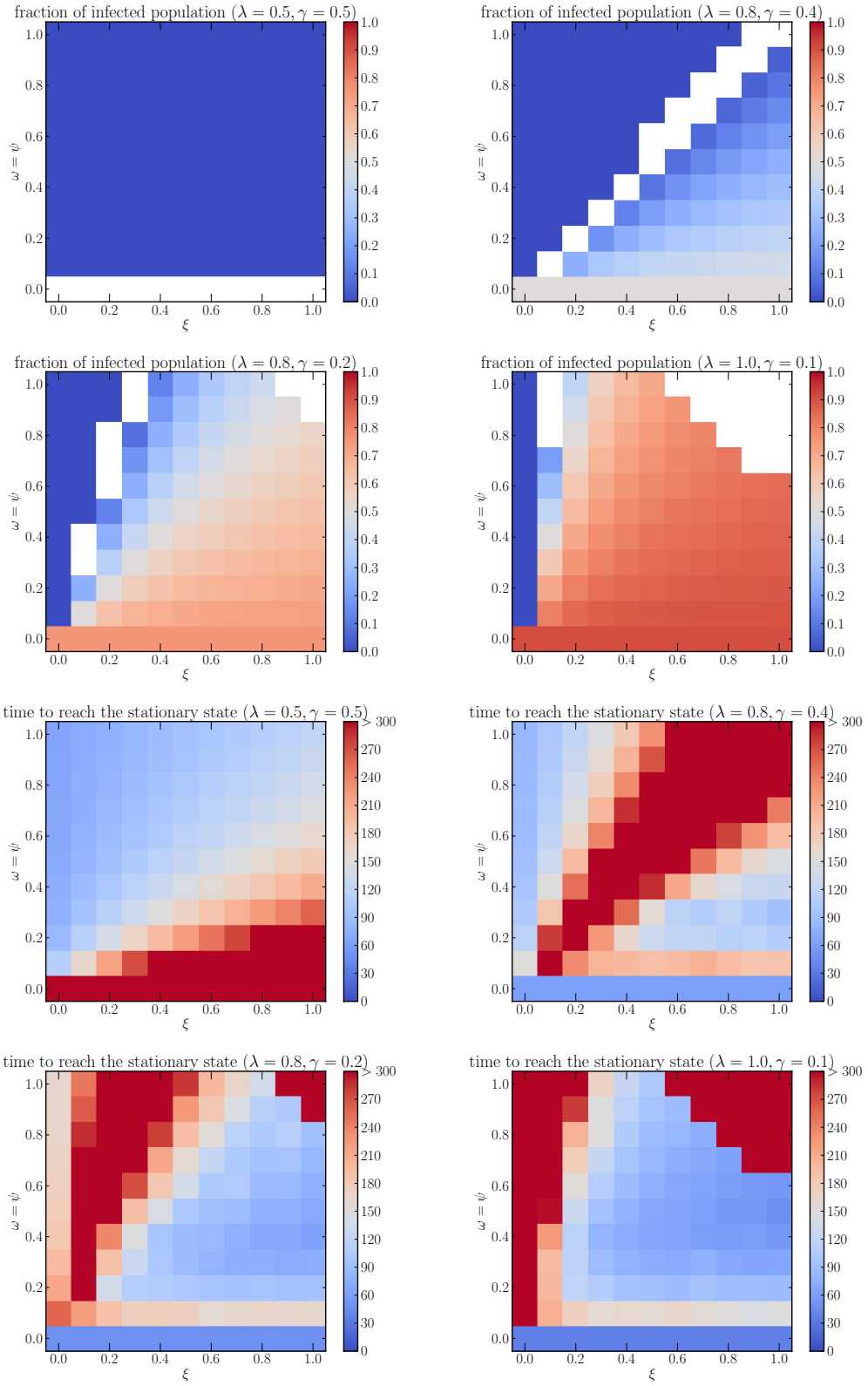


Figure 4.6: Fraction of infected population at the stationary state (first two rows) and time required to reach the stationary state (last two rows), for different values of $r_0 = \frac{\lambda}{\gamma}$ (1, 2, 4, 10, respectively). White spaces show that no stationary state is reached during the period of observation.

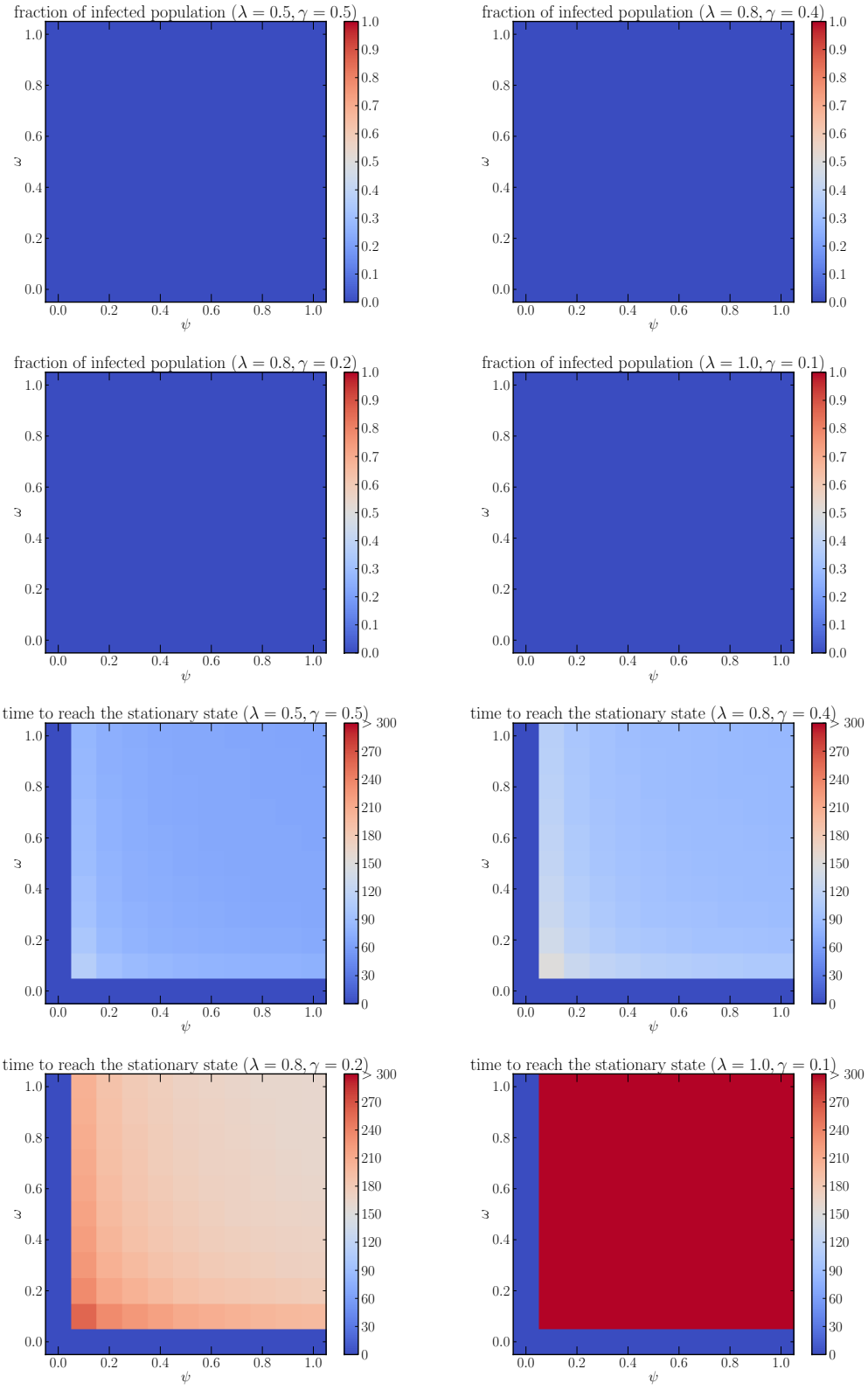


Figure 4.7: Fraction of infected population at the stationary state (first two rows) and time required to reach the stationary state (last two rows), for different combinations of $\frac{\lambda}{\gamma}$ (1, 2, 4, 10, respectively) and $\xi = 0$.

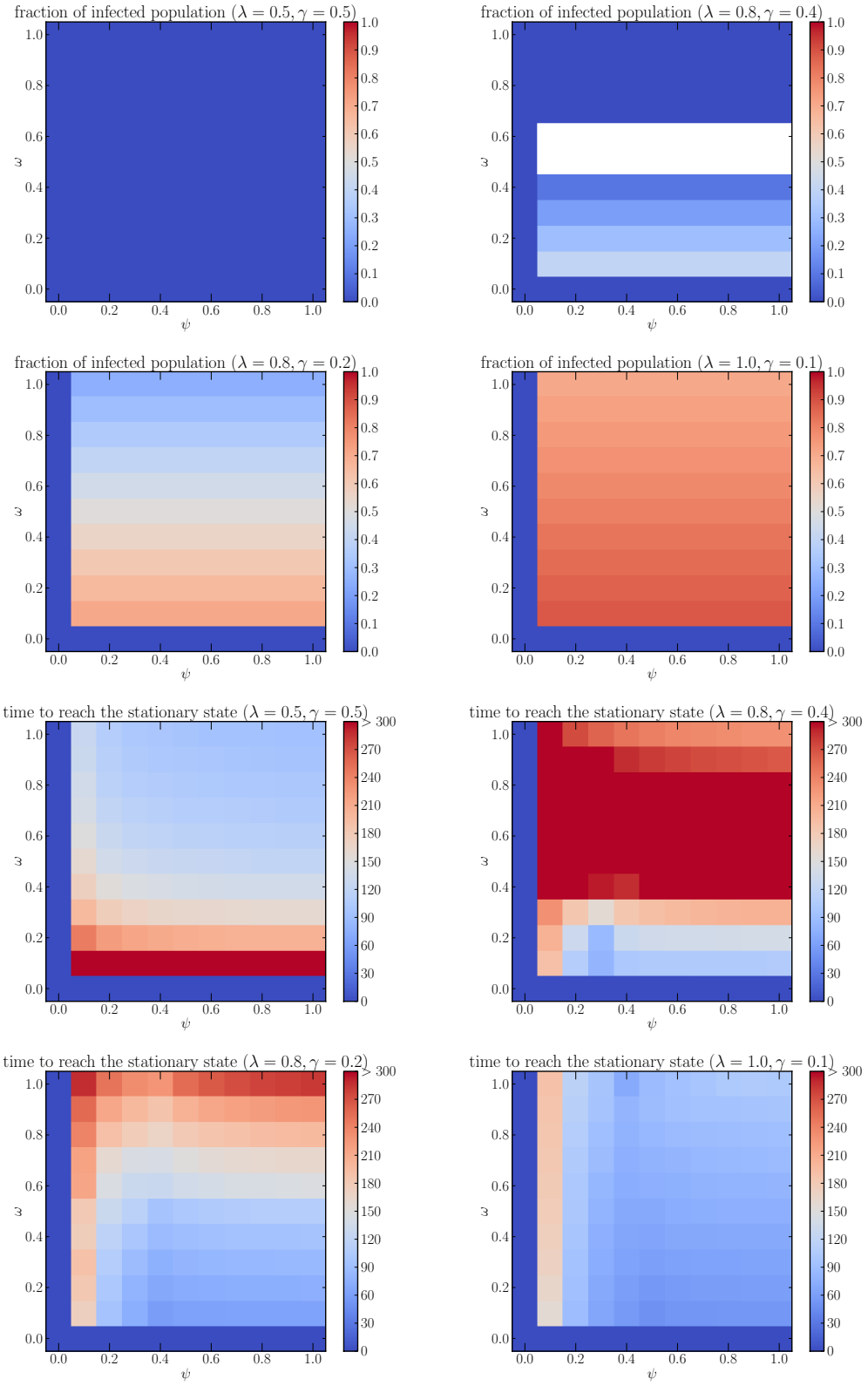


Figure 4.8: Fraction of infected population at the stationary state (first two rows) and time required to reach the stationary state (last two rows), for different combinations of $\frac{\lambda}{\gamma}$ (1, 2, 4, 10, respectively) and $\xi = 0.5$. White spaces show that no stationary state is reached during the period of observation.

by distributing the immunising information to 1% of the population, randomly chosen regardless of their location. These people will be informed and will be instructed to spread the information. In other words, I assume that they will contact their social connections, according to the call network.

Fig. 4.6 shows the density plots describing i^{∞} and τ for various values of r_0, for a subset of scenarios where $\omega = \psi$, i.e., when the information that spreads among the population has the same chance to immunise a person and to involve the person in the spreading process. This is consistent with a scenario where the same set of people who become aware also become immunised by the information they have received. Blank squares show that a stationary state was not reached for the corresponding set of parameters. The figure shows how contagious (ω=ψ) the immunising information has to be with respect to how often people "forget" (ξ) in order to slow down the disease considerably and to reduce the endemic cases. When $\omega = \psi = 0$ the model falls back to the case without information spreading, and the value of ξ does not affect i^{∞} and τ. For $\omega = \psi > 0$ and $\xi = 0$ the fraction of infected population goes to zero in all cases, because the number of people aware of the information does not decrease, thus increasing the number of new immunised individuals at each step. We can notice that even for low values of participation ω and for information that gives temporary immunisation ($\psi > 0$), the final fraction of infected individuals is considerably lower than in the case where no countermeasures are taken.

Figs. 4.7 and 4.8 show the density plots for ω and ψ when ξ is constant. In particular, the scenario for $\xi = 0$, shown in Fig. 4.7, might represent dissemination of information about vaccination campaigns (individuals who have been administered vaccination do not lose immunity). For every combination of parameters the endemic state is absent even with the highest considered value of r_0. The two parameters that represent how individuals are likely to get involved in the immunisation and in the information spreading (ω and ψ, respectively) seem to have similar impact on the delay of the infection.

The value $\xi = 0.5$ (Fig. 4.8) might describe a scenario in which the information is about a good practice (e.g., boiling water, using mosquito nets, etc.), which loses its

effectiveness or it is stopped being used by a person with rate ξ. For this case we can notice that the fraction of infected population is independent from ψ, as rows in the density plot are of the same colour. This suggests that, for this scenario, the rate at which people lose immunity does not affect the final size of the endemic state.

4.4 Evaluating the risk of individuals

While so far I have discussed epidemic mitigating measures taken at a collective level, I now move to the individual level. It has been shown that human mobility is predictable [dMHV+13; SQB+10], characterised by a high level of regularity [GHB08] and slow exploration [SKW+10]. Additionally, individuals show large variability in how much they move; recent work has successfully captured this aspect by classifying people into two main groups: "explorers" who visit many locations and "returners" who move mostly within a small group of places [PSR+15]. It is reasonable to expect that some individuals who travel a lot might play a more important role in spreading the disease than people who travel less.

On the other hand, regions also typically show a substantial heterogeneity, in terms of incidence of the disease. Especially at the first stages of the outbreak some regions might have a large number of infections while others might have a very low number or none at all. That makes us presume that also some regions play a more central role than others in spreading the disease.

These two heterogeneities of individual mobility patterns and of regional incidence of the disease effectively make some individuals much more central actors in the spreading dynamics. A group of people who move frequently between high incidence and low incidence areas will contribute much more than a group of people who move less and mostly within low incidence only (or high incidence only) areas. The remainder of this chapter is aimed at formalising such an intuitive concept into a risk measure and evaluating

through computer-based simulations whether targeted efforts based on this measure are really effective. This work is motivated by several reasons, briefly discussed here.

Individual spreading behaviour is highly heterogeneous. Existing epidemic models are based on analyses conducted at population level to assess how infectious a disease is, based on the basic reproductive ratio r_0, i.e., the average number of secondary cases generated by a single infected person. However, several studies have concluded that spreading processes are usually highly heterogeneous and that some individuals remain responsible for a large proportion of the spreading. The presence of these influential spreaders has been investigated for generic networks [KGH+10], as well as in epidemic processes. Superspreading seems to be a common feature of the spread of diseases and targeted individual-based control measures are much more effective than population-wide measures, as reported by Lloyd-Smith et al. [LSK+05]. For this reason, identifying superspreaders is extremely important in order to contain epidemics.

Existing techniques, such as contact tracing, are not sufficient. Moreover, efforts in fighting disease outbreaks mainly focus on contact tracing techniques, as it happened for Ebola [Mur14]. Contact tracing works by finding all the people who have been in contact with an infected person, and then interviewing, monitoring and isolating them when necessary. The process is repeated for everyone who is found to be infected. While contact tracing can be effective, it has some drawbacks. First of all, information provided by people might be subject to errors, due to fear, shame, faulty memory or other reasons. Secondly, contact tracing needs time: contact tracing only starts when a person is already diagnosed with the disease, or at least shows symptoms. Tracing the contacts also takes time: if the disease has an asymptomatic phase or is highly infective, the contacts might be likely to have infected others before they are traced.

Localisation techniques have been successful during critical scenarios. During the recent Ebola outbreak, Nigeria resorted to GPS technology to improve, scale up and speed up contact tracing, repurposing GPS devices used for polio vaccinations [FSL+14; Gat14]. The huge effort of the country resulted in eradication of Ebola and Nigeria was

declared "Ebola-free" by the WHO[1]. While this success story demonstrates how location tracking can be very useful during similar scenarios, the very same strategy could not have been used if the epidemic was in a more advanced state, i.e., if many more people had already been infected. For this reason, it is very important to investigate the use of alternative systems that can provide coarser location tracking but for a large number of individuals.

Medical treatment is scarce and costly. For example, in the case of Ebola, although the disease is seen as a serious challenge by the whole world, vaccinations have to face serious technical and financial issues before being administered[2]. When a commodity such as vaccinations is scarce, who should be given priority during a vaccination campaign? A risk-based measure might a good answer on how to prioritise such scarce resources in a fair and effective way, keeping in mind that the objective is to maximise the benefit for the whole community.

4.4.1 Risk model

Here I propose a method to quantify the risk associated with each person during an outbreak, depending on their mobility behaviour, inferred from their phone-activity. Here this method is referred to as the *risk model*. The goal of the model is not the estimation of the individual cost (i.e., the chance of getting infected), but the cost that an *entire community* faces by not treating a specific person. Early testing, medical treatment, vaccination and quarantine of specific individuals might reduce costs sustained by the community at a later time.

A general estimate of the total risk R associated to a set of events E is defined by:

$$R = \sum_{E} P_E \times L_E \tag{4.10}$$

[1] http://www.who.int/mediacentre/news/ebola/20-october-2014/en/

[2] http://news.sciencemag.org/health/2014/10/leaked-documents-reveal-behind-scenes-ebola-vaccine-issues

where P_E and L_E are the probability and the expected loss for each event, respectively [Vap98].

This definition can be applied to the epidemiology domain by considering a scenario in which several geographic areas are assigned different values of time-varying contagion risk. The risk measures how likely it is for an individual to get infected in a region. As in common models of infectious diseases, I assume it is directly proportional to the fraction of infected people in the region and I also assume homogeneous mixing within the region. Similarly, I assume that the risk to infect a healthy individual is directly proportional to the fraction of susceptible people in the region.

By staying in a geographic area with a non-zero risk, there is a chance a person will get infected; there is also a chance that same person will infect someone else, increasing the risk of the geographic area. When moving between two or more areas, the person will affect the risk of these areas. It is important to remark that the model does not determine whether each person is in a susceptible, infected or recovered state, it is agnostic of their state. Instead, the model assumes they might be in any state and it can be used to assess how risky their mobility behaviour is.

In general, the way people transmit disease across geographic areas has been extensively studied in literature [BPR+11; BCG+09; MA10]. Most of the studies dealing with the effects of mobility on epidemic spreading usually make the assumption that the mobility patterns of individuals in a subpopulation are homogeneous [CV07], while they are indeed highly heterogeneous [DPE13; MA10]. This is particularly true for developing countries, where highly irregular and temporally unstructured contact patterns have been observed [VBS+13].

Here I consider a disease that has contagion rate per contact β (i.e., given a friendship between an infected and a susceptible person, a contagion will happen with rate β). Assuming the user u spends $T_{u,l}$ fraction of his time in each location $l \in \mathcal{L}_u$ (hence,

$\sum_i T_{u,i} = 1$) I define a time-dependant contagion risk:

$$C_u(t) = \beta \sum_{l,m \in \mathcal{L}} T_{u,l} T_{u,m} [i_l(t) s_m(t) + i_m(t) s_l(t)]. \tag{4.11}$$

where $i_l(t)$ and $s_l(t)$ refer to the fraction of infected and susceptible population in location l at time t, respectively. Note that now the probability of the event occurring, in this case, is the probability that a person becomes infected in a region, according to the time fraction spent there, while the expected loss is the number of people expected to be infected in another region, according to the time fraction spent there. As the model does not make any assumption on the location of a possible infection, among those visited by a person, this formula accounts for all the combinations, which are assumed as equally likely. The maximum risk value, for a specific state of the network, is reached by an individual who equally spends his time in the region with the highest infected fraction of individuals and in the region with the highest susceptible fraction. The normalisation here is not necessary for ranking purposes, since it is a common factor among all the individuals considered; the rate β can also be ignored for the same reason. Furthermore, the model proposed here could be generalised by defining different risk classes depending on demographic indicators, which can be inferred from mobile data [ZTM+13] or other behavioural indicators, such as those provided with the D4D-Dataset [dMST+14].

4.4.2 Evaluation

Next, I evaluate the effectiveness of the risk identification and containment model proposed above. I set up a realistic epidemic scenario and perform stochastic simulations, following an approach similar to that used in the GLEaM model [BCG+09], while keeping track of the movement of individuals following the real traces found in the dataset. Each simulation "run" represents a distinct realisation of the epidemic process; plots that show the process evolution include 95% confidence intervals. I use the SEIR model, where each individual can be in one of the following discrete states at any given time instant:

susceptible (S), exposed (E), infected (I), permanently recovered or deceased (R). This model has been used for the 2002 seasonal influenza outbreak [BCG+09] and the 2014 Ebola outbreak [Alt14], among other outbreaks. It is described by the following set of equations:

$$\frac{dS}{dt} = -\beta S(t)I(t)/N \tag{4.12}$$
$$\frac{dE}{dt} = \beta S(t)I(t)/N - kE(t) \tag{4.13}$$
$$\frac{dI}{dt} = kE(t) - \gamma I(t) \tag{4.14}$$
$$\frac{dR}{dt} = \gamma I(t) \tag{4.15}$$

The spreading model is informed with the realistic parameters taken from estimates of the 2014 Ebola outbreak in Sierra Leone [Alt14], as reported in Tab. 4.2. Where σ^{-1} and γ^{-1} are the average durations of incubation and infectiousness, respectively. The transmission rate per day in absence of control interventions is β, and $r_0 = \beta/\gamma$ is the basic reproduction number.

σ^{-1}	5.3 [days]
γ^{-1}	5.61 [days]
r_0	2.53
β	0.45

Table 4.2: Parameters used for the simulations.

I simulate the epidemics in the following different contexts:

- in total absence of any treatment;
- when treatment is given with rate ξ per day and people given treatment are chosen randomly;
- when treatment is given with rate ξ per day to highest ranked people, according to the risk measure C_u.

For simplicity, in this study I focus only on treatment that takes the form of travel restrictions, not allowing high-risk individuals to travel outside the metapopulation they

are found when the treatment is applied. This is an extreme scenario, realistic only for diseases for which specific treatments or vaccinations are not available (e.g., Ebola virus). Without loss of generality, the same method can investigate the effects of vaccination and/or early treatment of people with higher-risk movement patterns. Since I use the same parameters for each metapopulation, and the treatment does not directly affect the epidemic process (i.e., it is not a vaccination or a cure) but only the movement of individuals, the local epidemic profiles will be similar and will be more or less shifted in time, depending on the travel fluxes.

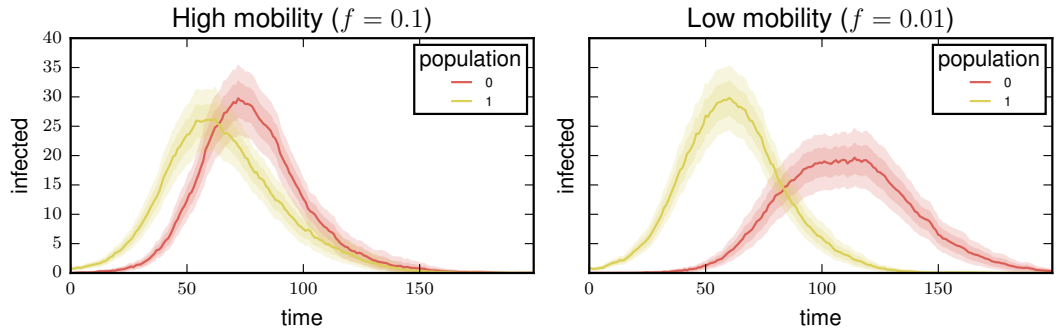


Figure 4.9: A simple example of two metapopulations composed of people who stay always in their own metapopulation and a fraction f of people who move between them randomly. Bands indicate 68 and 95% confidence intervals. In the left figure $f = 0.1$, while in the right figure $f = 0.01$. The outbreak dynamics in the second case are less synchronised.

A simple synthetic example

I first show how much the disease dynamics synchronisation can be reduced by restricting the travel of high-risk individuals in a simple example. As an illustrative case, I simulate a synthetic model composed of two metapopulations. People are assigned to either metapopulations and they belong to two classes: a fraction of "static" people $(1 - f)$ who never travel out of their metapopulation, and a fraction of "travellers" f who choose at each step, at random, in which population to be. The SEIR is set with the parameters mentioned before and the simulation is initialised with a single infected case in one

of the two metapopulations, chosen randomly. Two scenarios are simulated: a scenario where travel between the two populations is common (the fraction of people who travel is $f = 0.1$), and another where travel is more rare ($f = 0.01$). Each simulation is realised 100 times. Fig. 4.9 shows how the current number of infections evolves in time. The left plot ($f = 0.1$) shows a high level of synchronisation between the two populations, while the right plot ($f = 0.01$) displays a clear delay in the growth of the epidemic size.

This simple example shows how the fraction of people who are travelling plays a big role in spreading the disease to other regions. The more people travel between two regions, the more synchronised their epidemic evolution is. Stopping these high-travel individuals (or, equivalently, treating them) might be very effective to delay the onset of the disease in other regions.

Real-data evaluation

I then evaluate the disease dynamics on real-data, initialising the computer-based simulations so that a single randomly chosen region is the unique source of infection with

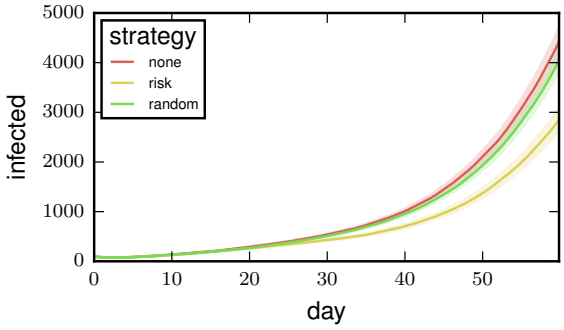


Figure 4.10: Evolution of the number of infected individuals in three cases: when no countermeasures are taken (none), when people are quarantined randomly (random) and when quarantined people are chosen among those with the highest risk (risk). The bands shown for each curve represent a 95% confidence interval, estimated from 50 independent simulations run for each strategy. The "random" strategy is indistinguishable from the case of absence of measures. Instead, the proposed risk-identification method reduces the number of infected individuals by 35% with fewer people in quarantine, using only aggregated information of the number of infection cases and mobility patterns from mobile phone data providers.

100 cases. I use the first six months of the coarse-grained long-term mobility dataset, from January to June 2013, to learn the movement habits of individuals. Then I use the following two months of the dataset (from July 2013 to August 2013) to perform simulations under the three scenarios mentioned above: no countermeasures, people quarantined randomly and people quarantined according to their risk rank. Each scenario is realised through 50 simulations. I set an adaptive quarantine rate of $\xi = \beta i(t)$ to match the countermeasure efforts with the speed of growth of the outbreak. Fig. 4.10 shows results for the months of July and August 2013, in terms of how the global prevalence of the disease changes with time in the three cases. Despite almost 40% of the population being quarantined through a "random" strategy by the end of the month, this does not have any impact on the spreading and the dynamics is indistinguishable from the case where no countermeasures have been taken. Instead, targeted quarantine based on risk manages to delay the spreading with fewer quarantined individuals: at the end of the two-month period there are 35% fewer infected individuals than in the baseline cases, with 27% of the population being quarantined. It is worth underlining the fact that the number of quarantined individuals is different in the two cases because it was set to be adaptive, depending on the number of cases. Although the "risk" strategy quarantines cumulatively fewer individuals it manages to be more effective than quarantining more people, chosen at random.

Fig. 4.11 shows the fraction of infected people in each region (estimated as a mean from all the realisations). The figure provides an insight into how the "risk" strategy manages to delay the spread of the disease. At day 45, as a result of the "risk" policy, most of the infected people are concentrated in the sub-prefecture where the infection started; the other regions, instead, have lower occurrence of the disease, compared to the "random" strategy. At day 90, while for both strategies the source region has a low occurrence (since the infected individuals became "recovered"), infected individuals are spread across several regions in the country in the "random" case. For the "risk" strategy they are concentrated around a few regions near the source.

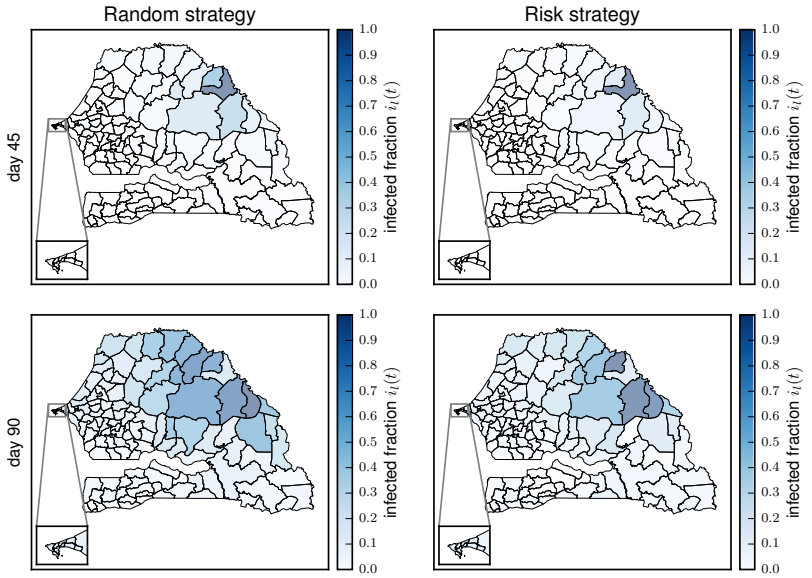


Figure 4.11: Evolution of the disease across the sub-prefectures. The shading of each region indicates the fraction of its population that is infected, estimated as a mean from the values obtained during all the simulations. In the first snapshot (top row) it is possible to notice that the source region (dark blue) has a high fraction of infected individuals. In regions that are close to the source region, the infection is starting, but there is lower occurrence of the disease when the "risk" strategy is used. This phenomenon is also visible in the second snapshot (bottom row): the "risk" strategy keeps spreaders in their source region, while the "random" strategy fails at identifying them and more nearby regions have a higher fraction of infected population.

4.5 Discussion and limitations

The models described here use data collected from mobile phones, hence it excludes people who do not use the mobile phones or share them with others. Since mobile penetration rates are already high and increasing in the vast majority of countries, including developing countries, it is safe to assume that the coverage problem will fade out as time goes on. Another potential problem when dealing with network-data is the sparsity of the call activity, but recent studies try to overcome this limitation [LLK+14] by interpolating information in space and time. Furthermore, for the second part of the chapter where I attempted at detecting individuals with higher disease-spreading risk, I would like to remark that the goal is not to find *every* high-risk individual, but a *large proportion* of them, given the data available. Moreover, it is worth noting that this method might also

be combined with other existing disease prevention and containment techniques already in use, such as contact tracing.

The models described in this chapter require access to sensitive data about individual call and mobility patterns. It is very important to take into account ethical and legislative issues arising from the use of these highly personal data. However, solutions based on the analysis of mobile data, such as that presented in this work, can play a critical role during emergencies. For this reason, it might be acceptable to use such a system when the benefits exceed the risks. The use of such systems might be beneficial only in well-defined circumstances, within specific time intervals and geographic boundaries, within the limits defined by the law and under user informed consent. The model could also be used to design a system that informs only the users themselves about their own behaviour, evaluating their risk level and, potentially, suggesting to them appropriate actions tailored to their risk profile (e.g., get tested, seek help, change lifestyle habits, etc.).

It is important to emphasise that a real-world system that would use these models would require access to two types of data: global information about the outbreak, which is already available (e.g., estimated number of infected people in various geographic regions); and individual information about user mobility, which is of a sensitive nature. For this reason, centralised deployments of the system might not be realisable, as user mobility might be inaccessible under local regulations and laws. On the other hand, decentralised deployments are possible and to be preferred. In such deployments the mobility data are only stored on the devices. The risk-profile is calculated on the phone and it is shown only to the user, who might optionally choose to follow tailored advice.

Moreover, it is worth stressing that the model proposed here simply gives a risk measure for each user. This measure can be used for less invasive measures than quarantine, for example to decide who gets access to scarce resources such as vaccinations or medical visits.

It is worth noting that technical and practical constraints might reduce the efficacy of mobility-based risk-assessment. In particular, I evaluate the model on mobility traces

that correspond to an epidemic-free case. People might significantly change their mobility behaviour once they are aware of the epidemic [MPA+11]. Users might not carry their device with them at all times, hence making mobility traces and risk-assessment less effective. Mobility containment and other individual-based strategies are difficult to enforce and they heavily depend on the ultimate choices of individuals in accepting the recommendations.

Finally, it is worth noting that epidemic simulations might deviate very much from reality. Predictions of the recent Ebola outbreak, based on state-of-the-art diffusion models and simulations were found to deviate significantly from real numbers [But14]. Scientists stressed that this is due to unreliable data, most notably about disease-control measures and about how they affect the spreading.

4.6 Summary

As introduced at the beginning of this chapter and detailed in Section 2, mobile network data have the peculiarity of describing the movement and communication of a large number (often millions) of people in country-wide areas. The high penetration rate of mobile devices suggests this kind of data is often an accurate description of the movement and communication happening in the whole country. This has made mobile network data very valuable for scientific analysis not only to uncover the salient traits of mobility and communication, but also for practical applications in several different fields (i.e., transportation, demographics, epidemic modelling, etc.) with encouraging results.

In this chapter, I have focussed on applications related to epidemic modelling and the design and simulation of mitigation strategies. Firstly, I expanded existing epidemic models, which rely mainly on mobility, to also make use of communication happening between individuals. I used the communication dynamics to initiate a concurrent counter-epidemic and I assessed its effectiveness through computer-based simulations. The results show that this mitigation strategy works in cases where the probability of successfully transmitting

the immunisation is low. Secondly, I used individual mobility traces generated from mobile network data to calculate a measure of personal risk, associated with the mobility profile of an individual and the disease incidence in the regions he or she typically visits. I proposed a mitigation strategy based on this measure and I also assessed it though computer-based simulations, finding that it has beneficial effects in a whole region.

While the results I have presented in this chapter rely on simplified assumptions and are a long way from being practically implemented, because of several technical issues, they successfully show how mobile network data can be used to simulate and devise innovative strategies and how they can also be used in decision making and in policy making. In the next chapter I will focus on GPS data, another source of mobility data which is geographically finer and temporally denser. At the end of the chapter I will show that mobile network data accuracy can be greatly improved through an analysis that uses GPS data and the street network.

CHAPTER 5

UNDERSTANDING AND ESTIMATING HUMAN PATHS WITH GPS DATA

Positioning devices, such as those based on Global Positioning System (GPS), are increasingly popular due to their high accuracy, low cost and because they can greatly improve services that are based on them. Many applications, from search to navigation, rely on the knowledge of real-time user positions to sense spatial context and to provide relevant results based on that.

On the other hand, traces collected from positioning services also represent a great research opportunity, as they open up the possibility of observing human movement with unprecedented detail. For this reason they have been used extensively to understand how people move [JJC+12; RSH+11; VBS+13] and to understand how to optimise urban transportation [SRS+14].

As previously discussed in Chapter 2, this data source has two main properties: it is very fine-grained, both in space and in time, and it typically describes a small number of individuals, usually much smaller than in a mobile data network dataset. One of the possible reasons for this is the high battery consumption of GPS receivers. While most recent smartphones carry GPS chips and use them to provide positioning, they tend to strictly limit its usage because it drains the battery quickly [PKG10; ZKS10]. Some recent approaches, used for example by Google Location History and Moves.app[2], combine

[2]https://www.moves-app.com

GPS with Wi-Fi and accelerometer sensors, to allow for continuous location logging and reasonable power consumption, albeit with reduced accuracy.

While the fine-granularity typically represents a desirable feature, it also complicates analyses of aggregate behaviours, for example when trying to structure them into a set of routine trips between relevant locations, and a set of recurring route choices.

This chapter will describe methodologies that deal with these negative points: first I will show how to extract routing behaviour patterns from GPS traces, then I will discuss how the analysis of GPS data can be used to increase spatial accuracy of cellular network data. This chapter is structured as follows.

- In the first section I analyse GPS traces to uncover salient traits of drivers' routing behaviour. I show how drivers tend to have a preferred route and how often it is not the optimal minimum-cost route. Then I explain how the deviation from the optimal route is often contained in an ellipsoidal spatial region which represents the virtual boundary of routing selection, regardless of the specific endpoints of the trip and its length.

- In the second section I show how GPS traces can also be used to increase the spatial resolution of data. In particular, I show how it is possible to accurately estimate the trajectories of mobile network subscribers from sparse and coarse data about their association to tower sectors. It is worth remarking that GPS traces were used only to develop the methodology and to evaluate its accuracy; additional GPS data may not be needed to estimate paths taken by other mobile network subscribers not represented in the data analysed here. The described methodology also makes use of information about the urban street network and manages to achieve a block-size accuracy. This represents a good example of how the combined analysis of mobility data with deeply different properties can produce data of higher quality.

5.1 Understanding drivers' routing behaviour

Many recent studies have uncovered the salient traits of visit patterns, such as in the first pioneering study by González et al. [GHB08]. However, there is little research about how people actually move between the places they regularly go to. This problem is important for several practical applications, most obviously to forecast traffic flow and optimise urban networks [TÇS+15], but also for more novel and unexpected applications, such as providing place-recommendations depending on the chosen route [HK12] or suggesting alternate paths that are less congested, something that is already done at a commercial level by services like Google Maps[1] and Waze[2].

While it might seem surprising, most of the transportation research studies that deal with traffic assignment and route choice rely on the widely accepted assumption that travellers take the minimum cost route, despite this assumption having limited empirical support. This assumption is sometimes referred to as Wardrop's first principle [War52], which says: *"the journey times in all routes actually used are equal and less than those which would be experienced by a single vehicle on any unused route".* In other words, travellers take the route that minimises their own travel times.

This seemingly reasonable assumption suffers from several drawbacks. First of all, while using the journey time as the cost function might seem sensible, the *actual* cost function minimised by drivers might involve several additional aspects (e.g., fuel consumption, trip length, trip duration, number of turns, reliability of the path, etc.). Secondly, it is reasonable to assume that each driver follows their own definition of cost: some people might travel on a tight budget, some others might want to reach their destination as soon as possible and some others might trade time with enjoyment and the possibility of choosing a picturesque route. The cost function is likely highly dependant on individual preferences. Finally, even if the cost function was perfectly known for each individual, drivers might still have incomplete or invalid information about the street network and

[1] `https://www.google.co.uk/maps`
[2] `https://www.waze.com`

its current state, so they might be *unable* to minimise it feasibly. This is reasonable to assume, as humans have cognitive limits to navigation [GPB16], and also they are not aware of real-time traffic conditions, unless they use some tailored external system, like GPS-aided navigation.

All the reasons detailed in the last paragraphs are part of the *behavioural heterogeneity* that characterises each traveller and the routes they choose: every person behaves in their own way, according to their beliefs, knowledge and habit. The goal of this study is to take a different approach aimed at understanding drivers' routing behaviour: instead of studying the reasons behind it, I try to capture and *measure* this heterogeneity.

By means of analysis of an extensive dataset containing trajectories of personal cars, in this section I show that 42% of the routes taken do not correspond to optimal routes, as generated by a leading navigation service. Nonetheless, the observed behaviour can be synthesised in a few rules. I find that individuals tend to have a preferred route when travelling frequently between two locations. In particular, most routes are fully contained within an ellipse, typically highly eccentric independently of the scale, which bounds how far individuals are usually willing to go from the ideal straight path.

Beyond the findings of fundamental importance, this work represents a starting point for future models of human mobility and route choice. Such models are widely used for traffic prediction to inform decisions and policies related to transportation infrastructure. The behavioural rules described here can be the basis of realistic routing models that capture individual behaviour and that are not based on the optimisation of traffic and travel times.

5.1.1 Dataset description

I analyse a dataset that contains 92,419 trajectories generated by GPS devices installed on personal cars. The data was collected in an undisclosed European country and the collection lasted for an 18-month period.

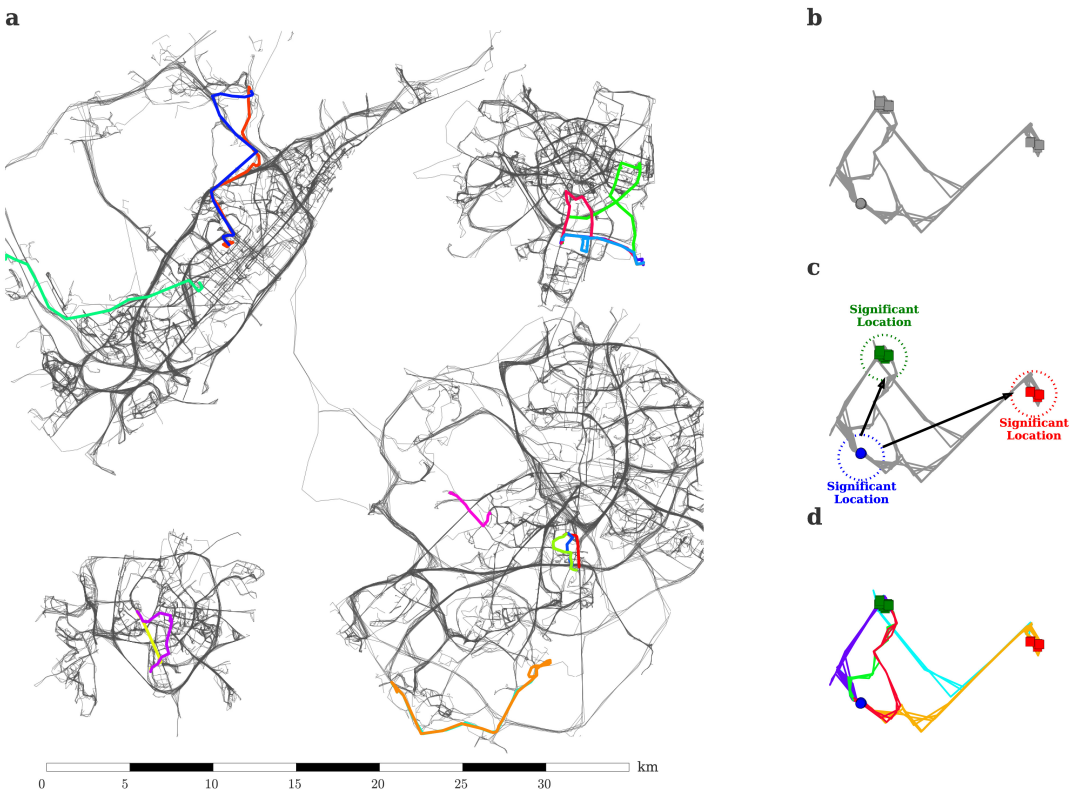


Figure 5.1: **From trajectories to route choices.** (a) A sample of the trajectories analysed from the four cities, shown in grey, outline their urban road networks. Coloured trajectories spanning between the same pair of points represent seven routine trips. In each routine trip, a coloured line represents a distinct route choice. (b) A set of trajectories belonging to a car. Each trajectory starts at the circle marker and ends at a square marker. (c) By clustering the endpoints of the trips three significant places are found. Two routine trips are shown with a solid black arrow. (d) The different route choices are finally discovered, for each routine trip performed by the driver. In this example one routine trip has three route choices (purple, green, red), the other has two (cyan, orange).

Each entry in the dataset contains: a unique anonymised user ID, a timestamp, a geographic location, encoded as latitude and longitude coordinates, the vehicle speed and a unique trip ID. Each trajectory is composed by periodic location updates, taken every 60 seconds, starting when the driver turns the engine on and ending when it is turned off. This means that the trips are already segmented depending on when the person starts and stops the engine.

I removed inconsistent data points, which are collected when the number of satellites available is lower than 4 or when the reported positions are inconsistent with average travel speeds higher than 110 km/h. I also removed trips that are too short in time (less

than 2 minutes) or space (less than 1 km). It is worth remarking that all user IDs were given in anonymised form.

I restrict this analysis to the 4 most popular cities in the country. By doing that, I focus on areas that have a reasonably high number of users and also a high density of roads, i.e., having several different routes to choose from. The filtered dataset contains information about the trajectories followed by 526 users.

5.1.2 Methodology

I describe a *trajectory* as a finite sequence of (t, x) tuples, where t represents a time value and x a location vector. The source and the destination of the trajectory are the first and last point of the sequence, respectively. A *significant place* is a geographic location that a person visits on a regular basis. In general, a significant place includes the neighbourhoods of the place the person is intending to go; in the case of car driving, it includes the region where the user might park, at walking distance to the destination. Several trips performed by a user from the same pair of significant places define together a *routine trip*. Finally, depending on how spatially similar are the trajectories of a routine trip, they can be grouped in one or more *route choices*. Fig. 5.1 shows some examples of route choices, highlighted on top of the four cities under consideration.

In order to analyse the routing behaviour of drivers, unstructured GPS trajectories must be organised into a structured set of significant places, routine trips and route choices [GNP+07; ZZ11]. The methodology used here and summarised in Fig. 5.1 follows these steps:

1. *Identifying significant places* by clustering the trajectories' endpoints. The output of this step is represented by the locations that are important for the driver, such as their home and work locations, and their favourite hang-outs.

2. *Grouping routes by their origin-destination into routine trips*, in terms of significant places (e.g., home to work, work to gym, etc.). Direction is taken into consideration

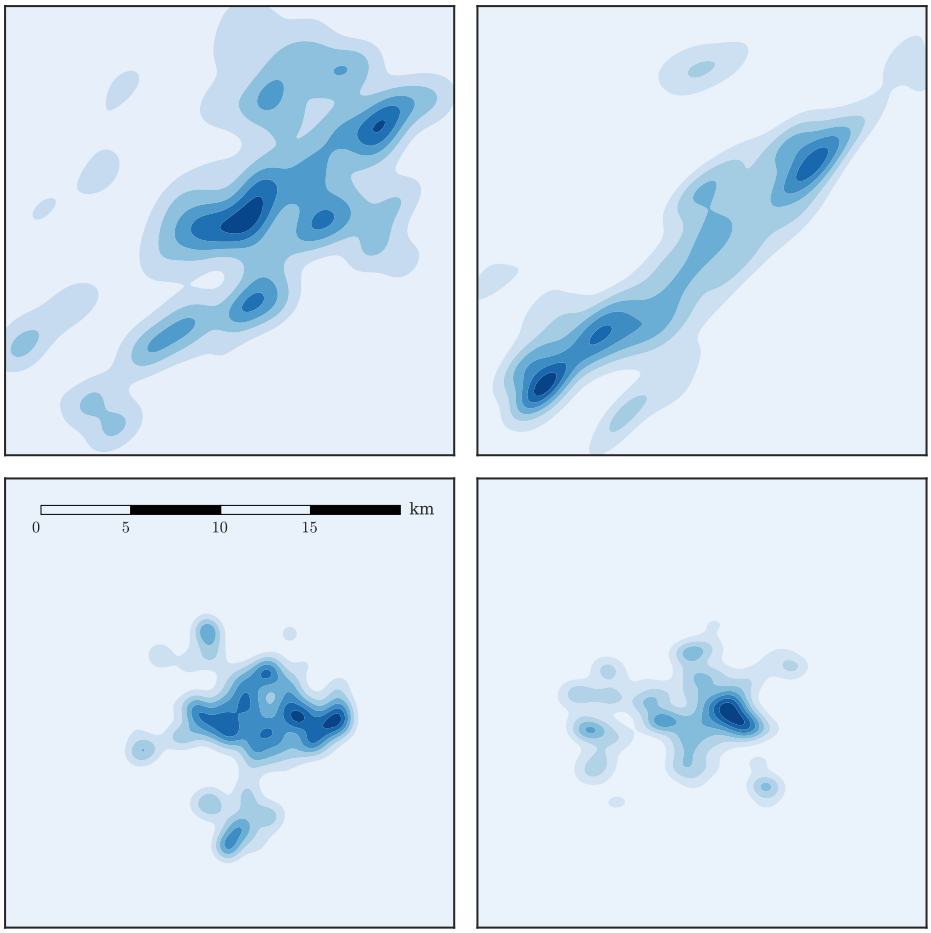


Figure 5.2: The spatial distributions of origin-destination points in the four cities. The scale is consistent across all four plots.

when identifying routine trips, as it might affect route choice (i.e., "A to B" is a different routine trip from "B to A").

3. *Clustering similar route choices* among trajectories belonging to the same routine trip. In order to compare trajectories that might be defined by a different number of points, Dynamic Time Warping distance [RCM+13] is used. Similar routes are then evaluated using a distance threshold, so that routes with minor detours, which might also be caused by sampling noise, are not counted as two distinct routes. More details are provided later on.

I firstly consider the starting and ending points of each trajectory; the spatial distribution of these points is shown in Fig. 5.2. A convenient clustering of these points will reveal

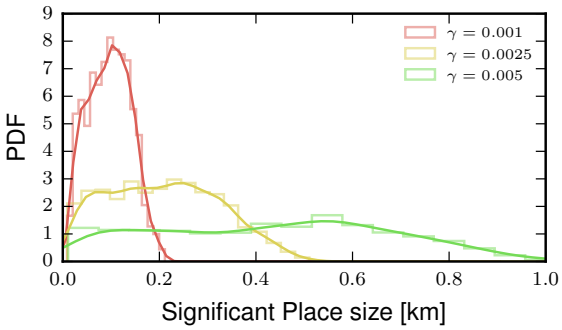


Figure 5.3: Probability density function of the significant place size, defined as the maximum distance between two points classified as belonging to the same significant place. As the bandwidth increases, the clusters of points defining significant places are larger.

the significant points of a person. Since the number of significant points of a person is not known in advance, it is appropriate to use a non-parametric clustering method, which does not need the number of clusters to be specified in advance. In particular, here I use the Mean Shift clustering algorithm [Che95] in conjunction with the Haversine distance, as it has already been successfully used to extract patterns from location data [KUB+11], and it also detects clusters that look reasonable on inspection. Mean Shift detects blobs of points with higher density and identifies their centroid. By choosing the bandwidth parameter $\gamma = 0.025$ I find clusters of points that are distant from each other at most by 600 m. These points can be reasonably different parking spots used to to reach the same final destination, located at walking distance. In Fig. 5.3 I show how the choice of γ influences the maximum distance between two points classified to be within the same significant place.

Once significant places are identified, I can group trajectories by their origin-destination, into a set of routine trips. Before moving on to the third step, I need to identify a method to compare trajectories. This task is not trivial because trajectories in this dataset are in general defined by an heterogeneous number of points. In this study I used the Dynamic Time Warping (DTW) algorithm, traditionally used in speech recognition and shape analysis. Given two paths $A = [a_1, a_2, ..., a_N]$ and $B = [b_1, b_2, ..., b_M]$, specified as sequences of geographic points of different length, this algorithm finds an alignment such that the

following recursive definition, for $i = 1 \dots N-1$, $j = 1 \dots M-1$, is minimised:

$$W(A_i, B_j) = d(a_i, b_j) + \min \begin{cases} W(A_{i+1}, B_{j+1}) \\ W(A_{i+1}, B_j) \\ W(A_i, B_{j+1}) \end{cases} \tag{5.1}$$

where A_i and B_j are subsequences containing all the elements $1 \dots i$ from A and $1 \dots j$ from B, respectively; the element-wise distance d is here considered to be the Haversine distance. The algorithm tries to match each point in A with a point in B, taking into consideration the sequence order. Initially, the two starting points are associated; then the algorithm advances one of the two trajectories, or both, depending on which pair of points minimises the element-wise distance; the algorithm proceeds until both end points are reached. Once the alignment is found, I consider the distance between the two trajectories to be the maximum distance between all the matched pairs of aligned points.

Now that I have defined a trajectory-distance, I can identify distinct route choices within a routine trip. To that end, I use the DBSCAN algorithm [EKS+96] on the maximum distance in the DTW-aligned trajectories. This clustering method has the advantage of not needing to specify the number of groups. However, it is necessary to choose two parameters: B, the minimum number of trajectories necessary to form a route, and ϵ, the maximum distance to consider an element part of the cluster. I set $B = 1$, so that a single different trajectory is considered as a distinct route choice. The best clustering results are obtained with a choice of $\epsilon = 0.5\,\text{km}$; such a value is reasonable, considering that a car travelling at an average speed of $30\,\text{km/h}$ covers that distance during the sample period of 60 seconds.

It is also worth mentioning that an alternative to the method used here is represented by map-matching combined with segment-by-segment comparison of routes. However, I decided not to use this option for several reasons. Firstly, map-matching requires full-knowledge of the urban network, making implementation and reproducibility of the results harder. Secondly, since map-matching performs best at higher sampling-rates, in this case

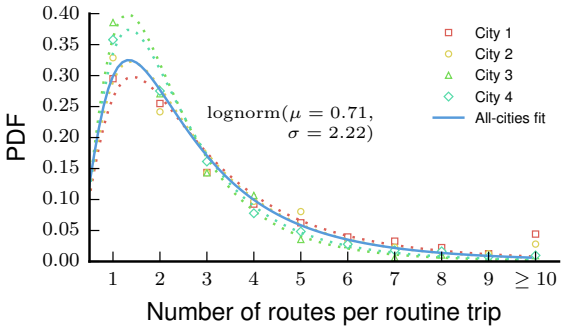


Figure 5.4: The distribution of the number of routes used for a routine trip. For most routine trips this number is low, despite the fact that these trips span over a period of up to 18 months. The markers show the empirical histograms about routine trips grouped by city. The solid curve shows the best lognormal fit, obtained on aggregated data generated in all four cities.

it will introduce additional undesired noise and bias. Finally and most importantly, I am not interested in achieving maximum precision: while the rate period of 60 seconds might seem high, the detour that a driver is able to make during this time period is quite limited, at most one block away ($\approx 150\,\mathrm{m}$) considering an average speed of $30\,\mathrm{km/h}$. Any smaller deviation is considered too small to be regarded as a different route choice.

5.1.3 Results

The first question I would like to answer is: how many different routes do drivers use in their routine trips? Fig. 5.4 shows the histogram of the number of routes used for each routine trip. The histograms are surprisingly similar among diverse cities. Independently of the urban settlement under consideration, individuals prefer to use a limited number of routes, and a third of them use only one route. This is a striking result, considering that these trips span over an 18-month period. I can safely conclude that users organise their routine trips only through a few preferred route options, where the number of choices follows a log-normal distribution with parameters $\mu = 0.71$ and $\sigma = 2.22$. The log-normal distribution, linked to a multiplicative random process, is ubiquitous in social science [LSA01] and has also been found in the distribution of single-mode distance

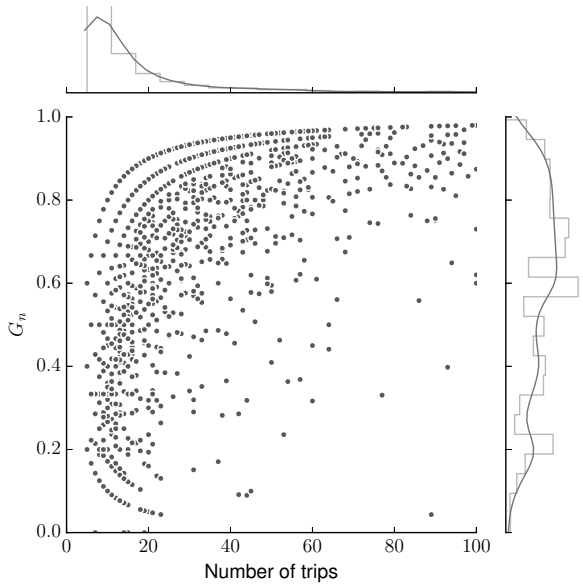


Figure 5.5: The number of trips performed during a routine journey versus the normalised Gini coefficient related to how many times each route choice is used. The two quantities show a weak correlation (Pearson $r = 0.48$, $p = 4.2e - 255$). The more a driver travels between two locations, the more likely it is for them to have a route of preference.

trips [KH03]. In this case, it may arise from the set of unknown random variables that determine individual route choices.

Next, for routine trips that have used more than one route, are some of them chosen more often than others? In order to answer this question, I use a normalised Gini coefficient G_n, corrected to have meaningful values when the number of routes is small. The Gini coefficient G is a statistical index of dispersion of values [Gin12], typically used in economics to quantify the inequality of income among people. Its value is bounded between $0 \leq G \leq 1 - \frac{1}{N}$, where N is the size of the population; the coefficient is null for perfect equality and maximum for complete inequality. I use the Gini coefficient to quantify, for a routine trip, how similar the usage frequencies are among all the routes employed at least once by the user. In order to compare this index on routine trips with a heterogeneous number of routes N, which is typically small, I consider a variant of the Gini coefficient, normalised by the maximum value of the Gini index G, obtainable with a number N of routes:

$$G_n = \frac{G}{1 - \frac{1}{N}} \tag{5.2}$$

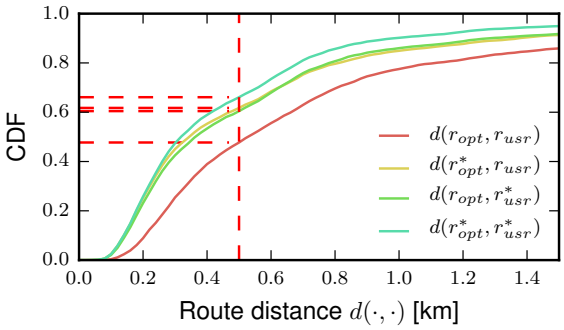
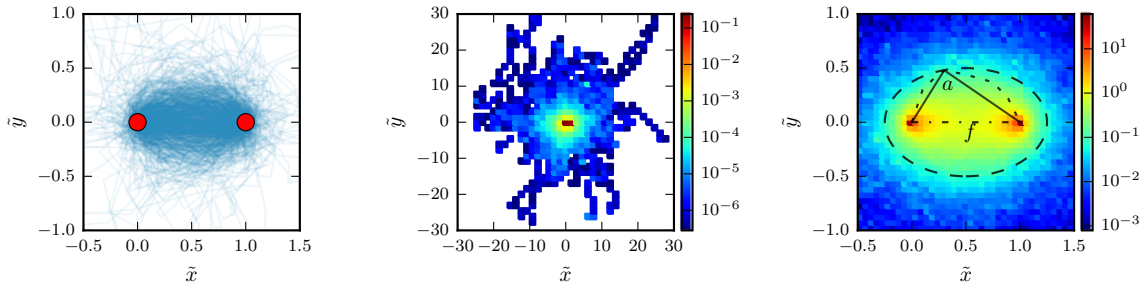


Figure 5.6: Maximum point distance between the optimal route r_{opt}, as suggested by the online routing service, and the favourite user route r_{usr}. For the other three curves I consider all the alternative routes returned by the service and all the routes ever used by the driver, choosing for each element the route that deviates the least from its counterpart, respectively r^*_{opt} and r^*_{usr}. Noticeably, 34% of the routes chosen are not any of the shortest paths and over 53% of the preferred routes are not optimal.

A value close to 0 (maximum equality) suggests that routes are evenly used. A value close to 1 (maximum inequality), suggests that the user is strongly biased towards one route for that routine trip, and that the alternate ways have been used seldom.

In Fig. 5.5 I plot, for routine trips that have at least two route choices, the normalised Gini, computed on the number of times that the route has been used. In general, routine trips have high values of the Gini coefficient with a median value of 0.6, suggesting that people tend to have a dominant route. Moreover, a mild correlation between the Gini values and the number of trips made suggests an adaptation process: when an individual repeats a journey more than 20 times, a preferred route tends to dominate their route choices. By contrast, I find both the number of routes and G_n to be uncorrelated with the most common time and day of the week of the routine trip.

Finally, what are the characteristics of a dominant route? Previous research assume that drivers prefer routes minimising some cost function, directly connected to travel time, fuel consumption or distance. I compare the routes taken by the user with the routes suggested by a popular online routing service. The service provides up to three alternative routes, accounting for expected travel times and traffic conditions. In order to compare these recommended optimal routes to the routes actually chosen, I measure the



(a) (b) (c)

Figure 5.7: **The boundaries of human routes.** Coordinates are projected to a Cartesian coordinate system using the spatial reference system EPSG:2062. Each trajectory (x, y) is roto-translated and scaled into $(\tilde{x}, \tilde{y})$ so that the source and destination of each trip are $(0, 0)$ and $(1, 0)$, respectively. (a) A 1% random sample of the trips, shown as partially transparent lines connecting consecutive $(\tilde{x}, \tilde{y})$ positions. (b) The probability density function $\Phi(\tilde{x}, \tilde{y})$ of the trajectory positions during a journey. Significant detours in all directions are uncommon but not unheard of. (c) 95% of the positions are within the elliptical region shown in the figure. The figure shows a sample trajectory, as a dotted line, the ellipse that fully contains it, as a dashed line, the focal distance as a dash-dot line and the major axis as a solid line.

maximum distance between GPS positions of a user's routine trip and the recommended path. In Fig. 5.6 I show the distribution of these distances, in four cases: when comparing only the top optimal route to the most used route; when comparing the optimal route to all a user's routes; when comparing the three suggested optimal routes to the dominant user route; and, finally, when comparing all suggested routes to all a user's routes. In the last three cases only the pairs of routes that deviate the least are considered. In about 53% of the cases, the dominant route chosen by the user is not the first optimal choice. For about 34% of the user routines none of the routes are compatible with the optimal choices, indicating that preferred routes do not minimise the travel cost. A previous study at a smaller scale had also found similar results that reject the shortest-path assumption [ZL15].

Next, my goal is to determine how far from the ideal route individuals are willing to go while undertaking their trip. I transform trajectories to a common reference frame of coordinates for all trajectories, so that each trajectory spans the same two points, regardless of the original actual trip length. The goal is to see how paths unfold and

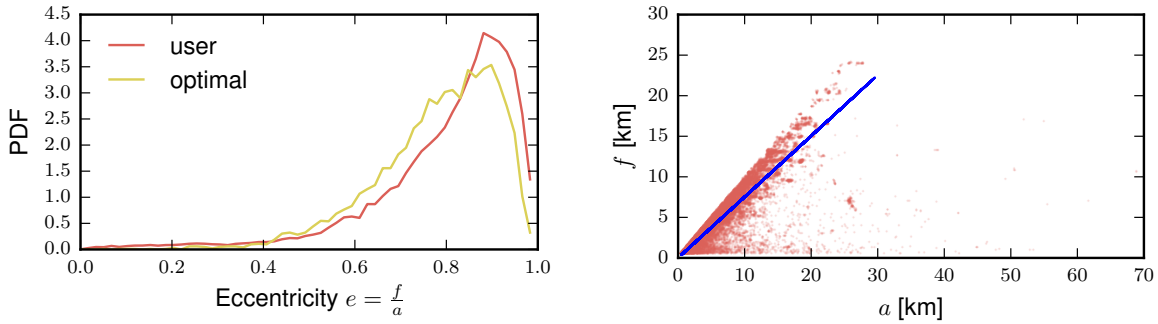
how far they usually go from their endpoints, regardless of their geographic position and of the trip length. Fig.5.7a shows a sample of the resulting trajectories. The figure vaguely resembles the field lines of a magnetic field and most lines appear to be fully contained in an elliptic region. I then proceed to study the probability density function $\Phi(x, y)$ of the route locations, normalised with respect to the source and the destination as explained. Fig. 5.7b shows how most of the deviations are small with respect to the source-destination endpoints. In particular, the majority of the positions recorded are contained within an area of elliptic shape, having as the two foci the first and last point of the trip, as Fig. 5.7c reveals. This result suggests that, while individuals commonly take detours due to personal preferences or characteristics of the street network [Bar11; PCL06], these detours are well bounded. The emergence of an elliptical shape is not surprising. Keeping in mind that an ellipse is the locus of the points P such that the sum of the distances to the two focal points F_1, F_2 is constant ($d(F_1, P) + d(F_2, P) = a$), this result shows the detour that people are willing to take is bounded. Trips that require larger detours are rare: they are unlikely to be undertaken, or they might be split into two distinct trips.

In order to further investigate this hypothesis and formally quantify the detours, I calculate two quantities for each trip: the geodesic distance between source and destination f; and a, the maximum value of the sum of the distance to the source and to the destination from any points along the path taken by the user. Finding these values is equivalent to identifying an idealised ellipse that fully contains all the paths taken by the driver. The eccentricity of the ellipse $e = f/a$ indicates how far from the geodesic this path goes. In the unlikely case where the endpoints lie on the same straight street and the driver takes the shortest route, $f = a$, the eccentricity takes the maximal value of 1, and the ellipse degenerates into a straight line. At the other extreme, a value of eccentricity close to 0 indicates that the path taken is very far from the endpoints, the ellipse tends to look like a circle in the target space and the two endpoints are close to each other compared to the path taken by the driver while moving between them.

Generally, the straight route is not a viable option, because of physical obstacles. Drivers deviate from that idealised shortest path according to the underlying street network and personal routing preferences. While these two phenomena are hard to treat, I find that routing detours are well approximated by an ellipse with high values of eccentricity (Fig. 5.8). This confirms the earlier observation that large deviations are rare; I speculate they are caused by intermediate destinations that the driver intends to reach before the final destination (e.g., giving a ride to somebody and dropping them off). Interestingly, the value of the eccentricity does not change considerably with distance between the endpoints (Fig. 5.9), suggesting that, in an urban setting, the space of the routing alternatives is proportional to the effective distance travelled. Whether this result also holds for trips at longer distances, such as inter-city journeys, is to be investigated in future analyses.

It is worth mentioning that ellipses have been previously used to understand the spatial extent of activity spaces and chained trips [BK06; NWS98; SA03]. To the best of my knowledge, instead, this is the first work that uses ellipses to quantify detours of single non-chained trips.

The results I have discussed here can be seen as a set of behavioural rules that capture individual behaviour in an urban environment. They are independent of the urban layout and were obtained by methods that are agnostic of the underlying street network. The rules establish the basic ingredients of realistic route-choice models. Once a travel plan is established for a user, a dominant route must be assigned. This choice should be spatially bounded within an elliptic shape of high eccentricity, as observed in the experimental distribution, opportunely scaled so that the origin and the destination are the foci of the ellipse. Although the choice can be driven by a distance/cost function from the main axis of the ellipse, it does not have to be deterministically chosen as the path that minimises a travel cost because, as I have explained, this does not typically reflect personal routing choices. Finally, individuals could choose alternate routes, within the ellipse, with probability inversely proportional to how often the person travels between the endpoints.



(a) (b)

Figure 5.8: (a) The probability density function of the eccentricity of the ellipse containing each trip, shown in red, and for comparison, the same quantity measured for the optimal trips, shown in yellow. While both groups of trajectories are characterised by high-eccentricity, optimal trips are slightly less eccentric than actual user trips, suggesting that the former deviate slightly more from the ideal origin-destination straight line. (b) The scatter plot of the endpoints distance f and the maximum sum-distance to the endpoints, calculated along the trajectory, a. For guidance only, the plot shows the best linear fit that goes through the origin ($f(x) = \alpha x$), with $\alpha = 0.75$.

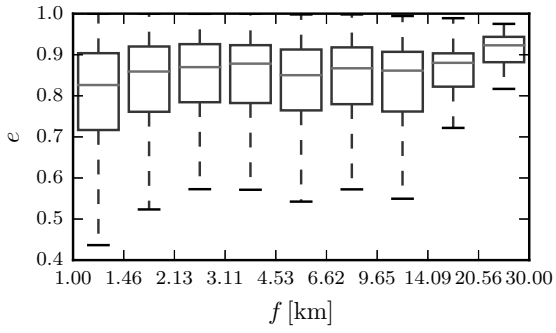


Figure 5.9: The boxplot shows the values of eccentricity e against the scale of the journey f. The median eccentricity value does not change considerably at different scales.

While in this section I have looked at how to make sense of fine-grained location data, structuring in routine trips and route choices, "zooming-out" from the particular to the general, in the following section I will focus on a process that might seem inverse, "zooming-in" from coarse-grained cellular network associations to more precise trajectories.

5.2 Estimating accurate paths from mobile data

In the last section I showed how the analysis of GPS data can uncover salient aspects of human routing behaviour. The goal of this section is to demonstrate it can also be used to increase data quality of datasets that have coarser temporal and spatial resolution.

As I described in Chapters 2 and 4 of this thesis, mobile network data is spatially coarse and temporally sparse. However, it has the big advantage of describing the mobility of much larger groups of individuals, frequently at national level, since mobile phones have a much higher penetration than navigation devices. A methodology that increases the spatial accuracy of mobile network data might make it possible to obtain mobility that is both very representative of the population and also accurate. This information might be very useful, for example to estimate travel demand and route preferences, as an alternative or complement to travel surveys [SG07]. Moreover, such methodology will not only be useful for data analysis purposes, but might also be used by service providers to provide positioning services to customers who do not use GPS, because of higher hardware cost or high battery consumption.

Contrary to common belief, in normal operation a cellular network cannot accurately localise every subscriber. Most cellular infrastructure keeps track of the tower associated to a mobile phone only during a call, when sending / receiving a text message, when switching tower, or while transmitting network data (e.g., while browsing the web). Most importantly, the signal strength of the tower-device connection is not typically logged by the infrastructure. While having access to the signal strength would make positioning

feasible through triangulation, logging this information would be a major departure from current operational practices, therefore here this possibility will not be considered. This means that popular localisation methods such as signal finger-printing [NBK10] or signal attenuation models [Ben07] cannot be used for this task. Therefore, this problem is significantly more challenging due to the sparse nature of the data under consideration.

The problem of estimating mobile paths from operational cellular data can be formalised as follows. The association of a mobile device m (also called in the following *subscriber*) to a network tower can be described in time by the sequence:

$$C_m = [(t_m[0], S_m[0]), (t_m[1], S_m[1]), \dots] \tag{5.3}$$

where each element $S_m[i]$ represents the tower sector the mobile m is associated with at the time $t_m[i]$. A graphical representation of such a sequence of observed sectors is shown on Fig. 5.10a. This sequence represents what can be collected from the cellular infrastructure as part of normal operation. Cellular networks do not presently log data at this granularity because of its significant volume over time, but it is certainly feasible to do so if it proves useful.

The sequence C_m is a proxy for the true location of the mobile m over time. A path estimation method that makes use of it would need to be tested against ground truth, ideally the true location of the subscriber. While ideal ground truth is continuous and accurate, in practice it must be collected at discrete times, using another localisation method; it can be described by the sequence:

$$L_m = [(t_m[0], l_m[0]), (t_m[1], l_m[1]), \dots] \tag{5.4}$$

where $l_m[i]$ is the location of the mobile device m at the time $t_m[i]$. Data coming from GPS sensors will be used as ground truth because they are the most accurate sensors available in this study.

The goal of this investigation is *to use the cellular observations C_m for each subscriber m to construct an estimate of the path of the mobile device $\hat{P}_m$ that is close to its actual path represented by L_m*. Observe that the formulation of this problem is simplified for convenience. Both C_m and L_m are discrete trajectories that give a sequence of mobile positions *at given times*. Here the final output of the estimation will be a continuous path $\hat{P}_m$ of the subscriber, which describes the sequence of positions but *not the times* by which it progressed along the path.

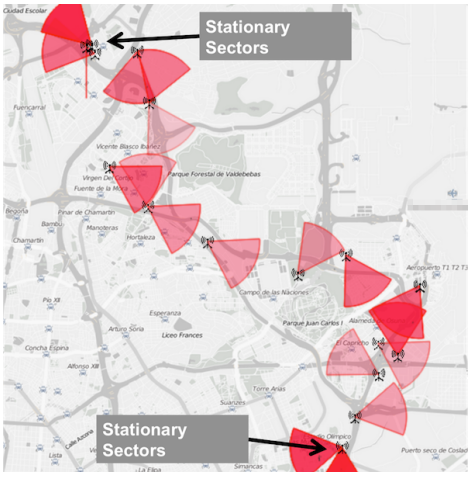
This simplification from trajectories to paths both meets typical path estimation needs, since many applications need the street-level path, and makes it more straightforward to assess whether the estimates are close to the ground truth, through a method that will be described in detail later.

There is much work on localisation [Ben07; NBK10], yet the problem explored here is significantly more difficult than traditional formulations because the input data is much more limited. The spatial granularity of the location information in cellular network events is at best that of a sector (*Base transceiver station*, or BTS), which is often quite large. In many cases the location is even less accurate as the mobile phone is associated to a nearby BTS but *not necessarily* the nearest one. Moreover, the temporal granularity of the location information heavily depends on subscriber activity in terms of either movement or use of the network. Arguably mobile subscribers use the network more or less continuously because of data services, but cell boundaries are carefully chosen to minimise the number of signalling messages for handovers, and inactive mobiles are paged only on the timescale of hours.

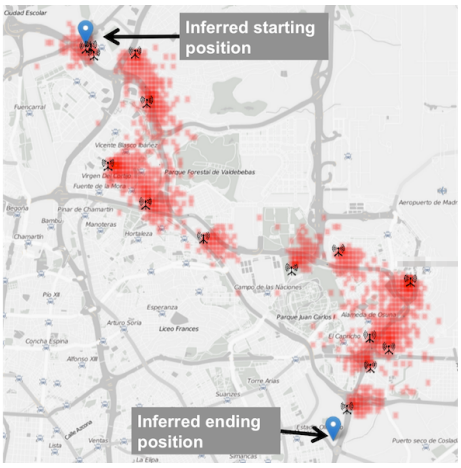
5.2.1 Solution

The path estimation procedure proposed here can be broken down into five steps.

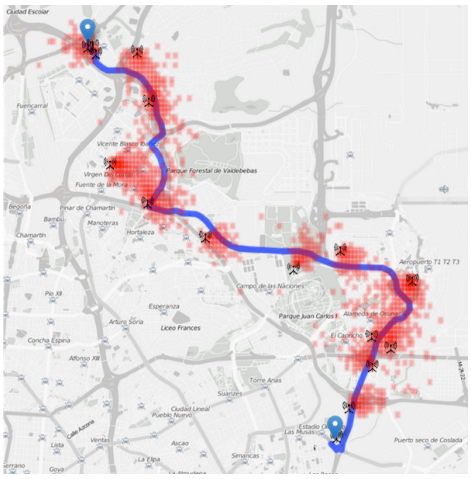
1. **Model the coverage area of the sectors.** First, it is necessary to understand where the subscriber is likely to be in space, given its association to a specific



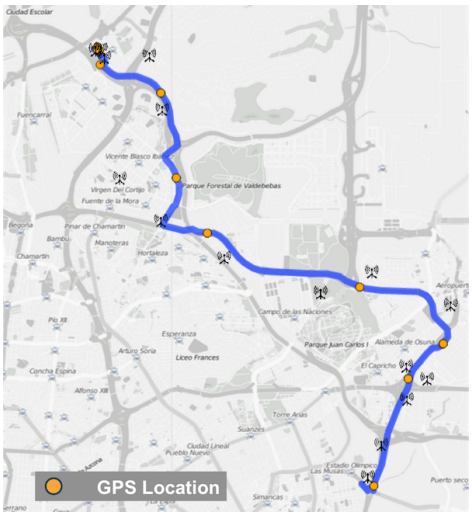
(a) Sequence of observed sectors.



(b) Endpoints extraction and intermediate spatial probabilities.



(c) Cell* through the weighted streets.



(d) Comparison with GPS ground-truth.

Figure 5.10: Steps of the Cell* path estimation method.

cell sector S. The goal of this step is to learn a spatial probability distribution which depends on characteristics of the antenna sector (i.e., orientation, beam width, spatial range). For each type of sector, a spatial probability distribution is built, using the position of the subscriber relative to the antenna location and orientation. Note that the expected coverage is not built for each specific tower, but for towers that share the same characteristics (e.g., femto, pico, macro cell, ...). The coverage models can be combined to identify *the overlap* between two sectors S_i and S_j, thus detecting likely handover locations. An example of a set of observed sectors C_m is

shown in Fig. 5.10a, while Fig. 5.12 shows two examples of coverage distributions of macro-sectored towers with different beam widths.

2. **Identify stationary and mobile segments.** In order to understand whether the subscriber is in a stationary or mobile state, C_m is partitioned into *stationary* and *mobile* sub-sequences or periods. In this work a stationary period is defined as a sequence of sector associations of a mobile device while it is in the same logical location (e.g., a building) for more than τ=15 minutes. A mobile sequence is defined as the sequence between two stationary sequences (Fig. 5.10b). Separating stationary and mobile segments is not trivial because devices may typically connect to multiple towers even when they are stationary. However, the coverage maps make it possible to understand whether a handover is more consistent with a single underlying location or a moving trajectory.

3. **Estimate location for a stationary sequence.** Each path is a movement between two stationary positions. The goal of this step is to estimate the stationary positions, which are used as the endpoints of each path. During each stationary sequence (derived from step 2), the device can connect to a number of distinct sectors (the median in this study is 4 sectors). Both the set of observed sectors and the duration of each association is used to estimate an accurate location of the device. Fig. 5.10b shows the estimated endpoints (stationary periods) of a path. In this example, at each endpoint 3 unique sectors were available to estimate the stationary location of the device.

4. **Identify high-probability areas of the path.** For each mobile sub-sequence (given from step 2), a number of intermediate sectors are observed (in this study the median is 12 unique sectors) (Fig. 5.10a). The coverage of the observed sectors indicates the areas that the mobile device travelled through with high probability. An illustration of these inferred coverage areas of the intermediate sectors is shown

in Fig. 5.10b. These areas typically constrain the path of the mobile subscriber, as shown in Fig. 5.10c.

5. **Estimate the path with map information.** The previous steps estimate the starting and ending locations of the path of a mobile subscriber (step 3), plus high-probability intermediate areas (step 4). The results I have shown in the last section suggest humans often do not take the shortest path, but the high-eccentricity of their trajectories still suggests that there are definite spatial and travel-time constraints in the route they choose. The main idea of this step is to combine shortest-path route selection with observed spatial probabilities, inferred from sector associations. The result is a biased route selection that does not necessarily return shortest paths, unless that is compatible with sector associations. Fig. 5.10d shows an example of a correctly estimated route that is not the shortest path.

In the following, I detail each of the steps.

Modelling the coverage of sectors

The goal of this step is to estimate the probability of a mobile being at a particular position relative to the sector S it is connected to. For convenience, polar coordinates relative to the sector location and orientation are used here. The probability for sectors with the same power P is then a function of two parameters: the distance d between the mobile device and the cell tower and the angle factor ϕ of the mobile relative to the antenna beam orientation. The angle factor is defined as the ratio of the polar angle of the mobile relative to the major axis of the sector and the beam width of the sector. Thus $\phi = 0$ represents the major axis of the sector while $\phi = 1$ represents the nominal extent of the antenna beam.

The coverage map is then approximated from ground truth data of associations at different locations. Details about how this data is collected and cleaned are given in

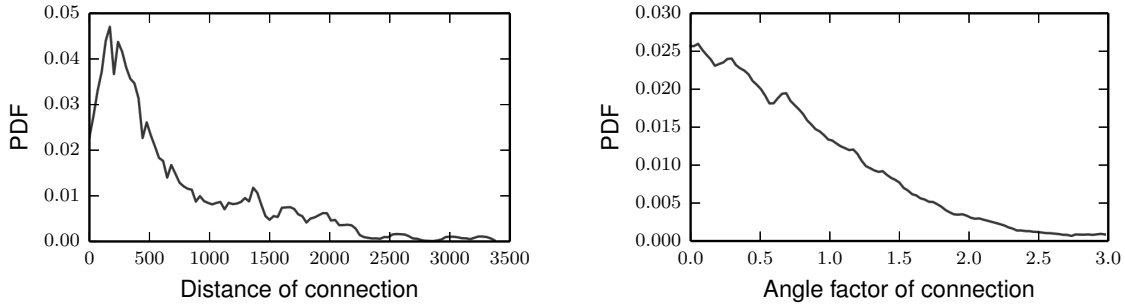


Figure 5.11: Projections of the coverage PDF for a macro-sectored tower on (a) distance of connection and (b) angle factor of connection.

Subsection 5.2.2. The result is spatial distributions denoted by $W_P(d, \phi)$, where P is the sector power.

To get a sense of coverage, the projection of $W_P(d, \phi)$ on both dimensions for a macro-sectored tower is shown in Fig. 5.11. It shows how the spatial distribution falls off roughly exponentially with distance d (left side) and more linearly with angle factor ϕ (right side). The coverage map is fixed for all the antennas of the same type, regardless of their location. Even though some local factors that affect radio-frequency (e.g., tall buildings, hills, etc) might also affect the spatial probability of the specific tower, this approximation is simple and works well in practice.

Fig. 5.12 shows the coverage probability density function (PDF) for sectors with beam widths of 60 and 120 degrees. For efficient computation, the space is discretised into a 15 m square grid. The spatial probability of being in square g while being connected to sector S is obtained by integrating the coverage PDF $W_P(d, \phi)$ over all points x in the square g.

$$Q(g, S) = \int_{x \in g} W_P(d_x, \phi_x).$$

$Q(g, S)$ are here referred to as the *probability grids*. The following steps will use these grids as the most convenient form of the coverage map.

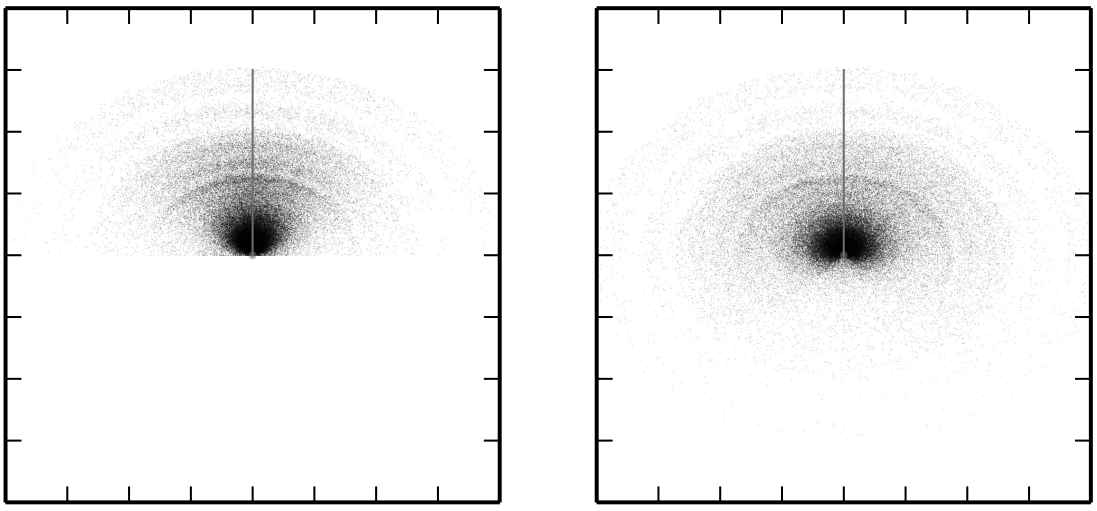


Figure 5.12: Coverage PDF for macro-sectored sectors having beam width of 60 degrees (left) and 120 degrees (right).

Identification of stationary and mobile segments

The goal of this step is to find all the stationary periods in C_m. Finding all stationary periods will also mean finding all the mobile periods.

Two sectors S_1 and S_2 are called *adjacent* if there is a square g in the grid such that $Q(g, S_1) > 0$ and $Q(g, S_2) > 0$, or, in other words, if a mobile can be in g and connect to either sectors, according to both their probability grids. Consequently, a subsequence $(t_m[i], S_m[i]), \ldots, (t_m[j], S_m[j])$ of the tower associations C_m is *stationary* if all the sectors $S_i, \ldots, S_j$ are *adjacent* to each other and also the time duration is greater than a threshold $t_j - t_i > \tau$. When this happens, the observed subsequence is compatible with a stationary device.

Maximal stationary subsequences define *stationary periods*. The subsequence between two stationary periods is a *mobile period*. The median number of connected sectors in a stationary period is 4. In contrast, during mobile periods the subscriber tends to associate to multiple non-adjacent sectors with frequent handovers. The median is 12 sectors per mobile period.

Now that the tower associations are partitioned into stationary and mobile periods, the method will proceed differently to estimate the location during each kind of period.

Estimation of stationary locations

If a stationary period consists of only one sector S, the most likely position is simply the square grid g that maximises $Q(g, S)$. In this case little can be done to improve the accuracy of the estimate. Instead, when the stationary period contains several sectors $S_i, \ldots, S_j$ then all of the sectors can be used to refine the location estimate. In particular, the location is estimated as the square grid g that maximises the sum of probabilities $Q(g, S_i)$ over the sectors weighted by the amount of time τ_i the subscriber was connected to the sector S_i, i.e.:

$$g^* = \arg\max_g \sum_{i=1}^{n} Q(g, S_i)\tau_i.$$

Identification of high-probability areas during moving periods

During an entire mobile period a subscriber's device typically connects to several sectors while traversing several areas. The goal of this step is to convert the association sequence into a set of scores, which gives higher values to areas that were most likely traversed during the trip. Recalling that probability grids $Q(g, S)$ describe the spatial probability density of locations from which a subscriber in the grid g is connected to sector S, the sum over all the sectors seen during a mobile segment $S_i, \ldots, S_j$ is here called *score* $\rho(g)$ of the mobile segment:

$$\rho(g) = \sum_{i=1}^{n} Q(g, S_i). \tag{5.5}$$

While this score is no longer a probability, the higher the score for a square g the more likely that the path of the subscriber passed through that square. Observe that summing the coverage PDFs also increases the weight on locations that overlap between sectors by "double counting" them. These are precisely the most likely handover locations. The score ρ is then used to bias the route search, as explained in the following section.

Evaluating a path from a mobile segment

Purely geometric approaches based on cell tower locations are unlikely to lead to good accuracy, as evaluated in Subsection 5.2.3. The location problem is simply too under-constrained. Since real paths are typically consistent with the underlying road infrastructure, using the road network to constrain the path between the stationary endpoints is likely to improve accuracy.

An interesting question is whether the use of shortest path routing on the street network is enough to correctly estimate the paths taken. If people simply followed the shortest (or fastest) path on the road network between two locations, it would be unnecessary to use information about which mobile towers the device was associated to during the journey. The results of Zhu and Levinson [ZL15] and those presented in *Section* 5.1 give supporting evidence that not all humans tend to follow short paths when travelling between two locations, but in fact they might use alternative routes, based on many factors such as personal preference, habit, knowledge of the street network, experience of traffic conditions, enjoyability of a given route, etc.

This suggests that a hybrid approach, which follows shortest path search, while biasing this search by the score $\rho(g)$ might lead to good accuracy. In fact, by using the scores, the search is preferentially constrained in a "corridor" formed by the coverage areas of all sectors in the mobile segment. High scores represent areas through which the route most likely passed.

The urban network of roads can be thought of as a graph $G = (\mathcal{V}, E)$, where nodes $\mathcal{V}$ represent the geographical locations of intersections or points along curved road segments, and edges represent a directional road-segment between two intersections. Information about intersections and roads is taken from OpenStreetMaps (OSM) [Hak10], a community-driven public repository that contains data about roads, points of interests (POIs), railway connections, etc. OSM is designed to be used for navigation purposes as it includes information for routing by many modes including driving, walking and cycling.

Here driving is assumed as the mode of transportation during the route search step; the problem of inferring the mode of transport is left for future work.

Initial edge weights. For each edge $e \in E$ an initial weight, $\mathcal{W}(e) \in \mathbb{R}$ is assigned that represents the *expected time* that is required to traverse it. This weight is available in the OSM data and depends on three factors: the length of the road segment, the type of the road segment (e.g., motorway, primary, secondary, footpath) and the mode of transport (e.g., walking, cycling, driving). Notice that for some means of transport certain roads might be restricted (e.g., motorways cannot be used for walking routes).

Modifying the weights based on mobile segment sectors $S_i, \ldots, S_j$. The default weights are then adjusted based on the scores of squares in the grid. Each road segment e crosses several square grids $g_j, \ldots, g_k$. In order to favour roads with high scores (i.e., through which the subscriber passed with high probability) the weight of the segment e is adjusted as follows:

$$\bar{W}(e) = \frac{W(e)}{\frac{1}{l}\sum_{i=j}^{k} \rho(g_i) + \epsilon \max \rho(g)}.$$

The use of ϵ allows the route to gracefully handle gaps in coverage due to low (zero) scores rather than require a major diversion. By tuning this parameter it is possible to balance the trade-off between the shortest routes and the routes that are close to the observed sectors. In this study as long as $\epsilon \leq 0.1$ the expected errors reported in the evaluation section are insensitive of the choice of ϵ.

Routing. After adjusting the weights, the standard A* algorithm is used on the adjusted weighted graph to search for the shortest path between the start and end points. For the path search, A* was preferred to Dijkstra because it was less computationally expensive. The resulting path in the graph is the final estimate for the path taken by the mobile subscriber, i.e $\hat{P}_m$. An example is shown in Fig. 5.10d.

5.2.2 Dataset

The algorithm was evaluated using a dataset that contains cellular-side observations and the corresponding ground truth for the paths taken by mobile subscribers. While the algorithm is designed to work solely with cellular-side data, in this small-scale evaluation both cellular data and location ground truth were collected using a smartphone. This dataset emulates the information that is available to the network operator (cellular-side data) and also collects the fine-grained position data, which can be used to evaluate the accuracy of the path estimation method.

A GPS location sample, including timestamp and geographic coordinates, is collected whenever there is any change to the *network-based* (WiFi or GSM trilateration) location of the device as reported by the smartphone operating system. In addition, location samples are also *periodically* collected every minute, unless the user is not continuously connected to the same Wi-Fi network and associated to the same cell (hence, assumed to be stationary).

A cellular network sample is generated every time there is a handover from one sector to another, every time there is a text/call and every time there is a data connection. Each sample includes the unique ID of the sector it is connected to, in the form of Cell ID (CID), Radio Network Controller ID (RNC) and Location Area Code (LAC). This information can be used to identify the tower in the database of the cellular operator. Other information that the phone might be able to collect, such as the signal-to-noise ratio, or the full list of towers available nearby, is not used in this study because it is not available to the cellular operator.

The full dataset includes information about 30 users for a period of 8 months. The users live in the proximity of different cities of the same country and they belong to different age groups (20-60). In total, the dataset contains more than 4.6 million samples over 19,000 hours of collection. A break-down of these samples is shown in Table 5.1. While this dataset represents a small number of people and on one hand is inevitably not representative of the general population, on the other it represents with a great level

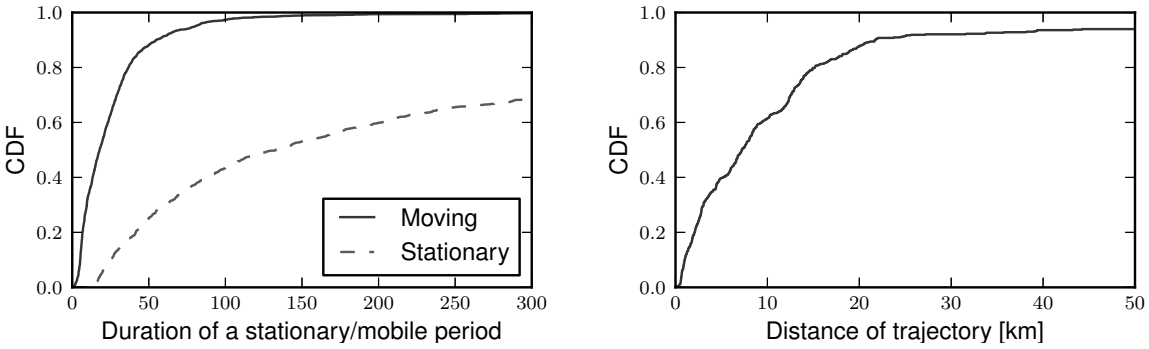
Type	Number
Total number of samples	4,669,077
Total number of network samples	1,718,504
Total number of handovers	433,031
Number of distinct sectors observed	15,455
Sectors in the operator's database	> 100,000
Number of location samples (netw. and GPS)	673,468
Number of GPS samples	259,032
Total time logged (all devices)	19,438 hours
Total time stationary	18,335 hours
Total time moving	1,102 hours
Number of trajectories (trips)	3,216
Total distance travelled	19.840 km

Table 5.1: The collected dataset. Each sample is a record of a sector association, handover, GPS/network location update, etc.

of detail their mobility patterns; it is hard to obtain such level of detail for much larger groups of people. As explained in Chapter 2, while GPS data are very spatially and temporally fine-grained, it is typically not possible to acquire them for large groups of people, because of practical constraints and privacy concerns.

The dataset was pre-processed as follows. The GPS sensor reports an estimate of its error L_{error}. Location measurements with reported error $L_{error} \geq 100\,\text{m}$, which account for 8% of the 673,468 location samples, are discarded.

Incomplete data encountered in the data collection (e.g., caused by battery depletion or system shutdown) is filtered out so that affected trajectories are removed. In the country where the data was collected, metro stations have connectivity through special sectors. Since metro travel is not the focus of this work, trajectories that go through metro segments are not considered. Consecutive GPS points are also merged using leader-based clustering, following a procedure similar to that described in [KUB+11]: each point is either associated to an existing cluster, if the distance to the median point of that cluster is below a chosen threshold $R_{cluster}$, or forms a new cluster. Using a $R_{cluster} = 100\,\text{m}$ produces reasonable clusters, and this value is also consistent with the L_{error} upper bound used to filter GPS entries.



(a) CDF of interval duration.

(b) Distance of a trajectory.

Figure 5.13: Ground-truth statistic.

After clustering locations, a user is determined to be stationary if staying in the cluster for a time interval longer than τ. This is similar to what was used for the estimation algorithm, and for consistency the same $\tau = 15$ minutes is used here. Similarly to what was defined before, users are mobile between consecutive stationary periods. This procedure identifies 3,216 distinct trajectories.

Fig. 5.13a shows the distribution of the stationary and mobile periods using ground-truth. Mobile trajectories are much shorter in duration than stationary periods. The median duration for a mobility segment is 23 minutes, and the majority of all recorded trajectories do not exceed one hour. Fig. 5.13b shows that most mobile trajectories are shorter than than 10 km. On the other hand, stationary periods have a median duration of 132 minutes, while 22% of them last for more than 8 hours. Such a result is expected because most stationary segments correspond to the user being at work during the day or at home during the night.

5.2.3 Evaluation

The dataset I described in the last section is then used to evaluate the quality of the estimated paths. In particular, the goal of the evaluation is to understand how closely the estimated path $\hat{P}_m$ matches the true paths of mobile subscribers. While the true path followed by the subscriber is not known, the best available estimate is the trajectory

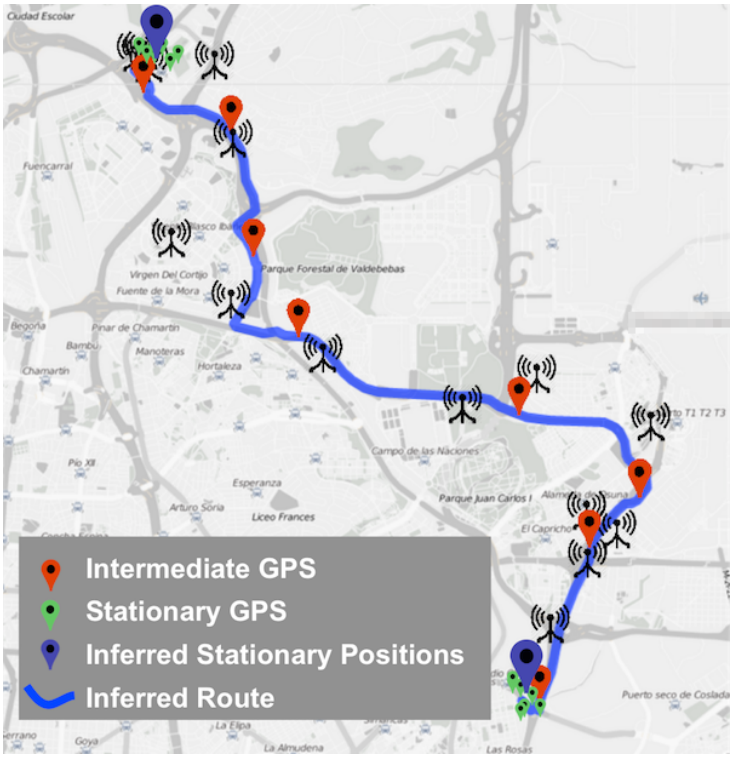
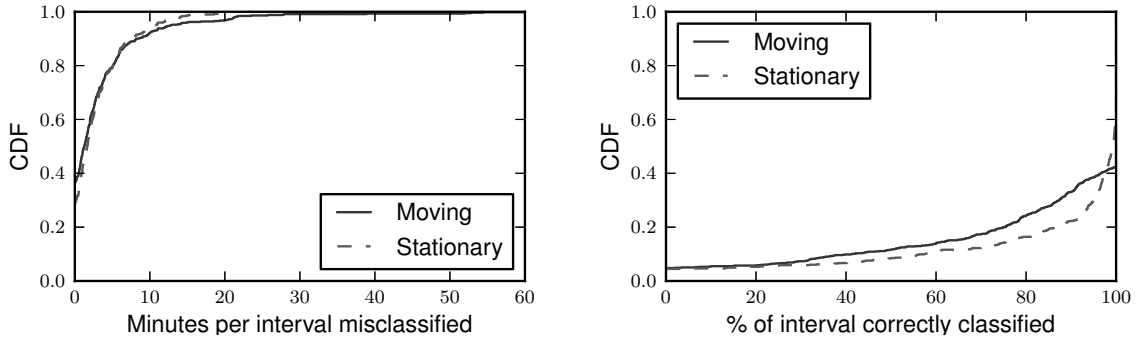


Figure 5.14: GPS ground-truth locations and estimated path.

inferred from GPS readings L_m, here used as ground truth. However, it is worth remarking that GPS samples are not true locations, and they are affected by an error L_{error} (median 19 m), which the sensor estimates depending on the signal quality. Even if the user is stationary the recorded location may slightly vary over time (especially when the user is indoors). This implies that, even if the estimated trajectory matches the true mobile path fully, a comparison with GPS-based ground truth will show a residual inaccuracy regardless.

Fig. 5.14 outlines the type of data that is available for evaluation, showing the GPS samples in green for stationary periods and red for mobile periods. The cellular trajectory C_m is denoted in the figure by the cell towers used for each handover. These observations are used to estimate the path of the subscriber $\hat{P}_m$, which is shown as a solid blue line. The accuracy of an estimated path is measured as the shortest distance between the GPS location and the estimated path. This quantity is denoted as e_i and it represents how close the path is to the GPS sample i. The sequence of errors e_i for an estimated path is denoted as E_m.



(a) CDF of time that was misclassified per period.

(b) Percentage of a period that was correctly classified.

Figure 5.15: Accuracy of detecting stationary/mobile periods.

Here I present the evaluation of the various steps of the estimating procedure. I begin by seeing how well *Cell** identifies the stationary and mobile segments of the trajectories. Then I compute the accuracy of the stationary locations compared to GPS locations. Finally, I look at the accuracy of the estimated paths.

To help understand why the steps in the proposed method lead to high accuracy, *Cell** is compared with two methods that estimate paths with less information: a *sector-only* method that incorporates the most likely location of the mobile while associated with the recorded sectors, but that does not use any GIS information (simply connecting the points); and a *routing-only* method that finds a route between the endpoints while ignoring all intermediate sector information. These comparisons suggest that a hybrid approach achieves higher accuracy. Finally, the accuracy of *Cell** is evaluated for different kinds of paths, in terms of speed and distance.

Accuracy of static and mobile segments

In Section 5.2.1, the sequence of sector observations C_m is partitioned into stationary and mobile periods, which then make it possible to identify trajectories. Here the ground-truth GPS trajectories L_m are used to evaluate the accuracy of this process. In particular, three quantities are measured: the duration of misclassified periods, the fraction of correctly classified periods and the number of trajectories that are not identified.

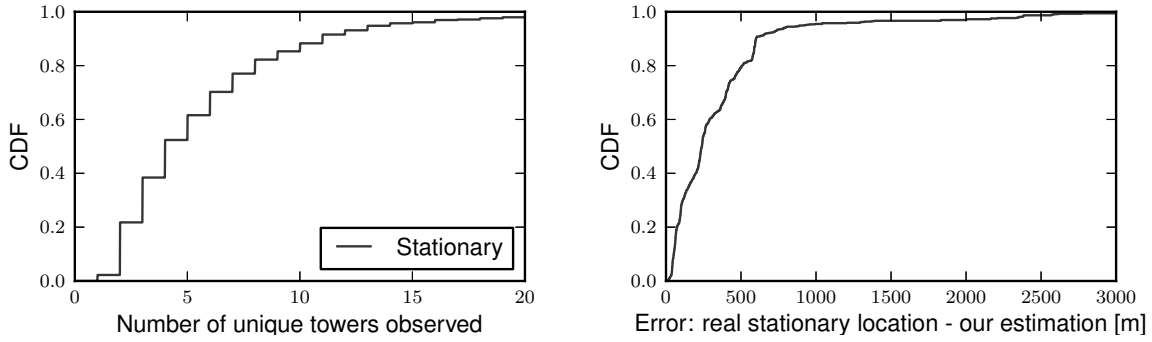
Fig. 5.15a shows the duration of misclassified periods, per ground-truth stationary/mobile period. The median error is 2 minutes and the 80th-percentile of errors is less than 5 minutes. These results are quite good given that estimation is performed on coarse-grained cellular-side data that only captures user movement when a handover occurs. As shown later, most of these errors occur during the beginning and the end of a mobile period when the user is still connected to the same sectors as in the stationary periods.

Fig. 5.15b shows which fraction of each period is correctly classified. A large fraction of stationary and mobile periods is correctly classified. 80% static periods and 70% mobile periods are correctly classified more than 90% of the time.

Finally, only 3.1% of the mobile trajectories are not identified at all by using cellular-side data. Further investigation showed that these are short trips within the same coverage area. This is caused by a limitation of this methodology, which cannot identify trips that do not trigger handovers.

Accuracy of stationary locations

Next, the accuracy of the stationary location estimation is evaluated. Stationary locations are important because they form the endpoints of the estimated path. In Section 5.2.1, I described how the algorithm uses multiple sector observations to increase the estimate accuracy during stationary position periods. Fig. 5.16a shows the distribution of the number of unique towers observed during a stationary period. In most of the cases (98%) a user is associated to more than 2 sectors while the median is 4 sectors. This suggests that combining multiple cellular observations can improve stationary location estimation in the vast majority of cases. The error in stationary location is simply defined as the geodesic distance between the estimated location and the spatial median of GPS points sampled during the stationary period. Fig. 5.16b shows the cumulative density function (CDF) for the values of this error. About 50% of the estimated locations have an error smaller than 230 m, and 80% of the estimated locations have an error smaller than 500 m. The accuracy is significantly higher than when considering the estimated location as the



(a) Number of unique sectors observed. (b) Error of estimated stationary location.

Figure 5.16: Accuracy of estimating stationary location.

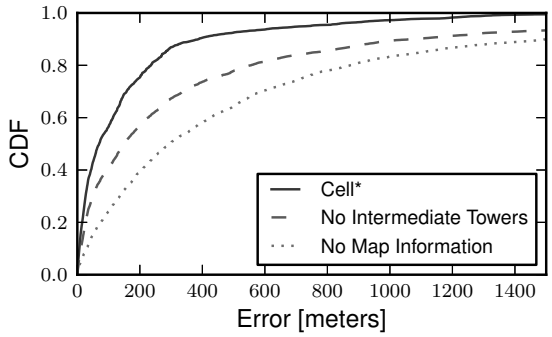


Figure 5.17: Error between GPS locations and estimated paths.

tower location, which has a median distance of 480 m from the ground truth location of the subscriber.

Accuracy of estimated paths

The accuracy of estimated paths is evaluated by measuring the error samples e_i, i.e., the shortest geodesic distance from each GPS location associated with a trajectory to the estimated path. Fig. 5.17 shows the distribution of the error samples across all trajectories. 75% of the errors fall below 180 m, and the median error is only 70 m. This positive result shows that *Cell** can accurately estimate paths given only the set of cellular handovers and without the assistance of GPS. For comparison, the median GPS accuracy for the same data is 19 m.

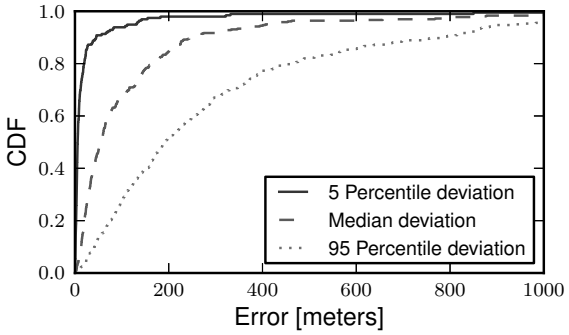
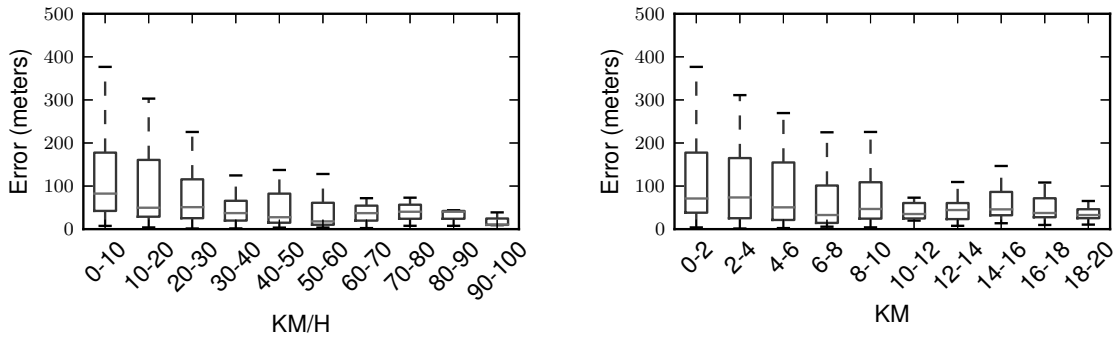


Figure 5.18: Per-path error statistics.

To understand how the different steps of this method contribute to the accuracy of paths, two comparisons are made. The first method, routing-only, ignores the sector associations and estimates the trajectory by running the route-search on the original urban network, with unbiased link weights. This process is equivalent to asking directions to any navigation system. Fig. 5.17 shows that, while the median error is doubled to 150 m, many of the errors are reasonably low. This implies that street-routing provides valuable semantic information for finding real paths. However, 10% of the errors have large deviations of more than 1 km. Routing is not sufficient by itself to estimate the path.

Second, to see the value of using the intermediate handovers alone, another comparison is made to a *sector-only* method, which simply connects the antenna locations with a straight line. The errors for this simple path are also shown in Fig. 5.17, which displays values larger than both *Cell** and *routing-only*. The median error is 298 m and in 10% of the deviations are greater than 2.5 km. This suggests that both routing and information about the cellular sectors is needed to estimate the path accurately.

Another evaluation is made on a per path basis. For each estimated path, the 5th, median, and 95th percentile of path error sequence E_m is shown in Fig. 5.18. 50% of the estimated paths have a median error of just 54 m, and a 95th percentile error of 200 m. Considering that the average block size of a city is approximately 200 m, the estimated paths are correct 95% of the time with uncertainty of one block.



(a) Path median error vs its speed.

(b) Path median error vs its distance.

Figure 5.19: Median error per route.

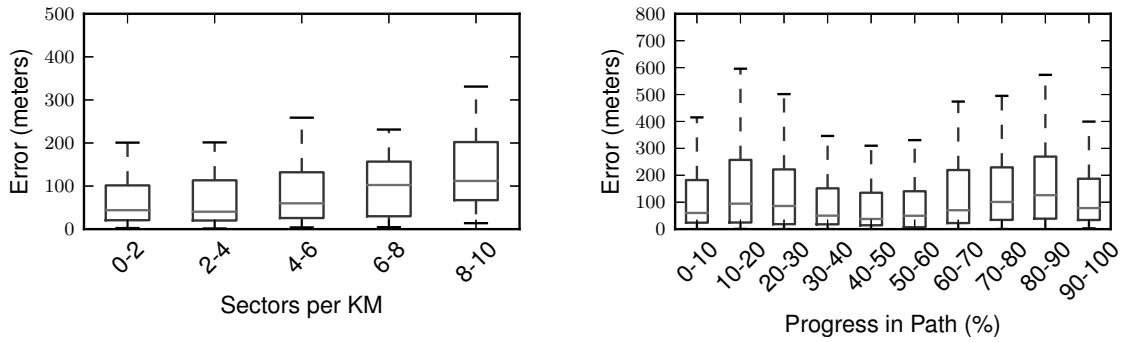
Accuracy by path properties

Finally, I show here how path properties affect estimation, making it more or less accurate. Speed and trip distance, for example, might affect accuracy, since they are expected to affect proportionally the number of handovers generated.

Fig. 5.19a shows the distribution of median errors per path when the paths are clustered according to their average speed. It is harder to infer low-speed paths, as these typically represent more chaotic walking routes in dense, urban environments. These routes might also require use of a different mode of transportation during route choice. Additionally, some of these routes might be very different from car trips. For instance, some of these routes do not have specific destinations, as is typical of leisurely strolls around the city. In contrast, high-speed paths are estimated with higher accuracy, probably due to the presence of highway segments that are easier to predict.

Similar behaviour is observed when looking into how distance affects error (Fig. 5.19b). Short path estimations are less accurate, as these include very few handovers and the route selection may deviate even further from the absolute shortest path.

It is also apparent that the errors in path estimation are impacted by the density of the cellular deployment in the vicinity of travel. For example, rural areas are expected to be covered with far fewer sectors than densely populated urban areas, and to cover a far greater distance with one sector. Indeed, the highest errors for stationary location



(a) Path median error vs density of deployment.

(b) Path median error vs stage within the route.

Figure 5.20: Median error per path (in metres).

estimation happen in rural areas. Fig. 5.20a shows the relationship between the path median error and the path sector density, defined as the number of sectors observed along the path divided by the path length. Contrary to what happens with stationary points, in areas with sparser tower installations (e.g., rural areas with less than 0-2 towers per km) the estimated paths are more accurate. This might happen because rural areas tend to have sparser road networks as well, constraining the road choice to fewer alternatives. Future work might explore the correlation between road density and cell infrastructure density.

Finally, Fig. 5.20b shows the accuracy of the estimated path as a function of the normalised elapsed time from the beginning of the path. Interestingly, accuracy is higher during the initial and final 10% of the route. This happens due to the fact that the estimation for the static points are very accurate, since they use information about several sectors. Furthermore, the accuracy is also good for the middle part of the trip as most people use primary roads (e.g., highways). Finally, accuracy is slightly lower while transitioning between these situations, likely due to less predictable secondary roads.

5.2.4 Discussion

In this section I demonstrated how data collected from cellular networks can be used to estimate paths taken by mobile subscribers. Cellular operators typically collect data

about user associations with base transceiver stations. This data is spatially coarse, temporally sparse and it is collected for billing and operational needs. It cannot be readily used to estimate subscriber paths at all times. The methodology I described here addresses this problem. Using a dataset containing more than 3,000 mobility trajectories, this methodology allows mobile paths to be estimated with a median error of 70 m. The use of this methodology might make it possible to provide location-based services also to devices that cannot use GPS sensors due to their cost or battery consumption. Finally, it might also be used to obtain real data about transportation and routing preferences, as an alternative or complement to user surveys.

5.3 Summary

In this chapter I have discussed how data coming from positioning services can be used to investigate drivers' routing behaviour and to estimate accurate travel paths from cellular network traces that are spatially coarse and temporally sparse.

In Section 5.1 I described how the analysis of GPS traces can lead to a set of rules that capture individual behaviour in an urban environment. The rules I presented are independent of the urban layout and were obtained by methods that are agnostic of the underlying street network. They establish the basic ingredients of a realistic route-choice models. Based on these rules, a viable route choice model might be designed as follows. Once a travel plan is established for a user, a dominant route must be assigned. This choice should be spatially bounded within an elliptic shape of high eccentricity, as observed in the experimental distribution, opportunely scaled so that the origin and the destination are the foci of the ellipse. Although the choice can be driven by a distance/cost function from the main axis of the ellipse, it does not have to be deterministically chosen as the path that minimises a travel cost, as we have seen this does not typically reflect personal routing choices. Finally, individuals could choose alternate routes, within the ellipse, with probability inversely proportional to how often the person travels between the endpoints.

In Section 5.1 I addressed a related problem of estimating routes from a less accurate source of data, such as mobile network data. The methodology that I described uses the probability of being at a certain distance and angle from the base station, while being connected, to model signal coverage for all the stations. Using this model in conjunction with shortest path routing on the urban network yields path estimates that are very accurate, down to one-block accuracy. This result can be used to provide location-based services for mobile devices without the use of GPS receivers and to extract more accurate user trajectories from mobile network datasets.

CHAPTER 6

CONCLUSIONS

The widespread popularity of mobile devices, internet connection and online social networks have deeply changed how people conduct their everyday lives. Every application used by an individual, from urban navigation to web browsing, from texting to tweeting, generates a vast deluge of data, as a result of how people interact with them.

This thesis was motivated by the fact that digital traces are not only useful in the original context they were produced and collected. On the contrary, the analysis of digital traces can be used to understand aspects of human behaviour, to build new services and improve existing systems.

This dissertation has explored this research space by analysing data coming from very different data sources (online social networks, cellular network, GPS), investigating several aspects of human behaviour (human mobility, interaction, collaboration) and modelling human behaviour in different areas (social media discussion, open-source collaboration, epidemic spreading and containment, urban routing behaviour and localisation).

6.1 Thesis summary and contributions

In Chapter 3 I firstly focussed on the analysis of online social networks data, which offers the opportunity to understand how people are connected to each to other by social links and how they move. I modelled the dissemination of a piece of information in the Twitter

network, noting that the dynamics is influenced by the social connections, the location of users and the external events. I then defined new measures that quantify a person's influence on a certain geographic area, discussing possible applications of such measures. Finally, I investigated social connections in GitHub, an online collaboration platform revolving around open-source projects. The study found that, while social ties show a long-tailed distribution like in other social networks, they also have specific properties, most notably they are much rarer, they have low reciprocity, and links are more strongly affected by distance.

I then continued exploring the effects of social interactions and mobility on epidemic spreading in Chapter 4, this time by analysing cellular network data. This data, compared to online social networks taken into consideration in the previous chapters, describes the location of a larger fraction of the population, usually with finer spatial and temporal granularity. The mobility and communication patterns of a whole country were used to inspire and inform a model where a disease spreads through mobility and immunisation spreads through communication, in the form of information about vaccination campaigns and prevention recommendations. I then revisited the goal of finding the most influential people according to their location, which I had discussed in Chapter 3, for the case of epidemic spreading. I devised a measure of risk that is based on the mobility patterns of an individual and the occurrence of a disease in a region. Through computer-based simulations based on these models, I demonstrated their potential effectiveness in containing a disease.

Finally, in Chapter 5, I analysed an even finer source of location data from people, namely positioning services (e.g., GPS devices). This data source was used for two different tasks, that is, to understand how people route between places and to increase the spatio-temporal accuracy of positioning derived from mobile network data. In particular, data collected for several months from hundreds of users in different cities showed how people do not take optimal paths when moving between the significant places that can be identified in their daily trajectories. Instead, they move in an ellipsoidal region that has

the source and destination points as foci. The unsuitability of pure shortest-paths also emerged during the evaluation of the second task: routing between source and destination leads to a poor path estimate. Instead, a combined approach that merges information coming from the street network and information coming from tower associations, managed to estimate paths down to one-block accuracy.

In summary, this thesis has provided evidence that digital traces coming from diverse sources, with different properties, spatial and time granularities, can be used to understand aspects of human behaviour and for practical applications. While in this thesis I used data from online social networks, cellular data and GPS devices, similar considerations might be found in new data sources as they emerge.

Although the studies presented in this thesis are quite varied in terms of the type of digital traces used and applications considered, a few common "lessons learnt" can be drawn from them. Most of the traces presented in this work were originally collected for different reasons (for example, mobile data traces were collected for billing, OSNs traces were collected to run the services themselves). This data was essentially collected for a different purpose. For this reason, it can be considered secondary data. The implications of this fact are double-edged: the availability of this data constitutes an exciting opportunity for researchers and practitioners in the area of computational social science, but at the same time, it is possible to identify ethical and privacy concerns, both for people who are working on them and for people who are the subjects represented in the data. Moreover, it is not clear what data released today might reveal at a later time, when more advanced algorithms and techniques might be developed.

Another lesson which was learnt through the preparation of this thesis relates to the validity and how each data source can be more or less suitable to various kinds of analysis. When working with digital traces, it is very important to understand the nature of the data, from which system it comes from, how it was collected, and how this affects the specific kind of analysis that is performed on the data themselves. For example, many studies that are based on mobile network data assume that every person is associated to

the closest tower [BLM+14; GHB08], but I have shown in Chapter 5 that it is not correct due to the mobile technology used for the selection of the tower. This assumption might be acceptable when evaluating the extent of human mobility at a coarser level; however, it might be inappropriate for detecting close proximity between two individuals, on the basis of the association to the same tower.

Finally, perhaps the most fascinating lesson learnt during the development of this thesis relates to how the analysis of human behaviour is always confronted with the duplicity of its nature, being at the same time individual and part of a collectivity. On one hand, as part of the society and social groups, people behave in a rather predictable way, along patterns that are quantifiable. This manifested itself in this thesis, for example, in how the models of information and disease dissemination treat each person in the same way and also in how the movements of people follow similar rules. On the other hand, though, seemingly paradoxically and in contrast with the previous statement, each individual is unique and manages to escape efforts of *strict* classification. This emerged, for example, in the identification of the most influential nodes in a spatial social network in Section 3.2, in the detection of superspreaders in an epidemic outbreak in Section 4.4, and in the analysis of the selection of sub-optimal routes by many individuals in Section 5.1. As if to say, at the same time: *we are all equal*, but also *everyone is unique*.

6.2 Future directions

This dissertation raises several interesting questions and opens up many possibilities for further investigations.

First, the fact that digital traces can be analysed to understand human behaviour is at the basis of many questions related to data ownership, i.e., about who has the rights to access and use this data. As the analysis of big datasets is fairly recent and quickly evolving, legislation across the world is lagging behind the new issues raised by the new possibilities that have been opened up [Sci16]. Such issues will be increasingly

important as technology becomes more pervasive and data flows become even bigger and more interrelated.

Moreover, the analysis and models I described in Chapter 3 could be used to design real systems that make use of them. Future work might look into using the models and measures I described for prediction and early detection of discussion trends, for identification of the key influencers in various spatial regions and to provide recommendations about project collaborations.

Further research is also needed to understand how viable is the implementation of disease containment techniques shown in Chapter 4, based on information dissemination through the social network and risk-assessment based on the mobility of people. While the results I described show that these strategies are effective, their implementation in the real world requires further research.

Finally, perhaps the most exciting direction for future research is represented by the possibility of combining multiple heterogeneous data sources, in a process that might be called *data fusion*. In fact, in Chapter 5 the combination of cellular data and data about the urban network was crucial to increase the spatio-temporal accuracy of cellular network data, hence reconstructing the trajectory followed by individuals from sparse and coarse-grained proxies of that information. This example shows how multiple data sources can be combined to increase data quality, both in terms of accuracy and validity.

Considering that our world is becoming increasingly characterised by the large availability of these new forms of data, often referring to and describing the same phenomena, there is a need for methods that fuse several data sources and are able to correct systematic and random errors. In the same way that error coding allows reliable delivery of digital data, data fusion techniques might make it possible to summarise data coming from multiple sources into a coherent output of data of minimal error and maximum information.

6.3 Outlook

Taken as a whole, this dissertation has attempted to make a step towards understanding human behaviour from digital datasets of varied nature, originally collected for different purposes from what they were used for in these studies. I hope that this thesis will be useful and inspiring to researchers and practitioners in the area. As digital traces become increasingly available in different domains of life, the possibilities for introspective analysis of human behaviour and behaviour-centred design of systems are expected to become more abundant. This projection opens up many new exciting opportunities for research and technology, but it also raises serious issues about personal rights, for example about data ownership and privacy issues [Pen09]. Some novel studies look at several ways to mitigate these problems, among which the assessment of how unique the traces are [RWM15; RWS+15], how to apply obfuscation techniques to them, and the use of personal data containers, which control access to the data [CCH+15; HDC14]. I hope that all future efforts in developing such opportunities will always be matched by consideration and respect towards personal rights.

LIST OF REFERENCES

[AAA+12] G. Aad, T. Abajyan, B. Abbott, J. Abdallah, S. Abdel Khalek, et al. "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC". *Physics Letters B* 716:1 (Sept. 17, 2012), pp. 1–29 (cit. on p. 32).

[ABL15] A. Acquisti, L. Brandimarte, and G. Loewenstein. "Privacy and human behavior in the age of information". *Science* 347:6221 (Jan. 30, 2015), pp. 509–514 (cit. on p. 61).

[AA05] E. Adar and L. A. Adamic. "Tracking Information Epidemics in Blogspace". In *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence*. WI '05. Compiègne, France. IEEE Computer Society, 2005, pp. 207–214 (cit. on pp. 25, 27).

[AJM+15] L. Alexander, S. Jiang, M. Murga, and M. C. González. "Origin–destination trips by purpose and time of day inferred from mobile phone data". *Transportation Research Part C: Emerging Technologies* (2015) (cit. on p. 22).

[Alt14] C. L. Althaus. "Estimating the Reproduction Number of Ebola Virus (EBOV) During the 2014 Outbreak in West Africa". *PLoS Current Outbreaks* (2014) (cit. on p. 106).

[ADB+05] J. Alvarez-Hamelin, L. Dall Asta, A. Barrat, and A. Vespignani. "Large scale networks fingerprinting and visualization using the k-core decomposition". In *Proceedings of Advances in Neural Information Processing Systems 18*. NIPS '05. Vancouver, BC, Canada, 2005, pp. 41–50 (cit. on pp. 42, 43).

[ADF91] U. Amaldi, W. De Boer, and H. Fürstenau. “Comparison of grand unified theories with electroweak and strong coupling constants measured at LEP”. *Physics Letters B* 260:3 (1991), pp. 447–455 (cit. on p. 30).

[AM91] R. M. Anderson and R. M. May. *Infectious diseases of humans: dynamics and control*. Oxford University Press, 1991 (cit. on p. 41).

[AS03] D. Ashbrook and T. Starner. “Using GPS to Learn Significant Locations and Predict Movement Across Multiple Users”. *Personal Ubiquitous Comput.* 7:5 (Oct. 2003), pp. 275–286 (cit. on p. 22).

[AL05] D. I. Ashby and P. A. Longley. “Geocomputation, Geodemographics and Resource Allocation for Local Policing”. *Transactions in GIS* 9:1 (2005), pp. 53–72 (cit. on p. 60).

[ASW+03] K. W. Axhausen, S. Schönfelder, J. Wolf, M. Oliveira, and U. Samaga. “80 weeks of GPS-traces: Approaches to enriching the trip information”. *Arbeitsbericht Verkehrs-und Raumplanung* 178 (2003) (cit. on p. 22).

[BBR+12] L. Backstrom, P. Boldi, M. Rosa, J. Ugander, and S. Vigna. “Four Degrees of Separation”. In *Proceedings of the 4th Annual ACM Web Science Conference*. WebSci ’12. Chicago, IL, USA. ACM, 2012, pp. 33–42 (cit. on pp. 2, 3, 23).

[BSM10] L. Backstrom, E. Sun, and C. Marlow. “Find me if you can: improving geographical prediction with social and spatial proximity”. In *Proceedings of the 19th International Conference on World Wide Web*. WWW ’10. Raleigh, NC, USA. ACM, 2010, pp. 61–70 (cit. on pp. 3, 29).

[BSU00] J. N. Bailenson, M. S. Shum, and D. H. Uttal. “The initial segment strategy: A heuristic for route selection”. *Memory & Cognition* 28:2 (2000), pp. 306–318 (cit. on p. 22).

[BPR+11] P. Bajardi, C. Poletto, J. J. Ramasco, M. Tizzoni, V. Colizza, and A. Vespignani. "Human Mobility Networks, Travel Restrictions, and the Global Spread of 2009 H1N1 Pandemic". *PLoS ONE* 6:1 (Jan. 31, 2011), e16591 (cit. on p. 104).

[BCG+09] D. Balcan, V. Colizza, B. Gonçalves, H. Hu, J. J. Ramasco, and A. Vespignani. "Multiscale mobility networks and the spatial spreading of infectious diseases". *Proceedings of the National Academy of Sciences* 106:51 (Dec. 22, 2009), pp. 21484–21489 (cit. on pp. 23, 26, 104–106).

[BGM07] S. Bansal, B. T. Grenfell, and L. A. Meyers. "When individual behaviour matters: homogeneous and network models in epidemiology". *Journal of The Royal Society Interface* 4:16 (Oct. 22, 2007), pp. 879–891 (cit. on p. 26).

[Bar05] A. Barabási. "The origin of bursts and heavy tails in human dynamics". *Nature* 435:7039 (2005), pp. 207–211 (cit. on p. 39).

[Bar11] M. Barthélemy. "Spatial networks". *Physics Reports* 499:1–3 (Feb. 2011), pp. 1–101 (cit. on p. 128).

[Bas69] F. M. Bass. "A New Product Growth for Model Consumer Durables". *Management Science* 15:5 (Jan. 1, 1969), pp. 215–227 (cit. on p. 25).

[BL12] F. Bauer and J. T. Lizier. "Identifying influential spreaders and efficiently estimating infection numbers in epidemic models: A walk counting approach". *Europhysics Letters* 99:6 (Sept. 1, 2012), p. 68007 (cit. on p. 23).

[BAB08] M. Beiró, J. Alvarez-Hamelin, and J. Busch. "A low complexity visualization tool that helps to perform complex systems analysis". *New Journal of Physics* 10:12 (2008), p. 125003 (cit. on pp. 42, 43).

[BS10] E. Ben-Elia and Y. Shiftan. "Which road do I take? A learning-based model of route-choice behavior with real-time information". *Transportation Research Part A: Policy and Practice* 44:4 (May 2010), pp. 249–264 (cit. on p. 22).

[Ben07] A. Bensky. *Wireless Positioning Technologies and Applications*. Artech House, Dec. 1, 2007. 317 pp. (cit. on pp. 2, 132, 133).

[BKH+12] C. A. Bliss, I. M. Kloumann, K. D. Harris, C. M. Danforth, and P. S. Dodds. "Twitter reciprocal reply networks exhibit assortativity with respect to happiness". *Journal of Computational Science*. Advanced Computing Solutions for Health Care and Medicine 3:5 (Sept. 2012), pp. 388–397 (cit. on p. 34).

[BDK15] V. D. Blondel, A. Decuyper, and G. Krings. "A survey of results on mobile phone datasets analysis" (Feb. 11, 2015). arXiv: 1502.03406 (cit. on p. 17).

[BEC+12] V. D. Blondel, M. Esch, C. Chan, F. Clerot, P. Deville, E. Huens, F. Morlot, Z. Smoreda, and C. Ziemlicki. "Data for Development: the D4D Challenge on Mobile Phone Data" (Sept. 29, 2012). arXiv: 1210.0137 (cit. on p. 84).

[BLM+06] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang. "Complex networks: Structure and dynamics". *Physics Reports* 424:4–5 (Feb. 2006), pp. 175–308 (cit. on p. 41).

[BPV03] M. Boguñá, R. Pastor-Satorras, and A. Vespignani. "Absence of Epidemic Threshold in Scale-Free Networks with Degree Correlations". *Physical Review Letters* 90:2 (Jan. 15, 2003), p. 028701 (cit. on pp. 42, 46, 49).

[BKL+09] U. Brandes, P. Kenis, J. Lerner, and D. van Raaij. "Network analysis of collaboration structure in Wikipedia". In *Proceedings of the 18th International Conference on World Wide Web*. WWW '09. Madrid, Spain. ACM, 2009, pp. 731–740 (cit. on p. 27).

[BHG06] D. Brockmann, L. Hufnagel, and T. Geisel. "The scaling laws of human travel". *Nature* 439:7075 (Jan. 26, 2006), pp. 462–465 (cit. on p. 21).

[BLM+14] C. Brown, N. Lathia, C. Mascolo, A. Noulas, and V. Blondel. "Group Colocation Behavior in Technological Social Networks". *PLOS ONE* 9:8 (Aug. 22, 2014), e105816 (cit. on p. 158).

[BAM+12] T. T. Brunyé, E. Andonova, C. Meneghetti, M. L. Noordzij, F. Pazzaglia, R. Wienemann, C. R. Mahoney, and H. A. Taylor. "Planning routes around the world: International evidence for southern route preferences". *Journal of Environmental Psychology* 32:4 (Dec. 2012), pp. 297–304 (cit. on p. 22).

[BK06] R. N. Buliung and P. S. Kanaroglou. "Urban Form and Household Activity-Travel Behavior". *Growth and Change* 37:2 (June 1, 2006), pp. 172–199 (cit. on p. 129).

[But14] D. Butler. "Models overestimate Ebola cases". *Nature* 515:7525 (Nov. 4, 2014), pp. 18–18 (cit. on p. 112).

[CMP+79] N. Cabibbo, L. Maiani, G. Parisi, and R. Petronzio. "Bounds on the fermions and Higgs boson masses in grand unified theories". *Nuclear Physics B* 158:2 (1979), pp. 295–305 (cit. on p. 30).

[Cai01] F. Cairncross. *The Death of Distance: How the Communications Revolution is Changing Our Lives*. Harvard Business Press, 2001. 342 pp. (cit. on p. 2).

[Cen12] Central Intelligence Agency. *The World Factbook*. Cote d'Ivoire. 2012. (Visited on 11/01/2012) (cit. on p. 93).

[CHB+10] M. Cha, H. Haddadi, F. Benevenuto, and P. K. Gummadi. "Measuring User Influence in Twitter: The Million Follower Fallacy." In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*. ICWSM '10. 2010, pp. 10–17 (cit. on pp. 17, 51).

[CMG09] M. Cha, A. Mislove, and K. P. Gummadi. "A Measurement-driven Analysis of Information Propagation in the Flickr Social Network". In *Proceedings of the 18th International Conference on World Wide Web*. WWW '09. Madrid, Spain. ACM, 2009, pp. 721–730 (cit. on pp. 25, 70).

[CKS+12] S. Chatrchyan, V. Khachatryan, A. M. Sirunyan, A. Tumasyan, W. Adam, et al. "Observation of a new boson at a mass of 125 GeV with the CMS

experiment at the LHC". *Physics Letters B* 716:1 (Sept. 17, 2012), pp. 30–61 (cit. on p. 32).

[CCH+15] A. Chaudhry, J. Crowcroft, H. Howard, A. Madhavapeddy, R. Mortier, H. Haddadi, and D. McAuley. "Personal Data: Thinking Inside the Box". *Aarhus Series on Human Centered Computing* 1:1 (Oct. 2015), p. 4 (cit. on p. 160).

[Che95] Y. Cheng. "Mean shift, mode seeking, and clustering". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 17:8 (Aug. 1995), pp. 790–799 (cit. on p. 122).

[CML11] E. Cho, S. A. Myers, and J. Leskovec. "Friendship and Mobility: User Movement in Location-based Social Networks". In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '11. San Diego, CA, USA. ACM, 2011, pp. 1082–1090 (cit. on p. 29).

[CP03] T. Choudhury and A. Pentland. "Sensing and Modeling Human Networks Using the Sociometer". In *Proceedings of the 7th IEEE International Symposium on Wearable Computers*. ISWC '03. White Plains, NY, USA. IEEE Computer Society, 2003, p. 216 (cit. on p. 24).

[CF07] N. A. Christakis and J. H. Fowler. "The Spread of Obesity in a Large Social Network over 32 Years". *New England Journal of Medicine* 357:4 (2007), pp. 370–379 (cit. on p. 52).

[CMF+14] B. Cici, A. Markopoulou, E. Frias-Martinez, and N. Laoutaris. "Assessing the Potential of Ride-sharing Using Mobile and Social Data: A Tale of Four Cities". In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. UbiComp '14. Seattle, WA, US. ACM, Sept. 2014, pp. 201–211 (cit. on p. 3).

[CSN09] A. Clauset, C. R. Shalizi, and M. E. Newman. “Power-law distributions in empirical data”. *SIAM review* 51:4 (2009), pp. 661–703 (cit. on p. 73).

[Cob54] A. Cobham. “Priority assignment in waiting line problems”. *Journal of the Operations Research Society of America* 2:1 (1954), pp. 70–76 (cit. on p. 39).

[Coh03] J. E. Cohen. “Human Population: The Next Half Century”. *Science* 302:5648 (Nov. 14, 2003), pp. 1172–1175 (cit. on p. 2).

[ÇAA+15] S. Çolak, L. P. Alexander, B. G. Alvim, S. R. Mehndiratta, and M. C. González. “Analyzing Cell Phone Location Data for Urban Travel”. *Transportation Research Record: Journal of the Transportation Research Board* 2526 (2015), pp. 126–135 (cit. on p. 22).

[ÇLG16] S. Çolak, A. Lima, and M. C. González. “Understanding congested travel in urban areas”. *Nature Communications* 7:10793 (Mar. 15, 2016) (cit. on pp. 8, 28).

[CBB+06] V. Colizza, A. Barrat, M. Barthélemy, and A. Vespignani. “The role of the airline transportation network in the prediction and predictability of global epidemics”. *Proceedings of the National Academy of Sciences of the United States of America* 103:7 (Feb. 14, 2006), pp. 2015–2020 (cit. on pp. 23, 26).

[CFS+06] V. Colizza, A. Flammini, M. A. Serrano, and A. Vespignani. “Detecting rich-club ordering in complex networks”. *Nature Physics* 2:2 (2006), pp. 110–115 (cit. on p. 71).

[CV07] V. Colizza and A. Vespignani. “Invasion Threshold in Heterogeneous Metapopulation Networks”. *Physical Review Letters* 99:14 (Oct. 5, 2007), p. 148701 (cit. on p. 104).

[CV08] V. Colizza and A. Vespignani. “Epidemic modeling in metapopulation systems with heterogeneous coupling pattern: Theory and simulations”. *Journal of Theoretical Biology* 251:3 (Apr. 7, 2008), pp. 450–467 (cit. on p. 26).

[DST+12] L. Dabbish, C. Stuart, J. Tsay, and J. Herbsleb. “Social coding in GitHub: transparency and collaboration in an open software repository”. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*. CSCW ’12. Seattle, WA, USA. ACM, Feb. 2012, pp. 1277–1286 (cit. on p. 74).

[DPE13] B. D. Dalziel, B. Pourbohloul, and S. P. Ellner. “Human mobility patterns predict divergent epidemic dynamics among cities”. *Proceedings of the Royal Society B: Biological Sciences* 280:1766 (2013) (cit. on p. 104).

[DLM+13] M. De Domenico, A. Lima, P. Mougel, and M. Musolesi. “The Anatomy of a Scientific Rumor”. *Scientific Reports* 3:2980 (Oct. 18, 2013) (cit. on pp. 5, 7, 9).

[DLM13] M. De Domenico, A. Lima, and M. Musolesi. “Interdependence and Predictability of Human Mobility and Social Interactions”. *Pervasive and Mobile Computing* 9:6 (Dec. 2013), pp. 798–807 (cit. on p. 8).

[dMHV+13] Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel. “Unique in the Crowd: The privacy bounds of human mobility”. *Scientific Reports* 3:01376 (Mar. 25, 2013) (cit. on pp. 21, 101).

[dMRS+15] Y.-A. de Montjoye, L. Radaelli, V. K. Singh, and A. Pentland. “Unique in the shopping mall: On the reidentifiability of credit card metadata”. *Science* 347:6221 (Jan. 30, 2015), pp. 536–539 (cit. on pp. 3, 14, 21).

[dMST+14] Y.-A. de Montjoye, Z. Smoreda, R. Trinquart, C. Ziemlicki, and V. D. Blondel. “D4D-Senegal: The Second Mobile Phone Data for Development Challenge” (July 18, 2014). arXiv: `1407.4885` (cit. on pp. 84, 105).

[DLM+14] P. Deville, C. Linard, S. Martin, M. Gilbert, F. R. Stevens, A. E. Gaughan, V. D. Blondel, and A. J. Tatem. “Dynamic population mapping using mobile phone data”. *Proceedings of the National Academy of Sciences* 111:45 (Nov. 11, 2014), pp. 15888–15893 (cit. on p. 83).

[DWS+14] P. Deville, D. Wang, R. Sinatra, C. Song, V. D. Blondel, and A.-L. Barabási. "Career on the Move: Geography, Stratification, and Scientific Impact". *Scientific Reports* 4 (Apr. 24, 2014) (cit. on p. 27).

[DWF03] C. Dewes, A. Wichmann, and A. Feldmann. "An analysis of Internet chat systems". In *Proceedings of the 3rd ACM SIGCOMM Conference on Internet Measurement*. IMC '03. Miami, Florida, USA. ACM, 2003, pp. 51–64 (cit. on p. 39).

[DW04] P. S. Dodds and D. J. Watts. "Universal Behavior in a Generalized Model of Contagion". *Physical Review Letters* 92:21 (May 24, 2004), p. 218701 (cit. on p. 23).

[DGM08] S. Dorogovtsev, A. Goltsev, and J. Mendes. "Critical phenomena in complex networks". *Reviews of Modern Physics* 80:4 (2008), p. 1275 (cit. on p. 41).

[Dug15] M. Duggan. *Mobile Messaging and Social Media 2015*. Pew Research Center, Aug. 2015 (cit. on p. 13).

[EGH+89] J. Ellis, J. Gunion, H. Haber, L. Roszkowski, and F. Zwirner. "Higgs bosons in a nonminimal supersymmetric model". *Physical Review D* 39:3 (1989), p. 844 (cit. on p. 30).

[EKS+96] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. "A density-based algorithm for discovering clusters in large spatial databases with noise." In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*. Vol. 96. KDD '96. Portland, OR, USA, 1996, pp. 226–231 (cit. on p. 123).

[FSL+14] F. O. Fasina, A. Shittu, D. Lazarus, O. Tomori, L. Simonsen, C. Viboud, and G. Chowell. "Transmission dynamics and control of Ebola virus disease outbreak in Nigeria, July to September 2014". *Eurosurveillance* 19:40 (Oct. 2014) (cit. on p. 102).

[Fin11] K. Finley. *Github Has Surpassed Sourceforge and Google Code in Popularity*. ReadWrite. June 2, 2011. (Visited on 09/09/2013) (cit. on p. 65).

[GPB16] R. Gallotti, M. A. Porter, and M. Barthelemy. "Lost in transportation: Information measures and cognitive limits in multilayer navigation". *Science Advances* 2:2 (Feb. 1, 2016), e1500445 (cit. on p. 118).

[GMC+14] M. Garcia-Herranz, E. Moro, M. Cebrian, N. A. Christakis, and J. H. Fowler. "Using Friends as Sensors to Detect Global-Scale Contagious Outbreaks". *PLoS ONE* 9:4 (Apr. 9, 2014), e92413 (cit. on p. 51).

[Gat14] B. Gates. *GIS Mapping & GPS Tracking for Polio in Nigeria*. gatesnotes. 2014. (Visited on 10/27/2014) (cit. on p. 102).

[GNP+07] F. Giannotti, M. Nanni, F. Pinelli, and D. Pedreschi. "Trajectory Pattern Mining". In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '07. San Jose, CA, USA. ACM, 2007, pp. 330–339 (cit. on p. 120).

[Gin12] C. Gini. "Variabilità e mutabilità". *Memorie di metodologica statistica* 1 (1912) (cit. on p. 125).

[GN02] M. Girvan and M. E. J. Newman. "Community structure in social and biological networks". *Proceedings of the National Academy of Sciences* 99:12 (Nov. 6, 2002), pp. 7821–7826 (cit. on pp. 14, 24).

[Gol95] R. G. Golledge. "Path selection and route preference in human navigation: A progress report". In *Spatial Information Theory A Theoretical Basis for GIS*. Ed. by A. U. Frank and W. Kuhn. Lecture Notes in Computer Science 988. Springer Berlin Heidelberg, Sept. 21, 1995, pp. 207–222 (cit. on p. 22).

[GSP+85] R. G. Golledge, T. R. Smith, J. W. Pellegrino, S. Doherty, and S. P. Marshall. "A conceptual model and empirical analysis of children's acquisition of spatial knowledge". *Journal of Environmental Psychology* 5:2 (June 1985), pp. 125–152 (cit. on p. 22).

[GAB+10] S. Gómez, A. Arenas, J. Borge-Holthoefer, S. Meloni, and Y. Moreno. "Discrete-time Markov chain approach to contact-based disease spreading in complex networks". *Europhysics Letters* 89:3 (Feb. 1, 2010), p. 38009 (cit. on p. 42).

[GGM+11] S. Gómez, J. Gómez-Gardeñes, Y. Moreno, and A. Arenas. "Nonperturbative heterogeneous mean-field approach to epidemic spreading in complex networks". *Physical Review E* 84:3 (Sept. 9, 2011), p. 036105 (cit. on pp. 42, 47).

[GHB08] M. C. González, C. A. Hidalgo, and A.-L. Barabási. "Understanding individual human mobility patterns". *Nature* 453:7196 (June 5, 2008), pp. 779–782 (cit. on pp. 3, 17, 18, 21, 101, 117, 158).

[GBR+11] S. González-Bailón, J. Borge-Holthoefer, A. Rivero, and Y. Moreno. "The dynamics of protest recruitment through an online network". *Scientific Reports* 1:197 (2011) (cit. on p. 43).

[Gra98] S. Graham. "The end of geography or the explosion of place? Conceptualizing space, place and information technology". *Progress in Human Geography* 22:2 (Jan. 4, 1998), pp. 165–185 (cit. on p. 2).

[GGA13] C. Granell, S. Gomez, and A. Arenas. "On the dynamical interplay between awareness and epidemic spreading in multiplex networks" (June 18, 2013). arXiv: `1407.4885` (cit. on p. 23).

[Gra78] M. Granovetter. "Threshold Models of Collective Behavior". *American Journal of Sociology* 83:6 (May 1978), pp. 1420–1443 (cit. on p. 25).

[Gra73] M. S. Granovetter. "The Strength of Weak Ties". *American Journal of Sociology* 78:6 (1973), pp. 1360–1380 (cit. on p. 23).

[Hak10] M. Haklay. "How Good is Volunteered Geographical Information? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets". *Envi-*

ronment and Planning B: Planning and Design 37:4 (Aug. 2010), pp. 682–703 (cit. on p. 140).

[HDC14] T. Hardjono, P. Deegan, and J. H. Clippinger. "Social Use Cases for the ID3 Open Mustard Seed Platform". *IEEE Technology and Society Magazine* 33:3 (Fall 2014), pp. 48–54 (cit. on p. 160).

[HHS+11] B. Hecht, L. Hong, B. Suh, and E. H. Chi. "Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles". In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '11. Vancouver, BC, Canada. ACM, 2011, pp. 237–246 (cit. on p. 59).

[HGH08] A. Hindle, D. M. German, and R. Holt. "What Do Large Commits Tell Us? A taxonomical study of large commits". In *Proceedings of the 2008 International Working Conference on Mining Software Repositories*. MSR '08. Leipzig, Germany. ACM, 2008, pp. 99–108 (cit. on p. 27).

[HLC12] B. Hofmann-Wellenhof, H. Lichtenegger, and J. Collins. *Global Positioning System: Theory and Practice*. Springer Science & Business Media, Dec. 6, 2012. 407 pp. (cit. on pp. 13, 20).

[HLC13] B. Hofmann-Wellenhof, H. Lichtenegger, and J. Collins. *Global Positioning System: Theory and Practice*. Springer Science & Business Media, Apr. 17, 2013. 370 pp. (cit. on p. 2).

[HLE+03] P. Holme, F. Liljeros, C. R. Edling, and B. J. Kim. "Network bipartivity". *Physical Review E* 68:5 (Nov. 7, 2003), p. 056107 (cit. on p. 17).

[HK12] E. Horvitz and J. Krumm. "Some Help on the Way: Opportunistic Routing Under Uncertainty". In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. UbiComp '12. Pittsburgh, PA, USA. ACM, 2012, pp. 371–380 (cit. on p. 117).

[HD14] L. F. Huntsinger and R. Donnelly. “Reconciliation of Regional Travel Model and Passive Device Tracking Data”. In *Proceedings of the 93rd Annual Meeting of the Transportation Research Board*. 2014 (cit. on p. 2).

[Jac61] J. Jacobs. *The Death and Life of Great American Cities*. 1961 (cit. on p. 1).

[JJC+12] T. Jia, B. Jiang, K. Carling, M. Bolin, and Y. Ban. “An empirical study on human mobility and its agent-based modeling”. *Journal of Statistical Mechanics: Theory and Experiment* 2012:11 (Nov. 1, 2012) (cit. on p. 115).

[JFY+13] S. Jiang, G. A. Fiore, Y. Yang, J. Ferreira Jr, E. Frazzoli, and M. C. González. “A review of urban computing for mobile phone traces: current methods, challenges and opportunities”. In *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*. UrbComp ’13. Chicago, USA. ACM, 2013, p. 2 (cit. on p. 2).

[KKB+12] M. Karsai, K. Kaski, A. Barabási, and J. Kertész. “Universal features of correlated bursty behaviour”. *Scientific Reports* 2:397 (2012), pp. 1–7 (cit. on p. 39).

[KE05] M. J. Keeling and K. T. D. Eames. “Networks and epidemic models”. *Journal of The Royal Society Interface* 2:4 (Sept. 22, 2005), pp. 295–307 (cit. on pp. 26, 41).

[KR08] M. J. Keeling and P. Rohani. *Modeling infectious diseases in humans and animals*. Princeton University Press, 2008 (cit. on pp. 44, 89, 95).

[KKT03] D. Kempe, J. Kleinberg, and É. Tardos. “Maximizing the Spread of Influence in a Social Network”. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’03. Washington, D.C., USA. ACM, 2003, pp. 137–146 (cit. on pp. 25, 27, 51, 54, 56).

[KM27] W. O. Kermack and A. G. McKendrick. “A Contribution to the Mathematical Theory of Epidemics”. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 115:772 (Aug. 1, 1927), pp. 700–721 (cit. on p. 24).

[KUB+11] A. Kirmse, T. Udeshi, P. Bellver, and J. Shuma. “Extracting Patterns from Location History”. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. GIS ’11. Chicago, IL, USA. ACM, 2011, pp. 397–400 (cit. on pp. 122, 143).

[KGH+10] M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley, and H. A. Makse. “Identification of influential spreaders in complex networks”. *Nature Physics* 6:11 (Nov. 2010), pp. 888–893 (cit. on pp. 23, 42, 102).

[Kle08] J. Kleinberg. “The convergence of social and technological networks”. *Communications of the ACM* 51:11 (Nov. 2008), pp. 66–72 (cit. on p. 52).

[KH03] R. Kölbl and D. Helbing. “Energy laws in human travel behaviour”. *New Journal of Physics* 5:1 (May 1, 2003), p. 48 (cit. on p. 125).

[KGA08] B. Krishnamurthy, P. Gill, and M. Arlitt. “A Few Chirps About Twitter”. In *Proceedings of the 1st Conference on Online Social Networks*. WOSN ’08. Seattle, WA, USA. ACM, 2008, pp. 19–24 (cit. on p. 29).

[KNT10] R. Kumar, J. Novak, and A. Tomkins. “Structure and Evolution of Online Social Networks”. In *Link Mining: Models, Algorithms, and Applications*. Springer, 2010, pp. 337–357 (cit. on p. 70).

[KLP+10] H. Kwak, C. Lee, H. Park, and S. Moon. “What is Twitter, a social network or a news media?” In *Proceedings of the 19th International Conference on World Wide Web*. WWW ’10. Raleigh, NC, USA. ACM, 2010, pp. 591–600 (cit. on pp. 29, 70).

[Lan81] P. Langacker. “Grand unified theories and proton decay”. *Physics Reports* 72:4 (1981), pp. 185–385 (cit. on p. 30).

[LLN+95] B. Latané, J. H. Liu, A. Nowak, M. Bonevento, and L. Zheng. “Distance Matters: Physical Space and Social Impact”. *Personality and Social Psychology Bulletin* 21:8 (Jan. 8, 1995), pp. 795–805 (cit. on p. 3).

[LM01] V. Latora and M. Marchiori. “Efficient behavior of small-world networks”. *Physical Review Letters* 87:19 (2001), p. 198701 (cit. on p. 57).

[LPA+09] D. Lazer, A. Pentland, L. Adamic, S. Aral, A. L. Barabasi, et al. “Life in the network: the coming age of computational social science”. *Science* 323:5915 (Feb. 6, 2009), pp. 721–723 (cit. on pp. 11, 52).

[Led93] L. Lederman. *The God particle: if the universe is the answer, what is the question?* Houghton Mifflin Company, 1993 (cit. on p. 31).

[Leh06] E. Lehmann. *Nonparametrics: Statistical Methods based on Ranks (POD)*. Prentice-Hall, 2006 (cit. on p. 72).

[LLK+14] I. Leontiadis, A. Lima, H. Kwak, R. Stanojevic, D. Wetherall, and K. Papagiannaki. “From Cells to Streets: Estimating Mobile Paths with Cellular-Side Data”. In *Proceedings of the 10th ACM International Conference on Emerging Networking Experiments and Technologies*. CoNEXT ’14. Sydney, Australia. ACM, 2014, pp. 121–132 (cit. on pp. 6, 8, 9, 110).

[LKG+07] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. “Cost-effective outbreak detection in networks”. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’07. San Jose, CA, USA. ACM, 2007, pp. 420–429 (cit. on p. 41).

[LZ13] D. Levinson and S. Zhu. “A portfolio theory of route choice”. *Transportation Research Part C: Emerging Technologies* 35 (Oct. 2013), pp. 232–243 (cit. on pp. 20, 22).

[LK07] D. Liben-Nowell and J. Kleinberg. “The link-prediction problem for social networks”. *Journal of the American Society for Information Science and Technology* 58:7 (2007), pp. 1019–1031 (cit. on pp. 14, 24).

[LNK+05] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins. “Geographic routing in social networks”. *Proceedings of the National Academy of Sciences of the United States of America* 102:33 (Aug. 16, 2005), pp. 11623–11628 (cit. on pp. 14, 24).

[LDP+15] A. Lima, M. De Domenico, V. Pejovic, and M. Musolesi. “Disease Containment Strategies based on Mobility and Information Dissemination”. *Scientific Reports* 5:10650 (June 2, 2015) (cit. on pp. 5, 7, 9).

[LDP+13] A. Lima, M. De Domenico, V. Pejovic, and M. Musolesi. “Exploiting Cellular Data for Disease Containment and Information Campaigns Strategies in Country-Wide Epidemics” (June 19, 2013). arXiv: `1306.4534 [physics]` (cit. on p. 23).

[LM12] A. Lima and M. Musolesi. “Spatial dissemination metrics for location-based social networks”. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. UbiComp ’12. Pittsburgh, PA, USA. ACM, Sept. 2012, pp. 972–979 (cit. on pp. 5, 7).

[LM14] A. Lima and M. Musolesi. “The Rebirth of Locality: Information, People and Places in a Connected World”. In *Proceedings of the 1st Workshop on Geographic Human Computer Interaction. Colocated with CHI.* Paris, France, 2014 (cit. on p. 8).

[LPR+15] A. Lima, V. Pejovic, L. Rossi, M. Musolesi, and M. C. González. “Progmosis: Evaluating Risky Individual Behavior During Epidemics Using Mobile Network Data” (Apr. 6, 2015). arXiv: `1504.01316` (cit. on pp. 5, 8, 9).

[LRM14] A. Lima, L. Rossi, and M. Musolesi. “Coding Together at Scale: GitHub as a Collaborative Social Network”. In *Proceedings of the 8th Interna-*

tional AAAI Conference on Weblogs and Social Media. Eighth International AAAI Conference on Weblogs and Social Media. ICWSM '14. Ann Arbor, MI, USA, May 2014 (cit. on pp. 5, 7).

[LSP+16] A. Lima, R. Stanojevic, D. Papagiannaki, P. Rodriguez, and M. C. González. "Understanding individual routing behaviour". *Journal of The Royal Society Interface* 13:116 (Mar. 2016) (cit. on pp. 6, 8).

[LSA01] E. Limpert, W. A. Stahel, and M. Abbt. "Log-normal Distributions across the Sciences: Keys and Clues". *BioScience* 51:5 (Jan. 5, 2001), pp. 341–352 (cit. on p. 124).

[LSK+05] J. O. Lloyd-Smith, S. J. Schreiber, P. E. Kopp, and W. M. Getz. "Superspreading and the effect of individual variation on disease emergence". *Nature* 438:7066 (Nov. 17, 2005), pp. 355–359 (cit. on p. 102).

[LZZ+09] Y. Lou, C. Zhang, Y. Zheng, X. Xie, W. Wang, and Y. Huang. "Map-matching for Low-sampling-rate GPS Trajectories". In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. GIS '09. Seattle, WA, USA. ACM, 2009, pp. 352–361 (cit. on p. 22).

[LLP+15] T. Louail, M. Lenormand, M. Picornell, O. García Cantú, R. Herranz, E. Frias-Martinez, J. J. Ramasco, and M. Barthelemy. "Uncovering the spatial structure of mobility networks". *Nature Communications* 6 (Jan. 21, 2015) (cit. on p. 22).

[MCM+12] A. Madan, M. Cebrian, S. Moturu, K. Farrahi, and A. Pentland. "Sensing the "Health State" of a Community". *IEEE Pervasive Computing* 11:4 (Oct. 2012), pp. 36–45 (cit. on p. 24).

[MAC15] E. J. Manley, J. D. Addison, and T. Cheng. "Shortest path or anchor-based route choice: a large-scale empirical analysis of minicab routing in London". *Journal of Transport Geography* 43 (Feb. 2015), pp. 123–139 (cit. on p. 22).

[MMW03] J. Masoliver, M. Montero, and G. Weiss. "Continuous-time random-walk model for financial distributions". *Physical Review E* 67:2 (2003), p. 021112 (cit. on p. 39).

[MKS+11] B. Meeder, B. Karrer, A. Sayedi, R. Ravi, C. Borgs, and J. Chayes. "We know who you followed last summer: inferring social link creation times in twitter". In *Proceedings of the 20th International Conference on World Wide Web*. WWW '11. Hyderabad, India. ACM, 2011, pp. 517–526 (cit. on pp. 33, 34).

[MPA+11] S. Meloni, N. Perra, A. Arenas, S. Gómez, Y. Moreno, and A. Vespignani. "Modeling human mobility responses to the large-scale spreading of infectious diseases". *Scientific Reports* 1 (Aug. 12, 2011) (cit. on pp. 23, 112).

[MA10] S. Merler and M. Ajelli. "The role of population heterogeneity and human mobility in the spread of pandemic influenza". *Proceedings of the Royal Society B: Biological Sciences* 277:1681 (2010), pp. 557–565 (cit. on p. 104).

[MMH13] D. Meshi, C. Morawetz, and H. R. Heekeren. "Nucleus accumbens response to gains in reputation for the self relative to gains for others predicts social media use". *Frontiers in Human Neuroscience* 7 (2013), p. 439 (cit. on p. 74).

[Mil67] S. Milgram. "The small world problem". *Psychology today* 2:1 (1967), pp. 60–67 (cit. on pp. 1, 2, 23).

[MGV11] B. Min, K. Goh, and A. Vazquez. "Spreading dynamics following bursty human activity patterns". *Physical Review E* 83:3 (2011), p. 036102 (cit. on p. 40).

[MLA+11] A. Mislove, S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, and J. N. Rosenquist. "Understanding the Demographics of Twitter Users". In *Fifth International AAAI Conference on Weblogs and Social Media*. ICWSM '11. July 5, 2011 (cit. on p. 15).

[MMG+07] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. "Measurement and analysis of online social networks". In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*. IMC '07. San Diego, CA, USA. ACM, 2007, pp. 29–42 (cit. on pp. 14, 24).

[Mit04] M. Mitzenmacher. "A brief history of generative models for power law and lognormal distributions". *Internet Mathematics* 1:2 (2004), pp. 226–251 (cit. on p. 40).

[MR07] E. Mok and G. Retscher. "Location Determination Using WiFi Fingerprinting Versus WiFi Trilateration". *Journal of Location Based Services* 1:2 (June 2007), pp. 145–159 (cit. on p. 22).

[MLR03] M. Montaner, B. López, and J. L. de la Rosa. "A Taxonomy of Recommender Agents on the Internet". *Artificial Intelligence Review* 19:4 (June 2003), pp. 285–330 (cit. on p. 3).

[MPV02] Y. Moreno, R. Pastor-Satorras, and A. Vespignani. "Epidemic outbreaks in complex heterogeneous networks". *European Physical Journal B* 26:4 (2002), pp. 521–529 (cit. on p. 41).

[Mur14] H. Murphy. "Contact Tracing Is Called Pivotal in Fighting Ebola". *The New York Times* (Oct. 2, 2014) (cit. on p. 102).

[MZL12] S. A. Myers, C. Zhu, and J. Leskovec. "Information diffusion and external influence in networks". In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. KDD '12. Beijing, China. ACM, 2012, pp. 33–41 (cit. on p. 41).

[New03a] M. Newman. "The Structure and Function of Complex Networks". *SIAM Review* 45:2 (Jan. 1, 2003), pp. 167–256 (cit. on pp. 14, 24, 41).

[New04a] M. E. J. Newman. "Coauthorship networks and patterns of scientific collaboration". *Proceedings of the National Academy of Sciences* 101 (suppl 1 June 4, 2004), pp. 5200–5205 (cit. on pp. 23, 27).

[New04b] M. Newman. “Who Is the Best Connected Scientist? A Study of Scientific Coauthorship Networks”. In *Complex Networks*. Ed. by E. Ben-Naim, H. Frauenfelder, and Z. Toroczkai. Vol. 650. Lecture Notes in Physics. Springer Berlin / Heidelberg, 2004, pp. 337–370 (cit. on p. 27).

[New02a] M. Newman. “Assortative mixing in networks”. *Physical Review Letters* 89:20 (2002), p. 208701 (cit. on pp. 35, 71).

[New02b] M. Newman. “Spread of epidemic disease on networks”. *Physical Review E* 66:1 (2002), p. 016128 (cit. on p. 41).

[New03b] M. Newman. “Mixing patterns in networks”. *Physical Review E* 67:2 (2003), p. 026126 (cit. on p. 35).

[New05] M. Newman. “Power laws, Pareto distributions and Zipf’s law”. *Contemporary Physics* 46:5 (2005), pp. 323–351 (cit. on p. 41).

[New10] M. Newman. *Networks: An Introduction*. Oxford University Press, 2010 (cit. on pp. 27, 51, 54, 56).

[NFB02] M. Newman, S. Forrest, and J. Balthrop. “Email networks and the spread of computer viruses”. *Physical Review E* 66:3 (2002), p. 035101 (cit. on p. 41).

[NWS98] T. H. Newsome, W. A. Walcott, and P. D. Smith. “Urban activity spaces: Illustrations and application of a conceptual model for integrating the time and space dimensions”. *Transportation* 25:4 (Nov. 1998), pp. 357–377 (cit. on p. 129).

[Nor98] J. R. Norris. *Markov Chains*. Cambridge University Press, July 28, 1998. 262 pp. (cit. on p. 87).

[NSL+12a] A. Noulas, S. Scellato, R. Lambiotte, M. Pontil, and C. Mascolo. “A Tale of Many Cities: Universal Patterns in Human Urban Mobility”. *PLoS ONE* 7:5 (May 29, 2012), e37027 (cit. on p. 22).

[NSL+12b] A. Noulas, S. Scellato, N. Lathia, and C. Mascolo. “Mining User Mobility Features for Next Place Prediction in Location-Based Services”. In *Proceedings of the 2012 IEEE 12th International Conference on Data Mining*. ICDM ’12. Brussels, Belgium. IEEE Computer Society, 2012, pp. 1038–1043 (cit. on p. 28).

[NBK10] P. Nurmi, S. Bhattacharya, and J. Kukkonen. “A Grid-based Algorithm for On-device GSM Positioning”. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*. Ubicomp ’10. Copenhagen, Denmark. ACM, Sept. 2010, pp. 227–236 (cit. on pp. 132, 133).

[ODA15] E. Omodei, M. D. De Domenico, and A. Arenas. “Characterizing interactions in online social networks during exceptional events”. *Interdisciplinary Physics* (2015), p. 59 (cit. on p. 17).

[PKG10] J. Paek, J. Kim, and R. Govindan. “Energy-efficient Rate-adaptive GPS-based Positioning for Smartphones”. In *Proceedings of the 8th International Conference on Mobile Systems, Applications, and Services*. MobiSys ’10. San Francisco, CA, USA. ACM, 2010, pp. 299–314 (cit. on p. 115).

[PMJ+14] V. Palchykov, M. Mitrović, H.-H. Jo, J. Saramäki, and R. K. Pan. “Inferring human mobility using communication patterns”. *Scientific Reports* 4 (Aug. 22, 2014) (cit. on p. 18).

[PSR+15] L. Pappalardo, F. Simini, S. Rinzivillo, D. Pedreschi, F. Giannotti, and A.-L. Barabási. “Returners and explorers dichotomy in human mobility”. *Nature Communications* 6 (Sept. 8, 2015), p. 8166 (cit. on p. 101).

[PV01a] R. Pastor-Satorras and A. Vespignani. “Epidemic dynamics and endemic states in complex networks”. *Physical Review E* 63:6 (2001), p. 066117 (cit. on pp. 41, 42).

[PV01b] R. Pastor-Satorras and A. Vespignani. “Epidemic spreading in scale-free networks”. *Physical Review Letters* 86:14 (2001), pp. 3200–3203 (cit. on pp. 41, 42).

[PMA+14] S. Pei, L. Muchnik, J. S. Andrade , Jr., Z. Zheng, and H. A. Makse. “Searching for superspreaders of information in real-world social media”. *Scientific Reports* 4 (July 3, 2014) (cit. on p. 3).

[Pen09] A. Pentland. “Reality Mining of Mobile Communications: Toward A New Deal On Data”. In *Social Computing and Behavioral Modeling*. Springer US, 2009, pp. 1–1 (cit. on pp. 17, 160).

[Per15] A. Perrin. *Social Media Usage: 2005-2015*. Pew Research Center, Oct. 2015 (cit. on pp. 2, 29).

[Pew14] Pew Research Center. *Emerging Nations Embrace Internet, Mobile Technology*. 2014 (cit. on p. 83).

[Pew15] Pew Research Center. *Internet Seen as Positive Influence on Education but Negative on Morality in Emerging and Developing Nations*. Mar. 2015 (cit. on p. 83).

[PTC12] C. Poletto, M. Tizzoni, and V. Colizza. “Heterogeneous length of stay of hosts’ movements and spatial epidemic spread”. *Scientific Reports* 2 (June 27, 2012) (cit. on p. 23).

[PCL06] S. Porta, P. Crucitti, and V. Latora. “The network analysis of urban streets: a primal approach”. *Environment and Planning B: Planning and Design* 33:5 (2006), pp. 705 – 725 (cit. on p. 128).

[Rai10] L. Rainie. *Internet, broadband, and cell phone statistics*. Pew Internet & American Life Project, Pew Research Center, 2010 (cit. on pp. 2, 29).

[RCM+13] T. Rakthanmanon, B. Campana, A. Mueen, G. Batista, B. Westover, Q. Zhu, J. Zakaria, and E. Keogh. "Addressing Big Data Time Series: Mining Trillions of Time Series Subsequences Under Dynamic Time Warping". *ACM Transactions on Knowledge Discovery from Data (TKDD) - Special Issue on ACM SIGKDD 2012* 7:3 (Sept. 2013), 10:1–10:31 (cit. on p. 121).

[Rav85] E. G. Ravenstein. "The Laws of Migration". *Journal of the Statistical Society of London* 48:2 (1885), pp. 167–235 (cit. on pp. 2, 21).

[RSH+11] I. Rhee, M. Shin, S. Hong, K. Lee, S. J. Kim, and S. Chong. "On the Levy-walk Nature of Human Mobility". *IEEE/ACM Transactions on Networking* 19:3 (June 2011), pp. 630–643 (cit. on p. 115).

[Rog62] E. M. Rogers. *Diffusion of innovations*. Free Press of Glencoe, 1962. 392 pp. (cit. on p. 23).

[RMK11] D. M. Romero, B. Meeder, and J. Kleinberg. "Differences in the Mechanics of Information Diffusion Across Topics: Idioms, Political Hashtags, and Complex Contagion on Twitter". In *Proceedings of the 20th International Conference on World Wide Web*. WWW '11. Lyon, France. ACM, 2011, pp. 695–704 (cit. on p. 41).

[RK10] D. Roongpiboonsopit and H. A. Karimi. "Comparative evaluation and analysis of online geocoding services". *International Journal of Geographical Information Science* 24:7 (June 2, 2010), pp. 1081–1100 (cit. on p. 15).

[RWM15] L. Rossi, J. Walker, and M. Musolesi. "Spatio-temporal techniques for user identification by means of GPS mobility data". *EPJ Data Science* 4:1 (Dec. 2015) (cit. on pp. 20, 160).

[RWS+15] L. Rossi, M. Williams, C. Stich, and M. Musolesi. "Privacy and the City: User Identification and Location Semantics in Location-Based Social Networks". In *Proceedings of the Ninth International AAAI Conference on Web and Social Media*. ICWSM '15. Oxford, UK, Apr. 2015 (cit. on p. 160).

[Sam12] Sample, Ian. *Higgs boson video leaks to CERN website*. 3rd July 2012. (Visited on 02/01/2016) (cit. on pp. 31, 32).

[SRS+14] P. Santi, G. Resta, M. Szell, S. Sobolevsky, S. H. Strogatz, and C. Ratti. "Quantifying the benefits of vehicle pooling with shareability networks". *Proceedings of the National Academy of Sciences* 111:37 (Sept. 16, 2014), pp. 13290–13294 (cit. on p. 115).

[SMM+11] S. Scellato, C. Mascolo, M. Musolesi, and J. Crowcroft. "Track globally, deliver locally: improving content delivery networks by tracking geographic social cascades". In *Proceedings of the 20th International Conference on World Wide Web*. WWW '11. Hyderabad, India. ACM, 2011, pp. 457–466 (cit. on pp. 3, 25, 28).

[SMM+10] S. Scellato, C. Mascolo, M. Musolesi, and V. Latora. "Distance matters: geo-social metrics for online social networks". In *Proceedings of the 3rd Workshop on Online Social Networks*. WOSN '10. Boston, MA, USA. USENIX Association, 2010, pp. 8–8 (cit. on pp. 3, 59).

[SNL+11] S. Scellato, A. Noulas, R. Lambiotte, and C. Mascolo. "Socio-Spatial Properties of Online Location-Based Social Networks". In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*. ICWSM '11. July 5, 2011 (cit. on p. 22).

[SNM11] S. Scellato, A. Noulas, and C. Mascolo. "Exploiting Place Features in Link Prediction on Location-based Social Networks". In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '11. San Diego, CA, USA. ACM, 2011, pp. 1046–1054 (cit. on pp. 14, 22, 24).

[Sch78] T. C. Schelling. *Micromotives and Macrobehavior*. Norton, 1978. 252 pp. (cit. on p. 25).

[SBC+13] C. M. Schneider, V. Belik, T. Couronné, Z. Smoreda, and M. C. González. "Unravelling daily human mobility motifs". *Journal of The Royal Society Interface* 10:84 (June 7, 2013), p. 20130246 (cit. on p. 22).

[SA03] S. Schönfelder and K. W. Axhausen. "Activity spaces: measures of social exclusion?" *Transport Policy*. Transport and Social Exclusion 10:4 (Oct. 2003), pp. 273–286 (cit. on p. 129).

[Sci16] Science and Technology Committee, House of Commons. *The big data dilemma: fourth report of Session 2015–16: report, together with formal minutes relating to the report*. 2016. (Visited on 03/13/2016) (cit. on p. 158).

[SCH+06] J. Scott, J. Crowcroft, P. Hui, and C. Diot. "Haggle: a Networking Architecture Designed Around Mobile Users". In *Proceedings of the Third Annual Conference on Wireless On-demand Network Systems and Services*. WONS '06. Les Ménuires, France, Jan. 2006, pp. 78–86 (cit. on p. 24).

[SGM+12] F. Simini, M. C. González, A. Maritan, and A.-L. Barabási. "A universal model for mobility and migration patterns". *Nature* 484:7392 (Apr. 5, 2012), pp. 96–100 (cit. on p. 18).

[Sno55] J. Snow. *On the Mode of Communication of Cholera*. John Churchill, 1855. 216 pp. (cit. on p. 23).

[SKW+10] C. Song, T. Koren, P. Wang, and A.-L. Barabási. "Modelling the scaling properties of human mobility". *Nature Physics* 6:10 (Oct. 2010), pp. 818–823 (cit. on pp. 17, 21, 101).

[SQB+10] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási. "Limits of Predictability in Human Mobility". *Science* 327:5968 (Feb. 19, 2010), pp. 1018–1021 (cit. on pp. 18, 21, 101).

[Sor03] O. Sorenson. "Social networks and industrial geography". *Journal of Evolutionary Economics* 13:5 (2003), pp. 513–527 (cit. on p. 52).

[SVB+11] J. Stehlé, N. Voirin, A. Barrat, C. Cattuto, L. Isella, et al. “High-Resolution Measurements of Face-to-Face Contact Patterns in a Primary School”. *PLoS ONE* 6:8 (Aug. 16, 2011), e23176 (cit. on p. 24).

[Ste11] R. A. Stein. “Super-spreaders in infectious diseases”. *International Journal of Infectious Diseases* 15:8 (Aug. 2011), e510–e513 (cit. on p. 23).

[SG07] P. R. Stopher and S. P. Greaves. “Household travel surveys: Where are we going?” *Transportation Research Part A: Policy and Practice*. Bridging Research and Practice: A Synthesis of Best Practices in Travel Demand Modeling 41:5 (June 2007), pp. 367–381 (cit. on pp. 21, 131).

[Ttt+12] The CDF Collaboration, the D0 Collaboration, the Tevatron New Physics, and Higgs Working Group. “Updated Combination of CDF and D0 Searches for Standard Model Higgs Boson Production with up to 10.0 fb-1 of Data” (July 2, 2012). arXiv: 1207.0449 (cit. on p. 31).

[TBD+14] M. Tizzoni, P. Bajardi, A. Decuyper, G. Kon Kam King, C. M. Schneider, V. Blondel, Z. Smoreda, M. C. González, and V. Colizza. “On the Use of Human Mobility Proxies for Modeling Epidemics”. *PLoS Computational Biology* 10:7 (July 10, 2014), e1003716 (cit. on p. 26).

[TÇS+15] J. L. Toole, S. Çolak, B. Sturt, L. P. Alexander, A. Evsukoff, and M. C. González. “The path most traveled: Travel demand estimation using big data resources”. *Transportation Research Part C: Emerging Technologies* 58, Part B (Sept. 2015), pp. 162–177 (cit. on pp. 22, 28, 117).

[VSH+12] N. Vallina-Rodriguez, S. Scellato, H. Haddadi, C. Forsell, J. Crowcroft, and C. Mascolo. “Los Twindignados: The Rise of the Indignados Movement on Twitter”. In *Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust*. SOCIALCOM-PASSAT ’12. Amster-

dam, The Netherlands. IEEE Computer Society, 2012, pp. 496–501 (cit. on p. 27).

[VS07] S. Valverde and R. V. Solé. "Self-organization versus hierarchy in open-source social networks". *Physical Review E* 76:4 (Oct. 26, 2007), p. 046118 (cit. on p. 27).

[Vap98] V. Vapnik. *Statistical learning theory*. Wiley New York, 1998 (cit. on p. 104).

[VBS+13] G. M. Vazquez-Prokopec, D. Bisanzio, S. T. Stoddard, V. Paz-Soldan, A. C. Morrison, J. P. Elder, J. Ramirez-Paredes, E. S. Halsey, T. J. Kochel, T. W. Scott, et al. "Using GPS technology to quantify human mobility, dynamic contacts and infectious disease dynamics in a resource-poor urban environment". *PLoS ONE* 8:4 (2013), e58802 (cit. on pp. 104, 115).

[WGZ13] L. Wang, P. D. Groves, and M. K. Ziebart. "GNSS Shadow Matching: Improving Urban Positioning Accuracy Using a 3D City Model with Optimized Visibility Scoring Scheme". *Navigation* 60:3 (Sept. 1, 2013), pp. 195–207 (cit. on p. 20).

[War52] J. G. Wardrop. "Some theoretical aspects of road traffic research". *Proceedings of the Institute of Civil Engineers* 1:3 (Jan. 6, 1952), pp. 325–362 (cit. on pp. 22, 117).

[WF94] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994 (cit. on pp. 23, 51, 54, 74).

[WS98] D. J. Watts and S. H. Strogatz. "Collective dynamics of 'small-world' networks". *Nature* 393:6684 (June 4, 1998), pp. 440–442 (cit. on p. 72).

[WH07] F. Wu and B. A. Huberman. "Novelty and collective attention". *Proceedings of the National Academy of Sciences* 104:45 (2007), pp. 17599–17601 (cit. on p. 50).

[ZZZ+13] Z. Zhang, L. Zhou, X. Zhao, G. Wang, Y. Su, M. Metzger, H. Zheng, and B. Y. Zhao. "On the Validity of Geosocial Mobility Traces". In *Proceedings of the Twelfth ACM Workshop on Hot Topics in Networks*. HotNets-XII. College Park, MD, USA. ACM, 2013, 11:1–11:7 (cit. on p. 16).

[ZZ11] Y. Zheng and X. Zhou. *Computing with spatial trajectories*. Springer Science & Business Media, 2011 (cit. on p. 120).

[ZTM+13] E. Zhong, B. Tan, K. Mo, and Q. Yang. "User demographics prediction based on mobile data". *Pervasive and Mobile Computing* 9:6 (Dec. 2013), pp. 823–837 (cit. on p. 105).

[ZM04] S. Zhou and R. J. Mondragón. "The Rich-club Phenomenon in the Internet Topology". *IEEE Communications Letters* 8:3 (2004), pp. 180–182 (cit. on p. 71).

[ZL15] S. Zhu and D. Levinson. "Do People Use the Shortest Path? An Empirical Test of Wardrop's First Principle". *PLoS ONE* 10:8 (Aug. 12, 2015), e0134322 (cit. on pp. 3, 22, 127, 140).

[ZKS10] Z. Zhuang, K.-H. Kim, and J. P. Singh. "Improving Energy Efficiency of Location Sensing on Smartphones". In *Proceedings of the 8th International Conference on Mobile Systems, Applications, and Services*. MobiSys '10. San Francisco, CA, USA. ACM, June 2010, pp. 315–330 (cit. on p. 115).