

**A Thesis Submitted for the Degree of PhD at the University of Warwick**

**Permanent WRAP URL:**

<http://wrap.warwick.ac.uk/88546>

**Copyright and reuse:**

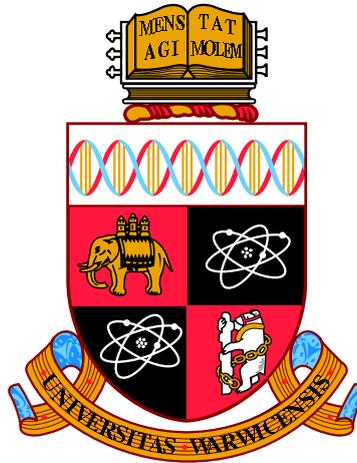
This thesis is made available online and is protected by original copyright.

Please scroll down to view the document itself.

Please refer to the repository record for this item for information to help you to cite it.

Our policy information is available from the repository home page.

For more information, please contact the WRAP Team at: [wrap@warwick.ac.uk](mailto:wrap@warwick.ac.uk)



Quantifying human behaviour using  
complex social datasets

by

**Federico Botta**

**Thesis**

Submitted to the University of Warwick

for the degree of

**Doctor of Philosophy**

**Centre for Complexity Science**

October 2016

THE UNIVERSITY OF  
**WARWICK**

CONTENTS
----------

<b>Acknowledgments</b>	<b>iii</b>
<b>Declarations</b>	<b>v</b>
<b>Abstract</b>	<b>viii</b>
<b>Chapter 1 Introduction</b>	<b>1</b>
<b>Chapter 2 Background</b>	<b>5</b>
2.1 Computational social science . . . . .	6
2.1.1 <i>Google, Yahoo!</i> and <i>Wikipedia</i> . . . . .	7
2.1.2 Social media data . . . . .	13
2.1.3 Mobile phone data . . . . .	19
2.1.4 Crowdsourced data . . . . .	21
2.1.5 Financial data . . . . .	22
2.1.6 Privacy issues . . . . .	23
2.2 Complex networks . . . . .	24
2.2.1 Communities in networks . . . . .	27
<b>Chapter 3 Quantifying stock return distributions</b>	<b>29</b>
3.1 Results . . . . .	31
3.1.1 Methods . . . . .	31
3.1.2 Changes in power law behaviour as $\Delta t$ increases . . . . .	33
3.1.3 Evidence of exponential decay at larger values of $\Delta t$ . . . . .	34
3.2 Conclusions . . . . .	37

---

<b>Chapter 4</b>	<b>Quantifying crowd size</b>	<b>38</b>
4.1	Data . . . . .	38
4.2	Results . . . . .	42
4.3	Conclusions . . . . .	49
<b>Chapter 5</b>	<b>Measuring crowd size using Instagram photos</b>	<b>50</b>
5.1	Data . . . . .	50
5.2	Results . . . . .	53
5.2.1	Selecting an appropriate spatial area for analysis . . . . .	56
5.2.2	Training models using only historic data . . . . .	57
5.2.3	Selecting an appropriate time windows for analysis . . . . .	61
5.2.4	Counting photos instead of users . . . . .	67
5.3	Conclusion . . . . .	71
<b>Chapter 6</b>	<b>Communities of a mobile phone network</b>	<b>72</b>
6.1	Time evolution of communities . . . . .	76
6.2	Period analysis of network structure . . . . .	79
6.3	Null model validation . . . . .	81
6.4	Weighted and multiplex analysis . . . . .	81
6.5	Conclusions . . . . .	83
<b>Chapter 7</b>	<b>Modularity density</b>	<b>87</b>
7.1	Traditional modularity and its limitations . . . . .	88
7.2	Modularity density . . . . .	91
7.3	A modularity density maximisation algorithm . . . . .	95
7.4	Implementation details . . . . .	99
7.1.1	Bisection . . . . .	99
7.1.2	Tuning steps . . . . .	105
7.1.3	Agglomeration . . . . .	111
7.1.4	Community detection algorithm . . . . .	113
7.2	Validation . . . . .	114
7.2.1	Disconnected communities and rings . . . . .	115
7.2.2	Random networks . . . . .	116
7.2.3	Benchmark networks . . . . .	118
7.3	Conclusions . . . . .	119
<b>Chapter 8</b>	<b>Conclusions</b>	<b>120</b>

## ACKNOWLEDGMENTS

Words cannot express how grateful I am, and will always be, to those who made all of this possible.

Dr. Suzy Moat and Dr. Tobias Preis have been an incredible source of support, ideas, discussion and help for the past three years. You have been truly inspiring people to work with. You have gone above and beyond to help me, and this has been an endless motivation for my work. With your help, I have learned what is the right research question to ask and how to answer it, how to present my work to make it engaging, how to make nice figures and use the correct font, and so many other skills which I will always carry with me. Dr. Charo I del Genio has been a unique source of knowledge on the topic of complex networks and computer programming. I would have not been able to do many of the things in here without your intuition on the problems.

And then there's the rest..

There are people who will never see this work, but have nonetheless given an immense contribution to it. All those moments shared with you will always be with me. Amongst many, there's Daniel Ek and Martin Lorentzon. Without them, *Spotify* would not exist. And without *Spotify*, I wouldn't have been able to focus and

my PhD wouldn't exist either. There's also Ludovico Einaudi. If you think the algorithm implemented in Chapter 7 is worth anything, believe me that it wouldn't have been possible without his music. The long hours spent coding were all with his company. And of course, an immense acknowledgement to all the anonymous contributors to the various websites that helped me solve all sort of coding issues (last but not least, the placement of figures in this very thesis. Believe me, it was not fun). What about the inventor of the Internet? Pretty sure that without him this work wouldn't exist. Literally. Even though, yes, some parts of the Internet did distract me "occasionally". Thanks, *Facebook*...

But wait, what sort of acknowledgements are these? Are you not going to thank all those people, friends and colleagues, that supported you in the past three years? The thing is, I do not like lists of names with just a short sentence next to each name. I wouldn't be who I am if it wasn't for every single person who ever happened to be part of my life. A short sentence wouldn't do you any justice. You all made a difference to me at some point, and this will never be forgotten, not even if I wanted (unless I start forgetting things because of my old age). Certain things cannot be expressed by words and are too personal. Better kept just between you and me. I hope all of these people will understand. I will make sure that my actions, rather than my words, will show you how grateful I am.

And then there's one person who will never even know about this work, but who has been the strongest source of inspiration. I like to think that you would have been proud of me, that you would have been happy to see me here. Every second of every day, every letter I typed, every bug in my code, you have always been with me. You were strong, and I am not. Your support, your lessons, our memories, they all mean the world to me. This journey would have not been possible without you at my side. You have taught me how to get to the top of our mountains step after step, and then enjoy the view from up there. I hope you will enjoy this one with me, and will be proud of it.

## DECLARATIONS

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree.

Parts of this thesis have been published by the author or are currently under review:

- (1) Chapter 2 has been accepted as a chapter in the book *Computational Social Science in the Age of Big Data* (Herbert von Halem, in *Neue Schriften zur Online-Forschung of the German Society for Online Research (DGOF)*; expected publication in 2017);
- (2) Chapter 3: Botta F, Moat HS, Stanley HE, Preis T. Quantifying Stock Return Distributions in Financial Markets. *PLOS ONE*, 10: e0135600 (2015). At the time of writing this thesis, it has received five citations and has been presented as a poster at two workshops. Part of this work (the analysis using all trading days and that at the 1% stress level) was performed as an MSc project. As such, it is only summarised in the corresponding chapter. The full discussion can be found in the published manuscript;
- (3) Chapter 4: Botta F, Moat HS, Preis T. Quantifying Crowd Size with Mobile Phone and *Twitter* data. *Royal Society Open Science*, 2: 150162 (2015). At the time of writing this thesis, it has received 22 citations and has been presented

in more than ten conference talks, two of which invited;

(4) Chapter 6 is currently under review;

(5) Chapter 7 is currently under review.

πρὸς ἑμαυτὸν δ' οὖν ἀπιὼν ἐλογιζόμην ὅτι τούτου μὲν τοῦ ἀνθρώπου ἐγὼ σοφώτερός εἰμι· κινδυνεύει μὲν γὰρ ἡμῶν οὐδέτερος οὐδὲν καλὸν καὶ ἀγαθὸν εἰδέναί, ἀλλ' οὗτος μὲν οἶεταί τι εἰδέναί οὐκ εἰδώς, ἐγὼ δέ, ὥσπερ οὖν οὐκ οἶδα, οὐδὲ οἴομαι· ἔοικα γοῦν τούτου γε σμικρῶ τινι αὐτῷ τούτῳ σοφώτερος εἶναι, ὅτι ἂ μὴ οἶδα οὐδὲ οἴομαι εἰδέναί.

*When I left him, I reasoned thus with myself: I am wiser than this man, for neither of us appears to know anything great and good; but he fancies he knows something, although he knows nothing; whereas I, as I do not know anything, so I do not fancy I do. In this trifling particular, then, I appear to be wiser than he, because I do not fancy I know what I do not know.*

The Apology of Socrates

## ABSTRACT

Being able to better understand and measure what is happening in the world is of great importance for a range of stakeholders, including policy makers. The recent explosion in the availability of data documenting our collective behaviour offers new opportunities to gain insights into our society.

Here, we focus on a series of case studies to demonstrate how new forms of data may be used to help us better understand human behaviour.

Data coming from financial transactions taking place in the stock market can help us better understand financial crises. We analyse a dataset comprising the stocks forming the *Dow Jones Industrial Average* at a second by second resolution. We investigate changes in stock market prices and how they arise at different time scales, showing a transition between power law and exponential decay in the tails of the distribution of logarithmic returns.

Accurate and quick estimates of the size of a crowd are crucial for the avoidance of crowd disasters. However, existing approaches rely on human judgement and can be slow and costly. Our findings suggest that data from mobile phone networks and social media platforms may allow us to estimate the size of a crowd. Such data could potentially be accessed in real time, leading to shorter delays than those experienced

with previous approaches to crowd size estimation.

We also show how communities on a network constructed from our social interactions through smartphones capture the temporal evolution of our behaviour in everyday life.

The complex datasets presented here also require complex methodologies to analyse them. Complexity science, and more specifically network science, has witnessed increasing attention within the scientific community in the last two decades. Here, we will present a new technique to analyse a common feature of many real world complex networks, namely community structure. We show how our methodology addresses many of the drawbacks of current techniques, and we also introduce an efficient algorithm which outperforms analogous methods on a set of standard benchmark networks.

Our findings suggest that the analysis of large complex social datasets coupled with methodological advances can allow us to gain valuable measurements of human behaviour.

# CHAPTER 1

## INTRODUCTION

*Imagine you just woke up. You stay in bed a few moments longer, checking the emails that your smartphone has automatically downloaded for you. You have a quick look at your favourite social media platforms. Maybe you like a post, or retweet a message. You get up, and scroll through some of the major newspapers on your tablet while having breakfast. An article about your favourite comedian catches your attention. She will be doing a show later on this week in your home town. You consult a search engine to find out more about it. With a few clicks, you fill in all your credit card details and buy a ticket. You leave the house in a rush, get to the underground station and walk in, swiping your contactless card. At lunch, you take a photo of your meal and upload it to a social media platform. You geolocate it, so that all your friends know where they can find that food. At the end of your long day, you do some exercise, tracking your run across the park with a dedicated fitness app. Later on, you watch a video online and decide that it's worth sharing. The information travels across your social network to all your friends, and then cascades through several social groups.*

We live in a digital world. Data on our behaviour, interests, hobbies and social interactions are constantly being generated. Never before have we had the opportunity to gain such a detailed picture of our lives and collective behaviour. Smartphones, social media platforms and the Internet have radically changed our lives in the last two decades.

The ability to better understand human behaviour and to gain insight into our society is vital for a range of decisions taken by governmental and commercial stake-

holders. State-of-the-art procedures to gather and process data to measure the state of our world — e.g. unemployment rates or gross domestic products — are often time consuming and costly, and can strongly rely on human judgement. Here, we highlight that existing approaches could potentially be complemented by new forms of data capturing a broad spectrum of human activities in a highly interconnected world. The variety of these new forms of data ranges from information seeking and dissemination behaviour online to mobile phone and crowdsourced data, opening up a new window for scientists to study complex social systems. In the following chapters, we will present a series of case studies where the wealth of data available allows us to gain valuable insight into our behaviour and our society.

In Chapter 2, we present a detailed discussion of the variety of studies that have exploited these new sources of data to measure the state of a range of complex social systems. We show how data from search engines, such as *Google* and *Yahoo!*, social media platforms, including *Twitter* and *Flickr*, and data derived from our interactions with smartphones have been used to study our collective behaviour.

The ability to quantify the probability of large price changes in stock markets is of crucial importance to understand financial crises that affect the lives of people worldwide. Large changes in stock market prices can arise abruptly, within a matter of minutes, or develop across much longer time scales. In Chapter 3, we analyze a dataset comprising the stocks forming the *Dow Jones Industrial Average* at a second by second resolution in the period from January 2008 to July 2010 in order to quantify the distribution of changes in market prices at a range of time scales. We find that the tails of the distributions of logarithmic price changes, or returns, exhibit power law decays for time scales ranging from 300 seconds to 3600 seconds. For larger time scales, we find that the distributions tails exhibit exponential decay. Our findings may inform the development of models of market behavior across varying time scales.

Being able to infer the number of people in a specific area is of extreme importance for the avoidance of crowd disasters and to facilitate emergency evacuations. In Chapter 4, using a football stadium and an airport as case studies, we present evidence of a strong relationship between the number of people in restricted areas and activity recorded by mobile phone providers and the online service *Twitter*. Our findings suggest that data generated through our interactions with mobile phone networks and the Internet may allow us to gain valuable measurements of the cur-

rent state of society.

Chapter 5 builds on and extends the results of the previous chapter. In particular, we show that publicly available data generated with our interactions with the social media platform *Instagram* can offer accurate measurements of the size of a crowd. We show that the number of users active on *Instagram* in a given place at a specific time can be used to infer the number of people in that location. We also present a detailed analysis that shows how changes in the behaviour, or number, of users need to be considered. Comparing the results we obtain in two different areas, we investigate how the relationship varies across locations. Our results provide further evidence that data derived from ordinary interactions with social media can be used to study our collective behaviour.

The analysis of increasingly complex data sets requires complex methodologies. A powerful tool which has gained increasing importance in the last two decades is that of networks. Networks are ubiquitous in society. Computers are connected together through the Internet; pages on the web have hyperlinks that allow users to navigate from page to page; cities are connected by airports and train stations; people are linked to each other on various levels, such as kinship, friendship, and work relationship; scientific discoveries build on previous work, thus creating links between scientists. Network thinking allows to develop a framework to model and understand the properties of these systems.

In Chapter 6, we present a detailed analysis of the community structure of the network of mobile phone calls in the metropolitan area of Milan revealing spatial and temporal patterns of communications between people. Our findings suggest that we can extract information about the behaviour of people from communication records and the interactions between social circles.

Identifying communities in a complex network is a key challenge for scientists. A common approach is to search for the network partition that maximizes a quality function. In Chapter 7, we present a detailed analysis of a recently proposed function, namely modularity density. We show that it does not incur in the drawbacks suffered by traditional modularity, and that it can identify networks without ground-truth community structure, deriving its analytical dependence on link density in generic random graphs. In addition, we show that modularity density allows an easy comparison between networks of different sizes. Finally, we introduce an

---

efficient community detection algorithm based on modularity density maximization, validating its accuracy against theoretical predictions and on a set of benchmark networks.

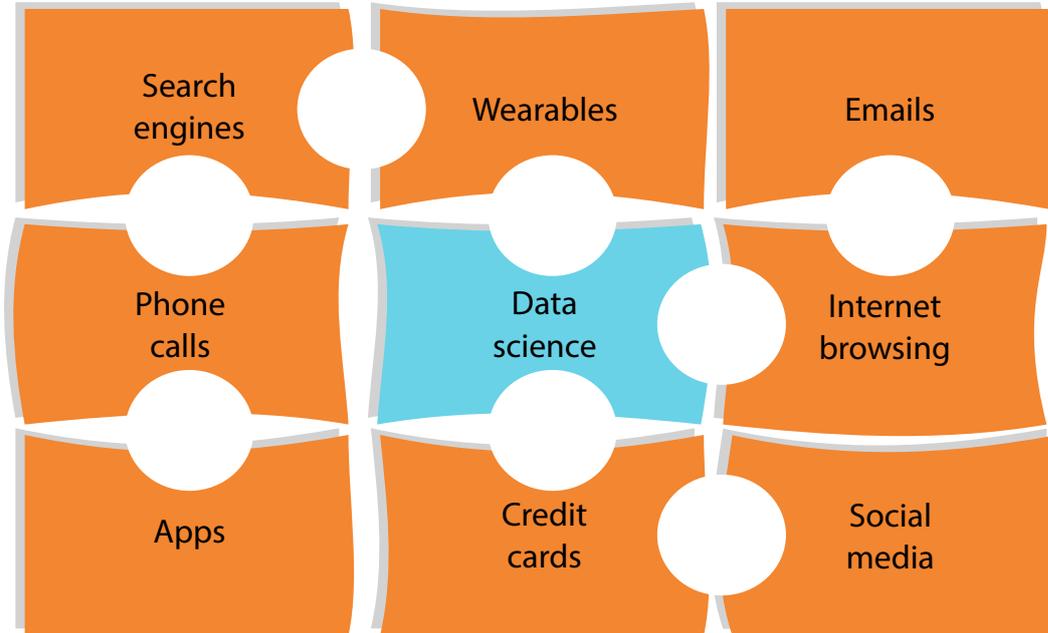
## CHAPTER 2

## BACKGROUND

Being able to better understand human collective behaviour is of fundamental importance for the shaping of a sustainable, efficient and smart society. Knowledge of what is happening in the world right now and of the current state of our society is crucial for a range of policy makers and stakeholders. Traditionally, obtaining such knowledge is a lengthy, costly and slow procedure that requires a great amount of human input and personal judgement. For instance, surveys, censuses, interviews and opinion polls gather data on samples of the population. Statistical agencies, researchers and private companies then analyse these data to extract estimates at a population level.

Recent years have witnessed an explosion in the availability of large and complex datasets encoding a vast amount of information on human behaviour in a readily accessible format. People interact with large technological systems, such as the Internet and mobile networks, to perform a range of actions, from information gathering to building social relations. This generates a large collection of digital traces that scientists from several disciplines can analyse to gain further insight into human behaviour and our society. Crucially, these new forms of data are, in principle, available immediately after their generation. In line with the traditional notion of forecasting, this has defined a new area of research called “nowcasting”, or the pursuit of using readily available sources of data to estimate the current state of society before other slower datasets become available.

The sudden availability of these large complex datasets also requires rigorous method-



**Figure 2.1: Data Science** | In recent years, large complex datasets containing detailed records of our collective behaviour have become increasingly available. Our interactions with technological systems, such as the Internet and mobile phone networks, generate huge volumes of data that can be analysed to help us better understand human behaviour.

ologies and tools to analyse and interpret them. Disciplines such as physics, mathematics, economics, computer science and statistics can all make a contribution to this emerging area of research, frequently referred to as **Data Science** or **Computational Social Science** [1–7]. Methods from all these disciplines, and many others, have been adapted to study social systems.

The combination of complex methodologies and the availability of large scale datasets is at the basis of models that can be used to measure human behaviour and may even be used to predict what is going to happen in the near future. In this Chapter, we will present a review of key results obtained through the analysis of large social datasets, as well as a review of some of the methodologies of interest for the rest of our work.

## 2.1 Computational social science

Analysis of large social datasets has potential to help us improve our understanding of our lives and societies. These new forms of data may come from a variety of

sources and be of different nature (Fig. 2.1). Search engines and online services such as *Wikipedia* collect a large corpus of information on what people are looking for; this may allow us to gain further insight into our decision making processes. Social media platforms offer further opportunities to investigate our social interactions, potentially shedding light on how a disease may spread through a population or how opinions may evolve.

We regularly check our emails, make phone calls and send messages through our smartphones. Datasets generated from these processes encode knowledge on our interactions and on the dynamics of our daily lives. The ubiquitous presence of smartphones in our lives also offers opportunities to conduct large scale social experiments. Compared to experiments or surveys, these studies have the potential to include numbers of participants orders of magnitude larger than what has been possible so far. The vast amount of personal information contained in these datasets also poses many challenges. Researchers and stakeholders alike may have detailed knowledge of individuals and their online behaviour. Privacy issues are an important feature of this new area of research and need to be considered carefully.

This section reviews the main results in the analysis of data derived from such sources, alongside some considerations of potential issues in the analysis of large social datasets.

### 2.1.1 *Google, Yahoo! and Wikipedia*

Decision making is a cognitive process in which people have to make a choice among many possibilities. People choose on the basis of their personal knowledge and the information available to them. In the digital era, the information gathering process often happens online. We use the Internet in a range of situations, including booking our holidays, looking up the weather forecast, buying goods, or searching for a job. We live in a connected world, where more often we collect information online. Search engines, such as *Google* and *Yahoo!*, are the starting point of our browsing activity, since they offer a quick method to find the most relevant websites on the topic under consideration. Every day, we generate a large amount of information on our interests through our search queries, and companies store these data to improve their algorithms and target their customers more accurately. However, these datasets also contain an enormous amount of information on the collective decision making process of people using the Internet [8–10].

**Search query data** One of the key sources of data in the emerging field of Computational Social Sciences are search engines. Due to its widespread use, *Google* has been the focus of several studies. *Google* provides access to search queries through *Google Trends*<sup>1</sup>. Here, a user can access a time series index of the volume of queries for keywords that people are typing into the search engine. Data on search queries can be requested since 2004, and queries can be restricted to various geographical areas. For privacy reasons, the index returned is normalised so that the highest value is equal to 100 and datasets for keywords with a very low search volume are not made available. Data derived from *Google Trends* can be used to gain insight into several aspects of social systems.

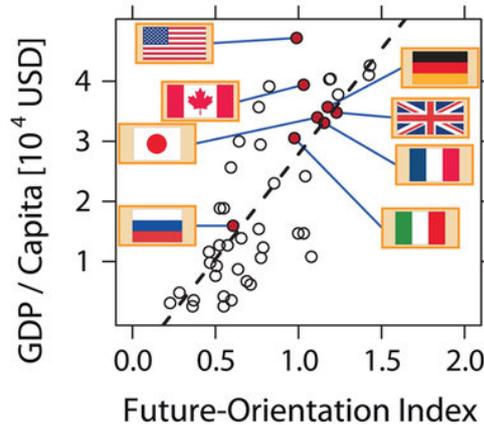
A team of researchers from *Google* itself joined forces with the *Center for Disease Control and Prevention* to exploit search query data to detect influenza epidemics [11]. Traditional surveillance systems have been developed on the basis of several sources of data, such as virological and clinical data, influenza-like illness (ILI) symptoms, or reports of visits to doctors. This collection process is slow, and the reports on levels of influenza are typically reported with a lag of one to two weeks. By processing hundreds of billions of individual searches from several years, the authors were able to build a statistical model, called *Google Flu Trends*, to perform near-to-real-time surveillance of influenza levels at different geographical scales. This study decreases the reporting lag to as little as one day. Other studies have highlighted the importance of information available in search query data to improve disease surveillance [12, 13].

Policy makers and governments have a great interest in accurate estimates of the status of the economy, the job market and several other indicators of the state of society. In [14], the authors analyse search query data in four specific case studies aimed at estimating various key indicators for a society, such as initial claims for unemployment benefits, travelling statistics and consumer confidence. This work uses autoregressive models to show that the inclusion of online data can outperform existing models by a margin ranging from 5% to 20%. A related study addresses the challenge of nowcasting unemployment rates in a specific group of countries in Eastern Europe, namely the Czech Republic, Hungary, Poland and Slovakia [15].

Search query data can also offer an interesting perspective on the collective interest in the future of a population. Previous works has found that countries with a higher

---

<sup>1</sup>[www.google.co.uk/trends](http://www.google.co.uk/trends)



**Figure 2.2: Future Orientation Index** | Countries with a higher tendency of searching for the future on *Google* show a larger GDP per capita. Figure taken from [16].

per capita gross domestic product (GDP) report a higher interest in future years, as opposed to previous years [16, 17]. The authors of this study construct an index based on search data to measure to what extent Internet users in a given country search for information about the future and show that this strongly correlates with the country's per capita GDP for several countries worldwide. Figure 2.2 depicts the relationship between the so-called *Future Orientation Index* and the GDP per capita.

Web search queries can also offer insight into the decision making process of investors in the financial market. Datasets on search queries on *Google* have been shown to bear a relationship to stocks listed in the S&P500 index [18]. An analogous relationship has been found between data derived from the search engine *Yahoo!* and stock market data for those companies listed on the NASDAQ stock exchange [19]. The daily number of search queries for a particular stock is strongly related to the volume of exchanges of the same stock, providing a link between the two. Interestingly, various statistical tests enable the authors to validate the directionality of the correlation and also the appearance of a *wisdom of crowd* effect: individual users typically search one stock only once rather than repeatedly, thus suggesting that the information gathering process is led by non expert investors.

Data derived from *Google Trends* can also be used to implement a trading strategy that can achieve significant returns [20]. This strategy tends to be more successful

for financially related keywords, such as “debt”, and buys or sells hypothetically stocks according to the dynamics of search behaviour. A related study has also shown that the semantic nature of the keywords provides insight in stock market movements [21]. For instance, data on searches for business or politics related keywords can be used to successfully trade stocks in the financial market, whereas data on music or movie related searches would achieve returns analogous to those of an entirely random trading strategy. *Google* search queries exhibit power law cross correlation properties for the *Dow Jones Industrial Average* (DJIA) stocks [22] and can also be used for portfolio diversification [23].

An interesting area of research is at the intersection between online data and digital currencies. *Bitcoin*<sup>2</sup> is a digital currency and open-software payment system that was introduced in 2009 and is the most popular of the virtual currencies introduced so far. These currencies are of interest because they are not issued by any specific central bank and are not related to any government. As such, their value has no particular connection with the real economy. However, some of these currencies, such as *Bitcoin*, experience wild fluctuations in their value. A strong relationship can be found between *Bitcoin* prices and searches for the currency both on the search engine *Google* and the online encyclopedia *Wikipedia* [24]. Sophisticated techniques, such as wavelet coherence analysis, have then identified correlations between various sources of price movements of *Bitcoin*, from online data to financial indices [25].

Our collective behaviour can often be predicted in cultural activities too. *Yahoo!* query data can be used to estimate the revenues of box office feature films in their opening weekend, as well as sales of video games or the ranking of songs [26].

**Wikipedia** *Wikipedia* is a free Internet encyclopedia where users can access and edit the articles themselves. It is vastly popular, and is one of the most visited websites. Due to its comprehensive open access summary of information, it is widely used worldwide. Data on page views and editing history are publicly available, thus offering an ideal source of knowledge on what people are looking for online.

The growth of the encyclopedia itself has attracted the interest of researchers [27]. Representing *Wikipedia* as a directed network, with topics being vertices and hyperlinks being edges, the authors find properties similar to those of the *World Wide Web*, despite a different growth mechanism. This suggests that the growth of

---

<sup>2</sup>bitcoin.com

*Wikipedia* can be described using local rules, such as preferential attachment [28]. Another study has focused on the hierarchical knowledge structure of the encyclopedia, trying to infer it from a network of related terms on *Wikipedia* [29].

The number of visits to financially related pages on *Wikipedia* can provide early signs of stock market moves [30], whereas views of pages in other categories, such as pages of actors, do not offer any information on stock market movements. This may provide evidence that data derived from *Wikipedia* can give an insight in the information gathering process of agents in the financial sector.

Since data on the editing dynamic of pages are available, researchers have also focused on measuring editorial activity on *Wikipedia* [31]. Analysing data from pages in 34 different languages, the authors investigate the geographical distribution of editors worldwide. This is of interest because the spatial distribution of editors may play a role in the biases present in certain pages, and it may explain the heterogeneity in topical coverage. The editorial activity seems to follow a universal circadian pattern for all pages, with a minimum at dawn and maximum later towards the end of the day. Interestingly, the majority of edits in English comes from Europe rather than North America. Controversial *Wikipedia* pages offer a fascinating case study on how editorial wars and social conflicts develop. An automated approach for detecting such conflicts has been developed [32], allowing for detailed studies of several editorial wars [33].

*Wikipedia* pages often provide biographies of important people across a range of disciplines and sectors. However, it is an open question whether the coverage provided by the online encyclopedia gives an accurate image of the situation in the real world. A first study found biases in the coverage of 400 academics on *Wikipedia*, showing that there was no statistical relationship between their *Wikipedia* pages and their academic performance [34].

In a similar fashion to search query data, information coming from *Wikipedia* can offer insight into our collective reaction to a new cultural product or change. The popularity of movies, for instance, can be predicted long before their release by analysing the activities of editors and viewers on the online encyclopedia [35]. *Wikipedia* data also offer a unique opportunity to study language and its complexity [36].

**Email data** Email data also contain a large amount of information on our activities and purchases. However, they are privately owned by companies that need to ensure the confidentiality of their users. As such, studies using these data are often performed by a team in the company itself. In [37], the authors analyse a large *Yahoo!* email dataset to predict the behaviour of consumers when purchasing goods online. A demographic analysis of users shows that the amount of money spent grows with the age of users, peaking in the late 30s. Patterns found in these data may help improve targeting systems for advertising companies.

**Issues** Online data coming from search engines and other platforms that share large scale information can be used in a variety of situations to understand our collective behaviour. However, large social datasets have to be exploited carefully. Despite them capturing a large amount of information, they cannot always replace traditional sources of data. Biases can always be present, and may be magnified by the huge volume of information available. The demographics of *Wikipedia* or *Google* users may not be representative of the whole population, for instance. Any result building on these sources of data may then not necessarily apply to the entire population. The sources of data themselves may also vary, both in availability and collection. For instance, algorithms behind search engines evolve to adapt to their users' needs, thus changing the way the data themselves are generated. The algorithms themselves are also privately owned, thus not allowing for transparency in the data gathering process. It is also important to bear in mind the difference between natural physical systems and social ones. Whereas particles or cells do not react to mathematical models analysing them, agents, such as people purchasing goods or investors trading stocks, respond and adapt to predictions made about their systems. This leads to several questions on the long-term predictability of complex social systems. An important case is that of the aforementioned *Google Flu Trends*. In 2013, the predictions of influenza cases given by the algorithm were more than double the real values. Several causes have been suggested for this, such as change in the behaviour of Internet users due to extensive media coverage of the 2013 flu that led to users wanting to know more about this [38]. A more transparent methodology and data-sharing practice would have also been beneficial, but this may be hard to achieve in situations where a company owns the data.

Big social datasets have to be analysed and interpreted carefully, and may not be able to substitute traditional sources. Indeed, a recent study has shown that online data coming from *Google Flu Trends* combined with more traditional data, such as

historic flu levels, can be used to estimate levels of influenza [39]. This suggests that online data should complement traditional data to increase the accuracy of predictive models, rather than replacing them. Similar results have also been found in studies using data derived from *Wikipedia* [40].

### 2.1.2 Social media data

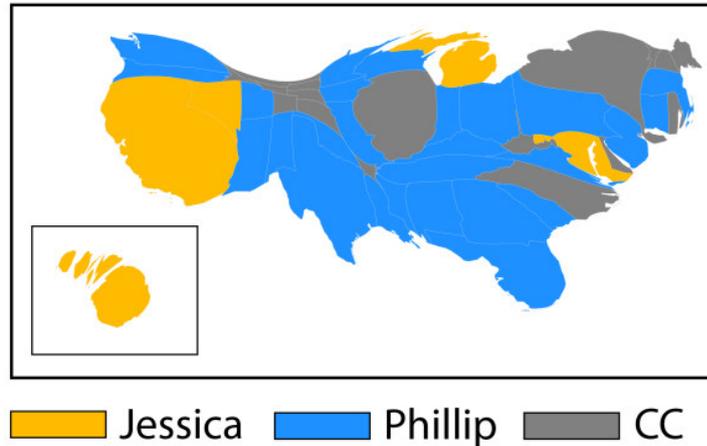
Social interactions form the basis of our social structure. Individuals create relations with each other leading to the emergence of the complex societal structure that we observe as a whole. Social norms and institution are established as a consequence. Traditionally, most interactions among individuals happen face-to-face and arise in a range of social contexts, such as family, friendship, business relations, geographical proximity or religious settings. Recent years have witnessed a shift in the way social interactions take place. People can now create new relationships in the online world, without the need of ever meeting face-to-face. Information flows through a highly connected social structure and interactions happen at an incredibly fast rate on social media platforms. It is probably not a surprise to know that computational social scientists have focused their attention on these new forms of social data. They provide an ideal setting for studying human behaviour at a large scale without having to conduct expensive and time consuming mass surveys.

**Twitter** *Twitter*<sup>3</sup> is among the most popular social networking websites. Registered users can send short messages, called tweets, of up to 140 characters. Users can follow other users and share their tweets, an activity called retweeting. *Twitter* immediately attracted the attention of researchers because the company makes part of their data available through the corresponding API. As was the case for search query data, an important research question concerns the relationship between the flow of information on *Twitter* and the behaviour of the stock market. Since our decision making process can be affected by emotions, researchers have investigated the collective mood of *Twitter* feeds and its correlation with the DJIA index [41]. Calmness of the general *Twitter* public was found to improve the accuracy in predicting the index. *Twitter* data have also been used to estimate the socio-economical status of specific geographical regions, with a particular focus on unemployment [42].

Since its onset, *Twitter* was used as a way to quickly share near-to-real time news. It is thus interesting to analyse the collective behaviour of people around large events that may trigger contrasting emotions in the population. Politics and protests, for

---

<sup>3</sup>twitter.com



**Figure 2.3:** *Twitter* popularity of the two final contestants across the US | Each US state is represented with an area proportional to the number of geotagged tweets coming from that state. Each state is then coloured according to the contestant who is more popular in that state. States coloured in gray cannot be assigned based on *Twitter* activity alone. Figure taken from [49].

instance, are prime examples of this. In the period leading up to the 2010 US congress elections, users have been shown to mostly retweet other users with a similar political view, thus reinforcing their opinions [43]. However, users engage in discussion with the entire network when mentioning other people in their tweets. The overall sentiment of the population is reflected by tweets in the German federal elections and the number of messages is a good indicator of the final outcome of the voting process [44]. Properties similar to those of physical systems close to a critical point have been found in the network of users involved in the Spanish anti-austerity movement during the May 2011 demonstrations [45]. Mechanisms of recruitment for the same protest have also been investigated and found to exhibit patterns analogous to those of complex disease contagion [46]. The information flow around terrorist attacks has also been studied [47], as well as the growth of communication network around the Occupy Wall Street movement [48].

Popular cultural events are also often discussed via *Twitter*. Using the activity of users on the platform, a team of researchers has analysed data related to American Idol, a popular TV show [49]. The results suggest that the rankings of the contestants in the show are significantly correlated with the activity on *Twitter*. This study provides evidence that social media data can be used to anticipate the votes of millions of people. Considering only tweets geolocated in the US, where the

voting for the show is restricted to, increases the predictive accuracy, thus showing the importance of the spatial information available in the dataset. Figure 2.3 depicts the popularity on *Twitter* of the two finalists. The area of each US state is proportional to the number of geotagged tweets coming from that state, and the colour indicates the contestant with the higher proportion of tweets (Fig. taken from [49]). Another study investigated more generally the dynamical properties of large collective social events on *Twitter* [50]. The authors focused on the release of a Hollywood blockbuster movie, protests, the discovery of the Higgs boson and other events. They use information theory techniques, such as symbolic transfer entropy analysis, to study how the dynamics of these systems change before, during and after the event. The main finding is that the characteristic time scales of the information transfer varies as you approach the event. More precisely, events which are mainly driven by an endogenous flow of information show a decrease in the time scale long before the onset of the actual event. Instead, events triggered by external factors show a constant flow of information until the event has taken place (Fig. 2.4). This study suggests that algorithms could be designed in order to analyse large collective events using open access data.

Similar results for the discovery of the Higgs boson have also been found in [51].

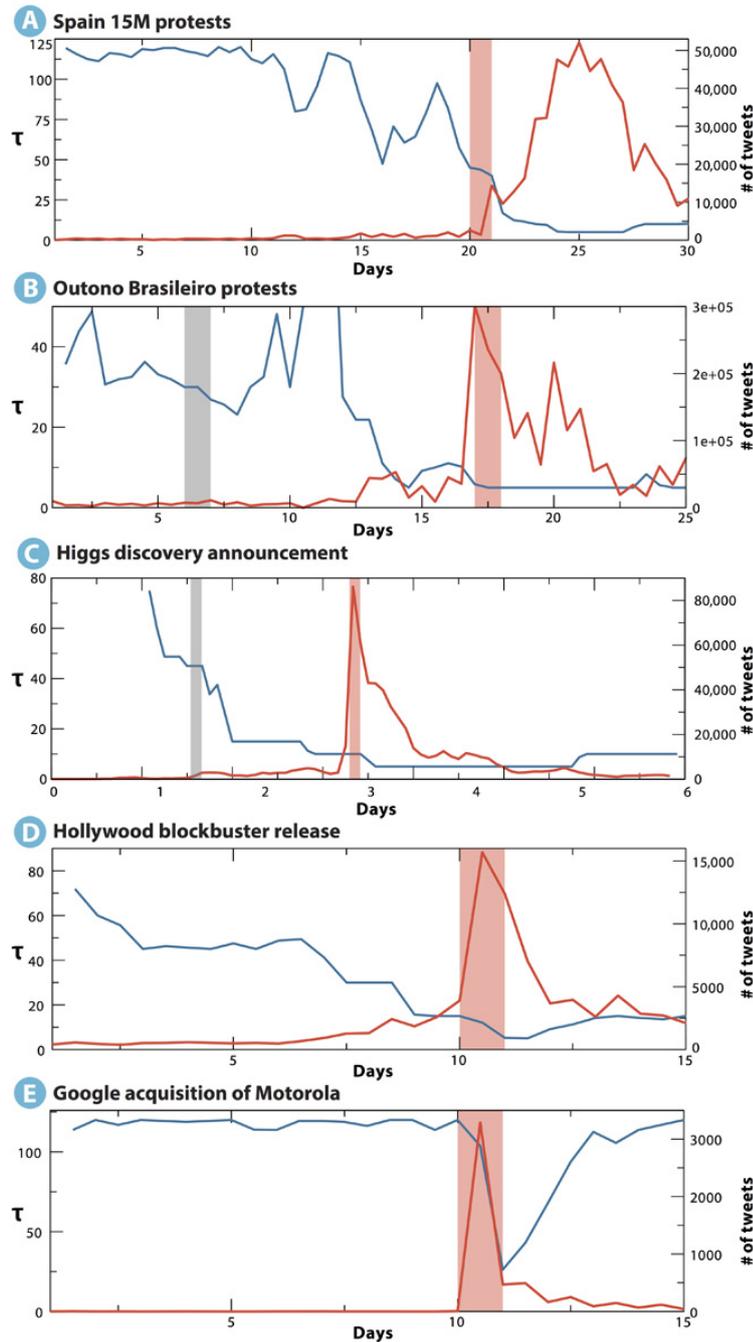
Languages and their dynamics can also be studied [52]. Their use can be mapped across the world [53] and differences in how they are used can also be analysed [54]. Language and the spatial distribution of users may affect how new social links are established on *Twitter* [55]. Human mobility can also play a role in how a social network grows, and the position of a user in the social network can also be used to predict their location [56].

Since *Twitter* users are not representative of the whole population, studies have also focused on investigating the demographic characteristics of people who tweet, such as their age, occupation, social class, and what factors affect the decision of sharing their location on the social media platform [57, 58]. Several other social features of *Twitter* have been investigated in recent years [59–61].

**Facebook** *Facebook*<sup>4</sup> is another widely popular online social networking website. Registered users can set up their own profile, add other users as friends, exchange messages with each other and more generally share information with their social ties. Despite its high penetration rate, *Facebook* has been at the centre of relatively

---

<sup>4</sup>facebook.com



**Figure 2.4: Time scale of events on *Twitter*** | The blue line depicts the characteristic time scale of the flow of information on *Twitter* for the different events considered. For the majority of events, the information flow gets faster as we approach the actual event, indicated with a vertical red bar, as is suggested by the gradual decrease of the time scale. The only exception is the last panel, where the activity is triggered by an external event, namely the media announcement of the acquisition. In this case, the information flows at a constant rate until the event takes place. The red lines indicate the number of tweets for reference. Figure taken from [50].

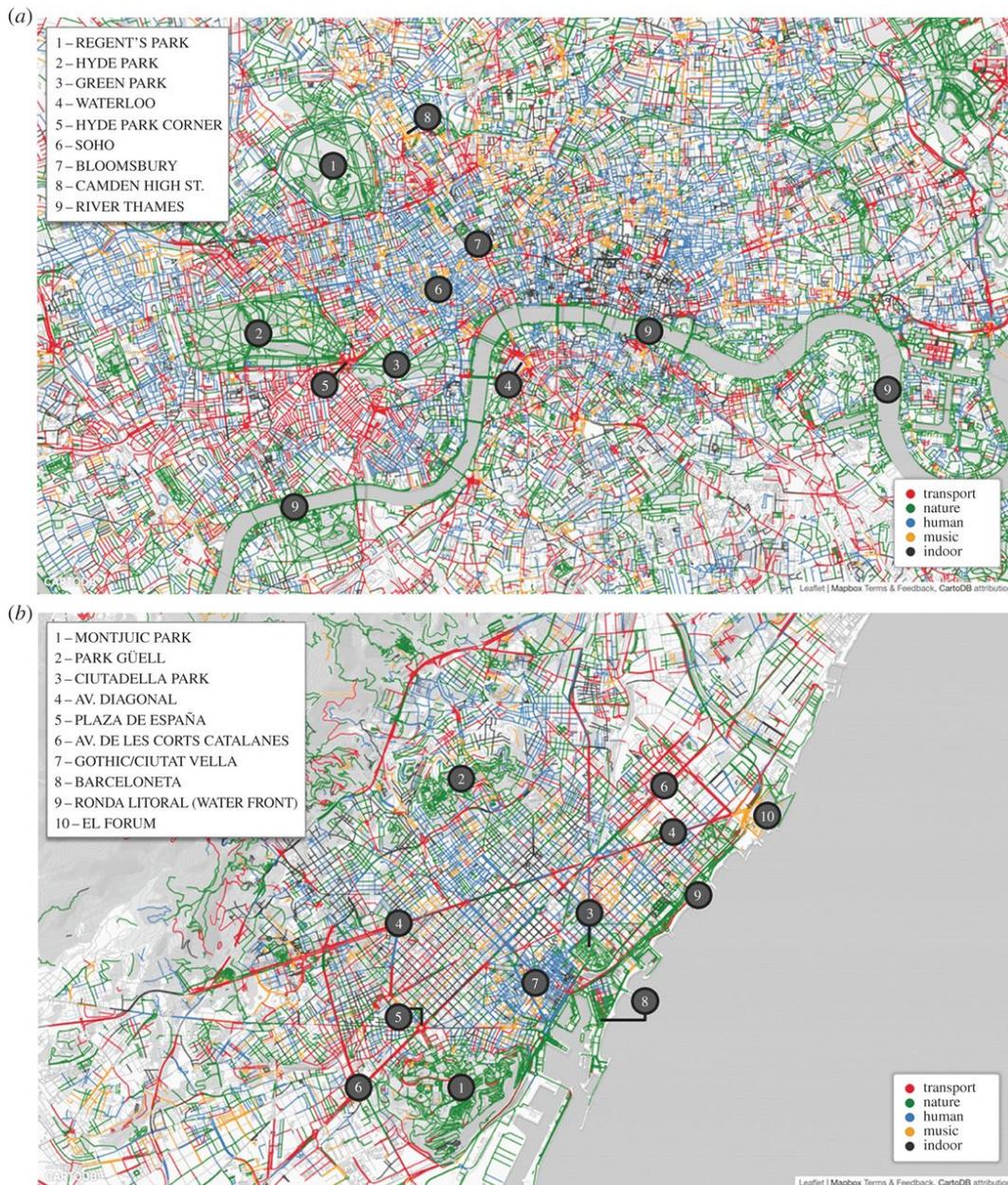
few studies because the company makes very little of their data publicly available. Studies have to be performed either by the research team at *Facebook* or on datasets that have been collected via *Facebook* apps.

A series of papers has focused on the problem of how misinformation spreads on *Facebook* [62–65]. Comparing information flow on pages about either scientific news or conspiracy theories, researchers have found that users selectively expose themselves to content which is in agreement with their views, thus creating so-called echo chambers and polarised communities in the social network structure. Other topics for which echo chambers appear are for pages related to environment, diet, health and geopolitics.

A team of researchers from *Facebook* has shown that users' choices play a significantly more important role in limiting exposure to challenging content, rather than this being an effect of the algorithms that decide what appears on *Facebook*'s News Feed [66].

The possibility of having direct access to the social media platform from within the company allows for large scale social experiments. A randomised controlled trial of 61 million *Facebook* users has shown that messages on the social media platform can be used to influence political self-expression, information seeking and voting behaviour in the 2010 US congressional elections [67]. Information diffusion on *Facebook* has also been studied in a large scale social experiment on 253 million individuals [68]. Interestingly, the authors find that weak ties may play a dominant role in how information is diffused on the social network, thus underlining their importance. Large scale experiments can also help researchers tackle questions on how emotions spread across a social network. A controversial study showed that emotions expressed by other users on *Facebook* may influence our own emotional status [69]. In particular, a reduction in positive posts shown to a user leads to an increased production of negative posts by that user, and vice versa. Interestingly, the authors show that emotional contagion can also happen online without the need of face-to-face interactions.

The emergence of social communities in the friendship network has also been studied [70].



**Figure 2.5: Maps of sounds in London and Barcelona** | Each street segment is assigned to a sound category based on how it is tagged on social media platforms. Different parts of London (a) and Barcelona (b) display differences in their sound, with natural sounds, for instance, being mostly observed in parks. Figure taken from [71].

**Flickr** *Flickr*<sup>5</sup> is a photo sharing platform where users can upload, tag and comment photos. *Flickr* has been the focus of several studies thanks to the public API which gives free access to the entire dataset. Human mobility patterns across the UK can be inferred by looking at the spatial and temporal trajectories of *Flickr* users [72]. International movements can also be estimated and the estimates have been found to correlate significantly with official estimates [73]. Interestingly, *Flickr* data can also be used to quantify the presence of art in a city such as London. In [74], the authors show that neighbourhoods with a higher proportion of art photographs also exhibit a greater relative gain in property prices. Another study has shown that the number of photos tagged with a keyword related to protest correlates with the number of news reports about protests in the corresponding country [75].

The large availability of geo-tagged photos and tweets has also been used to build detailed maps of how users experience the environment around them. In particular, using photos from *Flickr* and *Instagram*<sup>6</sup>, another photo-sharing platform, and posts from *Twitter*, researchers have been able to map the smells, sounds and emotional layers of cities [71, 76, 77]. Figure 2.5 depicts the urban sounds in two different cities. These studies suggest that social media data may provide a fundamental support to policy makers in the design of smart and sustainable cities which take into account how citizens perceive the environment they are living in.

Studies based on other social networks have explored the laws of human mobility using data from location-based platforms [78], how communication between users affects the growth of social networks [68], and what biases may be present in samples of data retrieved from publicly available APIs [79].

### 2.1.3 Mobile phone data

Smart mobile phones have had an enormous impact on our every day lives. We now have the opportunity to make phone calls, browse the Internet, read emails, update our social media, all just with our fingertips. We have GPS enabled devices which can give us directions on a map. We can even make transactions using dedicated applications on the phones. Due to their high penetration rate, which in certain countries is over 100%, smartphones can be used as sensors of most aspects of our daily lives. This has been a great stimulus for the scientific community to investigate our collective behaviour using the vast amount of data generated on our

---

<sup>5</sup>flickr.com

<sup>6</sup>instagram.com

smartphones. A detailed survey of some of the main results in this area has been recently published [80]. Here, we will only focus on specific key results.

Mobile phone datasets typically come in the form of *Call Detail Records* (CDRs). CDRs are recorded by mobile phone providers for billing purposes and contain a large amount of information on our communication patterns. In [81], the authors provide a detailed description of how CDRs are constructed in a specific example. CDRs contain our social interactions, but also spatial and temporal information. They can address some of the shortcomings exhibited by traditional surveys. Sample size is typically not an issue and will most likely be orders of magnitude larger than what is common in social science studies. Some of the self-reporting biases are not likely to occur, but new ones may be introduced. For instance, it is important to bear in mind that these data only capture a specific aspect of our social structure, namely that expressed via mobile phones. The demographic characteristics of users may also vary and may influence the usage patterns of smartphones.

A range of studies have focused on the geographic information stored in these datasets in order to estimate the density of population living in different regions worldwide [82–84]. Geographical distance has also been shown to play a role in the probability of communications occurring. In [85], the authors analyse data from more than two million users in Belgium and find that the probability of two users being connected through a mobile phone communication decreases with the square distance between them in a gravity-like fashion. An analogous relationship on the same dataset also holds between the communication duration and distance [86].

Temporal dynamics can also provide interesting insights into our social behaviour. The persistence of a link in a mobile phone network has been shown to follow a bimodal distribution [87]. Most links either appear just once in the network, or they always appear. This suggests that most phone calls either take place once or they happen regularly between two users.

Understanding how people move is of great importance for several reasons, such as infrastructure planning, public transport design and spreading of epidemics. Mobile phones offer a unique opportunity to study our movements, since they can combine both the spatial and temporal dimension of our behaviour. Interestingly, individual mobility has been shown to follow regular patterns that exhibit a high degree of temporal and spatial regularity [88]. Introducing a parameter that describes the

characteristic spatial length of each individual's trajectory, the authors show that human mobility follow simple and reproducible patterns.

#### 2.1.4 Crowdsourced data

Mobile phones, social media and more generally the Internet allow researchers to perform large-scale surveys and experiments with the potential of reaching out hundreds of thousands of people. This compares favourably with traditional social science surveys and experiments which typically cover a few hundred participants in the best scenario. Researchers can design applications for smartphones and use them as social sensors for their particular research interests. Traditional surveys are usually administered once or few times to each participant, making it difficult to gain granular data on within-individual variations over time. Since we constantly interact with our smartphones, a well designed application can track the relevant features more regularly. Scientists can also create dedicated websites that people can browse and interact with, and this can provide an additional source of data.

In [89], the authors use data from a crowdsourced platform where users had to compare streets in London and rate how beautiful, quiet and happy they were. From the ratings, they construct measures of how different routes are perceived and then construct a recommendation system that suggest routes that are not only short but also emotionally pleasant. Another study investigated the role of green spaces on happiness [90]. The authors designed an application for smartphones that would present a short questionnaire to their users at random moments during the day. The questions were designed to measure the momentary subjective wellbeing of the user. This study managed to collect over one million responses from its users and the results indicate that users are happier in green spaces or other natural habitats. A related question is that of the relationship between environmental scenicness and our health. Using crowdsourced data from the website *Scenic-Or-Not*, researchers have shown that people living in more scenic areas report better health in urban, suburban and rural areas [91]. The relationship holds also when taking other socioeconomic indicators of deprivation into account, such as income and employment.

Crowdsourced data have also been used to study human interactions in different environments and how they can affect the transmission of diseases. The *SocioPatterns* collaboration project has performed several data collection studies on physical proximity and face-to-face interactions of people. This has resulted in several publications investigating the patterns of human interactions and how the network of

contacts can influence the transmission of diseases [92–105].

### 2.1.5 Financial data

The financial market is a complex system for which detailed records of human decisions in form of financial transactions exist. The analysis of such systems has drawn the attention of many physicists, because the accessibility of this large amount of experimental data is a unique opportunity to study in detail the statistical properties of financial markets.

Market changes are of extreme importance since they affect the personal fortunes of people and may also have consequences at political levels. A vast number of studies have focused on many different aspects of financial markets, trying to unravel all the different facets of such a complex system. A first major effort dates back to the 1970s when Black and Scholes derived a first rational option-pricing formula. Since then, however, many different changes have happened in the financial world: the volume of transactions has quickly increased, the financial derivative market has grown exponentially and electronic trading has become standard (thus allowing electronic storage of data). The original proposal of Black and Scholes models the distribution of relative price changes as a log-normal distribution, but it is now known that this provides only a first approximation of what we actually observe in experimental data. In particular, it was first shown by Stanley and collaborators that distributions of returns found in empirical data are consistent with a power law decay [106–109]. These fatter tails assign higher probabilities, compared to Gaussian tails, to extreme events; therefore, the study of such datasets is of fundamental importance since rare events in stock market transactions are high risk situations in which investors want to avoid large losses. Many important features of financial markets have been discovered since then and power laws have been found to describe fluctuations in prices, trading volumes and number of trades. Interestingly, also the distribution of U.S. firm sizes can be consistently described by a power law [110]. The appearance of phase transitions in physical systems with many interacting elements leads to scaling and critical behaviour close to critical points and produce large fluctuations resulting in power law distributions; a similar analysis has shown that this phenomenon can be linked with the dynamics of a human system with many interacting elements (i.e. human participants in the financial market) where volatile market changes relate to the empirical power law distributions through the general frame of phase transitions [111]. Another interesting feature is that many of the scaling exponents that have been found are similar, even in the case of dif-

ferent size of market, trends and also countries [112]. This gives insights in possible universal phenomena that lead to these similarities.

The desire to understand market crashes, crisis and the appearance of financial bubbles has provided another rich area of research. Evidence of speculation, for example, has been found in the 2006-2008 oil bubble when oil prices had an incredible rise followed by an extreme crash [113]. Exploiting techniques from statistical physics to analyse the oil price time series, Sornette, Woodard and Zhou were able to predict the peak of the bubble that immediately preceded the crash in July 2008. Upward and downward trends are consequences of switching processes that appear at different time scales [114]. Such switches at extreme values which form the end of a trend have no scale and this provides evidence that large financial bubbles are inherent features of the scale-free behaviour of the market. Therefore information on microbubbles can be used to study the appearance of large financial crisis. The identification of states of the market through a similarity measure has shown that knowledge of previous states can help to forecast upcoming crises, thus allowing for an early warning and reaction [115]. During such crises, investors rely on diversification to avoid large losses. This effect should protect portfolios in high stress situations of the market and therefore a well chosen basket of stocks should have a smaller risk than each of the stock separately, under the assumption that correlation among stocks are constant in time. However, it has recently been shown that the average correlation scales linearly with market stress, implying that the diversification effect breaks down, precisely when it is most needed [116].

### **2.1.6 Privacy issues**

As we have just shown, the availability of large-scale datasets has had a large impact on several aspects of social science research. Digital technologies have rapidly changed the way we interact, purchase goods and communicate. The generation of high granularity data offers unprecedented opportunity for policy makers and stakeholders alike. However, most studies presented so far rely on the availability of the underlying datasets. Moreover, scientific research makes sharing these datasets necessary, so that studies can be reproduced and new ideas can build on previous analysis. This poses a serious challenge to the privacy of the individuals who are generating this information. Has the dataset been properly anonymised before it is shared? And how secure is this procedure? Individuals should not be identifiable when these datasets are shared.

Recent studies have shown that simple anonymisation procedures are highly ineffective in ensuring the privacy of individuals. Typically, large datasets on our collective behaviour, such as those coming from mobile phones for instance, do not contain information on the names, addresses or phone numbers of the users. However, our digital trajectories may be more unique than we think and the removal of those basic identifiers may not be sufficient to hide our identities. Indeed, recent research has shown that as little as four spatio-temporal points can be enough to identify the vast majority of mobile phone users [117]. The authors analyse several months of human mobility data for hundreds of thousands of individuals and show that coarsening the dataset is a poor way of anonymising it. An analogous result holds for credit card records of more than one million individuals [118]. These results lead to challenging questions about how our privacy can be protected in a robust fashion while making these large-scale datasets available to researchers and policy makers.

## 2.2 Complex networks

A powerful tool which has gained increasing importance in the last two decades is that of networks [119–121, 28, 122–129]. Networks are ubiquitous in nature. Computers are connected together through the Internet; web pages have hyperlinks that allow users to navigate from page to page; cities are connected by airports and train stations; cells, molecules and proteins all interact via biological or chemical reactions; people are linked to each other on various levels, such as kinship, friendship, and work relationship; neurons in our brain are constantly interacting with each other to give rise to the richness of behaviour that we observe in people; scientific discoveries build on previous work, thus creating links between scientists. Network thinking allows to develop a framework to model and understand the properties of these systems.

Traditionally, networks are described using the mathematical language of graphs and graph theory. Generally speaking, a graph is a set of objects, typically called nodes or vertices, in which some pairs of these objects may share a connection, known as a link or an edge. Nodes can have attributes, such as colours or labels, and links can have a direction. The relationship between two nodes can be binary in nature, giving rise to a so called unweighted graph, or can take values in a specific set, such as the natural or the real numbers, in which case we have a weighted graph. Edges can also be of different types, and this gives rise to more complicated

structures.

Mathematically, a graph can be represented as an ordered pair  $G = (V, E)$ , where  $V$  is the set containing the nodes, and  $E$  is a set detailing the list of connections. The size of  $V$  is given by the number of nodes in the graph and is usually denoted with  $N$ . Every element of  $E$  is formed by a pair of nodes contained in  $V$ ; an element of  $E$  establishes an adjacency relation between the corresponding nodes  $x$  and  $y$ , often denoted  $x \sim y$ . If the graph  $G$  is directed, then the order of the pair in each element of  $E$  contains the information about the directionality of the relationship. It is often the case that links between a node and itself are not of particular importance, but such cases can be considered when of interest.

A path is a sequence of adjacent nodes. Paths on graphs allow to connect nodes that are not adjacent, but are nevertheless connected through other nodes. The concept of path can be used to introduce the notion of connectedness: a graph is connected if, for every pair of nodes  $i$  and  $j$ , there exists a path from  $i$  to  $j$ . The most common representation of a graph is in form of a matrix  $A$ , of size  $N \times N$ , whose elements are:

$$a_{ij} = \begin{cases} 1 & \text{if } i \sim j \\ 0 & \text{otherwise} \end{cases}$$

For undirected graphs, this matrix is symmetric. If links between a node and itself are not allowed, the diagonal of  $A$  is all zero. For weighted graphs, the analogous definition is:

$$a_{ij} = \begin{cases} w_{ij} & \text{if } i \sim j \\ 0 & \text{otherwise} \end{cases}$$

where  $w_{ij}$  is the weight of the corresponding link.

From the adjacency matrix, we can easily introduce the degree  $k_i$  of node  $i$  as the number of edges to which  $i$  is attached:

$$k_i = \sum_{j \in V} a_{ij}$$

The complete list of degrees of all nodes in  $G$  gives rise to the degree sequence. From the degree sequence, we can then construct the degree distribution  $P(k)$  which is a central feature of a graph. It is defined as the probability that a randomly chosen node  $i$  will have degree equal to  $k$ . If the probability of a node of degree  $k$  being connected to another node of degree  $\tilde{k}$  is independent of  $k$ , then the network is said

to be uncorrelated; it is correlated otherwise. The correlation can give rise to an assortative structure, where nodes connect to other nodes of similar degrees, e.g. high degree nodes connect with high degree nodes, or to disassortative structure, where nodes with low degree connect with those with high degree.

In a graph, clustering refers to the presence of triangles, i.e. triplets of nodes all pairwise connected. A related definition introduces the concept of clustering coefficient, which measures the probability that two neighbours of node  $i$  are neighbours themselves. This is an important concept in social relations, where it simply measures the probability that two of my friends are friends themselves.

Moving to even larger structures, a community in a graph is a subset of nodes in  $V$  that share many connections within the subset. Communities play an important role in many applications, and we will explore this topic in more detail further in this chapter.

Complex networks differ from graphs in that they often display non-trivial behaviour in the topological features just presented. Networks derived from real-world data often exhibit heterogeneous degree distributions, large clustering coefficient, assortative behaviour and community structure. As such, they are rather different from traditional random graphs or regular lattices, which display a high degree of regularity.

Network science tools have already been used in hundreds of studies in a range of disciplines. Processes taking place on networks, such as random walks [130, 92], epidemic spreading [131–137] and rumour spreading [138, 139] have been thoroughly investigated. The network framework has been applied to various contexts, such as the emergence of social conventions [140], to understand political elections [141], the properties of the Internet [142], the social structure on Facebook [143], the structure of committees in the U.S. House of Representatives [144], and the importance of financial institutions in the network of financial exposures between them [145]. Network representations of complex systems have been used in biology [146–148], and in studies of technological systems [28] and communication systems [149]. Researchers have applied complex systems techniques to a wide range of disciplines, identifying and analyzing several defining features of complex networks, such as the small world property [150–152], heterogeneous degree distributions [28, 127], clustering [153, 154], degree-degree correlations [155, 156], assortativity [157], syn-

chronizability [158], and community structure [159].

In recent years, the wide availability of network data has given network scientists the opportunity to study the relationship and dependencies between different networks. In online social networks, for instance, where nodes represent people, users can be active on different platforms. On each platform, they will have a friendship network which may depend on what they use that specific social media for. This gives rise to a multilayer network, where each layer represents a different social media platform. The generalisation of this object has received extensive interest in recent years, due to the interesting properties of processes taking place on it [160–166].

### 2.2.1 Communities in networks

Communities were originally studied in the context of social networks, in which they are formed by groups of people that share close friendship relations. However, communities of densely connected modules have been observed in several real-world and model networks of diverse nature [167–178], where, in general, they are defined as groups of nodes whose internal connections are denser or stronger than those that link nodes belonging to different groups. In all these cases, the presence of communities directly influences the behaviour of the system, where there is often a correspondence between communities and functional units. Ever since the discovery of community structure in real-world networks, a plethora of techniques devoted to their detection has been introduced [179–187]. The challenge is both theoretical, in proposing a good mathematical definition of what constitutes a community, and computational, in developing good heuristics that can detect communities in a reasonable time.

A common way of investigating the community structure of networks starts with the definition of a quality function, which assigns a score to any network partition. Larger scores correspond to better partitions, and algorithms are created to find the partition with the largest score. By far, the most common and used of such quality functions is modularity [188], that compares the number of links inside each community to the number of links that would be expected if the nodes were connected at random, without any preference for links within or outside the community. A partition with a large modularity indicates that the communities have many internal links and few external ones, when compared to a randomized version of the network.

In its most general form, modularity can be written as follows:

$$Q = \frac{1}{2m} \sum_{ij} (A_{ij} - P_{ij}) \delta(C_i, C_j)$$

The sum is over all pairs of nodes,  $m$  is the overall number of links in the network, and  $P_{ij}$  represents the number of edges that we expect to see between node  $i$  and node  $j$  in a suitably defined randomised version of the same network. Despite the choice of  $P_{ij}$  being independent, it is commonly accepted that the null model should keep some of the topological properties of the original network. In particular, it is usually the case that we want to keep the degree distribution fixed. In this scenario, we need to calculate the probability  $p_i$  to pick at random a stub, or half-edge, incident on a node  $i$  of degree  $k_i$ . From this, we can calculate the probability of an edge linking node  $i$  to node  $j$  because an edge can only exist if two stubs incident with  $i$  and  $j$  are linked together. Since there are  $k_i$  stubs attached to vertex  $i$ , and there are  $2m$  overall stubs, the probability to pick one at random which is linked to  $i$  is given by  $\frac{k_i}{2m}$ . Then, the probability of having a link between the two nodes  $i$  and  $j$  is simply calculated as the product of the two, giving  $\frac{k_i k_j}{4m^2}$ . This leads to the expected number of links between the two vertices of  $P_{ij} = \frac{k_i k_j}{2m}$ . Using this expression for modularity, we obtain:

$$Q = \frac{1}{2m} \sum_{ij} \left( A_{ij} - \frac{k_i k_j}{2m} \right) \delta(C_i, C_j)$$

This is the most common expression for modularity that can be found in several studies. Any partition of a network will yield a score of  $Q$ , with larger scores indicating a stronger community structure, according to the implicit definition of communities used to define  $Q$ . The problem of identifying network communities becomes then an optimization problem over the space of all possible partitions of a network. An exhaustive procedure is not feasible because of the large number of possible ways of partitioning a network when both the size and the number of the communities are not fixed. Several techniques have been developed in the literature for this problem, and an extensive review can be found in [181].

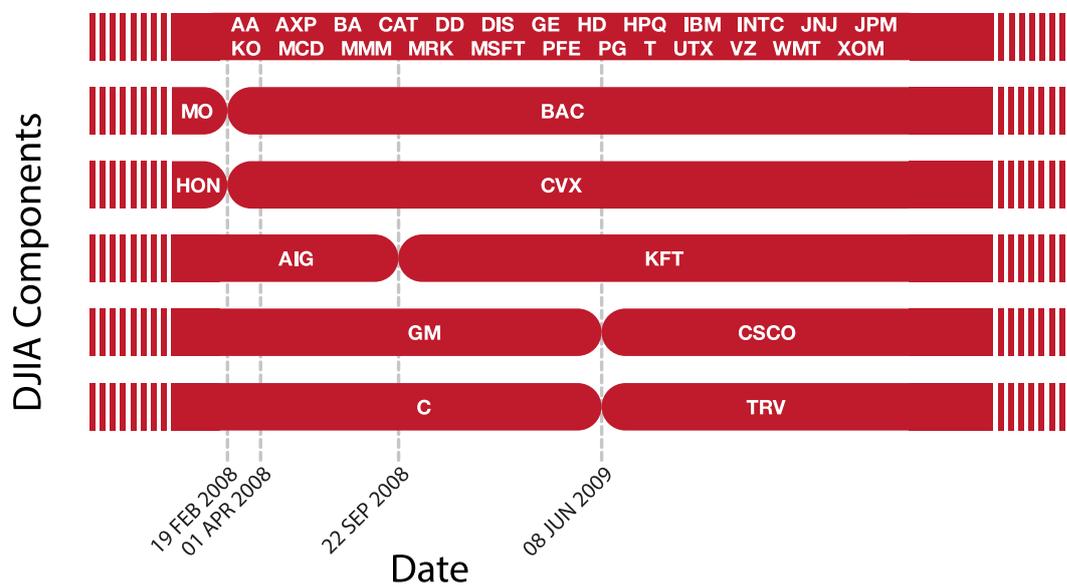
## CHAPTER 3

# QUANTIFYING STOCK RETURN DISTRIBUTIONS IN FINANCIAL MARKETS

Complex movements in stock market prices affect the personal fortunes of people around the globe [189–193]. An ability to more accurately quantify and predict such changes would allow us to gain more insights into how financial crises arise [194] and provide greater empirical basis for the development of theories of financial market behavior [195–199, 109, 200].

A vast amount of data on financial decisions made in stock markets is available [16, 20, 30, 21, 115]. Previous studies have shown that distributions of returns observed in empirical data are consistent with power law decay [107, 108, 201, 202, 112, 203–210], in contrast with widely used models that assume Gaussian behavior of these returns. Power law behavior has also been observed in other economical and financial sectors of society [110, 113].

Changes in stock market prices can occur at a range of different time scales. Here, we analyse a large dataset of stocks forming the Dow Jones Industrial Average (DJIA) at a second-by-second resolution for a range of different time scales in order to quantify the distribution of returns. We provide evidence that while the distribution of returns exhibits power law behaviour at small time scale, exponential behaviour is observed at larger time scales. We find analogous results when restricting our analysis to volatile trading periods. Our findings could help to gain insight into changes in stock market prices in shorter periods and longer periods and provide further empirical basis for the development of new models of market behaviour.



**Figure 3.1: Components of the DJIA.** Here we depict the components of the DJIA in the time period between 02 January 2008 to 30 July 2010. Dashed vertical lines correspond to changes in the stocks forming the DJIA. In our analysis, we focus on the 25 stocks that were part of the DJIA during the period of analysis. Stocks are labelled using ticker symbols that uniquely identify the company name, as used by the stock exchange.

### 3.1 Results

The DJIA is a U.S. benchmark index that consists of 30 different stocks. For all 30 stocks, we retrieve price time series with a second by second resolution from the Trade and Quote (TAQ) database provided by Wharton Research Data Services (WRDS). Our dataset covers the period from 2 January 2008 to 30 July 2010 comprising a total of 647 trading days. Figure 3.1 shows the various components of the DJIA. As five stocks were replaced during this period, we focus on the 25 components that were consistently part of the DJIA between 02 January 2008 and 30 July 2010.

We define returns as the relative logarithmic change in price of a given stock  $i$  at a given time  $t$ :

$$r_i(t) = \log(p_i(t + \Delta t)) - \log(p_i(t)) \quad i = 1, \dots, 25$$

where  $\Delta t$  is the time lag between price observations. As a trading day starts at 9:30 and ends at 16:00 local time,  $\Delta t$  is constrained to be at most 6 hours and 30 minutes.

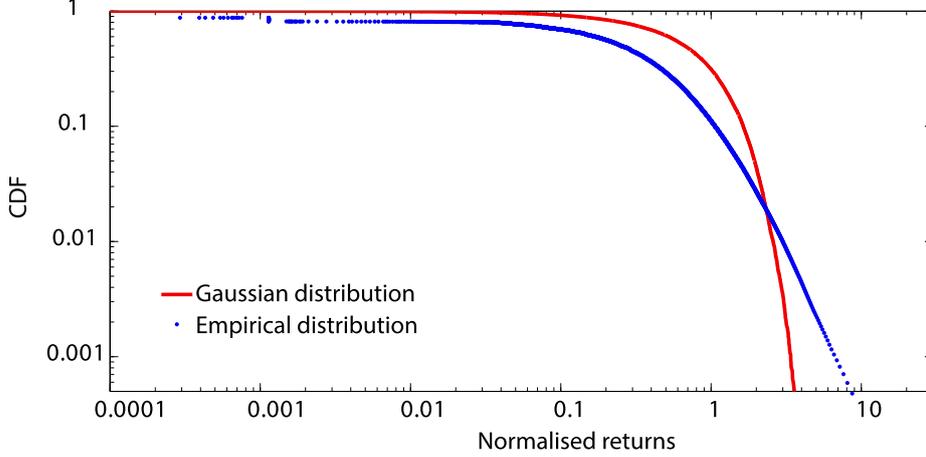
We compute the standardised distribution of the returns for the 25 components of the DJIA that we consider. We conduct separate analyses of the cumulative distribution function (CDF) of the positive and negative component of the distribution of returns.

Figure 3.2 depicts the positive CDF for *American Express* for  $\Delta t = 300$  seconds and compares this to a Gaussian distribution. Note that the empirical distribution strongly deviates from the Gaussian distribution and provides initial evidence for power law behaviour. We perform a statistical analysis to check the consistency of the tails of the empirical distributions with power law behaviour across different time scales, as proposed by Clauset, Shalizi and Newman [211] and detailed in the Methods section.

#### 3.1.1 Methods

A power law is a distribution of the form:

$$p(x) = \frac{\alpha - 1}{x_{\min}} \left( \frac{x}{x_{\min}} \right)^{-\alpha}$$



**Figure 3.2: Empirical distribution of normalised returns for *American Express*.** We build returns distributions for the 25 stocks of the DJIA for different time lags across the full period of analysis. We standardise each distribution by subtracting the mean return from each observation and dividing by the standard deviation. We depict in blue the cumulative distribution function of the positive component of the return distributions for *American Express* for a time lag of 300 seconds. We depict in red the positive tail of a Gaussian distribution with mean zero and standard deviation one. We observe a strong deviation of the empirical distribution from the Gaussian distribution. Instead, visual inspection of the distribution tail reveals consistency with a linear relationship on a log-log scale. This provides initial evidence for possible power law behaviour at this time scale.

where  $\alpha$  is the scaling exponent. We require  $\alpha > 1$  for this to be a Probability Distribution Function (PDF).  $x_{\min}$  is the lower bound of the power law behaviour. We estimate the scaling exponent  $\alpha$  using the maximum likelihood estimator (MLE). Assuming we have  $n$  observations of  $x_i (i = 1, \dots, n)$  which are independent and identically distributed random variables, the likelihood function, which represents the probability of observing the data given the parameter, is given by:

$$p(x|\alpha) = \prod_{i=1}^n \frac{\alpha - 1}{x_{\min}} \left( \frac{x_i}{x_{\min}} \right)^{-\alpha}$$

We then maximise this probability to find the MLE estimator for the scaling exponent:

$$\hat{\alpha} = 1 + n \left[ \sum_{i=1}^n \ln \frac{x_i}{x_{\min}} \right]$$

We measure distances between distribution using the Kolmogorov-Smirnov statistic (KS statistic):

$$D = \max_{x \geq x_{\min}} |E(x) - F(x)|$$

where  $E(x)$  is the empirical CDF and  $F(x)$  is the best fit of the data. We determine the lower bound  $x_{\min}$  by choosing the value that minimizes the distance between the empirical distribution and the fitted distribution as measured by the KS statistic. Once we have determined the lower bound  $x_{\min}$  and the scaling exponent  $\alpha$ , we then check the consistency of the hypothesis of power law behaviour in the observed empirical distributions. We construct the empirical tails choosing a bin size such that we have 1,000 data points in each tail. We then compare the KS statistic observed for the empirical data when compared to a fitted power law distribution with the KS statistic obtained for the synthetic data when compared to a fitted power law distribution. We obtain a  $p$ -value by counting the number of times that the synthetic KS statistic is larger than the empirical KS statistic. We generate 1,000 synthetic data sets and make the conservative choice of accepting our hypothesis of consistency with power law behaviour if the  $p$ -value is larger than 0.1.

To determine whether the distribution is consistent with exponential decay, we perform a parallel analysis fitting the data to an exponential distribution instead of a power law probability distribution. We then generate synthetic data from the fitted distribution in the same manner as previously described. We evaluate whether our data are consistent with exponential decay by comparing the empirical data to the synthetic data using KS statistics as described above.

### 3.1.2 Changes in power law behaviour as $\Delta t$ increases

A power law probability distribution is a probability distribution in which the probability of an event decays as a negative power of the event. The distribution function is characterised by a scaling exponent. Distributions of returns typically exhibit power law decay in the tail of the distribution. Here, we want to understand how the exact nature of power law behaviour depends on the time lag between price observations. We analyse all 25 stock price time series and use a time lag  $\Delta t$  ranging from 300 to 3,600 seconds. We investigate how the scaling exponent changes as a function of the time lag between price observations.

Our previous work has shown how the exponent for the tails of the positive (denoted as  $\alpha^+$ ; Fig. 3.3a) and negative (denoted as  $\alpha^-$ ; Fig. 3.3b) returns distributions in-

creases with the time lag  $\Delta t$  [212]. We have also shown that this finding holds for a subset of the price time series in which relatively extreme price movements occur. In particular, we have restricted the analysis to price observations recorded on trading days on which the corresponding stock gained or lost more than 1% on a daily basis. We refer to this as a stress level of 1%. Figures 3.3c and 3.3d depict the relationship between the power law exponents and the time lag  $\Delta t$  between price observations on trading days on which the market experienced a stress level of at least 1%.

Here, we extend our results by performing a parallel analysis and considering a 2% stress level (Fig. 3.3e and 3.3f). We find that the mean scaling exponent increases with the time lag  $\Delta t$  between price observations ( $\alpha^+$ : Adjusted  $R^2 = 0.782$ ,  $N = 12$ ,  $p < 0.001$ , ordinary least squares regression;  $\alpha^-$ : Adjusted  $R^2 = 0.836$ ,  $N = 12$ ,  $p < 0.001$ , ordinary least squares regression):

$$\alpha^+ = 0.022(\pm 0.003)\Delta t + 3.09(\pm 0.13)$$

$$\alpha^- = 0.017(\pm 0.002)\Delta t + 3.14(\pm 0.08)$$

At a stress level of 3%, we again observe that the scaling exponent increases as we increase the time lag  $\Delta t$  ( $\alpha^+$ : Adjusted  $R^2 = 0.573$ ,  $N = 12$ ,  $p < 0.05$ , ordinary least squares regression;  $\alpha^-$ : Adjusted  $R^2 = 0.458$ ,  $N = 12$ ,  $p < 0.05$ , ordinary least squares regression):

$$\alpha^+ = 0.066(\pm 0.017)\Delta t + 2.04(\pm 0.61)$$

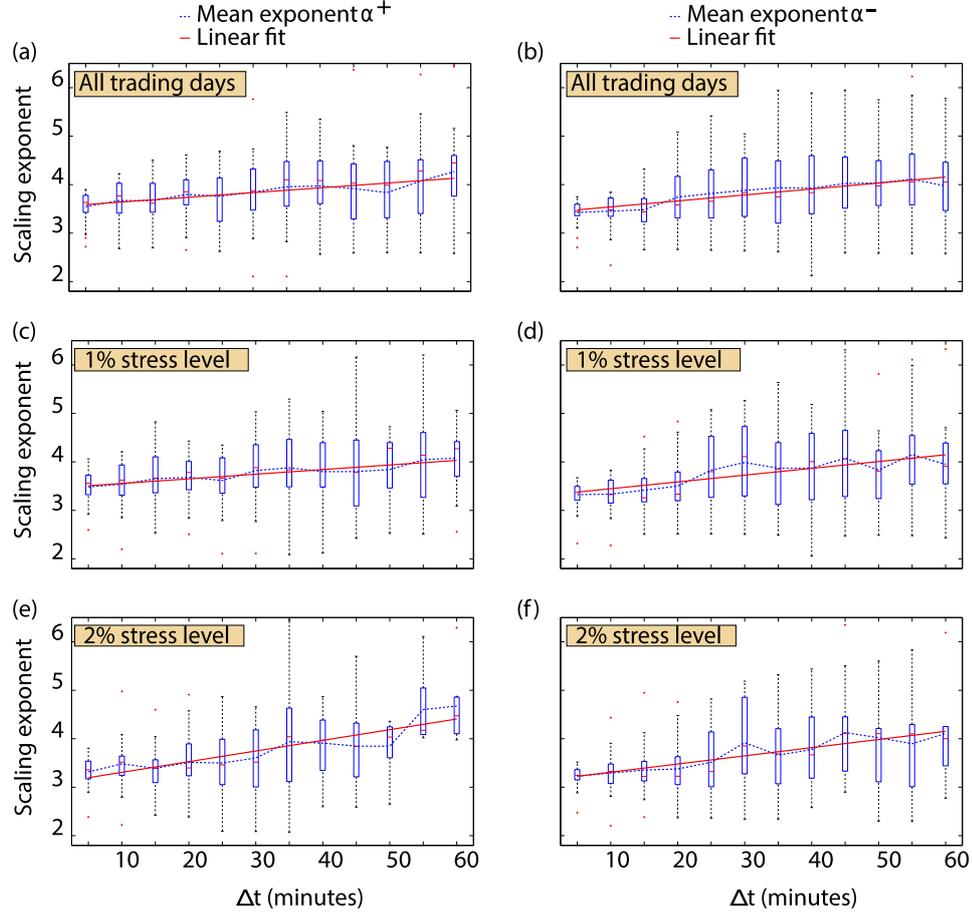
$$\alpha^- = 0.016(\pm 0.005)\Delta t + 2.91(\pm 0.19)$$

Our new results are consistent with our previous findings and provide further evidence on the relationship between the scaling exponent and the time lag  $\Delta t$ .

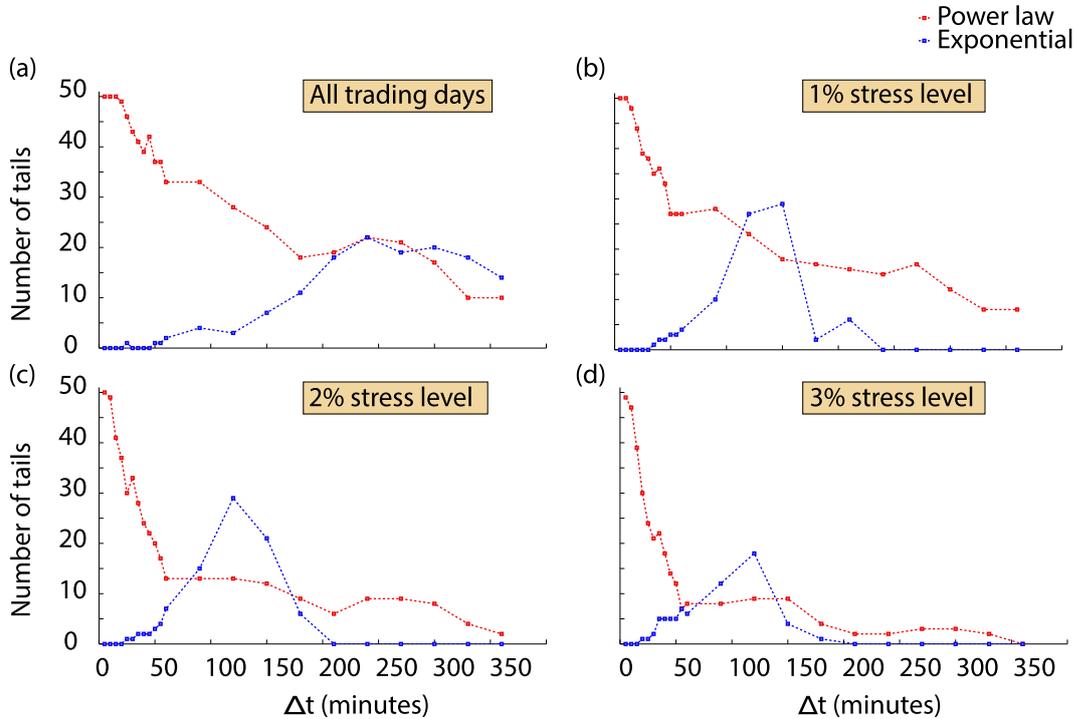
### 3.1.3 Evidence of exponential decay at larger values of $\Delta t$

Our previous work highlighted how for  $\Delta t > 60$  minutes the number of tails consistent with power law behaviour decreases (Fig. 3.4a). We then investigated this change in behaviour at a range of time scales and analysed whether we started to observe consistency with exponential decay. Exponential decay had already been observed in daily returns of stocks from the National Stock Exchange in the Indian stock market[213].

Figures 3.4a and 3.4b depict the number of distributions consistent with either power



**Figure 3.3: Relationship between  $\Delta t$  and the scaling exponent for the empirical tails of return distributions.** (a) We investigate the relationship between the time lag between price observations used to build the returns distribution and the scaling exponents of the tails of distributions. We consider here the tails of the positive component of the distributions obtained when analysing all trading days present in our dataset. We find that the mean scaling exponent increases as  $\Delta t$  increases (Adjusted  $R^2 = 0.802$ ,  $N = 12$ ,  $p < 0.001$ , ordinary least squares regression) (b) In a similar fashion, we observe that when analysing all trading days the mean scaling exponent for the tail of the negative component of the distributions increases with the time lag (Adjusted  $R^2 = 0.839$ ,  $N = 12$ ,  $p < 0.001$ , ordinary least squares regression) (c) We now restrict our analysis to trading days on which the prices of stocks have changed by more than 1%. We find that the mean scaling exponent of positive tails consistent with power law behaviour increases with  $\Delta t$  (Adjusted  $R^2 = 0.856$ ,  $N = 12$ ,  $p < 0.001$ , ordinary least squares regression) (d) Under 1% stress, an increase in the time lag  $\Delta t$  results again in an increase of the mean scaling exponent for the tails of the negative returns distributions (Adjusted  $R^2 = 0.729$ ,  $N = 12$ ,  $p < 0.001$ , ordinary least squares regression) (e) We now perform the same analysis for days on which the prices of stocks have changed by more than 2%. The mean scaling exponent for the tails of the positive component of the distributions again shows an increase with increasing  $\Delta t$  (Adjusted  $R^2 = 0.782$ ,  $N = 12$ ,  $p < 0.001$ , ordinary least squares regression) (f) Similarly, the mean scaling exponent for the tails of negative returns distributions at the 2% stress level increases as the time lag  $\Delta t$  between price observations increases (Adjusted  $R^2 = 0.836$ ,  $N = 12$ ,  $p < 0.001$ , ordinary least squares regression).



**Figure 3.4: Consistency of empirical returns distributions with power law and exponential decay.** (a) For  $\Delta t > 60$  minutes, we note a decrease in the number of tails consistent with power law decay. We investigate whether the tails of the returns distributions are consistent with power law behaviour or exponential decay using the Kolmogorov-Smirnov statistic, as described in the methods section. We first consider all trading days present in our dataset. At short time scales, we observe that the tails of most empirical distributions are consistent with power law behaviour. As we increase the time lag, the number of tails consistent with power law behaviour decreases and we see an increase in the number of tails of returns distributions that are consistent with exponential decay. We depict here the overall number of tails, both for the positive and negative returns distributions, for the 25 components of the DJIA. (b) We consider transaction days on which the prices of stocks have changed by more than 1%. We refer to this as a stress level of 1%. In this scenario, the number of tails consistent with power law decreases more sharply. Consistency with exponential decay appears when  $\Delta t$  is roughly 2 hours. (c) In a similar fashion, when we consider a stress level of 2%, we again observe a sharp decrease in the number of distributions consistent with power law behaviour. We also find an increase in the number of tails consistent with exponential decay again when  $\Delta t$  is roughly 2 hours. (d) Under a stress level of 3%, the number of empirical distributions consistent with power law behaviour decreases more quickly than in the other scenarios. The number of tails consistent with exponential decay peaks at a lower number than in other scenarios, but is again highest when  $\Delta t$  is roughly two hours, similar to other scenarios.

law behaviour or exponential decay. Using all trading days, the tail of most distributions is consistent with power law behaviour at small time scales. As we increase the time lag between price observations, we observe an increase in the number of tails consistent with exponential decay. At the 1% stress level, the decrease in the number of tails consistent with power law is sharper and we observe a peak in the number of tails consistent with exponential decay when  $\Delta t$  is roughly 2 hours.

As we increase the stress level, the number of tails consistent with power law behaviour decreases even more sharply. The number of tails consistent with exponential decay exhibits a peak at similar time scales, but peaks at a lower number than observed at the 1% stress level (Fig. 3.4c and 3.4d).

## 3.2 Conclusions

Large changes in stock market prices can occur at a range of time scales, arising within minutes or developing across longer time scales. Our findings provide evidence that in different scenarios the scaling exponent of those distributions consistent with power law behaviour increases with the time lag between price observations. As this time lag increases, we observe that the number of return distributions consistent with power law behaviour decreases sharply. At a time lag of roughly two hours, we also find an increase in the number of distributions which are consistent with exponential decay. Our results are consistent with the hypothesis that changes in stock market prices have different behaviours at different time scales. We observe that these results hold in different scenarios of the market, both when we consider all trading days, but also when restricting our analysis to scenarios with different stress levels. We suggest that our analysis may provide further empirical insights for the development of models of market behaviour.

## CHAPTER 4

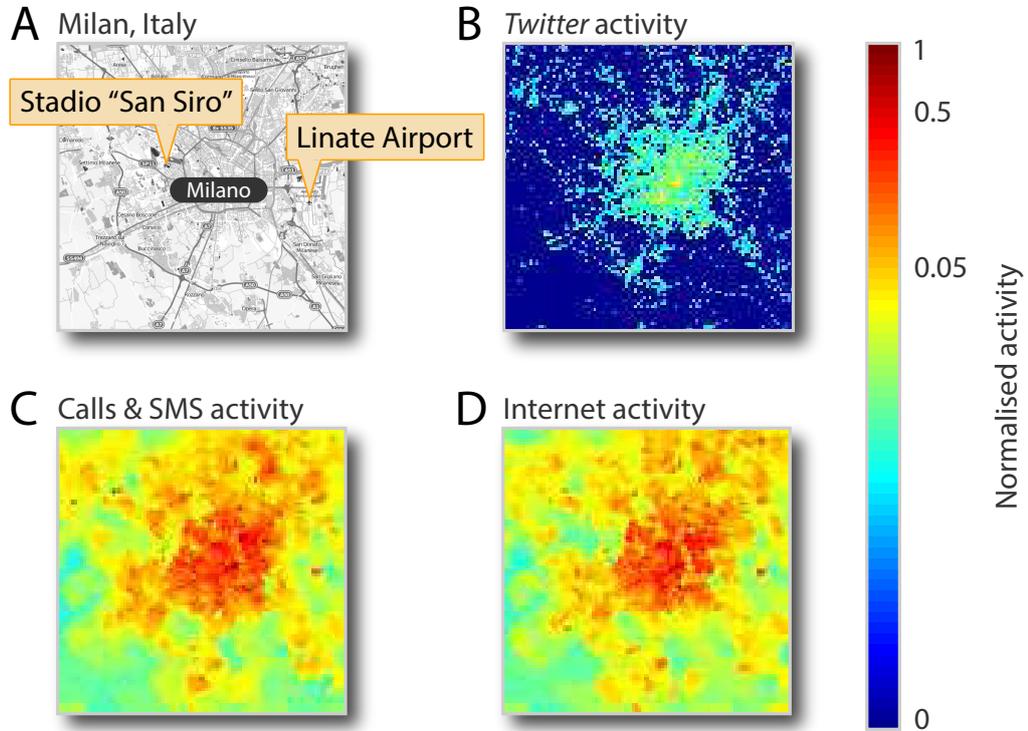
# QUANTIFYING CROWD SIZE USING MOBILE PHONE AND TWITTER DATA

The ability to quickly and accurately estimate the size of a crowd is crucial in facilitating emergency evacuations and avoiding crowd disasters [214]. However, existing approaches which rely on human analysts counting samples of the crowd can be time consuming or costly [215]. Similarly, image processing solutions require image data to be available in which members of the crowd can be identified and counted by an algorithm [216–218]. Here, we investigate whether data on mobile phone usage and usage of the online social media service *Twitter* can be used to estimate the number of people in a specific area at a given time. We consider data resulting from ordinary use of smartphones, without the need for users to install specific applications on their mobile phone.

### 4.1 Data

We retrieve data on mobile phone and *Twitter* activity recorded in the city of Milan and surroundings in a period covering two months from 1 November 2013 to 31 December 2013. Both datasets describe activity in the geographic area depicted in Fig. 4.1A. A detailed description of how the dataset was constructed can be found in [81].

**Twitter data** We retrieved the complete set of geo-localised tweets posted in Milan and surroundings between 1 November 2013 and 31 December 2013 from <http://www.telecomitalia.com/bigdatachallenge> as part of the *Big Data Chal-*



**Figure 4.1: *Twitter*, calls and SMS, and Internet activity in Milan.** (A) We analyse *Twitter*, calls and SMS, and Internet activity data recorded from mobile phones in the city of Milan and surroundings. The geographic area around Milan for which all these datasets are available is represented in this map, created using data from *OpenStreetMap*. The datasets cover the period from 1 November 2013 to 31 December 2013. We aim to determine whether such mobile phone data can be used to infer the number of people in a specific location at a specific time. To calibrate our model, we consider two case studies: *San Siro* football stadium and *Linate Airport*. (B) We depict the normalised number of tweets recorded during the first week of November 2013, for the geographic area shown in A. Tweet counts are extracted from the full set of geolocalised tweets sent during this period. We observe a higher density of tweets in the centre of Milan. (C) We depict normalised data on the total number of calls made and received as well as text messages (SMS) sent and received during the time interval between 08:20 and 08:30 of 1 November 2013, for the geographic area depicted in A. We again observe more activity in the centre of Milan. (D) We depict normalised data on the number of requests made by mobile phones to access the Internet during the time interval between 08:20 and 08:30 of 1 November 2013, for the geographic area shown in A. Visual inspection of this dataset provides further evidence that more mobile phone activity is recorded in locations where greater numbers of people would be expected. Colours in B, C and D are normalised to the maximum recorded activity level in each dataset.

lence set up by *Telecom Italia*.

The *Twitter* dataset consists of all messages sent via *Twitter* ("tweets"), with associated geographic coordinates located within the area shown in Fig. 4.1A. Tweets are also timestamped. Initial visual inspection of the *Twitter* data shows that greater numbers of tweets are recorded in the centre of Milan, where we would expect greater numbers of people to be found (Fig. 4.1B).

**Mobile phone data** The mobile phone activity dataset describes the volume of calls made and received, SMSs sent and received and Internet connections opened, closed and maintained. Mobile phone activity measurements are provided at ten minute granularity, for cells in a discrete grid superimposed on the area of Milan. This grid has 10,000 cells of size  $235\text{ m} \times 235\text{ m}$ . Visual inspection of the distribution of call and SMS activity (Fig. 4.1C) and Internet connection activity (Fig. 4.1D) again confirms mobile phone activity is highest in the city centre of Milan. Data on mobile phone call, SMS and Internet activity in Milan and surroundings from 1 November 2013 until 31 December 2013 were retrieved from <http://www.telecomitalia.com/bigdatachallenge> as part of the *Big Data Challenge* set up by *Telecom Italia*.

Interactions with the *Telecom Italia* mobile network generate Call Detail Records (CDRs). In the dataset we consider, *Telecom Italia* provides data on CDRs relating to the following activities:

- SMS: a CDR is generated for every SMS which is sent and every SMS which is received
- Calls: every incoming and outgoing call generates a CDR
- Internet access: a CDR is generated for each of the following events:
  - An Internet connection is opened
  - An Internet connection is closed
  - An Internet connection is open and 15 minutes has passed since the last CDR
  - An Internet connection is open and 5 MB have been transferred since the last CDR

For privacy reasons, the values which *Telecom Italia* provides are rescaled using an unknown factor. *Telecom Italia* specifies that counts of mobile phone call and SMS

**Table 4.1: Full names of football teams.** The full names of the football teams referred to in our analysis.

Abbreviation	Full Name
Milan	A.C. Milan
Inter	F.C. Internazionale Milano
Fiorentina	A.C.F. Fiorentina
Livorno	A.S. Livorno Calcio
Italy	Italy National Football Team
Germany	Germany National Football Team
Genoa	Genoa C.F.C.
Sampdoria	U.C. Sampdoria
Trapani	Trapani Calcio
Parma	Parma F.C.
Ajax	AFC Ajax
Roma	A.S. Roma

CDRs are rescaled using the same factor, and are therefore comparable. Counts of Internet activity CDRs are rescaled using a different factor.

**Football match attendees** We retrieved football match attendance figures from the following websites:

- Seven of the ten games which took place during the period of analysis were part of the Italian National Football League ‘Serie A’. We retrieved attendance figures from the official website of the ‘Serie A’: [www.legaseriea.it/it/lega-calcio/regolamenti-e-documenti/dati-statistici-su-incassi-e-spettatori](http://www.legaseriea.it/it/lega-calcio/regolamenti-e-documenti/dati-statistici-su-incassi-e-spettatori)
- Attendance figures for the three remaining games that took place during this period were retrieved from the following URLs of two online newspapers:

– <http://www.calciomercato.com/news/inter-trapani-3-2-il-tabellino-919259>

– <http://www.milannews.it/il-match/quasi-cinquantamila-spettatori-riempiono-san-siro-105483>

– <http://www.milannews.it/il-match/milan-ajax-superati-i-61mila-spettatori-a-san-siro-130840>

**Airport data** Flight schedule data for *Linate Airport* can be retrieved from <http://www.milanolate-airport.com/it> for the current date and the following four days. In May 2014, we retrieved the flight schedule for the week between Monday 5 May 2014 and Sunday 11 May 2014. We assume that weekly flight schedules

**Table 4.2: *San Siro* football match attendance figures.** Attendance figures for the football matches analysed.

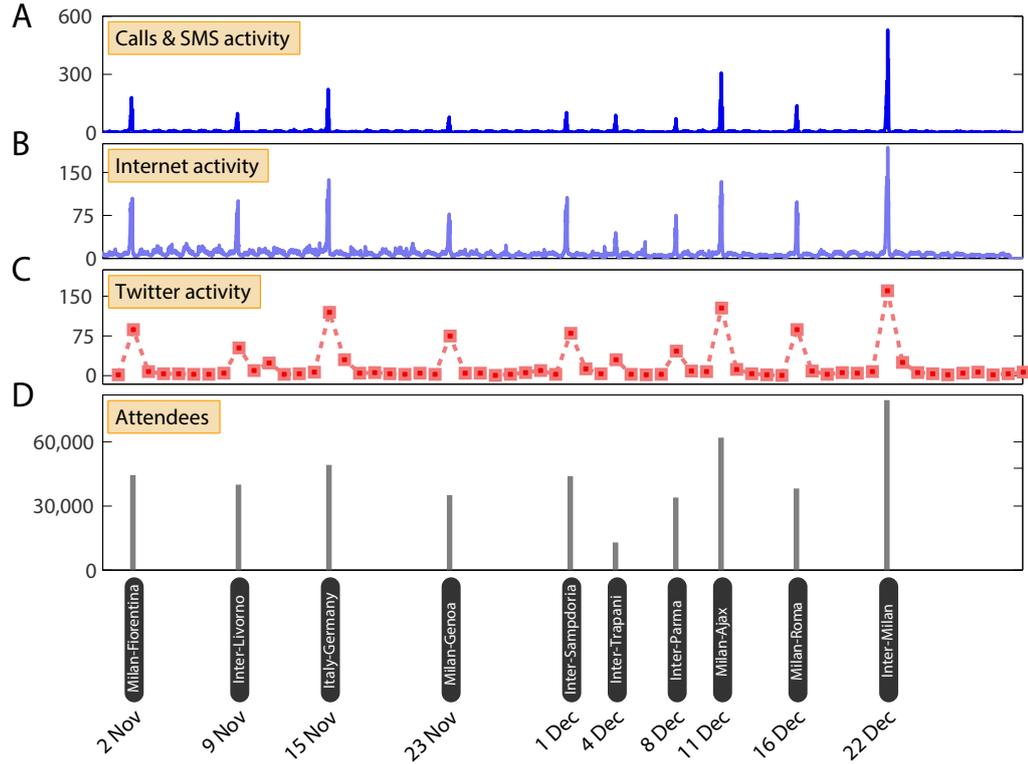
<b>Match</b>	<b>Attendees</b>
Milan-Fiorentina	44261
Inter-Livorno	39775
Italy-Germany	49000
Milan-Genoa	34848
Inter-Sampdoria	43607
Inter-Trapani	12714
Inter-Parma	33732
Milan-Ajax	61744
Milan-Roma	37987
Inter-Milan	79311

are reasonably constant across time, and use the schedule retrieved for this week as a proxy for the flight schedule in the weeks between 1 November 2013 and 31 December 2013.

## 4.2 Results

We investigate whether the information present in these datasets can be used to infer the number of people in specific areas of Milan at a given time. To calibrate our model, we consider two case studies of access restricted areas for which relevant data exist: *San Siro* football stadium, for which we have attendance counts for ten football matches which took place during the period of analysis, and *Linate Airport*, for which we use flight schedule data to create a proxy indicator for the number of people present in the airport at any given time.

We examine the time series of call and SMS activity (Fig. 4.2A), Internet activity (Fig. 4.2B) and *Twitter* activity (Fig. 4.2C) recorded in the vicinity of the football stadium *Stadio San Siro* during the period of analysis between 1 November 2013 and 31 December 2013. The coordinates of the area for which data were analyzed is given in Tables 4.3 and 4.4. In all three time series, we observe ten distinct spikes, which occur at the same times across all time series. We find that the dates on which these spikes occur coincide exactly with the dates on which the ten football matches took place in the stadium during this period (Fig. 4.2D). Furthermore, we note that the relative sizes of the spikes in the mobile phone and *Twitter* activity time series



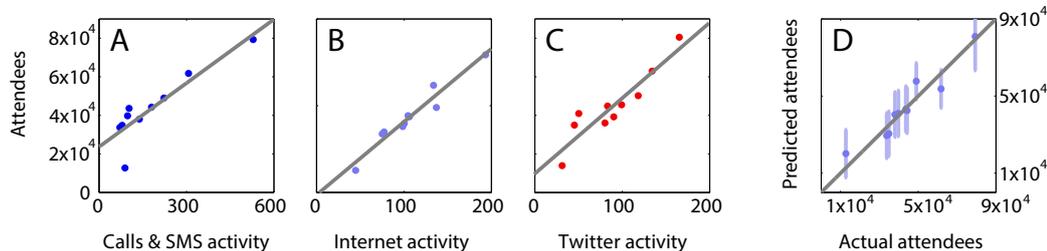
**Figure 4.2: Mobile phone and *Twitter* activity in football stadium *Stadio San Siro*.** (A) We depict the time series of mobile phone call and SMS activity recorded in the cell in which the football stadium *Stadio San Siro* is located, during the period of analysis between 1 November 2013 and 31 December 2013. The time series is plotted at 10 minute granularity. (B) Similarly, we depict the time series of Internet connection activity in the cell in which *Stadio San Siro* is located, at 10 minute granularity. (C) Finally, we depict the daily counts of tweets recorded within the vicinity of the stadium. (D) We determine the dates of football matches which took place during this period, and plot the number of attendees which were recorded at each of these matches. Visual inspection reveals a remarkable alignment between the spikes that can be observed in the communication activities and the dates on which football matches took place. The heights of the spikes bear a strong similarity to the number of attendees at each match.

**Table 4.3: Coordinates of the area around *San Siro* for which data on phone calls, SMS and Internet activity was retrieved.** This corresponds to one cell in the *Telecom Italia* dataset. Coordinates are specified using the WGS84 coordinate system. Note that the *Telecom Italia* cells do not appear precisely square using this system.

Corner	Latitude	Longitude
Top left	45.4793078474071	9.12276821006816
Top right	45.479304576233446	9.125775032395008
Bottom right	45.477189306362206	9.125770326481447
Bottom left	45.477192577295924	9.122763616655133

**Table 4.4: Coordinates of the area around *San Siro* for which *Twitter* data was retrieved.** This area constitutes a bounding box around the *San Siro* stadium, including the entrance gates. Coordinates are given using the WGS84 geographic coordinate system.

Corner	Latitude	Longitude
Top left	45.480084	9.121097
Top right	45.480084	9.125807
Bottom right	45.476308	9.125807
Bottom left	45.476308	9.121097



**Figure 4.3: Comparing football match attendance figures to mobile phone and *Twitter* activity.** (A) We investigate whether there is a relationship between the number of people attending each football match and the recorded mobile phone call and SMS activity inside the stadium. We find a linear relationship between these two variables (Adjusted  $R^2 = 0.771$ ,  $N = 10$ ,  $p < 0.001$ , ordinary least squares regression). (B) Similarly, we find a pattern consistent with a linear relationship between Internet connection activity in the stadium and the number of attendees at each match (Adjusted  $R^2 = 0.937$ ,  $N = 10$ ,  $p < 0.001$ , ordinary least squares regression). (C) We also observe a linear relationship between *Twitter* activity in the stadium and the number of match attendees (Adjusted  $R^2 = 0.855$ ,  $N = 10$ ,  $p < 0.001$ , ordinary least squares regression). (D) We explore whether this relationship could be exploited to infer the number of attendees from communication data if no other measurements were available. Using data on Internet activity, we build a linear regression model using only nine out of the ten football matches and then predict the attendance at the tenth match. We then repeat this leaving a different match out every time. Here, we plot the resulting estimates and their 95% prediction intervals. We find that the actual number of attendees falls within the 95% prediction interval for all ten matches.

(Fig. 4.2A-C) bear a strong similarity to the relative sizes of the attendance counts for these matches, as depicted in Fig. 4.2D.

We extract the maximum values of the spikes in calls and SMS activity, Internet activity and *Twitter* activity. We observe a linear relationship between the number of people attending the football matches and the volume of incoming and outgoing phone calls and SMS messages (Adjusted  $R^2 = 0.771$ ,  $N = 10$ ,  $p < 0.001$ , ordinary least squares regression; Fig. 4.3A). We find similar relationships between the number of attendees and both Internet activity (Adjusted  $R^2 = 0.937$ ,  $N = 10$ ,  $p < 0.001$ , ordinary least squares regression; Fig. 4.3B) and *Twitter* activity (Adjusted  $R^2 = 0.855$ ,  $N = 10$ ,  $p < 0.001$ , ordinary least squares regression; Fig. 4.3C). While Fig. 4.3A-C suggest a strongly linear relationship between mobile phone activity data and the number of attendees, we note that this relationship holds in a non-parametric analysis too (calls and SMS activity: Spearman's  $\rho = 0.927$ ,

$N = 10$ ,  $p < 0.001$ ; Internet activity: Spearman's  $\rho = 0.976$ ,  $N = 10$ ,  $p < 0.001$ ; *Twitter* activity: Spearman's  $\rho = 0.924$ ,  $N = 10$ ,  $p < 0.001$ ).

We investigate the possibility of using the information present in communication data to infer the number of attendees in situations where no other measurements are easily accessible. As an example, we consider data on Internet activity, for which the relationship with the number of recorded attendees was strongest. We carry out a *leave-one-out cross-validation* analysis as follows: for each of the ten attendance figures, we build a linear regression model based on the remaining nine attendance figures and the corresponding Internet activity data. We then use this model to generate an estimate of the attendance figure which was removed from the recorded Internet activity data. In Fig. 4.3D, we plot the resulting estimates and their 95% prediction intervals. We find that the actual attendance figure is always within the 95% prediction interval of our estimate.

We note that our analysis of mobile phone activity data may be affected by capacity constraints, such as signal truncation, on mobile phone communication in the stadium. Should data on such constraints become available in the future, the influence of these constraints on the relationship between communication data and crowd size may merit further analysis.

We perform a parallel analysis of the relationship between mobile phone and *Twitter* data and the number of passengers at *Linate Airport*. To estimate the number of people in *Linate Airport* at any given hour during the analysis period, we assume that passengers may arrive at the airport up to two hours before a departing flight, and depart within an hour following a flight arrival. For each hour, we therefore calculate the number of flights departing in the following two hours or arriving in the previous hour, and use this as a proxy indicator for the number of passengers in the airport. We base our calculations on one week of flight schedule data from May 2014, as explained in the data description above, and assume that weekly flight schedules are relatively constant. Our proxy indicator is therefore calculated for each of the 168 hours in a week. We omit the three initial days and two final days of the analysis period to create a period of exactly eight weeks. We compare this proxy indicator to the average mobile phone call and SMS activity and to the average Internet activity recorded for each hour in a week, in the cells in which the airport is located, as detailed in Table 4.5. We find that greater phone call and SMS activity corresponds to a greater estimated number of passengers (Adjusted  $R^2 = 0.175$ ,  $N = 168$ ,  $p < 0.001$ , ordinary least squares regression; Fig. 4.4A). Similarly, we

**Table 4.5: Coordinates of the area around *Linate Airport* for which data on phone calls, SMS and Internet activity was retrieved.** This corresponds to a square of nine cells in the *Telecom Italia* dataset, centered around the airport. Coordinates are given using the WGS84 geographic coordinate system.

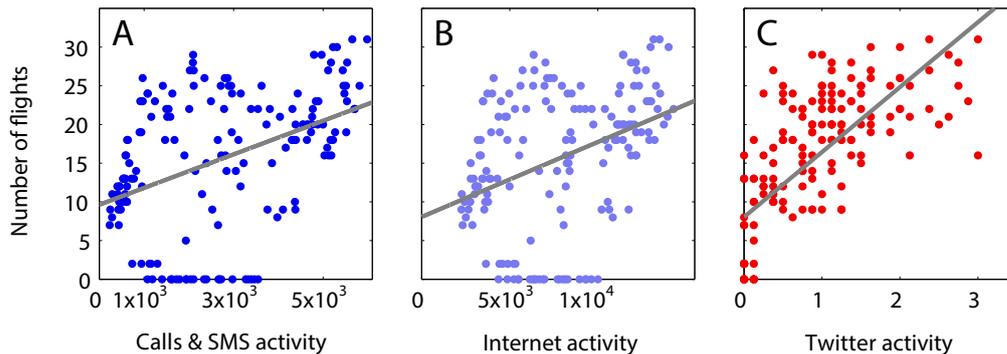
Corner	Latitude	Longitude
Top left	45.464233335498925	9.27604292987004
Top right	45.464211186534364	9.285060903904881
Bottom right	45.45786542106381	9.285028927452782
Bottom left	45.45788756515503	9.276011964969305

**Table 4.6: Coordinates of the area around *Linate Airport* for which *Twitter* data was retrieved.** This area corresponds to the square of nine cells around the airport, but corner coordinates are modified slightly to produce a square area under the WGS84 coordinate system.

Corner	Latitude	Longitude
Top left	45.464233335498925	9.276011964969305
Top right	45.464233335498925	9.285060903904881
Bottom right	45.45786542106381	9.285060903904881
Bottom left	45.45786542106381	9.276011964969305

find that greater Internet activity relates to a higher estimated number of passengers (Adjusted  $R^2 = 0.143$ ,  $N = 168$ ,  $p < 0.001$ , ordinary least squares regression; Fig. 4.4B). The relationships we find are weaker than those found in the previous case study, but remarkable given the coarse nature of our estimate of the number of passengers. We analyse *Twitter* activity in the area of the airport detailed in Table 4.6. In this case, we observe a stronger relationship between the estimated number of passengers and activity on *Twitter* (Adjusted  $R^2 = 0.510$ ,  $N = 168$ ,  $p < 0.001$ , ordinary least squares regression; Fig. 4.4C). We observe that mobile phone, SMS and Internet activity is still recorded when no flights take place, generally during night-time periods. In contrast, few tweets are logged at these times, potentially explaining the greater strength of this relationship.

We note that roughly 58% of the passengers travelling to and from *Linate Airport* are Italian [219]. Given the current costs of using mobile phone networks abroad, the mobile phone activity analysed here may reflect the behaviour of Italian passengers more strongly than the behaviour of international passengers.



**Figure 4.4: Parallel analysis of the relationship between mobile phone and *Twitter* data and the number of passengers at *Linate Airport*.** (A) We create a proxy indicator for the number of passengers at *Linate Airport* in each hour by calculating the number of flights departing in the following two hours or arriving in the previous hour. We compare this proxy indicator to the average mobile phone call and SMS activity recorded for each hour in a week, in the cells in which the airport is located. We find that greater activity corresponds to a greater estimated number of passengers (Adjusted  $R^2 = 0.175$ ,  $N = 168$ ,  $p < 0.001$ , ordinary least squares regression). The relationship we find is weaker than that found for the football attendance figures, but remarkable given the coarse nature of our estimate of the number of passengers. (B) We then explore the relationship between the proxy indicator of the number of passengers and Internet connection activity recorded in the cells in which the airport is located. Again, we find that greater Internet activity corresponds to a higher number of passengers (Adjusted  $R^2 = 0.143$ ,  $N = 168$ ,  $p < 0.001$ , ordinary least squares regression). (C) As a final example, we consider *Twitter* activity recorded in the cells in which the airport is located. Again, we consider the average number of tweets recorded during each of the 168 hours in a week, over the 8 week period of our analysis. Here, we find a stronger relationship between the estimated number of passengers and activity on *Twitter* (Adjusted  $R^2 = 0.510$ ,  $N = 168$ ,  $p < 0.001$ , ordinary least squares regression).

### 4.3 Conclusions

Our results provide evidence that accurate estimates of the number of people in a given location at a given time can be extrapolated from mobile phone or *Twitter* data, without requiring users to install further applications on their smartphones. As well as being of clear practical value for a range of business and policy stakeholders, our findings suggest that data generated through our interactions with mobile phone networks and the Internet may allow us to gain valuable measurements of the current state of society.

## CHAPTER 5

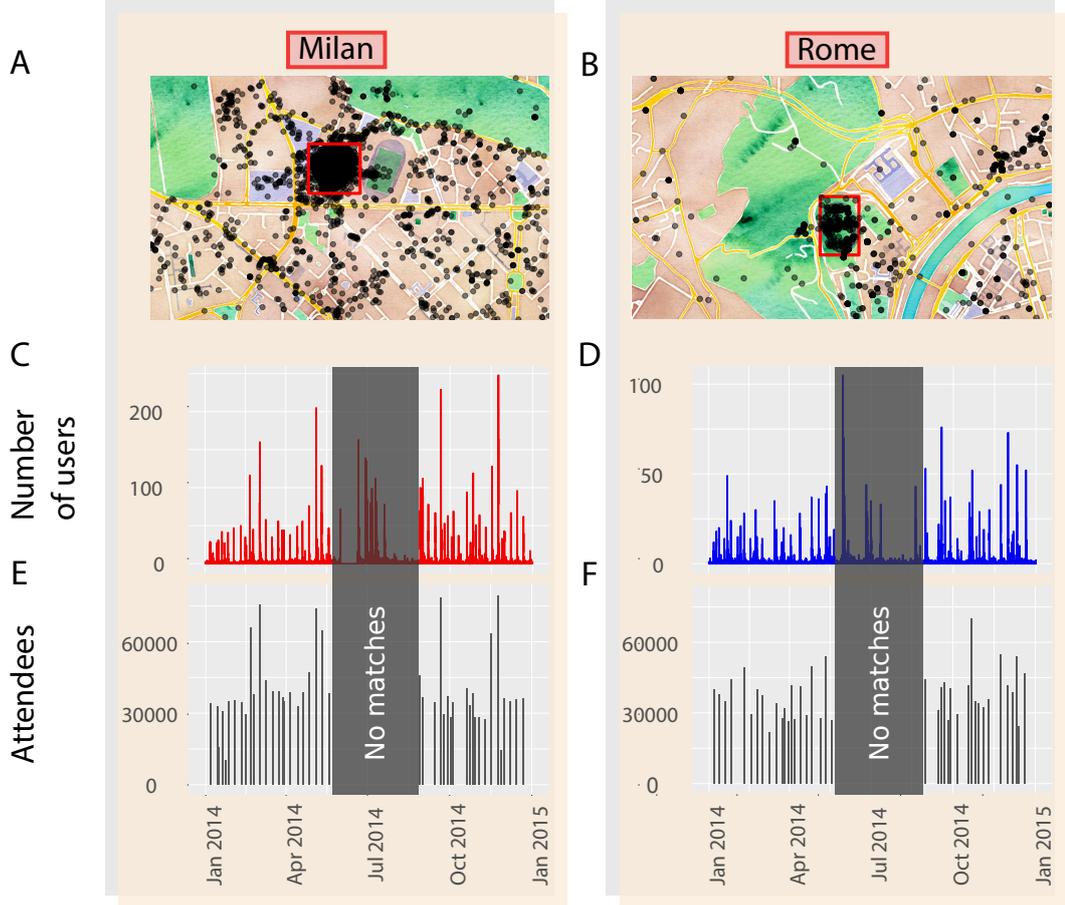
# MEASURING CROWD SIZE USING INSTAGRAM PHOTOS

In Chapter 4, we showed that activities derived from our usage of smartphones and the social media platform *Twitter* can be used to infer the size of a crowd in a given location at a given time. However, our analysis was restricted to a two months period and only one location. Mobile phone records are owned by mobile phone providers and, therefore, are not widely available. Here, we want to assess whether the results described in Chapter 4 hold for datasets that are more easily accessed. For this reason, we consider publicly available data derived from the usage of the photo sharing platform *Instagram*. We aim to investigate whether analogous results hold for this platform and we also aim to study how the relationship between social media activity and crowd size varies in different locations.

### 5.1 Data

We investigate whether the activity of users on *Instagram* can be used to estimate the number of people in a specific location at a given time. In order to calibrate our model, we need a case study where we have accurate figures for the number of people present in a specific area. As we have seen in Chapter 4, football stadiums are access restricted areas for which attendance figures during football matches are publicly available.

We retrieved data on photos uploaded to the social media platform *Instagram* in large areas around two football stadiums in Milan and Rome between 1 January 2014 and 31 December 2014 using the publicly available *Instagram* API. We consider the San Siro football stadium in Milan, and the Stadio Olimpico football stadium in



**Figure 5.1: Activity of *Instagram* users in football stadiums in Milan and Rome** | (A-B) We collected data on geolocalised photos uploaded to the photo sharing platform *Instagram* in the proximity of two Italian football stadiums in Milan and Rome. The dataset covers the period from 1 January 2014 to 31 December 2014. We depict here the location of all of the photos uploaded to *Instagram* within the vicinity of the two stadiums in the time interval ranging from one hour before the beginning of a football match to three hours after the beginning. Visual inspection reveals a higher activity in the proximity of the stadiums. We aim to determine whether such activities can be used to infer the number of attendees at football matches. We depict in red the bounding boxes that we use in the subsequent analysis. These maps were created using map data from *OpenStreetMap* and tiles from *Stamen Design*. (C-D) We depict the time series of unique active users on *Instagram* recorded within the vicinity of the San Siro football stadium in Milan at one hour granularity. Similarly, we present the analogous time series in the vicinity of Stadio Olimpico football stadium in Rome. (E-F) We plot the number of officially recorded attendees at the football matches taking place in the two stadiums. Visual inspection suggests that peaks in the number of users within the stadiums align perfectly with dates when a football match took place. The size of the spikes in number of users also seems to correspond to the number of attendees. Regions shaded in grey correspond to dates when there was no football match but other events took place in the stadiums, such as concerts. For these, no official attendance figures are available and the corresponding events will be discarded in the analysis.

Rome, for which we have official attendance figures for all football matches that took place during the period of analysis. Both stadiums are home stadiums for two different teams: AC Milan and FC Internazionale in San Siro, and AS Roma and SS Lazio in Stadio Olimpico. The *Instagram* dataset consists of all photos uploaded to *Instagram* for which the geographical coordinates of the photo are available. The photos are also timestamped. Figures 5.1A and 5.1B depict the location of photos uploaded to *Instagram* within the vicinity of the two stadiums in a time window of four hours beginning one hour before the official starting time of a football match. Initial visual inspection shows a higher *Instagram* activity within the bounding boxes defined around the two stadiums.

We also retrieve official attendance figures for all football matches taking place during the period of analysis in the two football stadiums through official reports available on the webpage of the major Italian sports newspaper *La Gazzetta dello Sport* ([www.gazzetta.it](http://www.gazzetta.it)).

**Table 5.1: Coordinates of the bounding box around San Siro football stadium. Coordinates are given using the WGS84 geographic coordinate system.**

Corner	Latitude	Longitude
Top left	45.479350	9.121881
Top right	45.479350	9.125776
Bottom right	45.476717	9.125776
Bottom left	45.476717	9.121881

**Table 5.2: Coordinates of the bounding box around Stadio Olimpico football stadium. Coordinates are given using the WGS84 geographic coordinate system.**

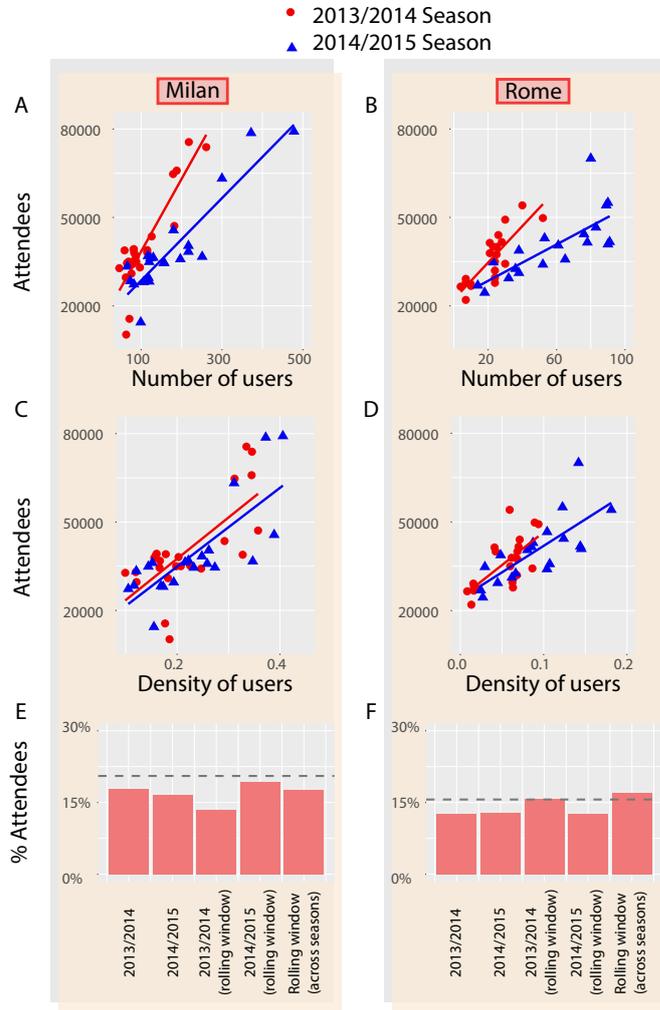
Corner	Latitude	Longitude
Top left	41.935546	12.453480
Top right	41.935546	12.456248
Bottom right	41.932417	12.456248
Bottom left	41.932417	12.453480

## 5.2 Results

We analyse the number of users who uploaded at least one photo to *Instagram* in the vicinity of the two football stadiums for the whole of 2014. The coordinates of the two areas used to extract the *Instagram* data are given in tables 5.1 and 5.2. Figures 5.1 C and 5.1 D depict the time series in the two stadiums at a granularity of one hour. In both time series we observe distinct spikes occurring throughout the year. Figures 5.1E and 5.1F present the number of attendees recorded at each football match taking place in the two stadiums. We note a strong similarity between the number of users on *Instagram* and the number of attendees at football matches. Regions shaded in grey represent dates when no football match took place but *Instagram* activity was recorded in the two stadiums. Further investigation of these periods shows that other events, such as concerts, took place in both stadiums during summer. For said events, no official attendance figures are available and for this reason they are not considered in the analysis.

We investigate the relationship between the number of active users on *Instagram* and the number of attendees at the corresponding football match. We consider a user to be active if they uploaded at least one photo on *Instagram* within a time window of four hours, starting one hour before the official starting time of a football match and within the geographical vicinity of the football stadium. The coordinates used to define the bounding boxes around the two stadiums are given in tables 5.1 and 5.2. Our analysis shows that a larger number of active *Instagram* users in the stadium corresponds to a larger number of attendees (San Siro:  $R^2 = 0.61$ ,  $N = 45$ ,  $p < 0.001$ ; Stadio Olimpico:  $R^2 = 0.47$ ,  $N = 40$ ,  $p < 0.001$ ; ordinary least-squares regression).

However, an analysis considering data for the whole year assumes that the number of active users on *Instagram* can be considered constant throughout the year. *Instagram*, however, is becoming an increasingly popular social media service and the number of registered users is constantly growing. We investigate this hypothesis by dividing the period of analysis in two parts: January 2014 to May 2014, corresponding to the last part of the 2013/2014 football season; and August 2014 to December 2014, corresponding to the first part of the 2014/2015 season. We investigate whether there is any difference between these two periods by fitting two separate models. Figures 5.2A and 5.2B depict the results of this analysis. We observe that a larger number of active users on *Instagram* corresponds to a larger



**Figure 5.2: Comparing football matches attendance figures to active users on *Instagram*** | We investigate the relationship between the number of people at football matches and the number of users uploading photos to *Instagram*.

We consider users that have uploaded at least one photo from within the stadium in a time window of four hours, starting one hour before the starting time of a football match. (A-B) In both stadiums and across seasons, an increase in number of users corresponds to a larger number of attendees (all  $p < 0.001$ , all  $R^2 \geq 0.57$ , all  $N > 19$ , ordinary least squares regression). We also observe that fewer *Instagram* users are found for each attendee at matches during the 2013/2014 season. This may be due to an overall increase in usage of the platform, or a change in the behaviour of the users. (C-D) We consider the number of *Instagram* users active in the football stadium normalised by the overall number of active users in a wide area around the football stadiums, and define this as the “density” of users. We find that a larger density of users inside the football stadiums corresponds to a linear increase in the number of attendees (all  $p < 0.001$ , all  $R^2 \geq 0.47$ , all  $N > 19$ ). (E-F) We investigate whether this relationship can be used to infer attendees from *Instagram* data alone, should no other measurements be available. We present the mean absolute percentage error of models built using all data from a given season (2013/2014; 2014/2015), a rolling window analysis for a given season, or a rolling window model that uses data from the whole year. The dashed line corresponds to the error found in a model that uses data from the whole period of analysis. We see that the rolling window analysis performs at least as well as that model.

**Table 5.3: Coordinates of the reference area around San Siro football stadium used to define the density of users inside the stadium. Coordinates are given using the WGS84 geographic coordinate system.**

Corner	Latitude	Longitude
Top left	45.527557	9.055555
Top right	45.527557	9.194536
Bottom right	45.427381	9.194536
Bottom left	45.427381	9.055555

**Table 5.4: Coordinates of the reference area around Stadio Olimpico football stadium used to define the density of users inside the stadium. Coordinates are given using the WGS84 geographic coordinate system.**

Corner	Latitude	Longitude
Top left	41.985084	12.387326
Top right	41.985084	12.521791
Bottom right	41.883495	12.521791
Bottom left	41.883495	12.387326

number of attendees in the stadium. This holds across the two stadiums and for both football stadiums (all  $R^2 \geq 0.57$ , all  $N > 19$ , all  $p < 0.001$ ; ordinary least squares regression). However, visual inspection reveals that a larger proportion of the attendees are active *Instagram* users at matches taking place in the 2014/2015 season (Fig. 5.2A and 5.2B), as can also be seen by the differences in the slopes of the fitted lines. This suggests that considering the number of users, or their behaviour, to be constant across the whole year may be inaccurate and that a more rigorous analysis should consider these variations.

If the number of users is increasing, this should also hold for areas other than the football stadium. This suggests that we could take this increase into account by considering the number of users that are inside the stadium divided by the number of users who are active in the same time window in an area used for reference.

We test this hypothesis by defining the *density of users* during a football match as the *number of active users active inside the bounding box divided by the number of active users in a much larger area around the stadium*. Specific coordinates for the reference areas in Milan and Rome are given in tables 5.3 and 5.4. As depicted

**Table 5.5:** We analyse the relationship between the density of users on *Instagram* active in the stadium and the number of attendees at football matches. We perform the analysis for two time periods to investigate whether the relationship between these quantities changes over the period of one year. We report here the 95% confidence interval for the estimated slopes of the regression models. We find that the slopes are consistent across seasons.

Stadium	Season	95% confidence interval for estimated slope
Milan	2013/2014	[110,883; 169,209]
Milan	2014/2015	[108,003; 159,217]
Rome	2013/2014	[168,201; 276,647]
Rome	2014/2015	[140,319; 218,019]

in Fig. 5.2C and 5.2D, we again find that a larger number of attendees corresponds to a larger density of users (all  $R \geq 0.47$ , all  $N > 19$ , all  $p < 0.001$ ; ordinary least squares regression). However, it is important to note that the parameters of the fitted models now change very little across seasons (table 5.5). Since the density of users takes does not change if the overall number of *Instagram* users varies, this supports our initial hypothesis that the number of users on *Instagram* is increasing over time.

### 5.2.1 Selecting an appropriate spatial area for analysis

Our analysis so far has considered a bounding box of fixed size around the football stadiums. However, at this stage it is not clear how this choice may affect the results and whether different considerations hold for different locations. If we consider a larger area, we may be able to capture more users that have been active within the proximity of the football stadium, but we may also introduce additional noise coming from users that are not attending the football match. Similarly, a smaller area would reduce the noise and only consider the users who are inside the stadium, but may not capture all the relevant information. We investigate how the strength of the relationship changes as we vary the size of the area considered to count users on *Instagram*. For each stadium, we define a circle of a given radius centred on the stadium, as depicted in Fig. 5.3A and 5.3B. Tables 5.6 and 5.7 report the coordinates used for the centre of the two stadiums. We then carry out the same analysis as before, counting the number of active users during football matches and comparing it to the number of attendees. For this analysis, we do not separate by season but we consider the entire period of analysis together. We vary the size of

**Table 5.6: Coordinates of the centre of San Siro football stadium. Coordinates are given using the WGS84 geographic coordinate system.**

Latitude	Longitude
45.478100	9.124000

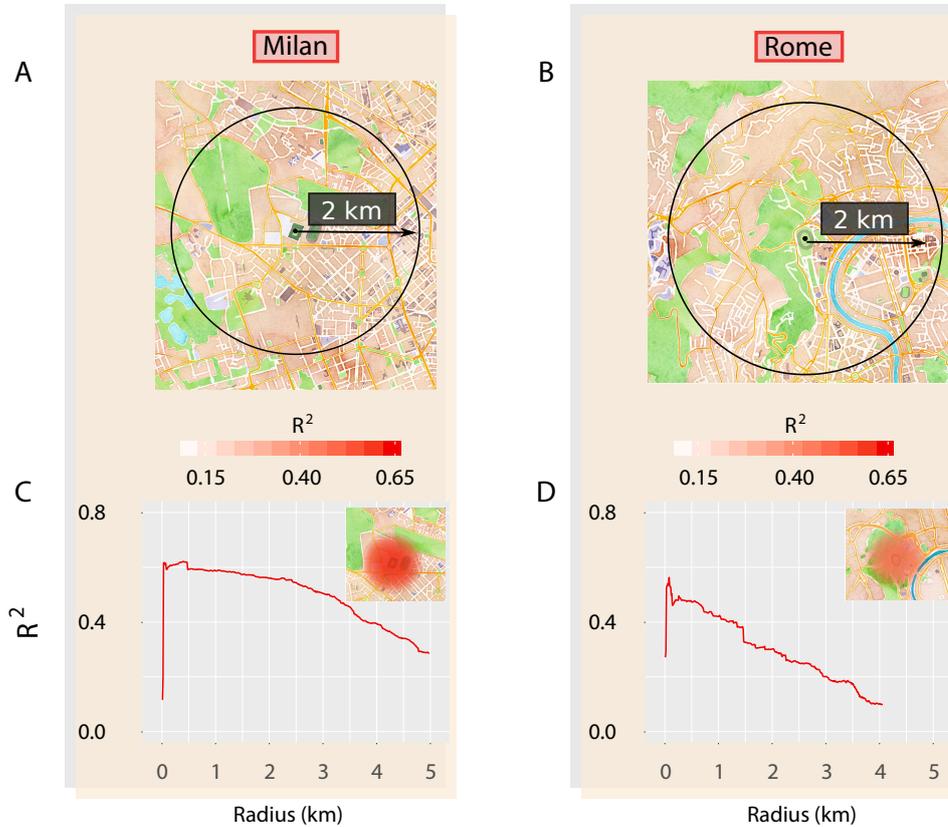
**Table 5.7: Coordinates of the centre of Stadio Olimpico football stadium. Coordinates are given using the WGS84 geographic coordinate system.**

Latitude	Longitude
41.934077	12.454730

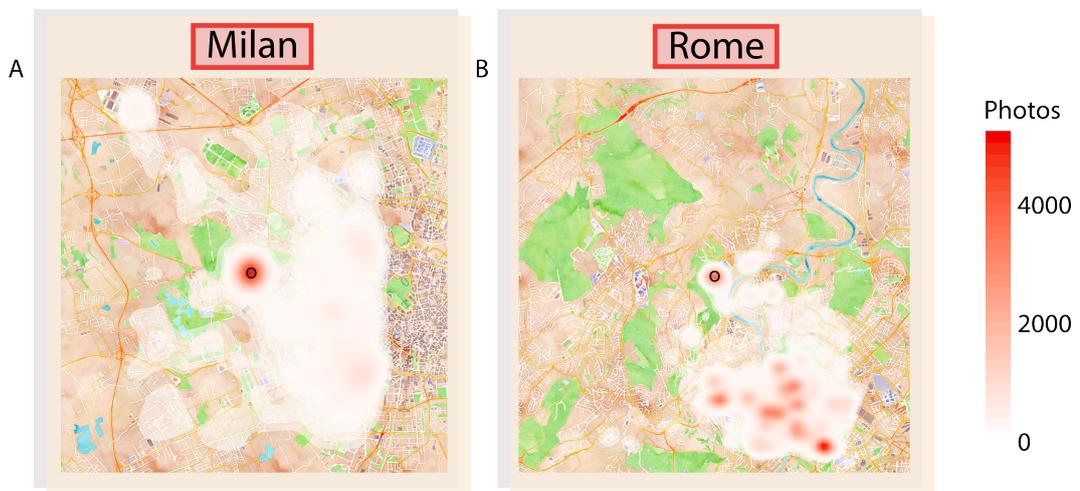
the radius from 10 metres to 5 kilometres. Figure 5.3 depicts the results of this analysis in the two stadiums. In Milan, the correlation (Fig. 5.3C) shows a smooth and slow decrease with the increasing radius. In Rome, the correlation exhibits a less smooth and faster change with the radius (Fig. 5.3D). This difference may arise from the different location in the corresponding cities of the two stadiums, with that in Rome being closer to areas with high density of tourists. Figure 5.4 depicts the spatial distribution of photos in a wide area centred on the two stadiums. Visual inspection provides a preliminary understanding of this difference. While in Milan we observe a rather homogeneous distribution of photos, with the football stadium being the only area with a large density of photos, in Rome we find several other locations with a large density. In particular, these locations correspond to important touristic sites in the city. This indicates that the choice of the bounding box should be carefully assessed by analysing the spatial location of the area when trying to infer crowd sizes from social media measures alone.

### 5.2.2 Training models using only historic data

In our previous analysis we have considered models fitted using all available data, either for the whole year or for a football season. In reality, however, we would only have access to data from matches that have already taken place. Here, we investigate whether we can use data from the last ten football matches to infer the number of attendees at the following match. We call this a rolling window analysis. For a given stadium and season, we fit a model using data from the last ten matches, and then predict the number of attendees at the following one based on the number of active *Instagram* users.



**Figure 5.3: Investigating the role of the bounding box** | (A-B) We investigate how the strength of the relationship varies as we change the size of the bounding box around the football stadiums. We consider concentric circles centred on the football stadiums and of increasing radius. For each radius, we consider users that have uploaded at least one photos to *Instagram* inside the corresponding circle and investigate the relationship between the number of users and the number of attendees at football matches. We examine radii varying from 10 metres to 5 kilometres in steps of 10 metres, and we only show results when the relationship is statistically significant ( $p < 0.05$ , ordinary least-squares regression). As before, the time window used to count users goes from one hour before the beginning of the football match, to three hours after the beginning. (C-D) We depict here how the coefficient of determination  $R^2$  varies when we increase the size of the circle around the two football stadiums. In the two insets, we present a map of how the correlation changes in the proximity of the two stadiums. We observe some differences in the two case studies: whereas in Milan the correlation decreases smoothly as we consider larger areas, in Rome we find a more fragmented change. This may be due to the different location of the two stadiums inside the city, with Rome’s stadium being close to tourist attractions from where *Instagram* users commonly upload photos. This analysis shows the importance of carefully assessing the location of the area in which the crowd is when calibrating the model.



**Figure 5.4: Spatial distribution of photos posted on *Instagram* during football matches** | We present here the spatial distribution of the photos posted on *Instagram* in a wide area around the football stadiums. Similarly to before, we consider photos uploaded in a time window extending from one hour before to three hours after the official starting time of the football matches. In Milan, we observe a distribution which is mostly concentrated around San Siro football stadium and is mostly homogeneous around it, with no other areas showing a large density of photos. In Rome, we find that the distribution around Stadio Olimpico football stadium is more fragmented and peaked around hotspots showing large densities of photos. Visual inspection shows that they correspond to several tourist attractions in Rome. Both maps were created using map data from *OpenStreetMap* and tiles from *Stamen Design*.

We measure the predictive accuracy using the symmetric mean absolute percentage error (SMAPE). We first define the mean absolute percentage error (MAPE) for the predicted values  $\hat{y}_i$  of a regression model with dependent variable  $y_i$  for  $n$  predictions:

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

However, the MAPE is not an ideal measure of predictive accuracy because it puts a heavier weight on negative errors. For this reason, it is often more useful to introduce its symmetric version, defined as:

$$\text{SMAPE} = \frac{1}{n} \sum_{i=1}^n \frac{|y_i - \hat{y}_i|}{(|y_i| + |\hat{y}_i|)/2}$$

We perform the rolling window analysis using data for the whole period of analysis, and we obtain a SMAPE of 17.5% in Milan and of 17% in Rome. If we carry out the rolling window analysis for each season separately, in Milan we obtain SMAPEs of 13.3% and 19.3% for the first and second season respectively, and of 15.7% and 12.6% in Rome, as is also depicted in Fig. 5.2E and 5.2F.

We want to compare this to models which using all available data in order to assess whether using only ten matches to calibrate the model significantly changes the predictive accuracy. To generate a comparable measure of predictive accuracy for models calibrated using all available data, we carry out a leave-one-out-cross-validation analysis as follows. First, we build a linear regression model leaving out one of the matches and considering all others. Then, we use this model to estimate the attendance figure at the match which was left out. We repeat this as many times as there are matches, so that each match is considered exactly once.

We find that models trained in this fashion, using all available match attendance data, exhibit a SMAPE of 20.5% in Milan and of 15.6% in Rome. We perform the analysis for each season separately, in Milan we obtain SMAPEs of 17.9% and 16.5% for the first and second season respectively, whereas in Rome we find errors of 12.5% and 12.9% (Fig. 5.2E and 5.2F). Comparing the results depicted in Fig. 5.2E and 5.2F, we observe that the prediction accuracies of the rolling window models are as good as those of models built using all data from the same football seasons. This is encouraging, since in reality we would not have access to data coming from matches that have not yet taken place.

Qualitatively similar results also hold if we consider the number of photos uploaded on *Instagram* instead of the number of active users. The use of the aggregated count of photos may be preferred in situations where privacy considerations would speak against counting individual *Instagram* users. More details on this can be found in Section 5.2.4.

### Further measures of predictive accuracy

Above, we presented one measure of predictive accuracy: the symmetric mean absolute percentage error (SMAPE). Other common measures of predictive accuracy are the root mean squared error (RMSE):

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n}}$$

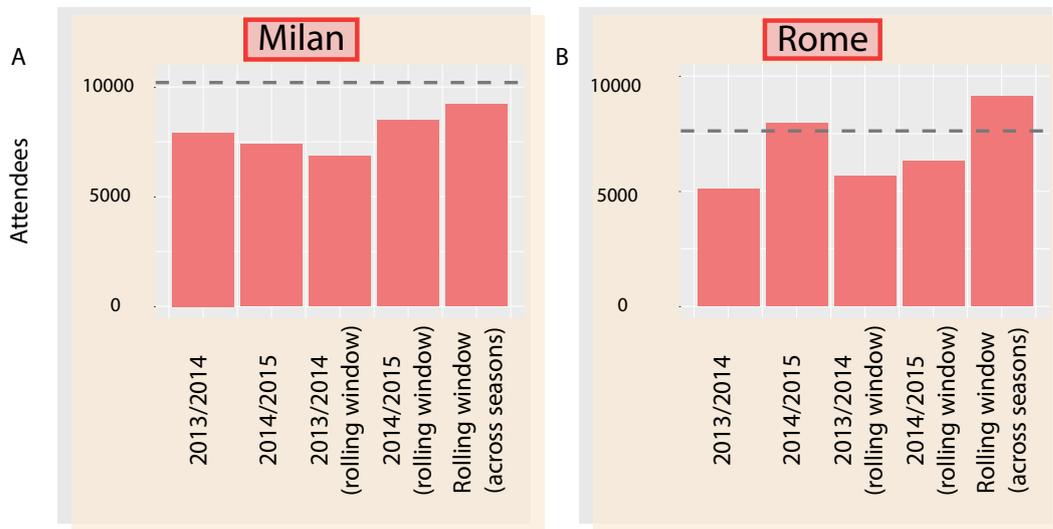
and the median absolute error (MAE):

$$\text{MAE} = \text{median}(|\hat{y}_i - y_i|)$$

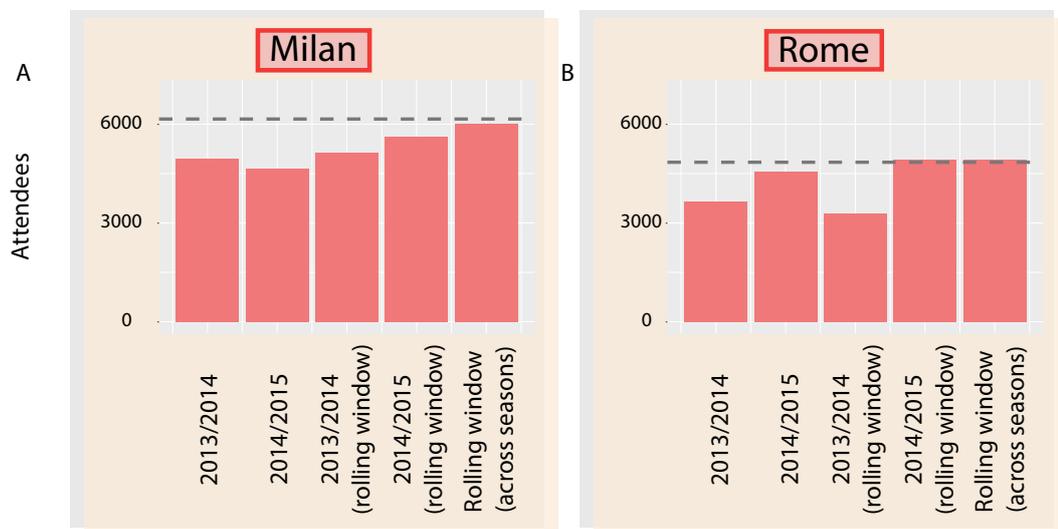
Figures 5.5 and 5.6 depict the RMSE and MAE for the analysis presented in Section 5.2. Comparing the results obtained with the RMSE (Fig. 5.5) in the different analyses, we again find that most prediction accuracies of the rolling window models are as good as those of models built using data from the whole period of analysis. Similarly, Fig. 5.6 depicts the results we find when using the MAE as measure of predictive accuracy. Results obtained with the rolling window analysis are comparable to those of other models.

### 5.2.3 Selecting an appropriate time windows for analysis

Our findings also depend on the time window used to count users who have been active on *Instagram* during a football match. A longer time window may capture users who are active before or after the match, but may also capture users who are in the proximity of the stadium for other reasons thus introducing additional noise. So far, we have counted users who were active at least once in a window of four hours starting one hour before the match. We now consider time windows of varying lengths, which start at different times. We pick a starting time and a length, count the number of users who are active on *Instagram* in that time interval, and then compare it to the number of attendees recorded at the match. Figure 5.7 depicts

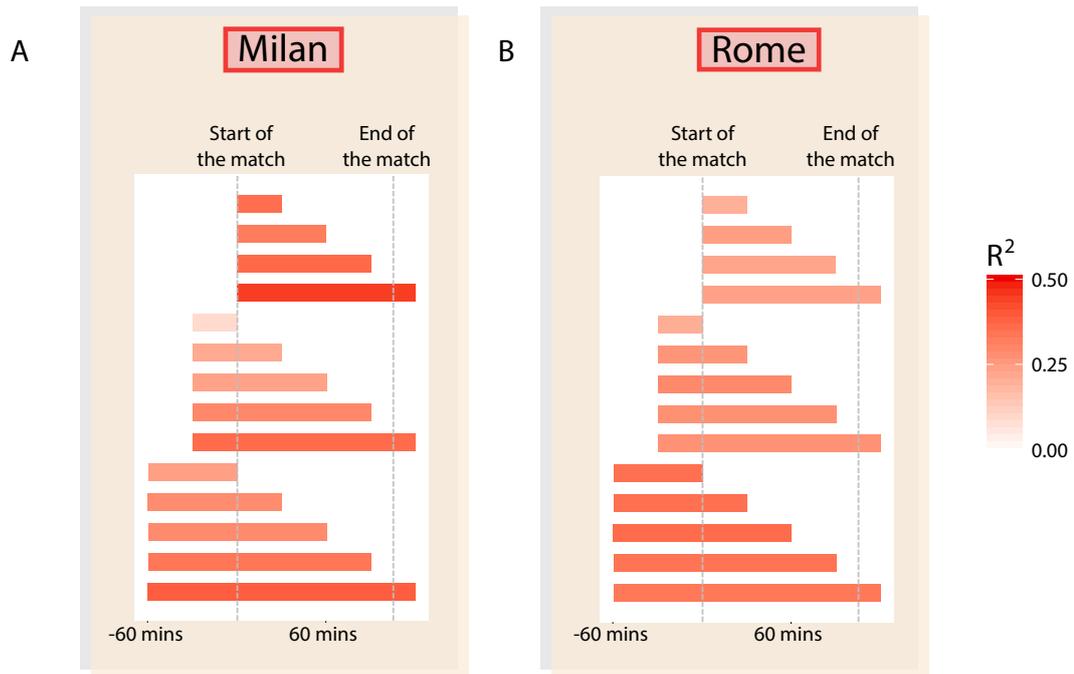


**Figure 5.5: Root mean squared error in the two stadiums** | We investigate whether the relationship between *Instagram* users counts and number of attendees can be used to infer the number of people attending a football match. We perform leave-one-out-cross-validation on models using all data from a given season (2013/2014; 2014/2015), a rolling window analysis for a given season, or a rolling window model that uses data from the whole year. Here, we present the root mean squared error of the various models considered. The dashed line corresponds to the error found in a model that uses data from the whole period of analysis. We see that the rolling window analysis performs at least as well as that model, with the exception of the rolling window analysis using data from both seasons in Rome.

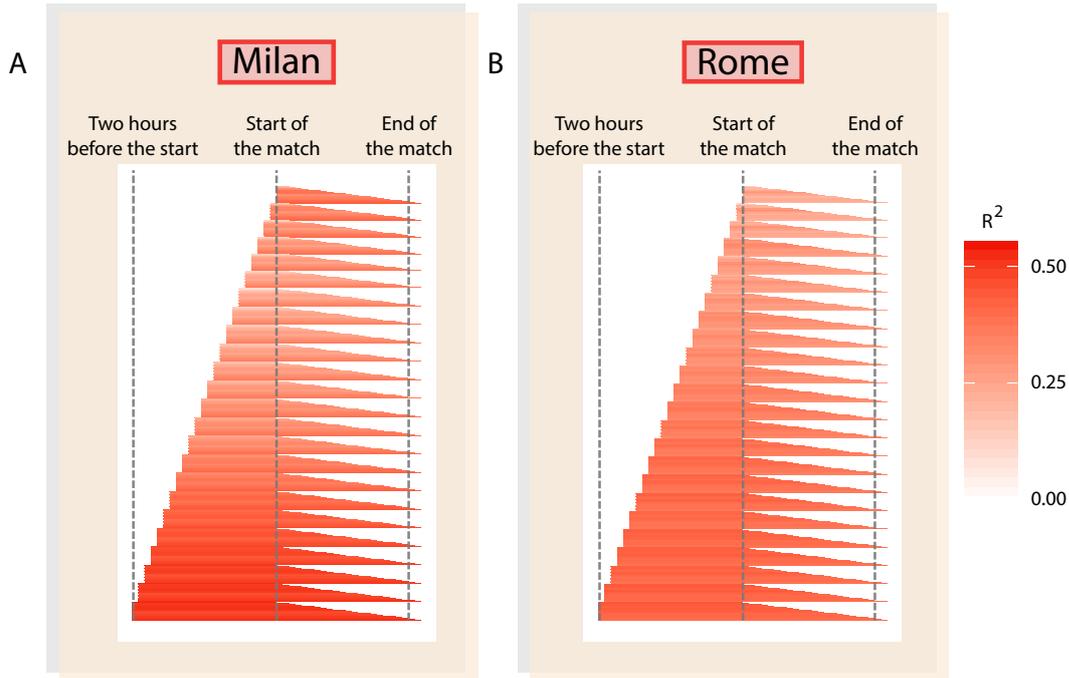


**Figure 5.6: Median absolute error in the two stadiums** | We investigate whether the relationship between *Instagram* users counts and number of attendees can be used to infer the number of people attending a football match. We perform leave-one-out-cross-validation on models using all data from a given season (2013/2014; 2014/2015), a rolling window analysis for a given season, or a rolling window model that uses data from the whole year. Here, we present the median absolute error of the various models considered. The dashed line corresponds to the error found in a model that uses data from the whole period of analysis. We see that the rolling window analysis performs at least as well as that model.

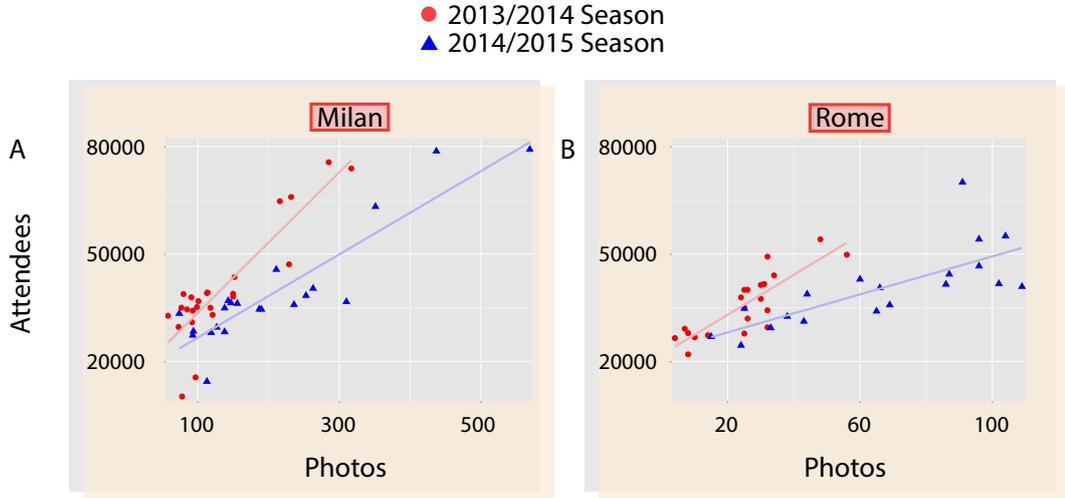
the results of this analysis. For each time window, we show its length and starting time, and the colour corresponds to the squared correlation between users' count and number of attendees. In Milan we find that the strength of the relationship increases when taking into account the entire match, whereas in Rome we observe that the time before the start of the match increases the strength of the relationship. Figure 5.8 depicts the same analysis at a higher temporal resolution. As before, we observe that in Milan the strength of the relationship increases when counting *Instagram* users active during the entire match. However, we also note that counting users who are active up to two hours before the match results in a stronger relationship with the number of attendees. In Rome, we again find that counting *Instagram* users active before the start of the match increases the strength of the relationship with the number of attendees in the stadium.



**Figure 5.7: Investigating the effect of the time window** | We want to investigate how the relationship changes as we change the size of the time window used to count users active on *Instagram* during a football match. In the figure, the bars extend from the starting point of the time window until the ending point. For instance, the top bar extends for 30 minutes starting at the beginning time of the match. The corresponding analysis counts all unique users who have been active on *Instagram* in that interval and compares it to the official number of attendees for that match. The colour of the bar indicates the squared correlation between the users' count and the number of attendees in the stadium. For this analysis, we consider a football match to be 105 minutes long, including a 15 minutes half-time break. In Milan the strength of the relationship increases if we consider *Instagram* users active during the entire match. In Rome, we find that counting *Instagram* users active before the start of the match results in a stronger relationship with the number of attendees in the stadium.



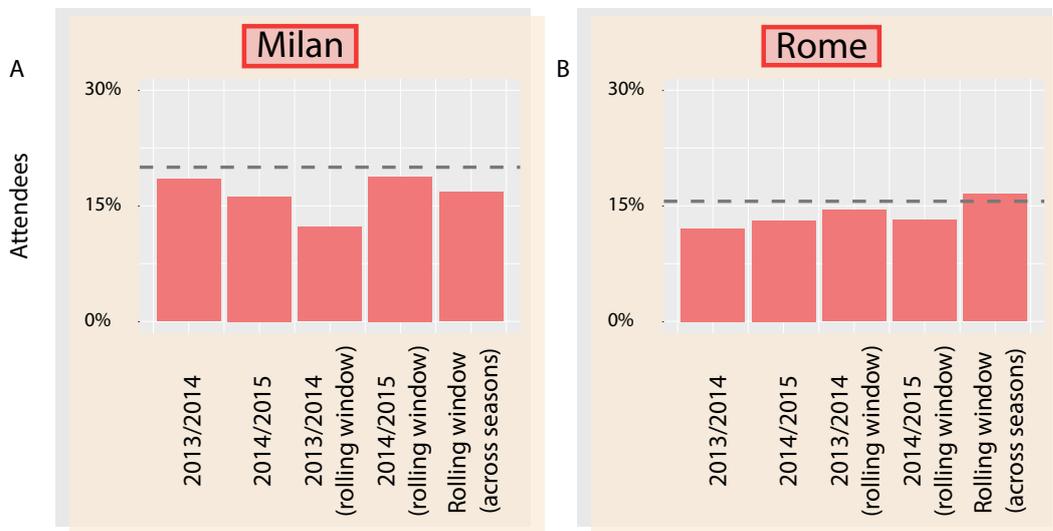
**Figure 5.8: High temporal resolution analysis of the effect of the size and starting point of the time window.** | We want to investigate how the relationship changes as we change the size of the time window used to count users active on *Instagram* during a football match. In the figure, the bars extend from the starting point of the time window until the ending point. The corresponding analysis counts all unique users who have been active on *Instagram* in that interval and compares it to the official number of attendees for that match. The colour of the bar indicates the squared correlation between the users' count and the number of attendees in the stadium. For this analysis, we consider a football match to be 105 minutes long, including a 15 minutes half-time break. In Milan, the strength of the relationship increases both when counting *Instagram* users active during the entire match, but also when considering users who are active up to two hours before the match. In Rome, our results suggest that taking into account *Instagram* users active before the start of the match increases the strength of the relationship with the number of attendees in the stadium.



**Figure 5.9: Comparing football matches attendance figures to number of photos posted on *Instagram*** | We investigate the relationship between number of people attending football matches and number of photos uploaded on *Instagram*. We consider photos uploaded in two football stadiums in a time window extending from one hour before the official starting time of a football match, to three hours after. In both stadiums and across seasons, we find that higher counts of *Instagram* users correspond to higher numbers of attendees (all  $p < 0.001$ , all  $R^2 \geq 0.55$ , ordinary least squares regression).

#### 5.2.4 Counting photos instead of users

We present here a parallel analysis to that presented above to show that we obtain qualitatively similar results if we use the number of photos posted on *Instagram* rather than the unique number of active users in the two stadiums. This may be of interest when privacy considerations suggest that aggregated information on the number of photos might be preferable to data on individual users. Figure 5.9 depicts the relationship between number of photos uploaded to *Instagram* and number of attendees in the stadium. Figures 5.10, 5.11 and 5.12 present a comparison between the predictive accuracies of a rolling window model fitted to the data and models built using all available data. As before, we find that in almost all cases the rolling window models perform as well as models built using data from the whole period of analysis. This further supports the use of a rolling window model, since it is only fitted to data which would actually be available in reality. Finally, Fig. 5.13 shows the effect of the size of the bounding box on the strength of the relationship. Similarly to before, we observe a smooth decrease around San Siro football stadium, whereas a faster decrease is observed around Stadio Olimpico football stadium.



**Figure 5.10: Predictive accuracy in the two stadiums when using number of photos for the analysis** | We present here the results of the leave-one-out-cross-validation analysis on models fitted using the number of photos inside the football stadiums. For the same measures of prediction accuracy, we again see that in most cases the rolling window analysis performs at least as well as the model considering data from the whole period of analysis. This suggests that even the aggregated number of photos posted inside the stadiums, without considering whether a user has uploaded more than one photo during a match, contains sufficient information to infer the number of attendees. This figure reports the symmetric mean percentage absolute error (SMAPE).

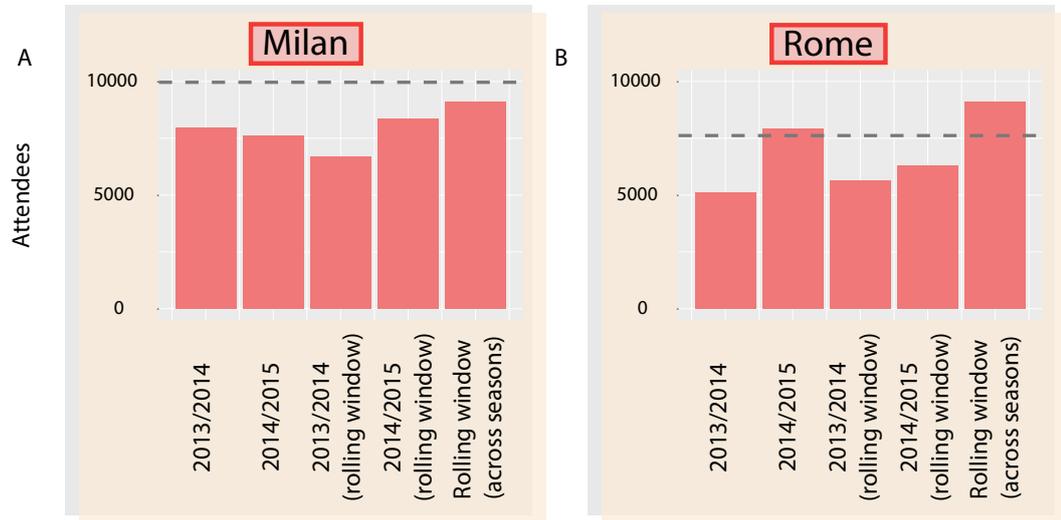


Figure 5.11: Root mean squared error in the two stadiums when using number of photos uploaded to *Instagram* as the predictor variable.

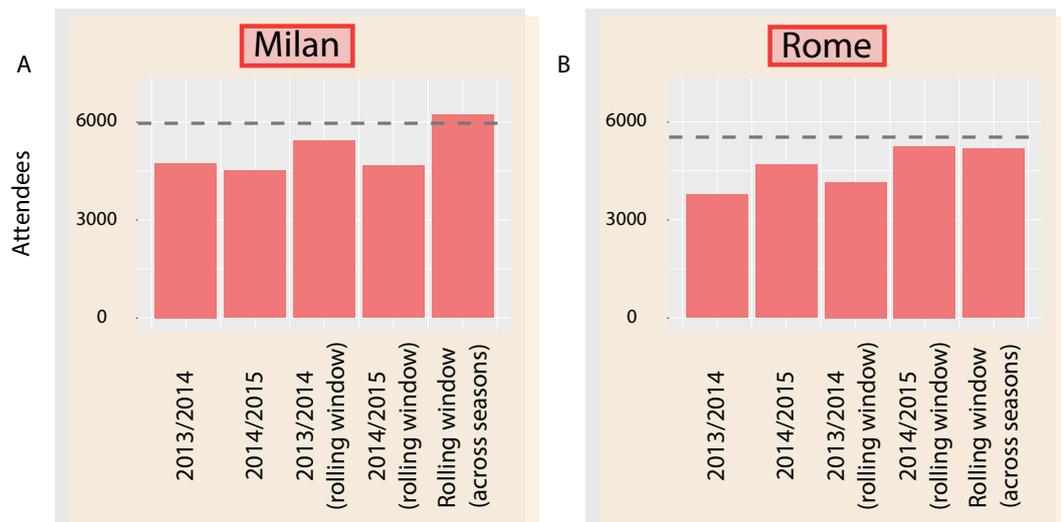
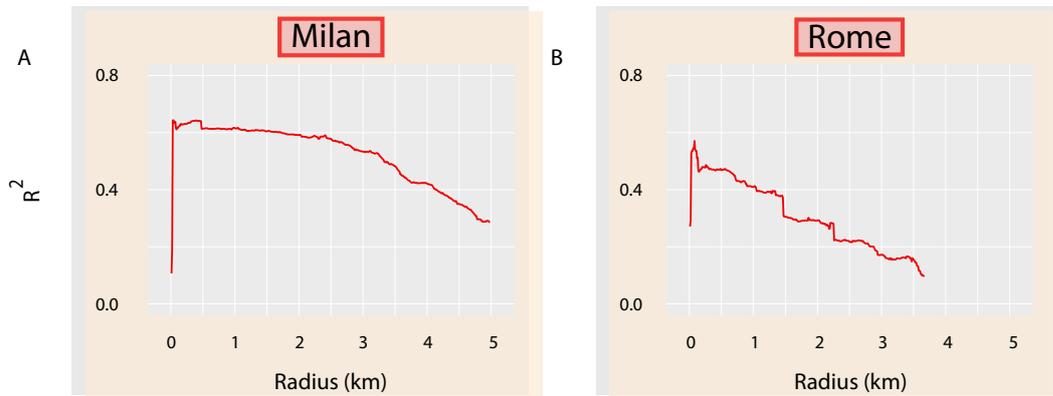


Figure 5.12: Median absolute error in the two stadiums when using number of photos uploaded to *Instagram* as the predictor variable.



**Figure 5.13: Spatial analysis** | We investigate how the relationship between number of photos uploaded to *Instagram* and number of attendees varies as we change the size of the bounding box around the football stadiums. We consider concentric circles centred on the football stadiums and of increasing radius. For each radius, we consider photos uploaded on *Instagram* inside the corresponding circle and investigate the relationship between them and the attendees at football matches. We examine radii varying from 10 metres to 5 kilometres, and we only show results when the relationship is statistically significant ( $p < 0.05$ , ordinary least squares regression). The time window used to count photos stretches from one hour before the starting time of the football match, to three hours after the starting time. As before, we again observe some differences in the two stadiums: in Milan the correlation decreases smoothly as we consider larger areas; however, in Rome we find a more fragmented change. This may be due to the different location of the two stadiums inside the city, with Rome’s stadium being close to tourist attractions from where *Instagram* users commonly upload photos. This further analysis again highlights the importance of carefully assessing the location of the area in which the crowd is when calibrating the model.

### 5.3 Conclusion

Being able to measure the size of a crowd can be crucial in emergency situations. However, this is a traditionally difficult task which is often performed manually, with human analysts counting samples of the crowd. In Chapters 4 and 5, we have presented evidence that data generated through our ordinary interactions with mobile phone networks and social media platforms, such as *Twitter* and *Instagram*, can be used to measure the size of a crowd. Our work highlights the importance of calibrating and testing methods to estimate crowd size on specific case studies where precise counts are available from other sources.

We have shown that there are several aspects to consider when analysing data derived from social media platforms to measure the size of a crowd. Changes in the number of active users of the service may affect the relationship and we have seen that they can be taken into account by considering temporally close data points. The location of the event where the crowd is gathered is also an important factor, and the area used to collect social media data should be carefully assessed.

Our findings hold potential value for a range of stakeholders and policy makers, who may need to generate quick and accurate estimates of the size of a crowd for a wide range of reasons, including the avoidance of crowd disasters and to facilitate emergency evacuations.

## CHAPTER 6

# ANALYSIS OF THE COMMUNITIES OF AN URBAN MOBILE PHONE NETWORK

In Chapter 4, we showed that aggregated data derived from our interactions with the mobile phone network, such as phone calls, can be used to estimate the size of a crowd. However, data derived from phone calls also contain information on interactions between social groups or geographical locations. Here, we analyse the community structure of the network induced by mobile phone calls placed and received within the Milan metropolitan area, in northern Italy, over a period of two months, revealing the spatial and temporal patterns in the local communications.

The dataset we retrieved for this study contains the anonymized records of phone calls between geographical areas in the city of Milan and surroundings, as presented in the left panel of Fig. 6.1. In a similar fashion to the dataset presented in Chapter 4, the mobile phone provider aggregated the data both spatially into a grid with 10000 cells, each cell being roughly square of side 235m, and also temporally at a ten minute granularity <sup>1</sup>. The period of analysis goes from 1 November 2013 to 31 December 2013. A more detailed description of how the dataset was constructed is presented in [81]. We study the cell activity by constructing a series of weighted networks. The nodes in these networks represent geographical locations, and the link strength is proportional to the volume of calls between the corresponding cells.

For a preliminary characterization of the networks structure, we build a single net-

---

<sup>1</sup>Data available at: Telecom Italia Big Data Challenge 2014, <https://dandelion.eu/datamine/open-big-data/>

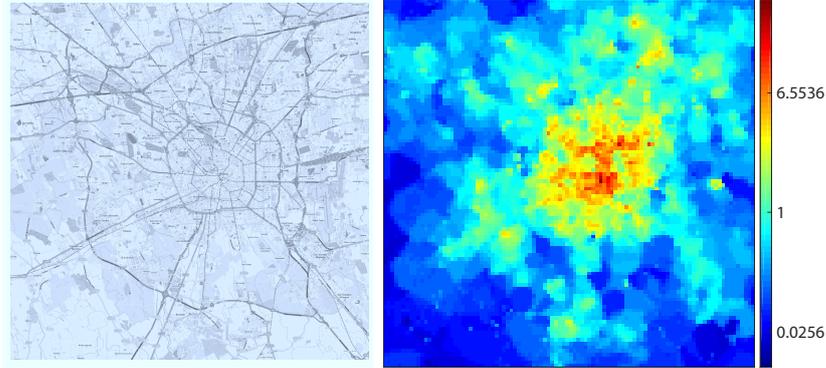
work aggregating all time intervals. As the whole period of analysis consists of 8784 time intervals, the edge weights are defined as:

$$\omega_{ij} = \frac{1}{Z} \left( \sum_{t=1}^{8784} \bar{w}_{ij} + \sum_{t=1}^{8784} \bar{w}_{ji} \right).$$

In the equation above,  $\bar{w}_{ij}$  is the volume of calls originating on node  $i$  and reaching node  $j$ . Thus, the edge weight  $\omega_{ij}$  is the normalized volume of phone calls between nodes  $i$  and  $j$ . The normalization constant  $Z = \max_{i,j} \left\{ \sum_{t=1}^{8784} \bar{w}_{ij} + \sum_{t=1}^{8784} \bar{w}_{ji} \right\}$  is chosen so that the strongest edge weight is 1. With these definitions, we assign to each node  $i$  an activity  $k_i$ , defined as:

$$k_i = \sum_{j=1}^{10,000} \omega_{ij}.$$

The activity is a weighted equivalent of the node degree, measuring the total strength of all the connections involving a given node. A geographical heat map of the activities, in the right panel of Fig. 6.1, shows that a higher call volume is recorded in downtown Milan, in agreement with our findings of Chapter 4 and with the intuitive notion that the centre is the busiest part of the territory. The study of



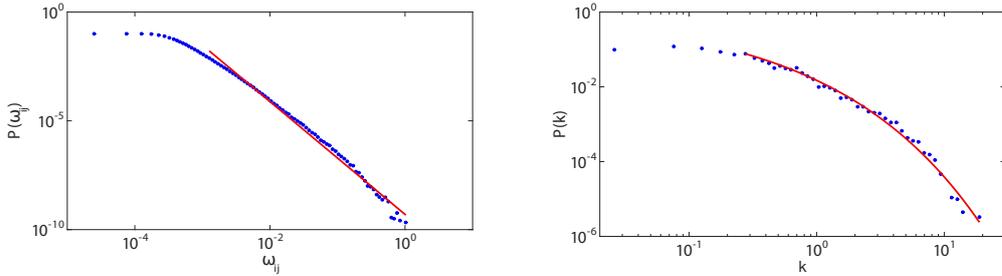
**Figure 6.1: Radially decreasing mobile phone activity.** The activities of the cells (heat map on the right) are highest in downtown Milan, and roughly decrease with distance from the city centre. Notable exceptions are the airport and residential suburbs. This map was generated with data from *OpenStreetMap* (© OpenStreetMap contributors).

the community structure could be performed, in principle, on the full aggregate network. However, this would have two drawbacks. First, it would not allow us to detect the hierarchy of the communities. Second, the analysis could be sensitive to the presence of noise, i.e., very weak links that may mask the underlying structural

character of the network. This is a particularly likely occurrence, given the slow-tail decay in the distributions of weights and activities (Fig. 6.2), which makes the weakest edge strength and the lowest node activity the most probable. More precisely, the distribution of weights exhibits a power-law tail with exponent  $-2.59$ , while the activity distribution follows a clear stretched exponential

$$P(k) \sim e^{-\left(\frac{k}{k^*}\right)^\alpha}, \quad (6.1)$$

with  $k^* = 0.023$  and  $\alpha = 0.383$ . Thus, we prefer to *threshold* the aggregate introducing a parameter  $\tau$ : for any chosen value of  $\tau$ , we create a network by removing from the aggregate any edge whose weight is less than  $\tau$ , and considering all other edges as unweighted. To analyze the networks thus created, we test the recently in-



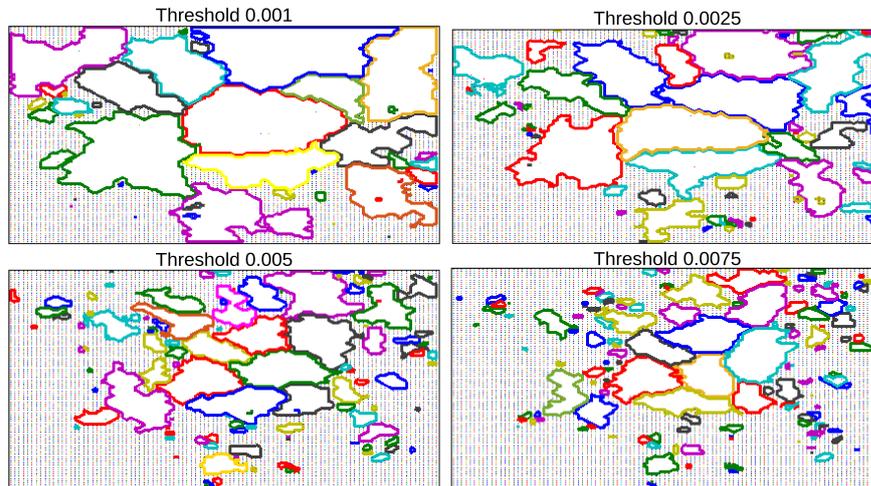
**Figure 6.2: Weights and activities of the aggregate network.** The distribution of the edge weights in the aggregate (left panel) shows a slow decay, with a tail that is well fitted by a power-law with exponent  $-2.59$ . The activities (right panel) follow instead a stretched exponential (Eq. 6.1), with  $k^* = 0.023$  and  $\alpha = 0.383$ .

roduced community detection algorithm described in Ref. [186]. This is a new fast spectral method that uses several refinement steps to identify the network partition that maximizes the modularity

$$q = \frac{1}{2m} \sum_{ij} \left( A_{ij} - \frac{d_i d_j}{2m} \right) \delta_{c_i, c_j}.$$

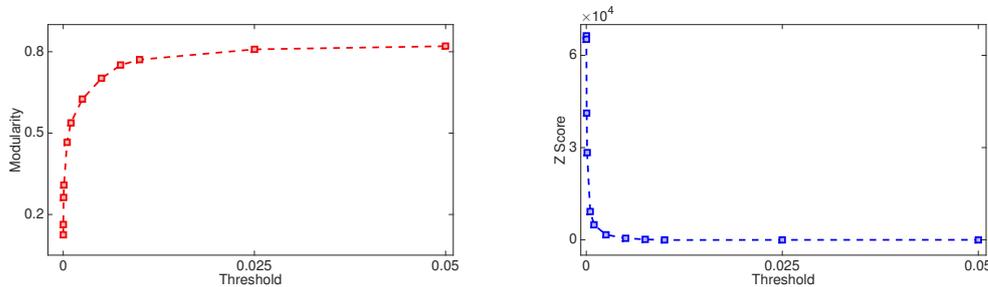
In the equation above, the sum runs over all pairs of nodes,  $m$  is the total number of edges in the network,  $d_i$  is the degree of node  $i$ ,  $c_i$  is the community to which node  $i$  is assigned,  $\delta$  is Kronecker's symbol, and  $A$  is the adjacency matrix, whose  $(i, j)$  element is 1 if there is an edge between nodes  $i$  and  $j$ , and 0 otherwise. The values of modularity are constrained between  $-1$  and  $1$ , with higher values corresponding to better partitions. The algorithm also provides the effect size of the detected partition in terms of a  $z$ -score, which is the number of standard deviations that separate the measured modularity from that of a random-graph null model. We

run the algorithm 100 times on each thresholded network, and select the partition with the highest value of modularity. As the values of  $\tau$  increase, we note that the



**Figure 6.3: Hierarchical backbone of communication communities.** For low values of the threshold  $\tau$  the noise still dominates the community structure detected. However, after the critical threshold of 0.005, increasing  $\tau$  only causes the communities to fragment into sub-modules.

evolution of the detected community structure undergoes a significant change when  $\tau$  reaches a “critical value”  $\tau^* \approx 0.005$ . At lower thresholds, the communities change significantly with  $\tau$ . Conversely, thresholds greater than  $\tau^*$  only result in fragmentation of the existing communities into smaller ones almost entirely contained within the parent module, without drastic changes in the overall structure. In addition, the individual communities correspond to connected areas of territory (Fig. 6.3). A second effect we note is that increasing thresholds correspond at the same time to higher values of the modularity, and lower  $z$ -scores (Fig. 6.4). Explaining this behaviour in detail is a complex problem, since, to a preliminary investigation, it appears to depend on the distribution of weights between modules, and it will be addressed in future publications. For the analysis of our data, we choose to work on the network corresponding to the critical threshold, as this provides a good balance between two necessities, namely that of a large enough threshold to remove the noise that might mask the community structure, and that of a small enough threshold to avoid excessive fragmentation. Even though this choice is arbitrary, our results are robust with respect to small threshold variations. Also, we show below that analogous results hold for weighted networks where we keep all weights unchanged. Thus, to take advantage of faster computational times, we use the unweighted network for further analysis.



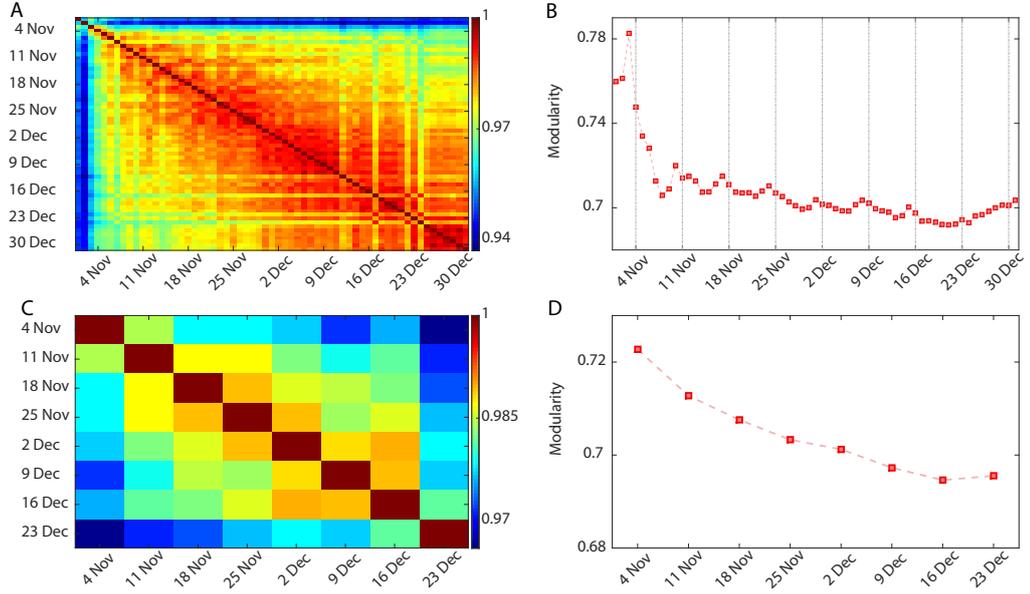
**Figure 6.4: Threshold evolution of network modularity.** For increasing values of the threshold, the modularity increases (panel A), apparently saturating at a value just above 0.8. For the same thresholds, the  $z$ -score, which is a measure of the effect size of a given modularity measurement, has a fast decay, indicating that the community structure quickly becomes similar to what would be found in a random network as more links are erased. The lines are guides for the eye.

## 6.1 Time evolution of communities

Our first goal is to investigate the communication patterns that appear over time at a community level, to gain insights in the emergent structures of human communication. We start by studying how the communities evolve on the time scale of single days. To do so, we create an aggregate network for each day over the period covered by our data, and perform community detection on each of them as described above, with the aim of quantifying the difference between the community structures in the different “daily” networks. One of the most widely used methods for the actual comparison and evaluation of such differences is to calculate the *Normalised Mutual Information* (NMI), a measure borrowed from information theory [220, 221, 179, 222, 180, 181, 183]. To find the NMI between two partitions  $C$  and  $\tilde{C}$ , first treat them as random variables and compute their mutual information:

$$I(C, \tilde{C}) = \sum_{i=1}^{n_C} \sum_{j=1}^{n_{\tilde{C}}} \frac{V_{ij}}{N} \log \left( \frac{V_{ij}N}{V_i V_j} \right),$$

where the  $V_{ij}$  are the elements of the *confusion matrix*  $V$ , whose entries are the numbers of nodes belonging to community  $i$  in partition  $C$  and to community  $j$  in partition  $\tilde{C}$ ,  $V_i$  denotes the sum over the elements of row  $i$  in  $V$ , and  $N$  is the total



**Figure 6.5: Determining the time-scale of social dynamics.** Panel A depicts the Normalised Mutual Information between partitions at different days, showing a strong similarity between all communities during the two months analyzed. Panel B presents the evolution of modularity during the period of analysis. Vertical dashed lines correspond to the beginning of the working week (Monday). The modularity has an unusual spike in the first days of November, probably due to a bank holiday long weekend, but only oscillates around a constant value for subsequent periods. We note that the modularity on weekends is consistently higher than it was during the working days of the corresponding week. The NMI analysis of partitions corresponding to different weeks, in Panel C, shows a strong similarity between all communities. Panel D illustrates the evolution of modularity of the weekly networks, with labels indicating the first day of each week. In agreement with the previous analysis, the modularity has a higher value in the first week of November.

number of nodes. Then the NMI between two partitions is defined as

$$\begin{aligned}
 NMI(C, \tilde{C}) &= \frac{-2I(C, \tilde{C})}{\sum_{i=1}^{n_C} \frac{V_i}{N} \log \frac{V_i}{N} + \sum_{j=1}^{n_{\tilde{C}}} \frac{V_j}{N} \log \frac{V_j}{N}} \\
 &= \frac{-2 \sum_{i=1}^{n_C} \sum_{j=1}^{n_{\tilde{C}}} V_{ij} \log \left( \frac{V_{ij} N}{V_i V_j} \right)}{\sum_{i=1}^{n_C} V_i \log \frac{V_i}{N} + \sum_{j=1}^{n_{\tilde{C}}} V_j \log \frac{V_j}{N}}.
 \end{aligned}$$

The normalised mutual information can assume values ranging from 0 to 1. Higher values indicate stronger similarity between the two partitions, with  $NMI(C, \tilde{C}) = 1$  found if the two partitions are identical. Conversely, partitions that are totally independent from each other have a normalised mutual information of 0.

The NMI values we find are always quite high (Fig. 6.5A), indicating a strong similarity in the community structure across different days. This provides evidence of the robustness of the structure of the mobile phone call network over the 24-hour time scale, with only minor changes between communities across the two months. Nonetheless, some days stand out as significantly different from the average. First, we observe an unusual structure in the first few days of November. This is most probably due to the particular nature of that period, which includes a bank holiday covering an important mandated Catholic holiday (1 November). In addition, in 2013, the holiday fell on a Friday, causing a “long weekend”. We also note that the community structure in these days had a substantially higher modularity than the average for the rest of the period (Fig. 6.5B).

Another remarkable difference in the structure appears on 12 December. This is likely caused by the combination of three major events happening in Milan on that day: 1) an annual demonstration in memory of the controversial *Piazza Fontana Bombing*, a terrorist attack that took place on 12 December 1969; 2) a second demonstration, part of ongoing protests against the Italian government; and 3) a major concert of One Direction, a highly popular pop boy band. Notably, both political demonstrations saw the occurrence of clashes between demonstrators and police forces, while the concert gathered thousands of people across the city for the whole day. The co-occurrence of these events clearly disrupted the usual patterns of communications in the city, causing the highly unusual community structure observed on that day. Finally, the changes in structure detected on 22 December and 24 December likely reflect the particular nature of this period of the year. In particular, 22 December was the last Sunday before Christmas, a day traditionally devoted to the final purchases before the start of the holiday period. Notice that these results provide direct evidence of how one can use mobile phone activity to extract information on people’s behaviour within social groups and directly detect socially relevant changes in their patterns.

The data also allow us to infer a strong similarity in the last week of our analysis period, which corresponds to Christmas and New Year’s holidays. This supports the idea that communities in the communication networks closely reflect our behaviour. In the holiday period, people traditionally spend more time with their families, and reduce the frequency of contacts with acquaintances and other people outside their close-friend circles. Thus, the structure of communications is better

defined, and links between different communities become less important, causing an increase in modularity. Also, this is an indication that the agents participating in communication tend to remain stable over this time period.

The analysis of the daily NMI also shows that days close to each other have a consistently higher similarity, suggesting that changes in the community structure happen over a longer time scale than just one day. To investigate this, we build aggregates for each entire week in the period of analysis and perform community detection as above. Our findings (Fig. 6.5) show that weeks close to each other are very similar, and the NMI exhibits a slower decay than what we observed in the daily structure. This suggests that the variability in the structure is due to a slow dynamics of the communities happening over different different days and repeating with the period of a week. In the next section, we present a detailed analysis of this two-time-scale behaviour. To verify the statistical significance of these results, we validated them against an appropriate null model. The results, confirming our findings, and are detailed in Section 6.3.

## 6.2 Period analysis of network structure

To investigate the periodic behaviour of the communication patterns, we employ the same NMI comparison approach introduced in the previous section, by building aggregates for each different day of the week. In other words, we construct seven different networks, the first aggregating the data collected on all Mondays, the second with the data from all Tuesdays, and so on up to the seventh network which corresponds to all the Sundays. Then, we build a daily NMI matrix where each element is the NMI between the structures detected on the corresponding aggregates.

The results, in Fig. 6.6A, show that different days are always very similar, with an NMI consistently greater than 0.95. However, a difference is still evident between working days and weekends, in agreement with the daily analysis. In fact, the NMI reaches its highest values when comparing either two working days or the two days of the weekend, while the smallest values are found when comparing a weekend day and a working day. This difference also corresponds to a higher value of modularity for weekend days than for the rest of the week (Fig. 6.6B), supporting the idea that on non-working days people tend to be active only within their closest social circles. Note that these results illustrate the ease with which one can extract quantifiable information about the behaviour of people in social contexts from

communication records, even if completely anonymized and already geographically aggregated in their raw form.

The results found so far show that we can clearly detect the difference in population behaviour over the different days of the week. However, human activities also change at the shorter time scale of hours. Thus, we investigate the changes in average community structure during a day by constructing 24 different networks, each aggregating the data collected during the same hour every day. The NMI matrix (Fig. 6.6C) shows a remarkable difference between daily and nightly communities. The structure of communities at night does not present particular patterns, but we find blocks of high similarity during the day. A first block corresponds to highly similar communities during morning hours, covering roughly the first part of a working day. A second block can also be observed in the afternoon hours, when the second part of a working day happens. Finally, a last block extends over the evening hours. We find this result remarkable, in that it confirms that mobile phone communications are closely related to human behaviour even at a community level.

Figure 6.6D shows the evolution of modularity for the hourly networks. We find that the waking hours correspond in general to stronger communities, with modularity dips in correspondence of the periods traditionally linked to lunch (12:00–13:00) and dinner (20:00).

Finally, to clearly show the periodic nature of the network, we analyze the data differentiating for given hours *and* days of the week. We create 168 networks, each aggregating the data corresponding to the same hour and the same day of the week, and perform an NMI analysis. The results, in Fig. 6.6E, show the emergence of a clear structure, where partitions obtained at daytime hours are strongly similar, and cluster in blocks with high values of NMI, separated by lower similarity partitions corresponding to the nights. Investigating this result more closely, we notice that higher similarities are observed between different daytime hours of the same day.

The evolution of modularity (Fig. 6.6F) displays again a similar pattern to the one previously observed with two peaks in the value of modularity in the morning and afternoon and a lower value during the night. However, we also find a peak in the middle of the night, particularly strong during weekends. This might reflect the fact that phone activity is naturally lower during the night. Thus, it is highly likely that someone placing a nighttime call will not call more than a few close con-

tacts, and will not receive a call back from people other than the persons originally called. This results in strong communities and a high modularity. Similarly, we also find a higher modularity during weekends than over weekdays, consistently with the social dynamics outlined before. In addition to validating these findings against a null model (Section 6.3), we also test their robustness using the method proposed by Mucha *et al.* [182], obtaining results that support our methodology (details in Section 6.4).

### 6.3 Null model validation

To validate the NMI analyses, we build a null model by fixing number and size of the communities detected in each instance, and randomising the community labels assigned to each node. We compute the NMI matrix averaged over 100 randomizations, and compare it with the one presented in the main text. This allows us to verify whether our results can be attributed to randomness, or they represent an effect present in the data. The results, shown in Fig.6.7, bear no resemblance to those in the main text. Also, in all null models we observe that the patterns characterizing the NMI matrices of the original data are absent. This shows that it is very highly unlikely that the structures detected arose due to random fluctuations.

### 6.4 Weighted and multiplex analysis

To test the robustness of the results presented in the main text, we use the approach described in [182]. This method provides a generalisation of the classical modularity to the case of time-dependent and multiplex networks:

$$Q_m = \frac{1}{2\mu} \sum_{ijsr} \left[ \left( A_{ijs} - \frac{k_{is}k_{js}}{2m_s} \right) \delta_{sr} + \omega \delta_{ij} \right] \delta_{g_{is}, g_{jr}}$$

where the indices  $i$  and  $j$  refer to nodes, the indices  $s$  and  $r$  refer to layers,  $\omega$  is a parameter that determines the strength of the coupling of a node to its copies in the neighbouring layers,  $\mu$  is the sum of all the strengths across all layers and  $g_{is}$  is the community of node  $i$  in layer  $s$ . This quality function is then maximised using a generalisation of the Louvain method [177]. Notice that when  $\omega = 0$ , the layers are independent. Conversely, high values of  $\omega$  increase the coupling to the point that all the replicas of each node are treated identically. This causes the communities found to be the same across layers, effectively neglecting the multiplex nature of the network. Thus, for this type of analysis, one needs to find an intermediate value

of  $\omega$  that offers a compromise between the two extremes. In our case, we choose  $\omega = 0.1$ , as the value above which the differences between layers start to smoothen.

We consider the same thresholded networks used in the main text for each day of the week and assign a different layer to each of them, thus creating a multiplex network with 7 different layers. Figure 6.8 shows the NMI analysis of the partitions of the individual layers for  $\omega = 0$  and  $\omega = 0.1$ . In both cases, we obtain results that are qualitatively similar to those presented in the main text. Moreover, the multiplex modularity  $Q_m$  in the two cases is 0.6935 and 0.6960, respectively, in agreement with the average modularity of the seven networks presented in the main text, which is 0.6934.

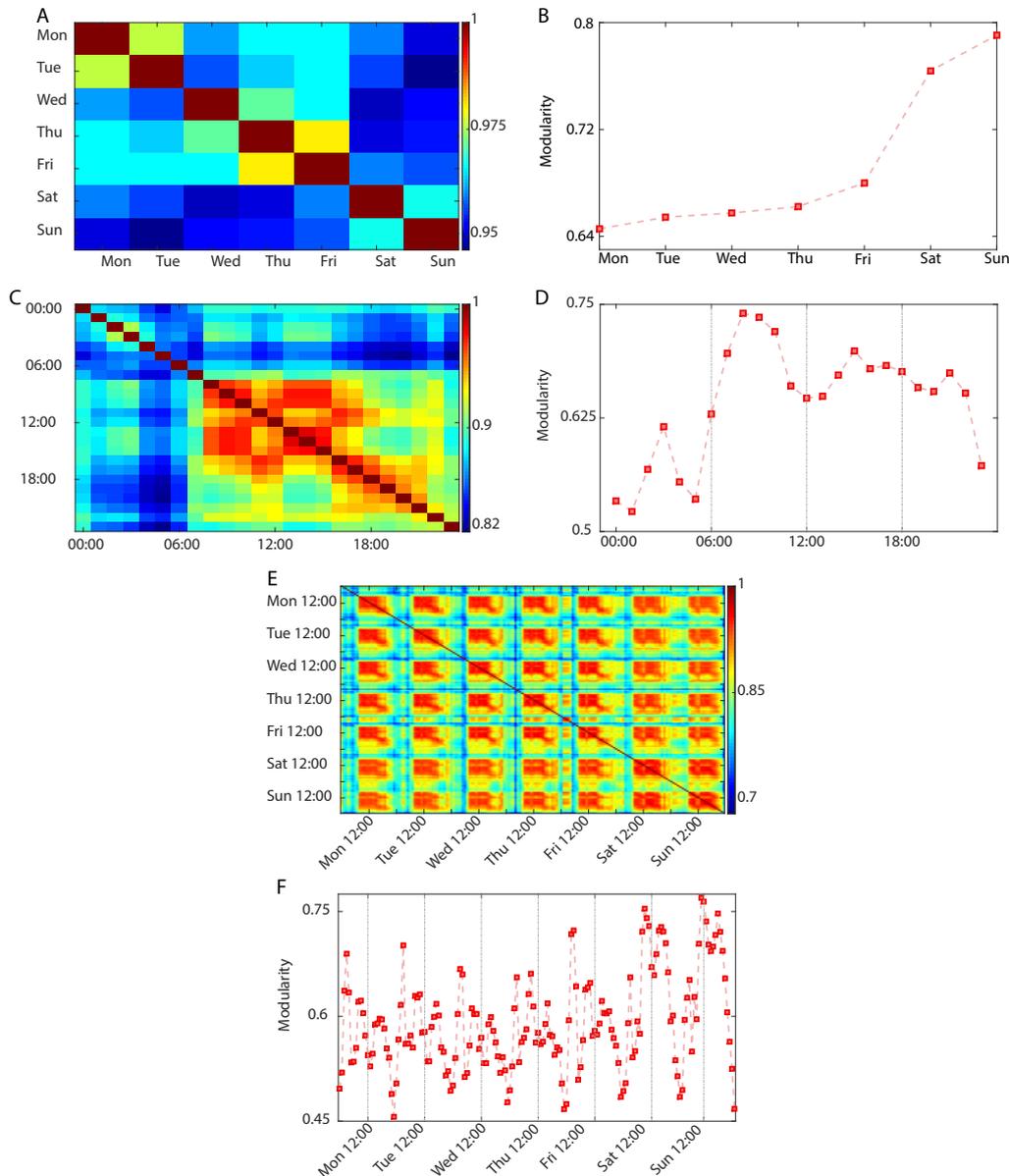
Studying the effect of preserving the link weights is also of great interest for the application of this methodology. In the main analysis presented above, we used a threshold parameter to remove weak links that may act as noise and mask the community structure. Here, we build aggregated networks for each day of the week keeping all links with their weights and analyze them with the method described above. As before, each layer in the multiplex corresponds to an aggregate network of a given day. Figure 6.9A shows that if we preserve all the links with their original weights and leave the layers uncoupled, the difference between weekdays and weekends is not remarkable. Figure 6.9B depicts the results when the coupling between the layers is  $\omega = 0.1$ . As before, we observe a structure similar to the one presented in the main text, with the exception of a difference in the typical Friday communities.

This last analysis supports our hypothesis that thresholding removes the noise in the network and allows us to uncover the underlying community structure, while leaving the relevant structural properties unchanged. We also see that, if we want to keep all the links with their weights, a coupling between the layers is essential. However, the size of the multiplex grows really quickly when considering several layers, such as the hourly-weekly routine, where we would have 168 networks with 10000 nodes each. The multiplex could possibly be even larger, depending on the granularity of the data, making its analysis not always feasible. Our methodology, instead, obtains valid results while analysing the networks separately, thus being faster and demanding less resources.

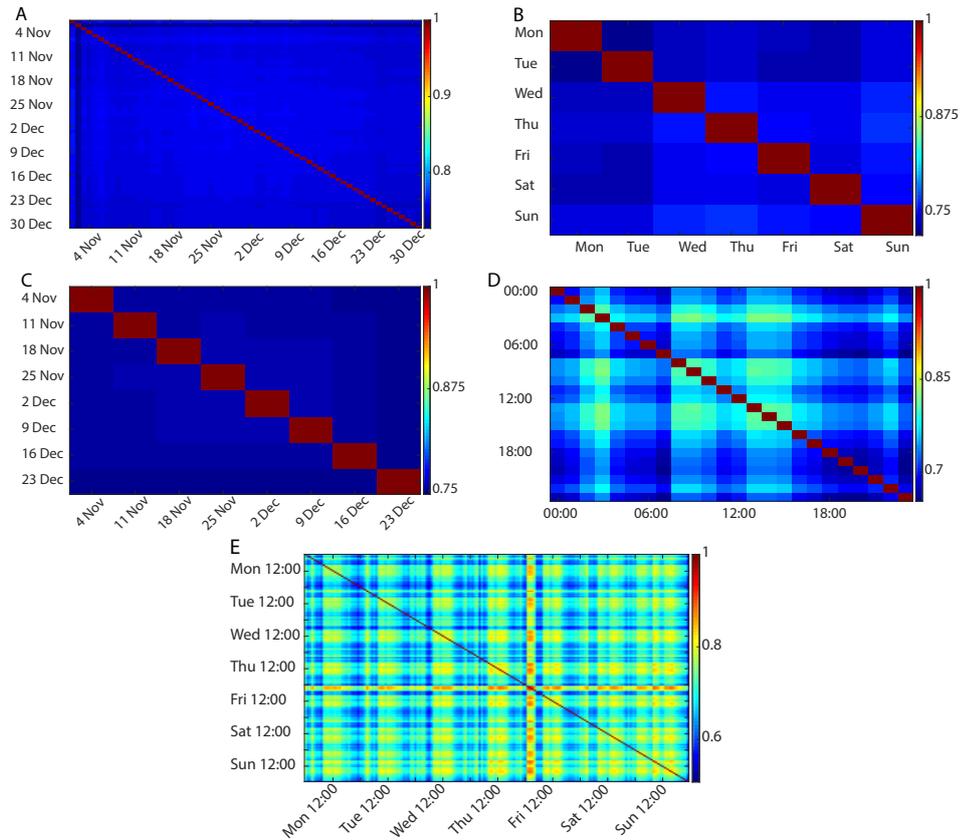
## 6.5 Conclusions

In conclusion, we have presented a complete characterization of the community structure of a mobile phone call network and discussed its evolution over time, revealing the spatial and temporal patterns in local communications. Our findings suggest that information about people's behaviour and their interactions in social groups can be easily extracted from the community structure of networks induced by communication records. In fact, our results provide direct evidence of how one can use mobile phone activity to point out the occurrence of socially relevant events.

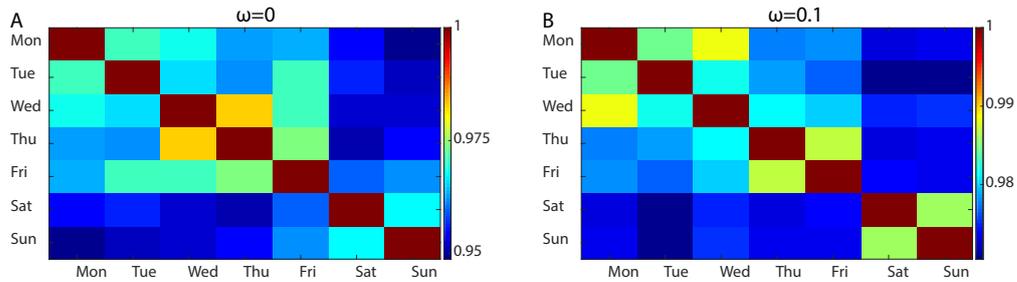
The ease with which our method can be applied, coupled to the high granularity of the data available to telecommunication companies, suggests that it may be useful even as a real-time tool to detect the occurrence of such events or activities, as evidenced by our results related to the day of 12 December. Our work supports the hypothesis that data generated through interactions with technological devices closely reflect human activity and can be used for quantitative studies of social systems. Moreover, it illustrates the ease with which one can extract valuable information even from completely anonymized and geographically aggregated data. In addition, the community structures detected can provide substantial support to telephone companies in the design and optimization of better, more efficient and flexible infrastructures, decreasing operating costs and increasing performances.



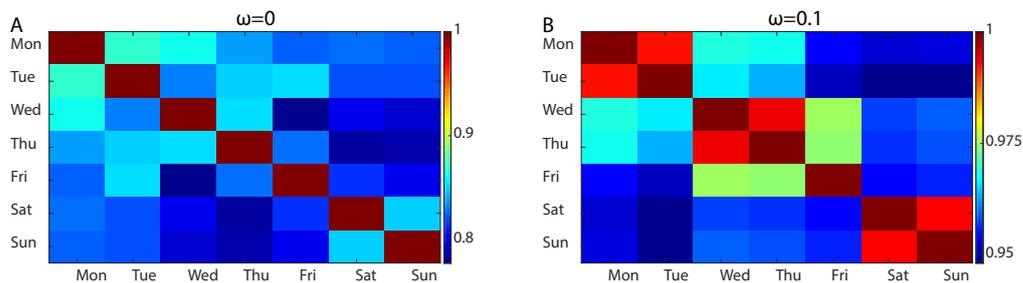
**Figure 6.6: Weekly, daily, and hourly-weekly routines.** Panels A, C and E show the Normalised Mutual Information between partitions of aggregates corresponding to different days, different hours, and different hours of each day, respectively. Communication communities on weekends are evidently different from those on working days. Also, waking hours are much more stable than the night, with two clear blocks corresponding to working hours and evening time. Moreover, the hourly-weekly analysis shows a striking structure corresponding to blocks of highly similar communities during the daytime. The modularities for the three types of networks (Panels B, D and F), show that communities are much tighter on weekends and during waking hours than they are on weekdays and during the night, with the exception of the weekend nights that are highly modular.



**Figure 6.7: Validation of NMI analyses.** Randomized NMI matrices for the daily (panel A), weekly (panel B), week aggregates (panel C), hourly (panel D) and hourly-weekly (panel E) show values that are roughly constant across the matrix, and always smaller than those observed in the original data. Also, we do not observe the patterns characterizing the NMI matrix presented in the main text, such as the separation between working days and weekends and the strong similarity between daytime communities. Times are reported in Central European Time (CET).



**Figure 6.8: Weekly community structure NMI using multiplex detection.** We observe results strongly similar to the results presented in the main text both with no coupling (Panel A) and with coupling between each node and its copies in the neighbouring layer (Panel B). The multiplex modularity value in the two cases is 0.6935 ( $\omega = 0$ ) and 0.6960 ( $\omega = 0.1$ ). These are compatible with the average value of modularity across the seven networks presented in the main text, which is 0.6934, thus compatible with this result.



**Figure 6.9: Weekly community structure NMI using weighted multiplex detection.** We observe results similar to the results presented in the main text. Here, we can also notice a smaller differentiation between weekday groups.

## CHAPTER 7

---

### FINDING NETWORK COMMUNITIES USING MODULARITY DENSITY

Network science provides a useful framework for analysing several of the new data sources that are becoming increasingly available. One of the key topological properties of many real-world networks is the community structure, as was discussed in chapter 2 and shown in chapter 6.

In this chapter, we first present a short review of the most common community detection methodology, show its limitations and then study a new quality function, *modularity density*, that was originally introduced in [223, 224]. This new technique has been shown to address some of the shortcomings of existing methods. We present a detailed analysis of its properties on synthetic networks typically used to evaluate quality functions, as well as on random graphs, which are a commonly used benchmark to test community detection methods. In addition, we describe a new community detection algorithm based on this metric, and validate it on synthetic and real-world networks, showing that it performs better than other currently available methods. Also, we argue that the nature of modularity density allows for a direct quantitative comparison of community structures across networks of different sizes.

## 7.1 Traditional modularity and its limitations

The modularity  $Q$  of a network with  $N$  nodes and  $m$  links is defined as:

$$Q = \frac{1}{2m} \sum_{ij} \left( A_{ij} - \frac{k_i k_j}{2m} \right) \delta_{C_i C_j},$$

where  $A$  is the adjacency matrix of the network,  $k_i$  is the degree of node  $i$ ,  $C_i$  is the community to which node  $i$  is assigned and  $\delta_{ij}$  is the Kronecker delta. The first term accounts for the presence or absence of a link between node  $i$  and node  $j$ ; the second term, instead, is the expected number of links between node  $i$  and node  $j$  in a random network with the same degree sequence as the original one.

A first limitation of modularity is that it is intrinsically dependent on the number and distribution of edges, rather than on the number of nodes. To see this, denote by  $m_C$  and  $e_C$  the number of internal and external links of community  $C$ , respectively. Moreover, let  $k_C = 2m_C + e_C$  be the sum of the degrees of the nodes in community  $C$ . With this notation, it is

$$Q = \sum_{C \in \mathcal{C}} \left[ \frac{m_C}{m} - \left( \frac{k_C}{2m} \right)^2 \right], \quad (7.1)$$

where  $\mathcal{C} = \{C_1, C_2, \dots, C_P\}$  denotes the set of all communities in the partition. In this expression, each term in the sum refers to a different community. The first factor of each term corresponds to the internal density of links in the community, whereas the second factor encodes the expected density of links in the random network null model. Now, introduce the positive parameter  $\alpha_C$ , representing the ratio of external links to internal ones:

$$e_C = \alpha_C m_C.$$

The value of  $\alpha_C$  is smaller for strong communities, and higher for weaker ones. Then, we can write

$$Q = \sum_{C \in \mathcal{C}} \left[ \frac{m_C}{m} - \left( \frac{2 + \alpha_C}{2m} \right)^2 m_C^2 \right]. \quad (7.2)$$

From this expression, it is clear that a community  $C$  gives a positive contribution to  $Q$  only if:

$$m_C < \frac{4m}{(\alpha_C + 2)^2}.$$

This implies that the condition for a community to give a positive contribution only depends on the number of edges in the community and on the total number of edges in the network, but not explicitly on the number of nodes.

A similar result can be obtained considering a network of  $\kappa$  communities disconnected from each other, along the lines of [179]. Under the assumption that all groups have the same number of links, we can write

$$\begin{aligned} m_C &= \frac{m}{\kappa}, \\ e_C &= 0, \\ k_C &= \frac{2m}{\kappa}. \end{aligned}$$

Then, from Eq. 7.1, it is

$$Q = \kappa \left[ \frac{1}{m} \frac{m}{\kappa} - \left( \frac{1}{2m} \frac{2m}{\kappa} \right)^2 \right] = 1 - \frac{1}{\kappa}. \quad (7.3)$$

This shows that modularity converges to 1 with the number of communities  $\kappa$  regardless of the internal properties of the communities, such as their size, or the number of internal edges. As long as  $\kappa$  is very large and all communities have the same number of edges  $m/\kappa$ , a network of disconnected trees has the same modularity of a network of disconnected cliques. As before, we also see that the number of nodes in each group does not explicitly contribute to  $Q$ , and, as an immediate consequence, a network composed of few cliques has a smaller modularity than a network composed of many disjoint trees.

In addition to these results, the effectiveness of modularity is not constant for all edge densities. To determine its dependence on this quantity, we follow [225] and connect the  $\kappa$  groups in a ring configuration, where each community is linked with exactly one edge to the next one, and one edge to the previous one in the ring, for a total of  $\kappa$  inter-community edges. In this scenario, we have

$$\begin{aligned} m_C &= \frac{m}{\kappa} - 1, \\ e_C &= 2, \\ k_C &= \frac{2m}{\kappa}. \end{aligned}$$

From Eq. 7.1, it follows that

$$Q = \kappa \left[ \frac{1}{\kappa} - \frac{1}{m} - \left( \frac{1}{2m} - \frac{2m}{\kappa} \right)^2 \right] = 1 - \frac{\kappa}{m} - \frac{1}{\kappa}.$$

For constant  $m$ , this expression reaches its maximum when  $\kappa = \sqrt{m}$ , for which it is

$$Q = 1 - \frac{2}{\sqrt{m}}.$$

Thus, the highest modularity corresponds to a partition in  $\sqrt{m}$  modules. Once again, the number of nodes in the communities does not affect its largest possible value. This major limitation of modularity is known as the *resolution limit*, and it indicates that modularity, as a quality function for community detection, has an intrinsic scale proportional to  $\sqrt{m}$ . The number and size of the communities that can be detected via modularity maximisation are bound to adhere to this limit, posing a serious question on the significance of results obtained with this method. In fact, in a more general framework, Fortunato and Barthélemy [225] have shown that, under some circumstances, the resolution limit can even force pairs of well-defined communities to be merged into a larger cluster, because this corresponds to a higher modularity.

Finally, it is worth noting that the trivial partition where all the nodes are put together in one single community, namely the whole network itself, has a modularity of 0. This can be easily seen from Eq. 7.2, since in this case the sum has only one term,  $\alpha_C = 0$  and  $m_C = m$ , so

$$Q = \frac{m}{m} - \frac{4m^2}{4m^2} = 0.$$

At first, this might seem a desirable property for a quality function, since, intuitively, the trivial partition should not have a positive modularity. However, this implies that any partition that achieves a modularity larger than 0 is retained as a valid community structure. Since community detection algorithms try to maximize modularity, it is often the case that such a positive value can be found even on Erdős-Rényi random graphs [186]. To stress this point, the trivial partition with  $Q = 0$  can always be considered, but since one is interested in the maximum value of  $Q$ , it is often discarded in favour of a clustering that achieves any positive value of modularity. This poses a serious limitation to the ability of modularity-based algorithms to partition random graphs correctly.

Several variants of modularity have been proposed to address the resolution limit. For instance, multi-resolution methods, such as the one described in [129], introduce an additional tunable parameter  $\eta > 0$  in the expression for  $Q$ :

$$Q_\eta = \sum_{C \in \mathcal{C}} \left[ \frac{m_C}{m} - \eta \left( \frac{k_C}{2m} \right)^2 \right].$$

Larger values of  $\eta$  cause  $Q_\eta$  to be larger for partitions with smaller modules, whereas smaller values favour larger communities. However, this approach suffers from similar limitations to those presented by the original modularity [226]. In particular,  $Q_\eta$  has two contrasting behaviours: small clusters tend to be merged together, while large communities tend to be split into subgroups. Networks in which all the communities are of comparable size are immune to this problem, and one can find a value of  $\eta$  for which they can all be resolved. However, the existence of an optimal  $\eta$  is not guaranteed in the general case. In particular, for networks whose community sizes are heterogeneously distributed, e.g., following a power law, it is not possible to find a value of  $\eta$  that avoids both problems. The reason for this is that the nature of the resolution limit is more general than the specific definitions of modularity and its multi-resolution extension. Several quality functions for community detection, including the one just mentioned, can be derived within the general framework of a first principle Potts model with Hamiltonian

$$H = - \sum_{ij} [a_{ij} A_{ij} - b_{ij} (1 - A_{ij})] \delta_{C_i C_j},$$

where  $a_{ij}$  and  $b_{ij}$  are non-negative weights. Different choices for the weights result in different quality functions. However, only those using non-local weights can be truly free from the resolution limit [227], while all others, including modularity, multi-resolution modularity and functions based on quantities such as betweenness, shortest paths, triangles and loops, can never avoid it.

## 7.2 Modularity density

Recently, a new quality function called *modularity density* has been proposed to overcome the issues outlined above [223, 224]. Given a network partition, modularity

density is defined as

$$Q_{ds} = \sum_{C \in \mathcal{C}} \left\{ \frac{2m_C^2}{mn_C(n_C - 1)} - \left[ \frac{2m_C + e_C}{2m} \frac{2m_C}{n_C(n_C - 1)} \right]^2 - \sum_{\tilde{C} \neq C} \frac{m_{C\tilde{C}}^2}{2mn_Cn_{\tilde{C}}} \right\}, \quad (7.4)$$

where  $n_C$  is the number of nodes in community  $C$ , the internal sum is over all communities different from  $C$ , and  $m_{C\tilde{C}}$  is the number of edges between community  $C$  and community  $\tilde{C}$ . This new metric brings two major improvements over traditional modularity. First, it contains an explicit penalty for edges connecting nodes in different communities. This addresses the problem of the splitting of large communities, since each split introduces external links and is thus penalized. Second, all terms, including the penalty for inter-community edges, are explicitly weighted by the community sizes. Therefore, a partition with many edges linking two small communities is penalized more than one with the same number of edges linking two large ones. Thus, modularity density introduces local dependencies that are not found in traditional of modularity. Additionally, it is not related to the Potts model Hamiltonian, thus avoiding the resolution limit problem. Note that Eq. 7.4 requires  $n_C > 1$ , which implies that partitions with communities consisting of an isolated node are not allowed.

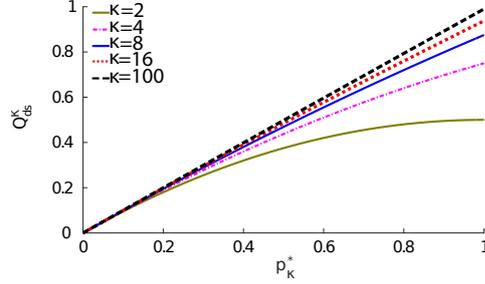
To investigate the properties of modularity density in more depth, rewrite the expression for  $Q_{ds}$  as

$$Q_{ds} = \sum_{C \in \mathcal{C}} \left[ \frac{m_C}{m} p_C - \left( \frac{2m_C + e_C}{2m} p_C \right)^2 - \sum_{\tilde{C} \neq C} \frac{m_{C\tilde{C}}^2}{2mn_Cn_{\tilde{C}}} \right], \quad (7.5)$$

where

$$p_C = \frac{2m_C}{n_C(n_C - 1)}.$$

The parameter  $p_C$  can assume values between 0 and 1, since it is the fraction of possible internal links actually present in community  $C$ . Thus, it measures the con-



**Figure 7.1:** Modularity density increases with number of communities and their edge density. In networks composed of  $\kappa$  disconnected modules, the modularity density  $Q_{ds}^{\kappa}$  depends only on  $\kappa$  and the edge density of the communities  $p_{\kappa}^*$ . Fixing one of the two parameters,  $Q_{ds}^{\kappa}$  always increases with the other.

nection density of the community, or, equivalently, the probability that two random nodes inside  $C$  are connected. From Eq. 7.5 it is clear that having many internal edges is not enough for a community to give a large contribution to modularity density. In fact, a strong community is one where the density of edges, rather than their number, is large. This also agrees with the intuitive notion that a community is a group of nodes that are densely connected amongst each other. Thus, a good partition is one that is characterized at the same time by a large number of intra-community links and a high density of edges within the communities. Modularity density achieves this by accounting for the number of nodes in each group and, in this sense, it has a more natural dependence on the local properties of the network and of the partition under consideration than does traditional modularity.

Next, it is instructive to study the behaviour of modularity density in the same cases described in the previous section. First, consider a network partitioned in just two communities,  $C$  and  $\tilde{C}$ . The contribution to  $Q_{ds}$  of community  $C$  is:

$$Q_{ds}^C = \frac{m_C}{m} p_C - \left( \frac{2m_C + e_C}{2m} p_C \right)^2 - \frac{m_{C\tilde{C}}^2}{2mn_C n_{\tilde{C}}}.$$

Introducing the proportionality constant  $\alpha_C$  as before, it is

$$Q_{ds}^C = \frac{m_C}{m} \left[ p_C - \frac{(2 + \alpha_C)^2}{4m} p_C^2 m_C - \frac{\alpha_C^2 m_C}{2n_C (N - n_C)} \right],$$

where we used  $n_{\tilde{C}} = N - n_C$  and  $m_{C\tilde{C}} = e_C$ . Unlike what happens with traditional modularity, the contribution of a single community depends explicitly on the number of internal links *and* on the size of the community itself. Consider now again a

network composed of  $\kappa$  disjoint communities. Assuming that each community has the same number of nodes  $N/\kappa$  and the same number of edges  $m/\kappa$ , the modularity density of such a network is:

$$Q_{ds}^\kappa = \kappa \left[ \frac{p_\kappa^*}{\kappa} - \left( \frac{p_\kappa^*}{\kappa} \right)^2 \right] = p_\kappa^* \left( 1 - \frac{p_\kappa^*}{\kappa} \right), \quad (7.6)$$

where

$$p_\kappa^* = \frac{2m}{N \left( \frac{N}{\kappa} - 1 \right)}$$

is the connection density of the communities. The first major difference between Eq. 7.3 and Eq. 7.6 is that  $Q_{ds}^\kappa$  depends not only on the number of communities, but also on their density of edges, unlike traditional modularity, which only depends on  $\kappa$ . Also, for a fixed value of  $\kappa$ ,  $Q_{ds}^\kappa$  increases with  $p_\kappa^*$  (see Fig. 7.1). This is remarkable, since it indicates that the strength of the partition increases as more links are added within each group, in striking opposition with the behaviour of traditional modularity. We also note that for a fixed value of  $p_\kappa^*$ , modularity density increases with the number of communities. Its theoretical maximum is reached in the limiting case of an infinite number of communities, with the special requirement that they are all cliques. Moreover, in one more substantial difference with traditional modularity, a network composed of few cliques in general has a higher modularity density than a network composed of an infinite number of sparse communities.

Finally, we study the test case of the ring of  $\kappa$  communities each linked by a single edge to the next community and a single edge to the previous one. As before, it is  $m_C = m/\kappa - 1$  and  $e_C = 2$ . In addition,  $m_{C\tilde{C}} = 1$  and  $n_C = N/\kappa$ . Introducing the variables

$$\beta_\kappa = \frac{\frac{m}{\kappa} - 1}{m}$$

and

$$p_\kappa^* = \frac{2 \left( \frac{m}{\kappa} - 1 \right)}{\frac{N}{\kappa} \left( \frac{N}{\kappa} - 1 \right)},$$

we can write the modularity density as

$$Q_{ds}^{\text{ring}} = \kappa \left[ \beta_\kappa p_\kappa^* - \left( \beta_\kappa + \frac{1}{m} \right)^2 (p_\kappa^*)^2 - \frac{\kappa^2}{mN^2} \right]. \quad (7.7)$$

The optimal number of communities is the one that maximizes this expression, or, equivalently, the one for which its derivative vanishes. Differentiating  $Q_{ds}^{\text{ring}}$  with

respect to  $\kappa$ , we obtain

$$\begin{aligned} \frac{\partial Q_{ds}^{\text{ring}}}{\partial \kappa} &= \kappa \beta_\kappa \partial_\kappa p_\kappa^* - 2 \left( \beta_\kappa + \frac{1}{m} \right) p_\kappa^* \partial_\kappa p_\kappa \\ &\quad + \left( \beta_\kappa + \frac{1}{m} \right)^2 (p_\kappa^*)^2 + \left( \beta_\kappa - \frac{1}{\kappa} \right) p_\kappa - \frac{3\kappa^2}{mN^2}, \end{aligned}$$

with

$$\partial_\kappa p_\kappa^* = \frac{2(\kappa^2 - 2\kappa N + mN)}{N(\kappa - N)^2}.$$

This expression does not have a simple general root in terms of  $\kappa$ . Rather, the solutions depend on the local and global properties of the network. Thus, the number of groups does not seem to be constrained by an intrinsic scale of order  $\sqrt{m}$ .

As briefly discussed above, a major drawback of traditional modularity is that algorithms based on its maximization often find supposedly viable partitions on graphs with no ground-truth community structure. In such cases, the correct partition is either the one where all nodes are placed together, or the one with  $N$  communities, each consisting of a single node. In either case, modularity vanishes. Thus, modularity-maximizing algorithms often suggest spurious community structures simply because they have a non-zero modularity. Conversely, from Eq. 7.5 it follows that the one-group partition has a modularity density

$$Q_{ds}^1 = p(1 - p),$$

where  $p$  is the network density. Note that this expression is a parabola, whose roots are  $p = 0$  and  $p = 1$ , which are the fully disconnected and fully connected graphs, respectively. Thus, a partition's  $Q_{ds}$  needs not only to be positive, but also to lie above the parabola for an algorithm based on modularity density maximization to accept it. We will see that this makes such algorithms not find communities on random graphs, as should be the case for a reliable community detection method.

### 7.3 A modularity density maximisation algorithm

Having discussed the advantages of modularity density as a quality function, we propose a community detection algorithm based on its maximization. Currently, the only published modularity density algorithm [224] is based on iterations of two steps, namely splitting and merging. The algorithm is divisive, starting from a partition where all the nodes are placed in a single community and then using bisections.

Each splitting is performed using the Fiedler vector of the network, which is the eigenvector of the graph Laplacian corresponding to the second smallest eigenvalue. The graph Laplacian  $L$  is defined as  $L = D - A$ , where  $D$  is the diagonal matrix of the node degrees. The merging steps try to merge pairs of communities together if doing so improves the current partition. The two steps are repeated until the partition cannot be improved any longer, and the algorithm is deterministic, meaning that the same initial network always yields the same partition. Here, we extend and adapt an existing modularity maximisation algorithm, originally proposed in [186], which achieves the largest published scores of traditional modularity. Along the lines of the original method, our algorithm consists of four main steps, which we describe below. Section 7.4 contains a fully detailed discussion of the algorithm implementation and its computational complexity.

### Bisection

In this step, we try to bisect the community under consideration. To do so, we use the leading eigenvector of the modularity matrix. Despite suffering from the limitations discussed above, modularity still provides a good initial guess for a partition that is then refined by the subsequent steps.

### Fine tuning

After every bisection, the partition can be often improved by using a variant of the Kernighan-Lin algorithm [228]. We consider moving every node  $i$  from the community into which it was assigned to the other. Every such move would result in a change  $\Delta Q_{ds}^i$  of the quality function, and we perform the move yielding the largest of such changes  $\Delta_{\max} Q_{ds}^i$ . Note that we introduce here a non-deterministic factor: given a tolerance parameter  $\tau_{acc}$ , we consider all moves achieving a change of modularity density within the interval  $[\Delta_{\max} Q_{ds}^i - \tau_{acc}, \Delta_{\max} Q_{ds}^i]$  to be equivalent. If more than one move falls within the acceptance interval, we randomly choose one to accept. This stochasticity allows the algorithm to explore the partition space without getting stuck on a local maximum, since it can accept moves that are not always optimal. Once a move has been performed, the corresponding node is flagged as blocked. Then, every non-blocked node is considered again and the procedure is repeated, until all nodes have been considered. At the completion of an iteration of this step, a decision tree is formed where each node of the tree represents a sequence of nodes in the network switching community, with an associated  $\Delta Q_{ds}$  equal to the sum of all the changes in modularity density along the branches leading to the tree

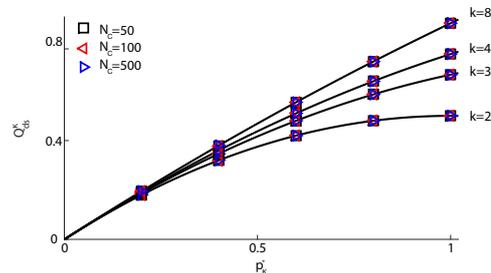
node. Then, we randomly choose a node in the decision tree amongst those achieving the largest positive increase in modularity density within an interval determined by the tolerance parameter  $\tau_{acc}$ , and perform all the moves corresponding to the chosen node. Finally, the whole step is repeated until no improvement in  $Q_{ds}$  can be obtained.

### Final tuning

A further refinement of a current partition can be achieved by performing an additional tuning step. In the final tuning, we consider every node  $i$  and try to move it to every other possible community  $C$  already present in the partition. The step is performed in a similar fashion to the fine tuning, repeatedly considering all the moves which result in an increase of modularity density in a small interval defined by the tolerance parameter  $\tau_{acc}$  until all nodes have been moved. As before, we build a decision tree of partial switches and then perform all the moves up to the level in the tree that has been selected amongst those yielding the largest increase in  $Q_{ds}$ . We repeat this step until no further refinements can be found.

### Agglomeration

A step that merges pairs of communities is fundamental. First of all, unlike both tuning steps, which are local because they only consider moving one node at a time, merging communities is a non-local step that allows one to better explore the landscape of modularity density [186]. For example, merging two entire communities can result in an increase of the quality function while partial mergers, i.e., moving only some nodes from one community to the other, could still have a lower score than the starting partition. Therefore, using only local moves, one could discard those partial mergers because they temporarily decrease the partition score, thus never achieving the beneficial complete merging of the two communities. In the case of modularity density, the agglomeration step is even more important, since no series of local moves could ever produce the full merging of two communities. This happens because modularity density does not allow communities of size 1. Thus, even if local steps had succeeded in moving all nodes except two from one community to another, any further move would be prohibited because it would result in a single-node community. This makes a global move essential for our algorithm. In the agglomeration step, we consider pairs of communities  $C$  and  $\tilde{C}$  and try to merge them. Each move results in a change in modularity density  $\Delta Q_{ds}^{C,\tilde{C}}$  and we randomly choose the move amongst those in the interval  $\left[ \Delta_{\max} Q_{ds}^{C,\tilde{C}} - \tau_{acc}, \Delta_{\max} Q_{ds}^{C,\tilde{C}} \right]$ , where



**Figure 7.2:** Modularity density for networks of  $\kappa$  disconnected communities. The predictions of Eq. 7.6 (solid lines) are confirmed by numerical simulations throughout the range of  $p_{\kappa}^*$  and for different values of  $\kappa$ . For each  $\kappa$ , we consider groups with 50, 100 and 500 nodes, respectively. Additionally, and as expected, we observe that the value of modularity density does not depend on the number of nodes in each community, but only on the number of communities and their internal density. Each point is the average over 100 network realizations.

$\Delta_{\max} Q_{ds}^{C, \tilde{C}}$  is the largest increase in modularity density achieved by any move. We build a decision tree by progressively merging pairs of communities, until there is only a single community left. We then look at the nodes in the tree corresponding to the largest increase in modularity density but, in difference from the previous steps, if more than one node results in the same increase, we select the one with the smallest number of communities. The whole step is repeated until the current partition cannot be improved further.

### Summary

With these four steps, the algorithm can be summarised as:

- Start with a single community containing all nodes.
- Try to bisect the network using the leading eigenvector of the modularity matrix.
- If the bisection was successful, then perform a fine tuning step.
- Iterate the bisection and fine tuning steps on each of the communities in the current partition, until no further splitting and refinement can be performed.
- Perform the final tuning step.
- Perform the agglomeration step.

- Repeat the sequence of steps until it is no longer possible to find an increase in modularity density.

As described in detail in the next Section 7.4, the worst-case computational complexity of the full algorithm is  $O(N^2)$ .

## 7.4 Implementation details

Here, we provide a detailed description of the implementation of the algorithm presented above. To describe how the different steps are carried out, first we introduce some notation. Let  $P = |\mathcal{C}|$  be the size of the current partition. Then, let  $M$  be the partition adjacency matrix of the network, i.e., the  $P \times P$  matrix whose elements  $m_{C\tilde{C}}$  are the number of links between community  $C$  and community  $\tilde{C}$ . Also, let  $X$  be the community spectra matrix, i.e., the  $N \times P$  matrix whose elements  $x_{iC}$  are the number of links between node  $i$  and nodes in community  $C$ . Finally, let  $\mathbf{S}$  be the  $P$ -dimensional community size vector, whose elements are the sizes of the communities.

Note that our implementation uses three tolerance parameters:

1. Power method tolerance  $\tau_{pwm}$ . This parameter determines the tolerance for the floating-point comparisons in the power method.
2. Bisection tolerance  $\tau_{bs}$ . Since a bisection with the leading eigenvector of the classical modularity matrix does not guarantee an increase in modularity density, we introduce a tolerance  $\tau_{bs}$ . After each bisection, we check the difference between the new and old values of modularity density. A bisection is accepted if modularity density increases or if it decreases by an amount smaller than  $\tau_{bs}$  (more details are given in Section 7.1.1).
3. Acceptance tolerance  $\tau_{acc}$ . This parameter defines the size of the tolerance range when finding the moves that maximally increase modularity density during tuning and agglomeration steps.

### 7.1.1 Bisection

The first step in the algorithm attempts to bisect a community, which can be either the whole network or a previously determined community, using the traditional modularity matrix. To do so, we use the spectral method, which we briefly review

here. The modularity matrix  $B$  is defined as

$$B = A_{ij} - \frac{k_i k_j}{2m},$$

and the expression for the modularity of a given partition is

$$Q = \frac{1}{2m} \sum_{ij} B_{ij} \delta_{C_i C_j}. \quad (7.8)$$

Since we are only considering a potential bisection,  $C_i$  can only assume two values. Thus, a partition can be represented by a vector  $\mathbf{s}$  whose entries  $s_i$  are 1 and  $-1$  if node  $i$  is assigned to the first or the second community resulting from the split, respectively. Then, substituting the expression

$$\delta_{C_i C_j} = \frac{1}{2}(s_i s_j + 1)$$

in Eq. 7.8, it is

$$Q = \frac{1}{4m} \sum_{ij} B_{ij} s_i s_j.$$

The vector  $\mathbf{s}$  can be expressed in terms of the normalized eigenvectors of  $B$  as

$$\mathbf{s} = \sum_{i=1}^N \vartheta_i \mathbf{v}_i,$$

where the  $\vartheta$  are linear combination coefficients, and  $\mathbf{v}_i$  is the  $i^{\text{th}}$  eigenvector of the modularity matrix, corresponding to the eigenvalue  $\lambda_i$ . substituting in Eq. 7.8, we obtain

$$Q = \frac{1}{4m} \sum_{i=1}^N \vartheta_i^2 \lambda_i.$$

If we label the eigenvalues so that  $\lambda_1 > \lambda_2 > \dots > \lambda_N$ , this expression is maximized when  $\mathbf{s}$  is parallel to the leading eigenvector  $\mathbf{v}_1$ . However,  $\mathbf{s}$  is a vector whose entries can only be  $\pm 1$ . Thus, we can only choose its elements to make it as parallel to  $\mathbf{v}_1$  as possible. One way of achieving this is to set  $s_i = 1$  if  $v_{1i} > 0$  and  $s_i = -1$  if  $v_{1i} < 0$ . Then, the bisection consists in finding the leading eigenvector of  $B$  and, if the corresponding eigenvalue is positive, dividing the nodes according to this rule. Several methods can be used to diagonalize  $B$ . Since we only need to find a single eigenvector, and this step only provides a starting guess, we choose to use the power method, which offers a good tradeoff between speed and accuracy.

Consider a matrix  $B$ . We want to solve the following equation:

$$Bv = \lambda_{max}v$$

where  $\lambda_{max}$  is the eigenvalue with the largest absolute value. Consider now to start with a vector  $v_0$ . The method iterates the following equation:

$$v_{k+1} = \frac{Bv_k}{\|Bv_k\|}$$

until it converges. The two steps in the iteration, therefore, involve multiplying the current approximate eigenvector  $v_k$  by the modularity matrix  $B$  and then normalising it.

In more detail, assume that the modularity matrix  $B$  has size  $N \times N$ . It then has  $N$  eigenvectors  $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_N$  and eigenvalues  $\lambda_1, \lambda_2, \dots, \lambda_N$ . Assume that the eigenvalues are ordered in decreasing absolute value, i.e.  $\lambda_{max} = |\lambda_1| > |\lambda_2| > \dots > |\lambda_N|$ . We can decompose our initial vector  $\mathbf{v}_0$  on the basis of the eigenvectors of  $B$ :

$$\mathbf{v}_0 = \sum_{k=1}^N a_k \mathbf{v}_k = a_1 \mathbf{v}_1 + a_2 \mathbf{v}_2 + \dots + a_N \mathbf{v}_N$$

We can now perform the first iteration step by multiplying this equation by the modularity matrix  $B$  on both sides:

$$\begin{aligned} B\mathbf{v}_0 &= a_1 B\mathbf{v}_1 + a_2 B\mathbf{v}_2 + \dots + a_N B\mathbf{v}_N \\ &= a_1 \lambda_1 \mathbf{v}_1 + a_2 \lambda_2 \mathbf{v}_2 + \dots + a_N \lambda_N \mathbf{v}_N \end{aligned}$$

Iterating a second time:

$$\begin{aligned} B(B\mathbf{v}_0) &= a_1 \lambda_1 B\mathbf{v}_1 + a_2 \lambda_2 B\mathbf{v}_2 + \dots + a_N \lambda_N B\mathbf{v}_N \\ &= a_1 \lambda_1^2 \mathbf{v}_1 + a_2 \lambda_2^2 \mathbf{v}_2 + \dots + a_N \lambda_N^2 \mathbf{v}_N \end{aligned}$$

It is clear, then, that repeated iterations yield the following:

$$\begin{aligned} B^k \mathbf{v}_0 &= a_1 \lambda_1^k \mathbf{v}_1 + a_2 \lambda_2^k \mathbf{v}_2 + \dots + a_N \lambda_N^k \mathbf{v}_N \\ &= \lambda_1^k \left[ a_1 \mathbf{v}_1 + a_2 \left( \frac{\lambda_2}{\lambda_1} \right)^k \mathbf{v}_2 + \dots + a_N \left( \frac{\lambda_N}{\lambda_1} \right)^k \mathbf{v}_N \right] \end{aligned}$$

Remembering that the eigenvalues are ordered in decreasing order, we can write that:

$$\left(\frac{\lambda_i}{\lambda_1}\right)^k \rightarrow 0 \quad \text{as } k \rightarrow \infty, \forall i > 1$$

From which:

$$B^k \mathbf{v}_0 \rightarrow a_1 \lambda_1^k \mathbf{v}_1 \quad \text{as } k \rightarrow \infty$$

Or, more formally:

$$\lim_{k \rightarrow \infty} \left(\frac{B^k \mathbf{v}_0}{\lambda_1}\right)^k = a_1 \mathbf{v}_1$$

We should note here that this method only works if  $a_1 \neq 0$ . In our initial choice for  $\mathbf{v}_0$ , therefore, we need to make sure that it has a non-zero value on the component parallel to the leading eigenvector. Typically, we choose the entries of  $\mathbf{v}_0$  to be random values and then normalise the vector to 1. In principle, we can then compute the leading eigenvalue and the corresponding eigenvector of the modularity matrix. However, the power method does not tell us anything about the sign of the eigenvalue. The iterations could converge to the eigenvector corresponding to the largest negative eigenvalue, which is of no interest for our purposes. To account for this, we need to introduce an adaptation of the method described so far.

Assume that the power method, after sufficient iterations, has converged to a negative eigenvalue denoted  $\Lambda$  ( $\Lambda < 0$ ). Let  $\mathbf{x}$  be the corresponding eigenvector:

$$B\mathbf{x} = \Lambda\mathbf{x}$$

We can then note the following:

$$(B - \Lambda I)\mathbf{x} = B\mathbf{x} - \Lambda\mathbf{x} = \Lambda\mathbf{x} - \Lambda\mathbf{x} = 0$$

which means that, if  $\mathbf{x}$  is the eigenvector corresponding to the largest negative eigenvalue of  $B$ , then  $\mathbf{x}$  will also be an eigenvector of  $(B - \Lambda I)$  with eigenvalue 0. Moreover, since  $\Lambda$  was the largest negative eigenvalue, by shifting the matrix  $B$  up by  $\Lambda$ , we ensured that all eigenvalues of the new matrix are now non negative. A new set of iterations of the power method on the shifted matrix, therefore, will converge to the largest (and positive) eigenvalue.

Assume now that  $\mathbf{y}$  is an eigenvector of  $B$  with eigenvalue  $|\lambda| < |\Lambda|$ :

$$B\mathbf{y} = \lambda\mathbf{y}$$

We can easily see that  $y$  is also an eigenvector of the shifted matrix:

$$(B - \Lambda)\mathbf{y} = B\mathbf{y} - \Lambda\mathbf{y} = \lambda\mathbf{y} - \Lambda\mathbf{y} = (\lambda - \Lambda)\mathbf{y} = (\lambda + |\Lambda|)\mathbf{y}$$

So that  $y$  is an eigenvector of the shifted matrix with eigenvalue  $\lambda + |\Lambda|$ .

Using the same trick, we can also solve a similar potential issue of the method. We could not only be facing the situation where the leading eigenvalue is negative, but also the scenario where there are two leading eigenvalues, one positive and one negative, of similar absolute value. In such a scenario, our estimate for the eigenvalue might start oscillating between these two eigenvalues, and corresponding eigenvectors. As before, a shift in the matrix would not change the eigenvectors and would only result in a constant shift in the eigenvalues. After the shift, we do not have eigenvalues of similar magnitude and, therefore, the power method can then converge to the correct answer.

At the end of the method, the absolute value of the shift is subtracted from the leading eigenvalue to give the correct final answer for the largest positive eigenvalue of the original modularity matrix.

In Algorithm 1 we provide a detailed implementation of the power method. Further consideration must be given to the fact that we are performing a bisection based on the modularity matrix, whereas our aim is to maximize modularity density. The potential problem is that a bisection based on modularity might not result in a larger value of modularity density. To avoid this, we introduce a tolerance parameter  $\tau_{bs}$ , whose role is to determine the largest possible decrease in modularity density that we want to accept when bisecting. In other words, if after the bisection the modularity density of the new partition has decreased by a value larger than  $\tau_{bs}$ , we do not accept the split, and keep the original partition. We consider only one exception to this rule, namely the first iteration of the bisection. At the start of the algorithm, all nodes are placed together and we try to bisect the whole network. At this point, we accept any bisection in order to allow at least a whole iteration of the whole algorithm. Indeed, if we didn't accept that, both the tuning and agglomeration steps could not be executed, thus leaving the network not partitioned. Note that not partitioning the network could be the correct answer, but we want to make sure that we have considered other partitions as well at least once. If not partitioning the network is the best answer, this will be found by the agglomeration step, that will merge all the communities together.

---

**Algorithm 1**

---

```

1: procedure POWER METHOD
2:    $g \leftarrow$  random normalised vector            $\triangleright$  initial guess for eigenvector
3:    $\Lambda \leftarrow 0$                               $\triangleright$  Initial value for eigenvalue
4:   while 1 do
5:     for  $i < N$  do  $g = g + B * g$                   $\triangleright B$  is the modularity matrix
6:     end for
7:      $\tilde{\Lambda} \leftarrow \Lambda$                         $\triangleright$  previous best guess for leading eigenvalue
8:      $\Lambda \leftarrow \sqrt{\|g\|}$                     $\triangleright$  update current best guess for leading eigenvalue
9:      $g \leftarrow \frac{g}{\Lambda}$                         $\triangleright$  normalise eigenvector
10:    if  $|\Lambda - \tilde{\Lambda}| < \tau_{pwm}$  then          $\triangleright$  if method is converging
11:      break
12:    end if
13:  end while
14:  if  $\Lambda < 0$  then                                $\triangleright$  if the leading eigenvalue found is negative
15:     $\text{shift} \leftarrow \Lambda$ 
16:     $B \leftarrow B + \Lambda \mathcal{K}$                   $\triangleright$  shift the modularity matrix
17:    repeat lines from 4 to 13
18:  end if
19:  if  $\text{shift} = 0$  then                                $\triangleright$  if the shift is zero, check for oscillations
20:    perform 100 iterations of lines from 4 to 13
21:    if leading eigenvalue oscillates then
22:      shift  $B$  and repeat lines from 4 to 13
23:    end if
24:  end if
25:   $\lambda \leftarrow \Lambda - |\text{shift}|$               $\triangleright$  largest positive eigenvalue
26:   $v \leftarrow g$                                     $\triangleright$  corresponding eigenvector
27: end procedure

```

---

Finally, we note that the previous expression for  $B$  is correct only when considering the whole network. When trying to partition a single community  $C$  which does not contain all the nodes, we need to construct an  $n_C \times n_C$  sub-modularity matrix  $B^C$  whose elements are

$$B_{ij}^C = A_{ij} - \frac{k_i k_j}{2m} - \delta_{ij} \left( k_i^C - k_i \frac{k_C}{2m} \right),$$

where  $k_i^C$  is the degree of node  $i$  within the community  $C$ . Using this matrix, we then perform the bisection step as described above.

In Algorithm 2, we present a detailed description of the implementation of this step. For each community, the computation of the leading eigenvalue through the power method requires  $O(m_c n_c)$  steps. Thus, the worst-case complexity of the the bisection step is  $O(mN)$ .

### 7.1.2 Tuning steps

The crucial part of both the fine tuning and final tuning steps is that they try to move individual nodes to different communities. Thus, we need to consider what happens to the current partition and how  $M$ ,  $X$  and  $\mathbf{S}$  change when we move a node  $i$  from community  $C$  to community  $\tilde{C}$ . Figure 7.3 provides an intuitive scheme to illustrate the changes that follow from such a move. In general, both the number of internal and external links of  $C$  will change, since node  $i$  is leaving this community. However, to correctly update the modularity density, we also need to keep track of the changes in all the specific numbers of links between  $C$  and every other community in the current partition. Similarly, we need to ensure that the internal and external links of  $\tilde{C}$  are updated correctly. Finally, the sizes of the two communities changes as well as a consequence of the move. Below, we describe how to efficiently perform these updates.

#### Updating the partition adjacency matrix

The partition adjacency matrix  $M$  keeps track of the number of edges between each pair of communities, as well as the internal number of edges of each community in its diagonal elements. Looking at Fig. 7.3, one can see that the following quantities change:

- The number of internal links of the community  $C$  that node  $i$  is leaving de-

**Algorithm 2**

Pseudocode for the bisection step.

---

```

1: procedure BISECTION STEP
2:   flag first bisection  $\leftarrow$  1 ▷ flag that this is the first bisection
3:    $w \leftarrow 1$  ▷  $w$  is the community under consideration
4:    $|S| \leftarrow 1$  ▷ Current number of communities
5:   while  $w \leq |S|$  do
6:     current number of nodes  $\leftarrow S[w]$  ▷  $S$  is the community size vector
7:     current nodes labels  $\leftarrow$  find nodes in  $S[w]$ 
8:      $B \leftarrow$  construct  $B$  ▷  $B$  is modularity matrix of the current nodes
9:     if current number of nodes  $> 2$  then
10:      leading  $\lambda$ , leading  $v \leftarrow$  power method( $B$ )
11:     end if
12:     if  $v$  has at least two negative and two positive components then
13:       flag bisection  $\leftarrow$  1
14:     end if
15:     if  $\lambda > 0$  & flag bisection then
16:       bisection( $v$ , current nodes labels, current number of nodes)
17:        $|S| \leftarrow |S| + 1$ 
18:       if old  $Q_{ds}$  - new  $Q_{ds} > \tau_{bs}$  and flag first bisection = 0 then
19:         cancel bisection
20:         flag[ $w$ ]  $\leftarrow$  1
21:         flag fine tuning  $\leftarrow$  0
22:       end if
23:       if  $S[w] > 2$  or  $S[w + 1] > 2$  and flag fine tuning then
24:         fine tuning(current number of nodes, current nodes labels)
25:       end if
26:       flag first bisection  $\leftarrow$  0
27:     else
28:       flag[ $w$ ]  $\leftarrow$  1 ▷ Flag  $w$  as blocked
29:     end if
30:     flag fine tuning  $\leftarrow$  1
31:     if flag[ $w$ ] then
32:        $w \leftarrow w + 1$ 
33:     end if
34:   end while
35: end procedure

```

---

**Algorithm 3**

Pseudocode for the fine tuning step.

---

```

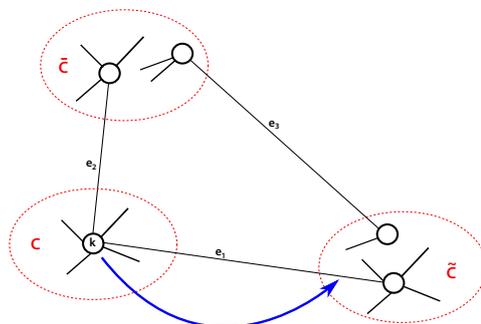
1: procedure FINE TUNING STEP
2:   flag increase  $\leftarrow$  1            $\triangleright$  Flag if there is an increase in modularity density
3:   while flag increase do
4:     flag increase  $\leftarrow$  0            $\triangleright$  Reset the flag
5:     for  $i_1 <$  current number of nodes do
6:       for  $i_2 <$  current number of nodes do
7:         if flag node[ $i_2$ ] = 0 then            $\triangleright$  if node  $i_2$  is not blocked
8:           if  $x_{i_2, \tilde{C}} > 0$  then
9:              $\Delta Q_{ds}[i_2] \leftarrow$  change in  $Q_{ds}$  if  $i_2$  changes community
10:          end if
11:         end if
12:       end for
13:       max  $\Delta Q_{ds} \leftarrow$  maximum increase in  $Q_{ds}$ 
14:       find all nodes within  $\tau_{acc}$  from max  $\Delta Q_{ds}$ 
15:       node to move  $\leftarrow$  pick randomly between nodes with max  $\Delta Q_{ds}$ 
16:       flag node[node to move]  $\leftarrow$  1
17:       fine tuning tree[ $i_1$ ]  $\leftarrow$  fine tuning tree[ $i_1 - 1$ ] + max  $\Delta Q_{ds}$ 
18:     end for
19:     max  $\Delta Q_{ds} \leftarrow$  max( fine tuning tree)
20:     if max  $\Delta Q_{ds} > 0$  then
21:       find all steps within  $\tau_{acc}$  of max  $\Delta Q_{ds}$ 
22:       step in fine tuning tree  $\leftarrow$  pick randomly step with max  $\Delta Q_{ds}$ 
23:       perform all updates in fine tuning tree until the chosen step
24:       flag increase  $\leftarrow$  1
25:     end if
26:   end while
27: end procedure

```

---

creases by the internal degree of node  $i$ , which is the number of links it has to other nodes in  $C$ .

- The number of internal links of the community  $\tilde{C}$  that node  $i$  is moving to increases by the number of links node  $i$  has with other nodes in  $\tilde{C}$ .
- The number of links between the old and the new community of node  $i$  increases by the number of links between  $i$  and its old community, and decreases by the number of links between  $i$  and its new community.
- The number of links between the old community  $C$  and all the other communities  $\tilde{C} \notin \{C, \tilde{C}\}$  decreases by the number of links between  $i$  and nodes in  $\tilde{C}$ .



**Figure 7.3:** Schematic illustration of node  $i$  moving from community  $C$  to community  $\tilde{C}$ .

- The number of links between the new community  $\tilde{C}$  and all the other communities  $\bar{C} \notin \{C, \tilde{C}\}$  increases by the number of links between  $i$  and nodes in  $\bar{C}$ .

In formulae:

$$\begin{aligned}
 m_C &\rightarrow m_C - x_{iC} \\
 m_{\tilde{C}} &\rightarrow m_{\tilde{C}} + x_{i\tilde{C}} \\
 m_{C\tilde{C}} &\rightarrow m_{C\tilde{C}} + x_{iC} - x_{i\tilde{C}} \\
 m_{C\bar{C}} &\rightarrow m_{C\bar{C}} - x_{i\bar{C}} & \forall \bar{C} \notin \{C, \tilde{C}\} \\
 m_{\tilde{C}\bar{C}} &\rightarrow m_{\tilde{C}\bar{C}} + x_{i\bar{C}} & \forall \bar{C} \notin \{C, \tilde{C}\},
 \end{aligned}$$

where we dropped the repeated index for the diagonal elements of  $M$  to keep the notation consistent.

### Updating the community spectra matrix

The rows of the matrix  $X$  are the community spectra of the nodes, containing the numbers of links that each node forms with nodes in all the individual communities in the current partition. When a node  $i$  changes community, its community spectrum does not change. However, every neighbour of  $i$  will experience a change in the number of connections it has to nodes in the old and new communities of  $i$ . In particular, in moving node  $i$  from  $C$  to  $\tilde{C}$ , the following changes happen:

- Since  $i$  is no longer in community  $C$ , all the nodes connected to  $i$  have one link less to  $C$ .
- Since  $i$  is now in community  $\tilde{C}$ , all the nodes connected to  $i$  have one connection

more to  $\tilde{C}$ .

In formulae:

$$\begin{aligned} x_{lC} &\rightarrow x_{lC} - 1 & \forall l \mid A_{il} = 1 \\ x_{l\tilde{C}} &\rightarrow x_{l\tilde{C}} + 1 & \forall l \mid A_{il} = 1. \end{aligned}$$

### Updating the community size vector

---

#### Algorithm 4

Pseudocode for the final tuning step.

---

```

1: procedure FINAL TUNING STEP
2:   flag increase  $\leftarrow$  1            $\triangleright$  Flag if there is an increase in modularity density
3:   while flag increase do
4:     flag increase  $\leftarrow$  0            $\triangleright$  Reset the flag
5:     for  $i_1 < N$  do
6:       for  $i_2 < N$  do
7:         if flag node[ $i_2$ ] = 0 then            $\triangleright$  if node  $i_2$  is not blocked
8:           for  $\bar{C} < |S|$  do
9:             if  $x_{i_2, \bar{C}} > 0$  then            $\triangleright$  if  $i_2$  has links to  $\bar{C}$ 
10:               $\Delta Q_{ds}[i_2][\bar{C}] \leftarrow$  change in  $Q_{ds}$  if  $i_2$  goes to  $\bar{C}$ 
11:            end if
12:          end for
13:        end if
14:      end for
15:      max  $\Delta Q_{ds} \leftarrow$  maximum increase in  $Q_{ds}$ 
16:      find all nodes within  $\tau_{acc}$  from max  $\Delta Q_{ds}$ 
17:      node to move  $\leftarrow$  pick randomly between nodes with max  $\Delta Q_{ds}$ 
18:      flag node[node to move]  $\leftarrow$  1
19:      final tuning tree[ $i_1$ ]  $\leftarrow$  final tuning tree[ $i_1 - 1$ ] + max  $\Delta Q_{ds}$ 
20:    end for
21:    max  $\Delta Q_{ds} \leftarrow$  max( final tuning tree)
22:    if max  $\Delta Q_{ds} > 0$  then
23:      find all steps within  $\tau_{acc}$  of max  $\Delta Q_{ds}$ 
24:      step in final tuning tree  $\leftarrow$  pick randomly step with max  $\Delta Q_{ds}$ 
25:      perform all updates in final tuning tree until the chosen step
26:      flag increase  $\leftarrow$  1
27:    end if
28:  end while
29: end procedure

```

---

**Algorithm 5**

Pseudocode for the agglomeration step.

---

```

1: procedure AGGLOMERATION STEP
2:   flag increase  $\leftarrow$  1       $\triangleright$  Flag if there is an increase in modularity density
3:   while flag increase do
4:     flag increase  $\leftarrow$  0       $\triangleright$  Reset the flag
5:     for  $\tilde{C} < |S|$  do
6:       for  $\hat{C} < |S|$  do
7:         if flag community[ $\tilde{C}$ ] = 0 then       $\triangleright$  if  $\tilde{C}$  is not blocked
8:           for  $\hat{C} < |S|$  do
9:             if flag community[ $i_3$ ] = 0 &  $m_{\tilde{C},\hat{C}} > 0$  then
10:               $\Delta_{Q_{ds}}[\tilde{C}][\hat{C}] \leftarrow$  change in  $Q_{ds}$  if we merge  $\tilde{C}$  and  $\hat{C}$ 
11:            end if
12:          end for
13:        end if
14:      end for
15:      max  $\Delta_{Q_{ds}} \leftarrow$  maximum increase in  $Q_{ds}$ 
16:      find pairs of communities within  $\tau_{acc}$  from max  $\Delta_{Q_{ds}}$ 
17:      communities to merge  $\leftarrow$  pick between those with max  $\Delta_{Q_{ds}}$ 
18:      flag community[ $\tilde{C}|\hat{C}$ ]  $\leftarrow$  1       $\triangleright$  Flag only the one with largest index
19:      agglomeration tree[ $i_1$ ]  $\leftarrow$  agglomeration tree[ $i_1 - 1$ ] + max  $\Delta_{Q_{ds}}$ 
20:    end for
21:    max  $\Delta_{Q_{ds}} \leftarrow$  max( agglomeration tree)
22:    if max  $\Delta_{Q_{ds}} > 0$  then
23:      step in agglomeration tree  $\leftarrow$  picks step with max  $\Delta_{Q_{ds}}$  and smallest
      number of communities
24:      perform all updates in agglomeration tree until the chosen step
25:      flag increase  $\leftarrow$  1
26:    end if
27:  end while
28: end procedure

```

---

The updates to this vector are straightforward:

$$S_C \rightarrow n_C - 1$$

$$S_{\tilde{C}} \rightarrow n_{\tilde{C}} + 1.$$

**Change in modularity density**

Since  $Q_{ds}$  is defined as a sum over all current communities, we consider the terms in its expression (Eq. 7.4) separately, and show how they change when node  $i$  moves from community  $C$  to community  $\tilde{C}$ . We first look at what happens to the contri-

butions of a community  $\bar{C}$  different from  $C$  and  $\tilde{C}$ . In this case, the only changes happen for two terms in the internal sum:

$$\sum_{\hat{C} \neq \bar{C}} \frac{m_{\bar{C}\hat{C}}^2}{2mn_{\bar{C}}n_{\hat{C}}} \rightarrow \sum_{\hat{C} \notin \{C, \bar{C}, \tilde{C}\}} \frac{m_{\bar{C}\hat{C}}^2}{2mn_{\bar{C}}n_{\hat{C}}} + \frac{(m_{\bar{C}C} - x_{i\bar{C}})^2}{2mn_{\bar{C}}(n_C - 1)} + \frac{(m_{\bar{C}\tilde{C}} + x_{i\bar{C}})^2}{2mn_{\bar{C}}(n_{\tilde{C}} + 1)}.$$

Then, we consider the contribution of community  $C$ :

$$\begin{aligned} \frac{2m_C^2}{mn_C(n_C - 1)} &\rightarrow \frac{2(m_C - x_{iC})^2}{m(n_C - 1)(n_C - 2)} \\ \frac{2m_C + e_C}{2m} \frac{2m_C}{n_C(n_C - 1)} &\rightarrow \frac{2(m_C - x_{iC}) + e_C + x_{iC} - \sum_{\bar{C} \neq C} x_{i\bar{C}}}{2m} \frac{2(m_C - x_{iC})}{(n_C - 1)(n_C - 2)} \\ \sum_{\hat{C} \neq C} \frac{m_{C\hat{C}}^2}{2mn_Cn_{\hat{C}}} &\rightarrow \sum_{\hat{C} \notin \{C, \bar{C}\}} \frac{(m_{C\hat{C}} - x_{i\hat{C}})^2}{2m(n_C - 1)n_{\hat{C}}} + \frac{(m_{C\tilde{C}} + x_{iC} - x_{i\tilde{C}})^2}{2m(n_C - 1)(n_{\tilde{C}} + 1)}. \end{aligned}$$

Finally, we consider the contribution of community  $\tilde{C}$ :

$$\begin{aligned} \frac{2m_{\tilde{C}}^2}{mn_{\tilde{C}}(n_{\tilde{C}} - 1)} &\rightarrow \frac{2(m_{\tilde{C}} + x_{i\tilde{C}})^2}{m(n_{\tilde{C}} + 1)n_{\tilde{C}}} \\ \frac{2m_{\tilde{C}} + e_{\tilde{C}}}{2m} \frac{2m_{\tilde{C}}}{n_{\tilde{C}}(n_{\tilde{C}} - 1)} &\rightarrow \frac{2(m_{\tilde{C}} + x_{i\tilde{C}}) + e_{\tilde{C}} - x_{i\tilde{C}} + \sum_{\bar{C} \neq \tilde{C}} x_{i\bar{C}}}{2m} \frac{2(m_{\tilde{C}} + x_{i\tilde{C}})}{(n_{\tilde{C}} + 1)n_{\tilde{C}}} \\ \sum_{\hat{C} \neq \tilde{C}} \frac{m_{\tilde{C}\hat{C}}^2}{2mn_{\tilde{C}}n_{\hat{C}}} &\rightarrow \sum_{\hat{C} \notin \{C, \bar{C}\}} \frac{(m_{\tilde{C}\hat{C}} + x_{i\hat{C}})^2}{2m(n_{\tilde{C}} + 1)n_{\hat{C}}} + \frac{(m_{\tilde{C}C} + x_{iC} - x_{i\tilde{C}})^2}{2m(n_{\tilde{C}} + 1)(n_C - 1)}. \end{aligned}$$

In Algorithm 3 and Algorithm 4, we present a detailed description of the implementation of the tuning steps. The complexity of computing the potential change in modularity density is  $O(P)$ , since we have to consider all the communities to update the split penalty term. For the fine tuning, this process is repeated  $N$  times per node, yielding a complexity of  $O(PN^2)$ . In the final tuning, instead, all communities are considered as potential targets, introducing an extra factor of  $P$  in the complexity, which becomes  $O(P^2N^2)$ . Note that these are worst case scenarios, since we typically do not have to consider all communities for the updates, because each node is only connected to a subset of them.

### 7.1.3 Agglomeration

The agglomeration step attempts the merger of pairs of communities. If a merger is carried out, a community is obtained whose size is the sum of the sizes of the

original ones. A delicate point is deciding the label of the new community. In our implementation, we always keep the smaller of the two labels. So, for instance, if we merge community 1 with community 4, the resulting community will be labelled 1 and community 4 will disappear. We then need to reassign the links of every node in the network to the new community, and also zero any link to the old community that disappeared. Below, we describe how to efficiently perform the required updates, assuming a merger between community  $C$  and community  $\tilde{C}$  in which the label of the resulting community is  $C$ .

### Updating the partition adjacency matrix

The following changes happen to the partition adjacency matrix:

- The number of internal links of the merged community is the sum of the internal links of the two original ones plus the number of links between the two.
- All the links of community  $\tilde{C}$  vanish, since it has been merged with community  $C$ .
- The number of links between the new community and any other community  $\bar{C}$  is the sum of the number of links between each of the two original communities and  $\bar{C}$ .

In formulae:

$$\begin{aligned}
 m_C &\rightarrow m_C + m_{\tilde{C}} + m_{C\tilde{C}} \\
 m_{\tilde{C}} &\rightarrow 0 \\
 m_{\tilde{C}\bar{C}} &\rightarrow 0 && \forall \bar{C} \in \mathcal{C} \\
 m_{C\bar{C}} &\rightarrow m_{C\bar{C}} + m_{\tilde{C}\bar{C}} && \forall \bar{C} \notin \{C, \tilde{C}\}.
 \end{aligned}$$

### Updating the community spectra matrix

The number of connections between every node  $i$  and the merged community is the sum of the number of links between  $i$  and each of the two original communities, and no node is connected to community  $\tilde{C}$  since it doesn't exist any more:

$$\begin{aligned}
 x_{iC} &\rightarrow x_{iC} + x_{i\tilde{C}} \\
 x_{i\tilde{C}} &\rightarrow 0.
 \end{aligned}$$

### Updating the community size vector

The changes to the Community Size Vector are once again straightforward:

$$\begin{aligned} S_C &\rightarrow n_C + n_{\tilde{C}} \\ S_{\tilde{C}} &\rightarrow 0 \end{aligned}$$

### Change in modularity density

As before, we consider the terms in the definition of modularity density separately, showing how they change for the merger considered. For the contribution of communities  $\tilde{C}$  other than  $C$  and  $\tilde{C}$ , the only changes happen in two terms in the internal sum:

$$\sum_{\hat{C} \neq \tilde{C}} \frac{m_{\tilde{C}\hat{C}}^2}{2mn_{\tilde{C}}n_{\hat{C}}} \rightarrow \sum_{\hat{C} \notin \{C, \tilde{C}, \bar{C}\}} \frac{m_{\tilde{C}\hat{C}}^2}{2mn_{\tilde{C}}n_{\hat{C}}} + \frac{(m_{\tilde{C}C} + m_{\tilde{C}\tilde{C}})^2}{2mn_{\tilde{C}}(n_C + n_{\tilde{C}})}.$$

Then, we consider the contribution of community  $C$ :

$$\begin{aligned} \frac{2m_C^2}{mn_C(n_C - 1)} &\rightarrow \frac{2(m_C + m_{\tilde{C}} + m_{C\tilde{C}})^2}{m(n_C + n_{\tilde{C}})(n_C + n_{\tilde{C}} - 1)} \\ \frac{2m_C + e_C}{2m} \frac{2m_C}{n_C(n_C - 1)} &\rightarrow \frac{2(m_C + m_{\tilde{C}} + m_{C\tilde{C}}) + e_C + e_{\tilde{C}} - 2m_{C\tilde{C}}}{2m} \frac{2(m_C + m_{\tilde{C}} + m_{C\tilde{C}})}{(n_C + n_{\tilde{C}})(n_C + n_{\tilde{C}} - 1)} \\ \sum_{\bar{C} \neq C} \frac{m_{C\bar{C}}^2}{2mn_Cn_{\bar{C}}} &\rightarrow \sum_{\bar{C} \notin \{C, \tilde{C}\}} \frac{(m_{C\bar{C}} + m_{\tilde{C}\bar{C}})^2}{2m(n_C + n_{\tilde{C}})n_{\bar{C}}}. \end{aligned}$$

Finally, the contribution of community  $\tilde{C}$  entirely vanishes.

In Algorithm 5, we present a detailed description of the implementation of the agglomeration step. The computational complexity is  $O(P^4)$ . Analogously to the tuning steps, this is the worst case scenario. In a typical situation, a community is only connected to a few others, and thus one does not need to update all the terms in the partition adjacency matrix.

#### 7.1.4 Community detection algorithm

Finally, in Algorithm 6 we provide a detailed description of how the steps presented above are linked together in our community detection algorithm. The overall complexity of the algorithm is dominated by the final tuning step, which is the most computationally expensive, with a complexity  $O(P^2N^2)$ . Along the lines

**Algorithm 6**

Pseudocode for the community detection method.

---

```

1: procedure COMMUNITY DETECTION
2:    $w \leftarrow 1$  ▷ Community under consideration
3:   flag repetition  $\leftarrow 1$  ▷ Flag if there is an increase in modularity density
4:   while flag repetition do
5:     flag repetition  $\leftarrow 0$ 
6:      $\tilde{Q}_{ds} \leftarrow Q_{ds}$ 
7:     Bisection
8:     if Current number of communities  $> 1$  then
9:       Final Tuning
10:      Agglomeration
11:       $\Delta Q_{ds} \leftarrow Q_{ds} - \tilde{Q}_{ds}$  ▷ Change in  $Q_{ds}$ 
12:      if  $\Delta Q_{ds} > 0$  then
13:         $w \leftarrow 1$  ▷ Restart from the first community
14:        flag repetition  $\leftarrow 1$  ▷ Repeat the whole algorithm
15:      end if
16:    else
17:      flag repetition  $\leftarrow 0$ 
18:    end if
19:  end while
20: end procedure

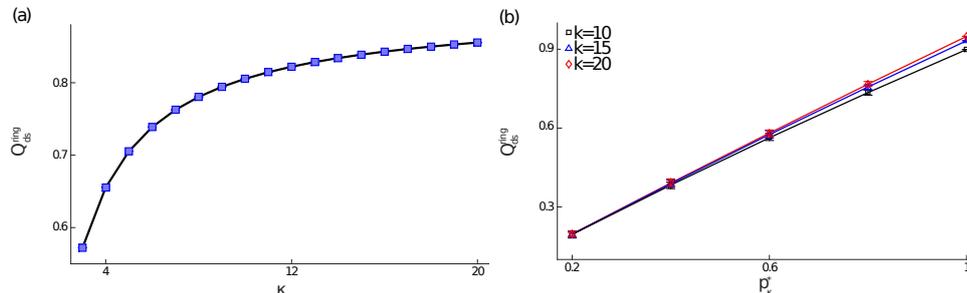
```

---

of [223, 224], we consider  $P$  a constant, and thus the worst-case complexity reduces to  $O(N^2)$ . To minimize running times, we take advantage of the independence of the incremental computing steps. Both the fine tuning and final tuning try to move nodes from one community to a different one. The calculations of the potential change in modularity density are independent of each other and thus can be performed in parallel, rather than serially. This task is fairly straightforward, and our implementation exploits the widely used C library *Open MP* to allow an efficient parallelization using multiple threads on each computing node during the tuning and agglomeration steps.

## 7.2 Validation

To validate our algorithm, we test it on several synthetic and real-world networks. First, we verify that it reproduces the theoretical predictions on networks of disconnected communities and on rings of modules, discussed in Section 7.1 and Section 7.2. Then, we analyse its behaviour on random networks belonging to different ensembles. Finally, we run it on a set of benchmark networks, comparing the results



**Figure 7.4: Modularity density for rings of communities.** We simulate ring networks composed by with a varying number of communities  $\kappa$  and compare the theoretical values of modularity density with the results of our algorithm. In panel (a) we consider networks of  $\kappa$  fully connected cliques, finding a perfect agreement between theoretical value (solid line) and simulations (squares). In panel (b), we build networks with different fixed values of  $\kappa$  and vary their internal density. Note that, differently from (a), here the groups are not fully connected. The theoretical values (lines) and simulation results match precisely. In both panels, each point is the average over 100 realizations of the same network.

with the best ones currently published.

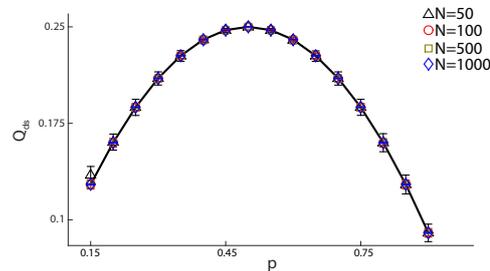
### 7.2.1 Disconnected communities and rings

First, we consider networks formed by  $\kappa$  disconnected communities. Equation 7.6 indicates that the modularity density of such networks depends only on the connection probability  $p_{\kappa}^*$  and on  $\kappa$  itself, but not on the size of each community. We find an exact agreement between the simulation results and the theoretical prediction for all the values of  $\kappa$  (Fig. 7.2). We also note that the values of modularity density found in the simulations do not depend on the number of nodes in the communities.

As a second test, we simulate two types of ring networks of communities. We start by making the communities cliques of 5 fully connected nodes, and vary  $\kappa$  from 3 to 20. From Eq. 7.7, the expected modularity density of these networks is

$$Q_{ds}^{\text{ring}} = \kappa \left[ \beta_{\kappa} - \left( \beta_{\kappa} + \frac{1}{m} \right)^2 - \frac{\kappa^2}{25m} \right].$$

The comparison between the modularity density predicted by this expression and the values obtained in our simulations is shown in Fig. 7.4(a). We find a precise agreement between the two, showing that our algorithm correctly identifies the cliques without splitting them, and finds the right value of modularity density. Next, we build ring networks in which we fix  $\kappa$  and vary the community density  $p_{\kappa}^*$ . Each

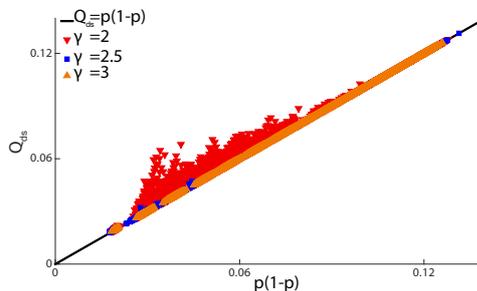


**Figure 7.5: Modularity density on Erdős-Rényi graphs.** We build ensembles of random networks, with different sizes and different densities, comparing the theoretical modularity density (solid line) and the one found by our algorithm. Up to finite-size effects for the smallest and least dense networks, we find a perfect agreement between theoretical prediction and simulation results, with all the results collapsing on the same curve. Each point is the average over 1000 realizations of the same network parameters.

community contains 50 nodes, and we vary  $p_\kappa^*$  from 0.2 to 1, perform the test for  $\kappa = 10$ ,  $\kappa = 15$  and  $\kappa = 20$ . The results, in Fig. 7.4(b), show a perfect agreement in all cases, again indicating that our algorithm correctly partitions the networks.

## 7.2.2 Random networks

As we argued in the previous sections, a desirable feature of a community detection algorithm is that it does not propose a complex partition of graphs without ground-truth community structure. To verify that our algorithm satisfies this requirement, we test it on Erdős-Rényi random graphs. For graphs in this ensemble, every possible edge between  $N$  nodes exists independently with probability  $p$ . Thus, the average number of edges is  $\frac{1}{2}Np(N-1)$ . These networks do not have any true community structure, since all their edges are fully random, and thus they are one of the benchmarks against which community detection algorithms are often tested. For our simulations, we create networks with values of  $p$  from 0.15 to 0.90 and number of nodes 50, 100, 500 and 1000. The results, in Fig. 7.5, show that for all network sizes, the average modularity density matches almost perfectly the theoretical prediction. Even for small networks, where finite-size effects are largest, the values lie in close proximity to the theoretical parabola and we can only observe a small deviation for the smallest networks at low values of  $p$ . Also note that all the results collapse on the theoretical curve, which does not depend on network size. These results represent a major improvement over modularity-based algorithms, that typically detect communities even on Erdős-Rényi networks.



**Figure 7.6: Modularity density for LFR networks.** We run our algorithm on random LFR networks without community structure, with  $N = 500$  nodes and varying parameters. In particular, we let the mean degree  $\langle k \rangle$  assume the values 15, 25, 35, 44 and 55, and the largest degree  $k_{\max}$  be 150, 200 and 250. For each combination of the parameters, we generate 100 networks and for each we record the edge density  $p$  and the largest modularity density our algorithm finds. The plot shows considerable agreement between the theoretical modularity density (solid line) and the one found by the algorithm. The only deviations appear for  $\gamma = 2$  and low  $p$ , and they are probably due to the breakdown of the model for this limiting value of the degree distribution exponent.

However, it is well known that most real-world networks are not well represented by Erdős-Rényi graphs. Rather, they are characterized by heterogeneous degree distributions. Thus, to further verify the performance of our algorithm, we test it on LFR networks [180]. These constitute a set of widely-used benchmark networks, whose distributions of degrees and community sizes follow a power-law  $P(k) \sim k^{-\gamma}$ . For our tests, we fix the network size to  $N = 500$  and vary the other parameters, namely the exponent  $\gamma$  of the degree distribution, the mean degree  $\langle k \rangle$  and the largest degree  $k_{\max}$ . Also, we ensure that the networks thus created contain a single community, so that no actual community structure is present. We run our algorithm on the networks thus generated and compare its results with the theoretical expectations. The results, presented in Fig. 7.6, show that for  $\gamma = 2.5$  and  $\gamma = 3$ , the modularity density found by the algorithm closely follows the predicted value for networks of all densities. We do observe, however, some deviations from the predicted values at  $\gamma = 2$ . This is probably due to the fact that, asymptotically, no networks exist with a pure power-law degree distribution for  $\gamma < 2$  [127]. Thus, in the limit of  $\gamma = 2$ , and particularly for low densities, a spurious structure of stars with bridges appears, effectively introducing communities in the networks.

**Table 7.1: Accuracy validation.** The comparison between the published results and the ones obtained with our algorithm on real-world and synthetic benchmark networks shows that our algorithm always performs better than the current best one. All the already published results are found in [224].

Benchmark	$Q_{ds}$	$Q_{ds, pub}$	$p$	$p(1 - p)$
Karate Club	0.235	0.231	0.139	0.120
Football Club	0.490931	0.4909	0.0935	0.0848
LFR, $\mu = 0.05$	$0.5220 \pm 0.0039$	0.4979	$0.0156 \pm 0.0001$	$0.0154 \pm 0.0001$
LFR, $\mu = 0.10$	$0.4638 \pm 0.0033$	0.4522	$0.0154 \pm 0.0001$	$0.0152 \pm 0.0001$
LFR, $\mu = 0.15$	$0.4249 \pm 0.0030$	0.4013	$0.0157 \pm 0.0002$	$0.0155 \pm 0.0002$
LFR, $\mu = 0.20$	$0.3982 \pm 0.0054$	0.384	$0.0156 \pm 0.0001$	$0.0154 \pm 0.0001$
LFR, $\mu = 0.25$	$0.3465 \pm 0.0085$	0.3347	$0.0156 \pm 0.0001$	$0.0154 \pm 0.0001$
LFR, $\mu = 0.30$	$0.2986 \pm 0.0034$	0.2619	$0.0156 \pm 0.0001$	$0.0154 \pm 0.0001$
LFR, $\mu = 0.35$	$0.2546 \pm 0.0101$	0.2377	$0.0156 \pm 0.0001$	$0.0154 \pm 0.0001$
LFR, $\mu = 0.40$	$0.2340 \pm 0.0069$	0.199	$0.0156 \pm 0.0001$	$0.0154 \pm 0.0001$
LFR, $\mu = 0.45$	$0.2029 \pm 0.0064$	0.169	$0.0156 \pm 0.0001$	$0.0154 \pm 0.0001$
LFR, $\mu = 0.50$	$0.1579 \pm 0.0027$	0.1385	$0.0156 \pm 0.0001$	$0.0154 \pm 0.0001$

### 7.2.3 Benchmark networks

We now verify the performance of our algorithm on some well known networks, for which results of the maximum modularity density obtained so far are available. The first is Zachary’s Karate Club network [229]. This is a friendship network between 34 members of a karate club in a U.S. university during the 1970s and it has become one of the most standard benchmarks to test community detection algorithms. The interest in this network lies in the fact that, not long after it was recorded, the club split into two subgroups due to internal problems between two members, namely the manager and the coach. Thus, a traditional challenge is to be able to detect these two groups based only on the friendship data available in the network topology, under the assumption that the members would decide to follow whichever leader they were more strongly related to between the coach and the manager. Of the 561 possible edges in the network, only 78 of them are present, making the network fairly sparse, with an effective connection probability  $p \approx 0.139$ .

A second benchmark network we consider is the American College Football Club network [159]. Here, the nodes represent different college football clubs and an edge connects two teams if there has been a regular-season game between them during the 2000 season. This network is known to have a natural community structure because the teams are divided into different leagues, thus making matches between teams more or less likely depending on the group they belong to. Finally, we con-

sider again some LFR benchmark networks, choosing a set of parameters for which already published results exist. Table 7.1 presents a comparison between the results obtained using our algorithm and the best results available in the literature. Because of the stochasticity within our method, for each value of the mixing parameter  $\mu$ , we create 10 realizations of the network and run the algorithm 100 times on each, reporting the average maximum modularity density found. In all cases considered, our algorithm finds a partition with higher modularity density than the best one currently published.

### 7.3 Conclusions

Communities are a fundamental structure that is often present in real-world complex networks. Thus, the ability to accurately and efficiently detect them is of great relevance to the analysis of complex data sets. Despite their success, traditional methods based on modularity have been shown to suffer from limitations. We have presented a detailed analysis of the properties of modularity density, an alternative quality function for community detection, showing that it does not suffer from the drawbacks that affect traditional modularity. In particular, modularity density does not depend separately on the size of the network or the number of edges, but only on the combination of these two properties in terms of the density of links within the communities. As a consequence, it allows a direct quantitative comparison of the community structure across networks of different sizes and number of edges. At the light of these considerations, we have introduced a new community detection algorithm based on modularity density maximization. Investigating its performance on Erdős-Rényi and heterogeneous random networks, we showed that it correctly identifies them as containing no actual communities. Moreover, our algorithm outperforms the other existing modularity-density-based method on every benchmark network that we tested. The high level of accuracy it reaches, its low computational complexity, and the ability to properly identify networks with no ground-truth communities make it a powerful tool to investigate complex systems and extract meaningful information from the network representation of large data sets, giving it a broad range of application throughout the physical sciences.

## CHAPTER 8

## CONCLUSIONS

Our interactions with technological systems, such as mobile phone networks and the Internet, generate a vast amount of data at an incredible pace. These data can be used to create an extremely detailed picture of our collective behaviour, our decision making processes, our interests, hobbies and several other aspects of our daily lives. These insights have the potential to help a range of stakeholders, including policy makers, who have an interest in understanding our collective behaviour in order to design a smarter and more sustainable society.

The studies described in this thesis provide insights into how these new forms of data may offer an unprecedented opportunity to improve our understanding of human behaviour. Previous work has demonstrated how the sudden availability of large and complex datasets has encouraged scientists to study our collective behaviour. Researchers have studied human mobility, how users experience the places where they live, and more generally the relationship between our behaviour on the Internet and in the real world. Here, we have shed light on several new aspects of our interactions with technological systems.

Financial transactions, such as buying or selling a stock, were historically one of the first sources of data containing detailed records of our decision making processes. In the first study presented in Chapter 3, we focussed on the behaviour of stocks in the *Dow Jones Industrial Average* index. Our analysis shows that the distribution of changes in stock market prices exhibits power law decay for short time scales, and exponential decay for larger time scales. Our findings may inform

the development of models of market behavior across varying time scales.

However, this is only one specific aspect of our decision making processes. Our interactions with technological systems, such as the Internet or mobile phone networks, are generating data which can offer insights into several other aspects of our lives. Indeed, in Chapters 4 and 5, we have shown how we can use the information contained in geolocated data derived from our ordinary interactions with smartphones and social media platforms, such as *Twitter* and *Instagram* to estimate the size of a crowd. Our findings suggest that these datasets could be used to infer the number of people in a given location at a given time. Such insights could be of great importance in emergency situations where crowd dynamics are critical and can lead to crowd collapses, such as evacuations or mass gatherings.

Our mobile phone records contain information not only on individual users and where they are, but also on our social interactions through phone calls. The spatial and temporal structure of how we interact with other people offers the opportunity to investigate yet another aspect of our behaviour. In Chapter 6, we have presented an analysis of the community structure of the network of mobile phone calls in the metropolitan area of Milan revealing spatial and temporal patterns of communications between people. Our findings suggest that the evolution of communities on a network induced by mobile phone calls capture the temporal evolution of our behaviour in everyday life.

The complexity of these new forms of data requires an increasingly complex set of methodologies. The results of research in the area of network science have led to the creation of a rich set of techniques to extract meaningful information from complex data sets. Here, we have studied in great detail the properties of a function, called modularity density, that can be defined on the partition of any network to measure its community structure. We have shown how it addresses drawbacks of existing methods, and we have introduced a sophisticated, yet efficient, algorithm based on this function. We have also shown how our algorithm outperforms analogous methods on a set of standard benchmark networks. Since modularity density has only recently been introduced, its properties were still mostly unknown. Our work helped to gain a greater understanding of modularity density and how it can be used to detect communities on complex networks.

Our work has laid the foundations for several novel streams of research. We have

shown how geolocated information, such as that we share on social media platforms, can be used to infer the size of a crowd in a given location at a given time. However, emergency situations, such as protests, might require more dynamic estimates. Further research can build on our findings and investigate whether data from social media platforms can be used to study how the crowd may move or how its size may vary. Researchers could also look at early signs of a large event taking place, and analyse whether the information flow on social media services can provide early estimates of the number of people attending it.

Recent methodological advancements in the area of multilayer networks will also be crucial in the analysis of data derived from different social media platforms. Each platform is only able to provide information on a specific subset of the population, which may be biased or not representative of the overall population. Multilayer networks offer the opportunity to integrate information coming from different sources in a unified framework.

Our analysis of the community structure of a mobile phone network has given insight into the evolution of our communication patterns. However, a plethora of techniques for analysing communities on networks is available, including the novel algorithm presented here. Further work could explore the robustness of our results when using different methodologies for finding communities on complex networks. This is a question of interest since the wide range of techniques, including recent techniques based on Bayesian statistical inference, provides a challenge for researchers interested in exploring network properties in their data.

Several extensions to our novel community detection algorithm could also be investigated. A major computational improvement could come through the use of computing on graphics processing units to perform some of the most computationally intensive steps, such as the two tuning steps. This could significantly reduce running times, thus making our algorithm feasible for larger networks. Further theoretical work could extend our work to different types of networks, such as weighted, directed and multilayered.

All in all, we have seen how the analysis of complex social datasets coupled with complex methodologies can offer unprecedented insights into human behaviour. We have seen how our interactions with technological systems can help us gain insight into different aspects of our society, from studying the properties of a crowd, to the

interactions between people making phone calls. The increasing availability of complex datasets requires constant development of the methodologies needed to analyse them. Here, we proposed developments to algorithms which can help us increase our understanding of community structure.

Together, the increasing availability of novel datasets alongside rapid development of new methodologies in data science represents an opportunity to dramatically improve our understanding of our collective behaviour and our society.

## BIBLIOGRAPHY

- [1] David Lazer, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Jebara Tony, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. Computational social science. *Science*, 323:721–723, 2009.
- [2] Alessandro Vespignani. Predicting the behavior of techno-social systems. *Science*, 325:425–428, 2009.
- [3] Alex Sandy Pentland. Reality mining of mobile communications: toward a new deal on data. *The Global Information Technology Report 2008–2009*, page 1981, 2009.
- [4] Gary King. Ensuring the data-rich future of the social sciences. *Science*, 331:719–721, 2011.
- [5] Jim Giles. Computational social science: making the links. *Nature*, 488:448–450, 2012.
- [6] Fosca Giannotti, Dino Pedreschi, Alex Pentland, Paul Lukowicz, Donald Kossmann, James Crowley, and Dirk Helbing. A planetary nervous system for social mining and collective awareness. *European Physical Journal: Special Topics*, 214:49–75, 2012.
- [7] Rosaria Conte, Nigel Gilbert, Giulia Bonelli, Claudio Cioffi-Revilla, Guillaume Deffuant, Janos Kertesz, Vittorio Loreto, Suzy Moat, J-P Nadal, Anxo Sanchez, Andrzej Nowak, Andreas Flache, Maxi San Miguel, and Dirk Hel-

- bing. Manifesto of computational social science. *European Physical Journal: Special Topics*, 214:325–346, 2012.
- [8] Helen Susannah Moat, Tobias Preis, Christopher Y Olivola, Chengwei Liu, and Nick Chater. Using big data to predict collective behavior in the real world. *Behavioral and Brain Sciences*, 37:92–93, 2014.
- [9] Helen Susannah Moat, Christopher Y Olivola, Nick Chater, and Tobias Preis. Searching choices: quantifying decision-making processes using search engine data. *Topics in Cognitive Science*, 8:685–696, 2016.
- [10] Adrian Letchford, Tobias Preis, and Helen Susannah Moat. Quantifying the search behaviour of different demographics using Google Correlate. *PLOS ONE*, 11:e0149025, 2016.
- [11] Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457:1012–1014, 2009.
- [12] John S Brownstein, Clark C Freifeld, and Lawrence C Madoff. Digital disease detection-harnessing the Web for public health surveillance. *New England Journal of Medicine*, 360:2153–2157, 2009.
- [13] Benjamin Althouse, Samuel V Scarpino, Lauren A Meyers, John W Ayers, Marisa Bargsten, Joan Baumbach, John S Brownstein, Lauren Castro, Hannah Clapham, Derek A T Cummings, Sara Del Valle, Stephen Eubank, Geoffrey Fairchild, Lyn Finelli, Nicholas Generous, Dylan George, David R Harper, Laurent Hebert-Dufresne, Michael A Johansson, Kevin Konty, Marc Lipsitch, Gabriel Milinovich, Joseph D Miller, Elaine O Nsoesie, Donald R Olson, Michael Paul, Philip M Polgreen, Reid Priedhorsky, Jonathan M Read, Isabel Rodriguez-Barraquer, Derek J Smith, Christian Stefansen, David L Swerdlow, Deborah Thompson, Alessandro Vespignani, and Amy Wesolowski. Enhancing disease surveillance with novel data streams: challenges and opportunities. *EPJ Data Science*, 4:17, 2015.
- [14] Hyunyoung Choi and Hal Varian. Predicting the present with Google Trends. *Economic Record*, 88:2–9, 2012.
- [15] Jaroslav Pavlicek and Ladislav Kristoufek. Nowcasting unemployment rates with Google searches: evidence from the Visegrad Group countries. *PLOS ONE*, 10:e0127084, 2015.

- 
- [16] Tobias Preis, Helen Susannah Moat, H Eugene Stanley, and Steven R Bishop. Quantifying the advantage of looking forward. *Scientific Reports*, 2:350, 2012.
- [17] Takao Noguchi, Neil Stewart, Christopher Y Olivola, Helen Susannah Moat, and Tobias Preis. Characterizing the time-perspective of nations with search engine query data. *PLOS ONE*, 9:e95209, 2014.
- [18] Tobias Preis, Daniel Reith, and H Eugene Stanley. Complex dynamics of our economic life on different scales: insights from search engine query data. *Philosophical Transactions of the Royal Society A*, 368:5707–5719, 2010.
- [19] Ilaria Bordino, Stefano Battiston, Guido Caldarelli, Matthieu Cristelli, Antti Ukkonen, and Ingmar Weber. Web search queries can predict stock market volumes. *PLOS ONE*, 7:e40014, 2012.
- [20] Tobias Preis, Helen Susannah Moat, and H Eugene Stanley. Quantifying trading behavior in financial markets using Google Trends. *Scientific Reports*, 3:1684, 2013.
- [21] Chester Curme, Tobias Preis, H Eugene Stanley, and Helen Susannah Moat. Quantifying the semantics of search behavior before stock market moves. *Proceedings of the National Academy of Sciences*, 111:11600–11605, 2014.
- [22] Ladislav Kristoufek. Power-law correlations in finance-related Google searches, and their cross-correlations with volatility and traded volume: evidence from the Dow Jones Industrial components. *Physica A*, 428:194–205, 2015.
- [23] Ladislav Kristoufek. Can Google Trends search queries contribute to risk diversification? *Scientific Reports*, 3:2713, 2013.
- [24] Ladislav Kristoufek. BitCoin meets Google Trends and Wikipedia: quantifying the relationship between phenomena of the Internet era. *Scientific Reports*, 3:3415, 2013.
- [25] Ladislav Kristoufek. What are the main drivers of the Bitcoin price? Evidence from wavelet coherence analysis. *PLOS ONE*, 10:e0123923, 2015.
- [26] Sharad Goel, Jake M Hofman, Sébastien Lahaie, David M Pennock, and Duncan J Watts. Predicting consumer behavior with Web search. *Proceedings of the National Academy of Sciences*, 107:17486–17490, 2010.

- 
- [27] Andrea Capocci, Vito DP Servedio, Francesca Colaiori, Luciana S Buriol, Debora Donato, Stefano Leonardi, and Guido Caldarelli. Preferential attachment in the growth of social networks: the internet encyclopedia Wikipedia. *Physical Review E*, 74:036116, 2006.
- [28] Reka Albert and Albert Laszlo Barabasi. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.
- [29] Lev Muchnik, Royi Itzhack, Sorin Solomon, and Yoram Louzoun. Self-emergence of knowledge trees: extraction of the Wikipedia hierarchies. *Physical Review E*, 76:016106, 2007.
- [30] Helen Susannah Moat, Chester Curme, Adam Avakian, Dror Y Kenett, H Eugene Stanley, and Tobias Preis. Quantifying Wikipedia usage patterns before stock market moves. *Scientific Reports*, 3:1801, 2013.
- [31] Taha Yasseri, Robert Sumi, and Janos Kertesz. Circadian patterns of Wikipedia editorial activity: a demographic analysis. *PLOS ONE*, 7:e30091, 2012.
- [32] Róbert Sumi, Taha Yasseri, András Rung, András Kornai, and János Kertész. Edit wars in Wikipedia. *SocialCom/PASSAT*, pages 724–727, 2011.
- [33] Taha Yasseri, Robert Sumi, András Rung, András Kornai, and János Kertész. Dynamics of conflicts in Wikipedia. *PLOS ONE*, 7:e38869, 2012.
- [34] Anna Samoilenko and Taha Yasseri. The distorted mirror of Wikipedia: a quantitative analysis of Wikipedia coverage of academics. *EPJ Data Science*, 3:1, 2014.
- [35] Márton Mestyán, Taha Yasseri, and János Kertész. Early prediction of movie box office success based on Wikipedia activity big data. *PLOS ONE*, 8:e71226, 2013.
- [36] Taha Yasseri, Andras Kornai, and Janos Kertesz. A practical approach to language complexity: a Wikipedia case study. *PLOS ONE*, 7:e48386, 2012.
- [37] Farshad Kooti, Kristina Lerman, Luca Maria Aiello, Mihajlo Grbovic, Nemanja Djuric, and Vladan Radosavljevic. Portrait of an online shopper: understanding and predicting consumer behavior. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pages 205–214, 2016.

- [38] David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. The parable of Google flu: traps in big data analysis. *Science*, 343:1203–1205, 2014.
- [39] Tobias Preis and Helen Susannah Moat. Adaptive nowcasting of influenza outbreaks using Google searches. *Royal Society Open Science*, 1:140095, 2014.
- [40] Kyle S Hickmann, Geoffrey Fairchild, Reid Priedhorsky, Nicholas Generous, James M Hyman, Alina Deshpande, and Sara Y Del Valle. Forecasting the 2013–2014 influenza season using Wikipedia. *PLOS Computational Biology*, 11:e1004239, 2015.
- [41] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2:1–8, 2011.
- [42] Alejandro Llorente, Manuel Garcia-Herranz, Manuel Cebrian, and Esteban Moro. Social media fingerprints of unemployment. *PLOS ONE*, 10:e0128692, 2015.
- [43] Michael Conover, Jacob Ratkiewicz, Matthew R Francisco, Bruno Goncalves, Filippo Menczer, and Alessandro Flammini. Political polarization on Twitter. In *ICWSM*, pages 89–96, 2011.
- [44] Andranik Tumasjan, Timm O Sprenger, Philipp G Sandner, and Isabell M Welpe. Predicting elections with Twitter: what 140 characters reveal about political sentiment. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, pages 178–185, 2010.
- [45] Javier Borge-Holthoefer, Alejandro Rivero, Iñigo García, Elisa Cauhé, Alfredo Ferrer, Darío Ferrer, David Francos, David Iñiguez, María Pilar Pérez, Gonzalo Ruiz, Francisco Sanz, Fermín Serrano, Cristina Viñas, Alfonso Tarancón, and Yamir Moreno. Structural and dynamical patterns on online social networks: The Spanish May 15th movement as a case study. *PLOS ONE*, 6:e23883, 2011.
- [46] Sandra González-Bailón, Javier Borge-Holthoefer, Alejandro Rivero, and Yamir Moreno. The dynamics of protest recruitment through an online network. *Scientific Reports*, 1:197, 2011.
- [47] Pete Burnap, Matthew L Williams, Luke Sloan, Omer Rana, William Housley, Adam Edwards, Vincent Knight, Rob Procter, and Alex Voss. Tweeting the

- terror: modelling the social media reaction to the Woolwich terrorist attack. *Social Network Analysis and Mining*, 4:206, 2014.
- [48] Michael D. Conover, Clayton Davis, Emilio Ferrara, Karissa McKelvey, Filippo Menczer, and Alessandro Flammini. The geospatial characteristics of a social movement communication network. *PLOS ONE*, 8:e55957, 2013.
- [49] Fabio Ciulla, Delia Mocanu, Andrea Baronchelli, Bruno Gonçalves, Nicola Perra, and Alessandro Vespignani. Beating the news using social media: the case study of American Idol. *EPJ Data Science*, 1:8, 2012.
- [50] Javier Borge-Holthoefer, Nicola Perra, Bruno Gonçalves, Sandra González-Bailón, Alex Arenas, Yamir Moreno, and Alessandro Vespignani. The dynamic of information-driven coordination phenomena: a transfer entropy analysis. *Science Advances*, 2:e1501158, 2016.
- [51] Manlio De Domenico, Antonio Lima, Paul Mougel, and Mirco Musolesi. The anatomy of a scientific rumor. *Scientific Reports*, 3:2980, 2013.
- [52] Andrea Baronchelli, Vittorio Loreto, and Francesca Tria. Language dynamics. *Advances in Complex Systems*, 15:1203002, 2012.
- [53] Delia Mocanu, Andrea Baronchelli, Nicola Perra, Bruno Gonçalves, Qian Zhang, and Alessandro Vespignani. The Twitter of babel: mapping world languages through microblogging platforms. *PLOS ONE*, 8:e61981, 2013.
- [54] Christian Alis, May Lim, Helen Susannah Moat, Daniele Barchiesi, Tobias Preis, and Steven Bishop. Quantifying regional differences in the length of Twitter messages. *PLOS ONE*, 10:e0122278, 2015.
- [55] Yuri Takhteyev, Anatoliy Gruzd, and Barry Wellman. Geography of Twitter networks. *Social Networks*, 34:73–81, 2012.
- [56] Przemyslaw A Grabowicz, Jose J Ramasco, Bruno Goncalves, and Victor M Eguiluz. Entangling mobility and interactions in social media. *PLOS ONE*, 9:e92196, 2014.
- [57] Luke Sloan and Jeffrey Morgan. Who tweets with their location? Understanding the relationship between demographic characteristics and the use of geoservices and geotagging on Twitter. *PLOS ONE*, 10:e0142209, 2015.

- [58] Luke Sloan, Jeffrey Morgan, Pete Burnap, and Matthew Williams. Who tweets? Deriving the demographic characteristics of age, occupation and social class from Twitter user meta-data. *PLOS ONE*, 10:e0115545, 2015.
- [59] Bruno Gonçalves, Nicola Perra, and Alessandro Vespignani. Modeling users' activity on Twitter networks: validation of Dunbar's number. *PLOS ONE*, 6:e22656, 2011.
- [60] Przemyslaw A Grabowicz, José J Ramasco, Esteban Moro, Josep M Pujol, and Victor M Eguiluz. Social features of online networks: the strength of intermediary ties in online social media. *PLOS ONE*, 7:e29358, 2012.
- [61] Javier Borge-Holthoefer, Alejandro Rivero, and Yamir Moreno. Locating privileged spreaders on an online social network. *Physical Review E*, 85:066123, 2012.
- [62] Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, 113:554–559, 2016.
- [63] Fabiana Zollo, Petra Kralj Novak, Michela Del Vicario, Alessandro Bessi, Igor Mozetič, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. Emotional dynamics in the age of misinformation. *PLOS ONE*, 10:e0138740, 2015.
- [64] Alessandro Bessi, Mauro Coletto, George Alexandru Davidescu, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. Science vs conspiracy: collective narratives in the age of misinformation. *PLOS ONE*, 10:e0118093, 2015.
- [65] Alessandro Bessi, Fabiana Zollo, Michela Del Vicario, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. Trend of narratives in the age of misinformation. *PLOS ONE*, 10:e0134641, 2015.
- [66] Eytan Bakshy, Solomon Messing, and Lada Adamic. Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348:1130–1132, 2015.
- [67] Robert M Bond, Christopher J Fariss, Jason J Jones, Adam DI Kramer, Cameron Marlow, Jaime E Settle, and James H Fowler. A 61-million-person experiment in social influence and political mobilization. *Nature*, 489:295–298, 2012.

- [68] Eytan Bakshy, Itamar Rosenn, Cameron Marlow, and Lada Adamic. The role of social networks in information diffusion. In *WWW 2012 - Session: Information Diffusion in Social Networks April 16-20, 2012, Lyon, France*, pages 519–528, 2012.
- [69] Adam DI Kramer, Jamie E Guillory, and Jeffrey T Hancock. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111:8788–8790, 2014.
- [70] Emilio Ferrara. A large-scale community structure analysis in Facebook. *EPJ Data Science*, 1:9, 2012.
- [71] Luca Maria Aiello, Rossano Schifanella, Daniele Quercia, and Francesco Aletta. Chatty maps: constructing sound maps of urban areas from social media data. *Royal Society Open Science*, 3:150690, 2016.
- [72] Daniele Barchiesi, Tobias Preis, Steven Bishop, and Helen Susannah Moat. Modelling human mobility patterns using photographic data shared online. *Royal Society Open Science*, 2:150046, 2015.
- [73] Daniele Barchiesi, Helen Susannah Moat, Christian Alis, Steven Bishop, and Tobias Preis. Quantifying international travel flows using Flickr. *PLOS ONE*, 10:e0128470, 2015.
- [74] Chanuki Illushka Seresinhe, Tobias Preis, and Helen Susannah Moat. Quantifying the link between art and property prices in urban neighbourhoods. *Royal Society Open Science*, 3:160146, 2016.
- [75] Merve Alanyali, Tobias Preis, and Helen Susannah Moat. Tracking protests using geotagged Flickr photographs. *PLOS ONE*, 11:27–30, 2016.
- [76] Daniele Quercia, Luca Maria Aiello, Kate Mclean, and Rossano Schifanella. Smelly maps: the digital life of urban smellscapes. In *AAAI Publications*, pages 327–336, 2015.
- [77] Daniele Quercia, Luca Maria Aiello, and Rossano Schifanella. The emotional and chromatic layers of urban smells. In *ICWSM*, pages 309–318, 2016.
- [78] Eunjoon Cho, Seth A Myers, and Jure Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1082–1090, 2011.

- [79] Sandra González-Bailón, Ning Wang, Alejandro Rivero, Javier Borge-Holthoefer, and Yamir Moreno. Assessing the bias in samples of large online networks. *Social Networks*, 38:16–27, 2014.
- [80] Vincent D Blondel, Adeline Decuyper, and Gautier Krings. A survey of results on mobile phone datasets analysis. *EPJ Data Science*, 4:10, 2015.
- [81] Gianni Barlacchi, Marco De Nadai, Roberto Larcher, Antonio Casella, Cristiana Chitic, Giovanni Torrisi, Fabrizio Antonelli, Alessandro Vespignani, Alex Pentland, and Bruno Lepri. A multi-source dataset of urban life in the city of Milan and the Province of Trentino. *Scientific Data*, 2:150055, 2015.
- [82] Pierre Deville, Catherine Linard, Samuel Martin, Marius Gilbert, Forrest R Stevens, Andrea E Gaughan, Vincent D Blondel, and Andrew J Tatem. Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences*, 111:15888–15893, 2014.
- [83] Harald Sterly, Benjamin Hennig, and Kouassi Dongo. “Calling Abidjan” – Improving population estimations with mobile communication data. *Mobile Phone Data for Development - Analysis of mobile phone datasets for the development of Ivory Coast*, pages 1–7, 2013.
- [84] Rex W Douglass, David A Meyer, Megha Ram, David Rideout, and Dongjin Song. High resolution population estimates from telecommunications data. *EPJ Data Science*, 4:4, 2015.
- [85] Renaud Lambiotte, Vincent D Blondel, Cristobald de Kerchove, Etienne Huens, Christophe Prieur, Zbigniew Smoreda, and Paul Van Dooren. Geographical dispersal of mobile communication networks. *Physica A*, 387:5317–5325, 2008.
- [86] Gautier Krings, Francesco Calabrese, Carlo Ratti, and Vincent D Blondel. Urban gravity: a model for inter-city telecommunication flows. *Journal of Statistical Mechanics: Theory and Experiment*, 2009:L07003, 2009.
- [87] Cesar A Hidalgo and C Rodriguez-Sickert. The dynamics of a mobile phone network. *Physica A*, 387:3017–3024, 2008.
- [88] Marta González, Cesar A Hidalgo, and Albert Laszlo Barabasi. Understanding individual human mobility patterns. *Nature*, 453:779–782, 2008.

- [89] Daniele Quercia, Rossano Schifanella, and Luca Maria Aiello. The shortest path to happiness. In *Proceedings of the 25th ACM conference on Hypertext and social media - HT '14*, pages 116–125, 2014.
- [90] George MacKerron and Susana Mourato. Happiness is greater in natural environments. *Global Environmental Change*, 23:992–1000, 2013.
- [91] Chanuki Illushka Seresinhe, Tobias Preis, and Helen Susannah Moat. Quantifying the impact of scenic environments on health. *Scientific Reports*, 5:16899, 2015.
- [92] Michele Starnini, Andrea Baronchelli, Alain Barrat, and Romualdo Pastor-Satorras. Random walks on temporal networks. *Physical Review E*, 85:056115, 2012.
- [93] Valerio Gemmetto, Alain Barrat, and Ciro Cattuto. Mitigation of infectious disease at school: targeted class closure vs school closure. *BMC Infectious Diseases*, 14:695, 2014.
- [94] Mark C Pachucki, Emily J Ozer, Alain Barrat, and Ciro Cattuto. Mental health and social networks in early adolescence: a dynamic study of objectively-measured social interaction behaviors. *Social Science & medicine*, 125:40–50, 2015.
- [95] Nicolas Voirin, Cecile Pavet, Alain Barrat, Ciro Cattuto, Nagham Khanafer, Corinne Regis, Bveul-a Kim, Brigitte Comte, Jean-Sebastien Casalegno, Bruno Lina, and Philippe Vanhems. Combining high-resolution contact data with virological data to investigate influenza transmission in a tertiary care hospital. *Infection Control & Hospital Epidemiology*, 36:254, 2015.
- [96] Mark Kibanov, Martin Aztmueller, Jens Illig, Christoph Scholz, Alain Barrat, Ciro Cattuto, and Gerd Stumme. Is web content a good proxy for real-life interaction? A case study considering online and offline interactions of computer scientists. In *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 697–704, 2015.
- [97] Mathieu Génois, Christian L Vestergaard, Ciro Cattuto, and Alain Barrat. Compensating for population sampling in simulations of epidemic spread on temporal contact networks. *Nature Communications*, 6:8860, 2015.
- [98] Mathieu Genois, Christian L Vestergaard, Julie Fournet, Andre Panisson, Isabelle Bonmarin, and Alain Barrat. Data on face-to-face contacts in an office

- building suggest a low-cost vaccination strategy based on community linkers. *Network Science*, 3:326–347, 2015.
- [99] Rossana Mastrandrea, Alberto Soto-Aladro, Philippe Brouqui, and Alain Barrat. Enhancing the evaluation of pathogen transmission risk in a hospital by merging hand-hygiene compliance and contact data: a proof-of-concept study. *BMC Research Notes*, 8:426, 2015.
- [100] Rossana Mastrandrea, Julie Fournet, and Alain Barrat. Contact patterns in a high school: a comparison between data collected using wearable sensors, contact diaries and friendship surveys. *PLOS ONE*, 10:e0136497, 2015.
- [101] Julie Fournet and Alain Barrat. Epidemic risk from friendship network data: an equivalence with a non-uniform sampling of contact networks. *Scientific Reports*, 6:24593, 2016.
- [102] Moses C Kiti, Michele Tizzoni, Timothy M Kinyanjui, Dorothy C Koech, Patrick K Munywoki, Milosch Meriac, Luca Cappa, Andre Panisson, Alain Barrat, Ciro Cattuto, and D James Nokes. Quantifying social contacts in a household setting of rural Kenya using wearable proximity sensors. *EPJ Data Science*, 5:21, 2016.
- [103] Rossana Mastrandrea and Alain Barrat. How to estimate epidemic risk from incomplete contact diaries data? *PLoS Computational Biology*, 12:e1005002, 2016.
- [104] Christian L Vestergaard, Eugenio Valdano, Mathieu Genois, Chiara Poletto, Vittoria Colizza, and Alain Barrat. Impact of spatially constrained sampling of temporal contact networks on the evaluation of the epidemic risk. *European Journal of Applied Mathematics*, page in print, 2016.
- [105] Timo Smieszek, Stefanie Castell, Alain Barrat, Ciro Cattuto, Peter J White, and Gérard Krause. Contact diaries versus wearable proximity sensors in measuring contact patterns at a conference: method comparison and participants’ attitudes. *BMC Infectious Diseases*, 16:341, 2016.
- [106] Rosario N Mantegna and H Eugene Stanley. Stochastic process with ultraslow convergence to a Gaussian: the truncated Levy flight. *Physical Review Letters*, 73:2946, 1994.
- [107] Rosario N Mantegna and H Eugene Stanley. Scaling behaviour in the dynamics of an economic index. *Nature*, 376:46–49, 1995.

- 
- [108] Parameswaran Gopikrishnan, Vasiliki Plerou, Luis A Nunes Amaral, Martin Meyer, and H Eugene Stanley. Scaling of the distribution of fluctuations of financial market indices. *Physical Review E*, 60:5305, 1999.
- [109] Rosario N Mantegna, H Eugene Stanley, and Neil A Chriss. An introduction to econophysics: correlations and complexity in finance. *Physics Today*, 53:70, 2000.
- [110] Robert L Axtell. Zipf distribution of U.S. firm sizes. *Science*, 293:1818–1820, 2001.
- [111] Vasiliki Plerou, Parameswaran Gopikrishnan, and H Eugene Stanley. Econophysics: two-phase behaviour of financial markets. *Nature*, 421:130, 2003.
- [112] Xavier Gabaix, Parameswaran Gopikrishnan, Vasiliki Plerou, and H Eugene Stanley. A theory of power-law distributions in financial market fluctuations. *Nature*, 423:267–270, 2003.
- [113] Didier Sornette, Ryan Woodard, and Wei Xing Zhou. The 2006-2008 oil bubble: evidence of speculation, and prediction. *Physica A*, 388:1571–1576, 2009.
- [114] Tobias Preis, Johannes J Schneider, and H Eugene Stanley. Switching processes in financial markets. *Proceedings of the National Academy of Sciences*, 108:7674–7678, 2011.
- [115] Michael C Münnix, Takashi Shimada, Rudi Schäfer, Francois Leyvraz, Thomas H Seligman, Thomas Guhr, and H Eugene Stanley. Identifying states of a financial market. *Scientific Reports*, 2:644, 2012.
- [116] Tobias Preis, Dror Y Kenett, H Eugene Stanley, Dirk Helbing, and Eshel Ben-Jacob. Quantifying the behavior of stock correlations under market stress. *Scientific Reports*, 2:752, 2012.
- [117] Yves-Alexandre de Montjoye, César A Hidalgo, Michel Verleysen, and Vincent D Blondel. Unique in the crowd: the privacy bounds of human mobility. *Scientific Reports*, 3:1376, 2013.
- [118] Yves-Alexandre de Montjoye, Laura Radaelli, Vivek Kumar Singh, and Alex Sandy Pentland. Unique in the shopping mall: on the reidentifiability of credit card metadata. *Science*, 347:536–539, 2015.

- 
- [119] Stefano Boccaletti, Vito Latora, Yamir Moreno, Martin Chavez, and D-U Hwang. Complex networks: structure and dynamics. *Physics Reports*, 424:175–308, 2006.
- [120] Albert Reka and Albert Laszlo Barabasi. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74:47, 2002.
- [121] Mark E J Newman. The structure and function of complex networks. *SIAM Review*, 45:167–256, 2003.
- [122] Ginestra Bianconi and Albert Laszlo Barabasi. Bose-Einstein condensation in complex networks. *Physical Review Letters*, 86:5632, 2001.
- [123] Alain Barrat, Marc Barthélemy, Romualdo Pastor-Satorras, and Alessandro Vespignani. The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences*, 101:3747–3752, 2004.
- [124] Ginestra Bianconi. Entropy of network ensembles. *Physical Review E*, 79:036114, 2009.
- [125] Guido Caldarelli, Andrea Capocci, Paolo De Los Rios, and Miguel A Muñoz. Scale-free networks from varying vertex intrinsic fitness. *Physical Review Letters*, 89:258702, 2002.
- [126] Charo I Del Genio, Hyunju Kim, Zoltán Toroczkai, and Kevin E Bassler. Efficient and exact sampling of simple graphs with given arbitrary degree sequence. *PLOS ONE*, 5:e10012, 2010.
- [127] Charo I Del Genio, Thilo Gross, and Kevin E Bassler. All scale-free networks are sparse. *Physical Review Letters*, 107:178701, 2011.
- [128] Mark E J Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98:404–409, 2001.
- [129] Alex Arenas, Albert Diaz-Guilera, Jurgen Kurths, Yamir Moreno, and Changsong Zhou. Synchronization in complex networks. *Physics Reports*, 469:93–153, 2008.
- [130] Nicola Perra, Andrea Baronchelli, Delia Mocanu, Bruno Goncalves, Romualdo Pastor-Satorras, and Alessandro Vespignani. Random walks and search in time-varying networks. *Physical Review Letters*, 109:238701, 2012.

- 
- [131] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic dynamics and endemic states in complex networks. *Physical Review E*, 63:066117, 2001.
- [132] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic spreading in scale-free networks. *Physical Review Letters*, 86:3200, 2001.
- [133] Romualdo Pastor-Satorras and Alessandro Vespignani. Immunization of complex networks. *Physical Review E*, 65:036104, 2002.
- [134] Yamir Moreno, Romualdo Pastor-Satorras, and Alessandro Vespignani. Epidemic outbreaks in complex heterogeneous networks. *European Physical Journal B*, 26:521–529, 2002.
- [135] Vittoria Colizza, Alain Barrat, Marc Barthélemy, and Alessandro Vespignani. The role of the airline transportation network in the prediction and predictability of global epidemics. *Proceedings of the National Academy of Sciences*, 103:2015–2020, 2006.
- [136] Michele Starnini, Andrea Baronchelli, and Romualdo Pastor-Satorras. Modeling human dynamics of face-to-face interaction networks. *Physical Review Letters*, 110:168701, 2013.
- [137] Su Yu Liu, Andrea Baronchelli, and Nicola Perra. Contagion dynamics in time-varying metapopulation networks. *Physical Review E*, 87:032805, 2013.
- [138] Yamir Moreno, Maziar Nekovee, and Amalio F Pacheco. Dynamics of rumor spreading in complex networks. *Physical Review E*, 69:066130, 2004.
- [139] Maziar Nekovee, Yamir Moreno, Ginestra Bianconi, and Matteo Marsili. Theory of rumour spreading in complex social networks. *Physica A*, 374:457–470, 2007.
- [140] Damon Centola and Andrea Baronchelli. The spontaneous emergence of conventions: an experimental study of cultural evolution. *Proceedings of the National Academy of Sciences*, 112:1989–1994, 2015.
- [141] Arda Halu, Kun Zhao, Andrea Baronchelli, and Ginestra Bianconi. Connect and win: the role of social networks in political elections. *EPL*, 102:16002, 2013.
- [142] Alexei Vázquez, Romualdo Pastor-Satorras, and Alessandro Vespignani. Large-scale topological and dynamical properties of the Internet. *Physical Review E*, 65:066130, 2002.

- 
- [143] Amanda L Traud, Peter J Mucha, and Mason A Porter. Social structure of Facebook networks. *Physica A*, 391:4165–4180, 2012.
- [144] Mason A Porter, Peter J Mucha, Mark E J Newman, and Casey M Warmbrand. A network analysis of committees in the U.S. House of Representatives. *Proceedings of the National Academy of Sciences*, 102:7057–7062, 2005.
- [145] Stefano Battiston, Michelangelo Puliga, Rahul Kaushik, Paolo Tasca, and Guido Caldarelli. DebtRank: too Central to fail? Financial networks, the FED and systemic risk. *Scientific Reports*, 2:541, 2012.
- [146] Richard J Williams and Neo D Martinez. Simple rules yield complex foodwebs. *Nature*, 404:180–183, 2000.
- [147] Hawoong Jeong, Balint Tombor, Réka Albert, Zoltán N Oltvai, and Albert-László Barabási. The large-scale organization of metabolic networks. *Nature*, 407:651–654, 2000.
- [148] Samuel Johnson, Virginia Domínguez-García, Luca Donetti, and Miguel A Muñoz. Trophic coherence determines food-web stability. *Proceedings of the National Academy of Sciences*, 1:17923–17928, 2014.
- [149] Esteban Moro and Jari Saramäki. From seconds to months: multi-scale dynamics of mobile telephone calls. *The European Physical Journal B*, 88:164, 2015.
- [150] Stanley Milgram. The small world problem. *Psychology Today*, 2:60–67, 1967.
- [151] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1998.
- [152] Luis A Amaral, Antonio Scala, Marc Barthelemy, and H Eugene Stanley. Classes of small-world networks. *Proceedings of the National Academy of Sciences*, 97:11149–11152, 2000.
- [153] Mark E J Newman, Steven H Strogatz, and Duncan J Watts. Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, 64:026118, 2001.
- [154] Charo I Del Genio and Thomas House. Endemic infections are always possible on regular networks. *Physical Review E*, 88:040801, 2013.

- 
- [155] Samuel Johnson, Joaquín Torres, J Marro, and Miguel A Muñoz. Entropic origin of disassortativity in complex networks. *Physical Review Letters*, 104:108702, 2010.
- [156] Oliver Williams and Charo I Del Genio. Degree correlations in directed scale-free networks. *PLOS ONE*, 9:e110121, 2014.
- [157] Mark E J Newman. Assortative Mixing in Networks. *Physical Review Letters*, 89:208701, 2002.
- [158] Charo I Del Genio, Miguel Romance, Regino Criado, and Stefano Boccaletti. Synchronization in dynamical networks with unconstrained structure switching. *Physical Review E*, 92:062819, 2015.
- [159] Michelle Girvan and Mark E J Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99:7821–7826, 2002.
- [160] Sergey V Buldyrev, Roni Parshani, Gerald Paul, H Eugene Stanley, and Shlomo Havlin. Catastrophic cascade of failures in interdependent networks. *Nature*, 464:1025–1028, 2010.
- [161] Jianxi Gao, Sergey V Buldyrev, H Eugene Stanley, and Shlomo Havlin. Networks formed from interdependent networks. *Nature Physics*, 8:40–48, 2011.
- [162] Jianxi Gao, Sergey V Buldyrev, Shlomo Havlin, and H Eugene Stanley. Robustness of a network of networks. *Physical Review Letters*, 107:195701, 2011.
- [163] Manlio De Domenico, Albert Solé-Ribalta, Emanuele Cozzo, Mikko Kivela, Yamir Moreno, Mason A Porter, Sergio Gómez, and Alex Arenas. Mathematical formulation of multilayer networks. *Physical Review X*, 3:041022, 2014.
- [164] Stefano Boccaletti, Ginestra Bianconi, Regino Criado, Charo I Del Genio, Jesus Gómez-Gardeñes, Miguel Romance, Irene Sendiña Nadal, Zhen Wang, and Massimiliano Zanin. The structure and dynamics of multilayer networks. *Physics Reports*, 544:1–122, 2014.
- [165] Mikko Kivela, Alex Arenas, Marc Barthélemy, James P Gleeson, Yamir Moreno, and Mason A Porter. Multilayer networks. *Journal of Complex Networks*, 2:203–271, 2014.
- [166] Federico Battiston, Vincenzo Nicosia, and Vito Latora. Structural measures for multiplex networks. *Physical Review E*, 89:032804, 2014.

- [167] Stuart L Pimm. The structure of food webs. *Theoretical Population Biology*, 16:144–158, 1979.
- [168] Geoffrey P Garnett, James P Hughes, Roy M Anderson, Bradley P Stoner, Sevgi O Aral, William L Whittington, H Hunter Handsfield, and King K Holmes. Sexual mixing patterns of patients attending sexually transmitted diseases clinics. *Sexually Transmitted Diseases*, 23:248–257, 1996.
- [169] Gary William Flake, Steve Lawrence, C Lee Giles, and Frans M Coetzee. Self-organization and identification of web communities. *Computer*, 35:66–71, 2002.
- [170] Kasper Eriksen, Ingve Simonsen, Sergei Maslov, and Kim Sneppen. Modularity and extreme edges of the Internet. *Physical Review Letters*, 90:148701, 2003.
- [171] Ann E Krause, Kenneth A Frank, Doran M Mason, Robert E Ulanowicz, and William W Taylor. Compartments revealed in food-web structure. *Nature*, 426:282–285, 2003.
- [172] David Lusseau and Mark E J Newman. Identifying the role that animals play in their social networks. *Proceedings of the Royal Society B*, 271:477–481, 2004.
- [173] Roger Guimera and Luis A Amaral. Functional cartography of complex metabolic networks. *Nature*, 433:895–900, 2005.
- [174] Gregory Palla, Imre Derenyi, Illes Farkas, and Tamas Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435:814–818, 2005.
- [175] Alex Arenas, Albert Díaz-Guilera, and Conrad J Pérez-Vicente. Synchronization reveals topological scales in complex networks. *Physical Review Letters*, 96:114102, 2006.
- [176] Juan G Restrepo, Edward Ott, and Brian R Hunt. Characterizing the dynamical importance of network nodes and links. *Physical Review Letters*, 97:094102, 2006.
- [177] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008:P10008, 2008.

- 
- [178] Charo I Del Genio and Thilo Gross. Emergent bipartiteness in a society of knights and knaves. *New Journal of Physics*, 13:103038, 2011.
- [179] Leon Danon, Albert Díaz-Guilera, Jordi Duch, and Alex Arenas. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, 09008:219–228, 2005.
- [180] Andrea Lancichinetti, Santo Fortunato, and Filippo Radicchi. Benchmark graphs for testing community detection algorithms. *Physical Review E*, 78:046110, 2008.
- [181] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486:75–174, 2010.
- [182] Peter J Mucha, Thomas Richardson, Kevin Macon, Mason A Porter, and Jukka-Pekka Onnela. Community structure in time-dependent, multiscale, and multiplex networks. *Science*, 328:876–878, 2010.
- [183] Karsten Steinhaeuser and Nitesh V Chawla. Identifying and evaluating community structure in complex networks. *Pattern Recognition Letters*, 31:413–421, 2010.
- [184] Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborova. Inference and phase transitions in the detection of modules in sparse networks. *Physical Review Letters*, 107:065701, 2011.
- [185] Tiago P Peixoto. Parsimonious module inference in large networks. *Physical Review Letters*, 110:148701, 2013.
- [186] Santiago Treviño, Amy Nyberg, Charo I Del Genio, and Kevin E Bassler. Fast and accurate determination of modularity and its effect size. *Journal of Statistical Mechanics: Theory and Experiment*, 2015:P02003, 2015.
- [187] Mark E J Newman and Tiago P Peixoto. Generalized communities in networks. *Physical Review Letters*, 115:088701, 2015.
- [188] Mark E J Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103:8577–82, 2006.
- [189] Didier Sornette. *Why stock markets crash: critical events in complex financial systems*. Princeton University Press, Princeton, NJ, 2004.

- [190] J Doyne Farmer and Shareen Joshi. The price dynamics of common trading strategies. *Journal of Economic Behavior and Organization*, 49:149–171, 2002.
- [191] Johannes Voit. *The statistical mechanics of financial markets*. Springer, Heidelberg, 2005.
- [192] Wolfgang Paul and Jorg Baschnagel. *Stochastic processes: from physics to finance*. Springer, Heidelberg, 2013.
- [193] Frederic Abergel, Bikas K Chakrabarti, Anirban Chakraborti, and Manupushpak Mitra, editors. *Econophysics of order-driven markets*. Springer, Milan, 2011.
- [194] Thomas Lux and Frank Westerhoff. Economics crisis. *Nature Physics*, 5:2–3, 2009.
- [195] J Doyne Farmer and Duncan Foley. The economy needs agent-based modelling. *Nature*, 460:685–686, 2009.
- [196] Ling Feng, Baowen Li, Boris Podobnik, Tobias Preis, and H Eugene Stanley. Linking agent-based models and stochastic models of financial markets. *Proceedings of the National Academy of Sciences*, 109:8388–8393, 2012.
- [197] Alexander M Petersen, Fengzhong Wang, Shlomo Havlin, and H Eugene Stanley. Market dynamics immediately before and after financial shocks: quantifying the Omori, productivity, and Bath laws. *Physical Review E*, 82:036114, 2010.
- [198] Cars H Hommes. Modeling the stylized facts in finance through simple nonlinear adaptive systems. *Proceedings of the National Academy of Sciences*, 99 Suppl 3:7221–7228, 2002.
- [199] Tobias Preis, Sebastian Golke, Wolfgang Paul, and Johannes J Schneider. Multi-agent-based order book model of financial markets. *EPL*, 75:510, 2006.
- [200] Anna Carbone, Giuliano Castelli, and H Eugene Stanley. Time-dependent Hurst exponent in financial time series. *Physica A*, 344:267–271, 2004.
- [201] Xavier Gabaix. Power laws in economics and finance. In *Annual Review of Economics*, volume 1, pages 255–294. 2009.
- [202] Thomas Lux and Michele Marchesi. Scaling and criticality in a stochastic multi-agent model of a financial market. *Nature*, 397:498–500, 1999.

- 
- [203] Xavier Gabaix, Parameswaran Gopikrishnan, Vasiliki Plerou, and H Eugene Stanley. A theory of power-law distributions in financial market fluctuations. *Nature*, 423:267–270, 2003.
- [204] Boris Podobnik, Davor Horvatic, Alexander M Petersen, and H Eugene Stanley. Cross-correlations between volume change and price change. *Proceedings of the National Academy of Sciences*, 106:22079–22084, 2009.
- [205] Gao Feng Gu, Wei Chen, and Wei Xing Zhou. Empirical distributions of Chinese stock returns at different microscopic timescales. *Physica A*, 387:495–502, 2008.
- [206] Danuta Makowiec and Piotr Gnacinski. Fluctuations of WIG - the index of Warsaw stock exchange preliminary studies. *Acta Physica*, 32:1487–1500, 2001.
- [207] William K Bertram. An empirical investigation of Australian stock exchange data. *Physica A*, 341:533–546, 2004.
- [208] HF Coronel-Brizio and AR Hernández-Montoya. On fitting the Pareto-Levy distribution to stock market index data: selecting a suitable cutoff value. *Physica A*, 354:437–449, 2005.
- [209] Vasiliki Plerou and H Eugene Stanley. Tests of scaling and universality of the distributions of trade size and share volume: evidence from three distinct markets. *Physical Review E*, 76:046109, 2007.
- [210] Guo-Hua Mu and Wei-Xing Zhou. Tests of nonuniversality of the stock return distributions in an emerging market. *Physical Review E*, 82:066103, 2010.
- [211] Aaron Clauset, Cosma Rohilla Shalizi, and Mark E J Newman. Power-law distributions in empirical data. *SIAM Review*, 51:661–703, 2009.
- [212] Federico Botta, Helen Susannah Moat, H Eugene Stanley, and Tobias Preis. Quantifying stock return distributions in financial markets. *PLOS ONE*, 10:e0135600, 2015.
- [213] Kaushik Matia, Mukul Pal, H Salunkay, and H Eugene Stanley. Scale-dependent price fluctuations for the Indian stock market. *EPL*, 66:909, 2004.
- [214] Dirk Helbing, Illes Farkas, and Tamas Vicsek. Simulating dynamical features of escape panic. *Nature*, 407:487–490, 2000.

- [215] Paul S F Yip, Ray Watson, K S Chan, Eric H Y Lau, Feng Chen, Ying Xu, Liqun Xi, Derek Y T Cheung, Brian Y T Ip, and Danping Liu. Estimation of the number of people in a demonstration. *Australian & New Zealand Journal of Statistics*, 52:17–26, 2010.
- [216] Antoni B Chan, Zhang Sheng John Liang, and Nuno Vasconcelos. Privacy preserving crowd monitoring: counting people without people models or tracking. In *26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2008.
- [217] Dan Kong, Doug Gray, and Hai Tao. A viewpoint invariant approach for crowd counting. In *Proceedings - International Conference on Pattern Recognition*, volume 3, pages 1187–1190, 2006.
- [218] Marcos Cruz, Domingo Gómez, and Luis M Cruz-Orive. Efficient and unbiased estimation of population size. *PLOS ONE*, 10:e0141868, 2015.
- [219] Enac. Dati di Traffico. Technical report, 2012.
- [220] Ana L N Fred and Anil K Jain. Robust data clustering. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages II–128, 2003.
- [221] Ludmila I Kuncheva and Stefan T Hadjitodorov. Using diversity in cluster ensembles. In *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics*, volume 2, pages 1214–1219, 2004.
- [222] Marina Meilua. Comparing clusterings-an information based distance. *Journal of Multivariate Analysis*, 98:873–895, 2007.
- [223] Mingming Chen, Tommy Nguyen, and Boleslaw K Szymanski. A new metric for quality of network community structure. *ASE Human Journal*, 2:226–240, 2013.
- [224] Mingming Chen, Konstantin Kuzmin, and Boleslaw K Szymanski. Community detection via maximization of modularity and its variants. *IEEE Transactions on Computational Social Systems*, 1:46–65, 2014.
- [225] Santo Fortunato and Marc Barthélemy. Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104:36–41, 2007.
- [226] Andrea Lancichinetti and Santo Fortunato. Limits of modularity maximization in community detection. *Physical Review E*, 84:066122, 2011.

- [227] Vincent A Traag, Paul Van Dooren, and Yurii Nesterov. Narrow scope for resolution-limit-free community detection. *Physical Review E*, 84:016114, 2011.
- [228] Brian W Kernighan and Shen Lin. An efficient heuristic procedure for partitioning graphs. *Bell Labs Technical Journal*, 49:291–307, 1970.
- [229] Wayne W Zachary. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33:452–473, 1977.