University of Pisa

Department of Computer Science

PhD Thesis

Big Data Analytics for Nowcasting and Forecasting Social Phenomena

Ioanna Miliou

Supervisor
Dino Pedreschi
University of Pisa

Supervisor
Salvatore Rinzivillo
ISTI-CNR

October 19, 2018

To happiness....

"Don’t cry because it’s over, smile because it happened."
Dr. Seuss

Abstract

One of the most pressing, and fascinating challenges of our time is understanding the complexity of the global interconnected society we inhabit. This connectedness reveals in many phenomena: in the rapid growth of the Internet and Web, in the ease with which global communication and trade now takes place, and in the ability of news and information as well as epidemics, trends, financial crises and social unrest to spread around the world with surprising speed and intensity. Ours is also a time of opportunity to observe and measure how our society intimately works: *Big Data* originating from the digital breadcrumbs of human activities promise to let us scrutinize the ground truth of individual and collective behavior at an unprecedented detail in real time. Multiple dimensions of our social life have Big Data proxies nowadays. We can use Big Data, as signals, as proxies for forecast and nowcast different phenomena, and even more social phenomena. We can manage to describe and predict how humans and society works. We can use geolocated data to observe and measure the behavior of a population, to build better cities tailored to the movement of the population, with lower commuting times and lower pollution. We can exploit medical data to build classifiers able to help in diagnosing and curing diseases. We can use industrial data to improve the production processes, and create smarter and more secure factories. We can do a lot of other incredible and useful things with the support of data and analytical tools able to extract useful knowledge from raw data.

In this thesis we introduce data-driven as well as model-driven approaches to predict different phenomena, from epidemics to socio-economic attraction. We use Big Data deriving from our everyday life as external proxies to nowcast and forecast the evolution of phenomena whose study relies only on historical data or data that come only with a significant lag. We use supermarket retail data as an external signal in order to predict the curve of an internal time series, the influenza one. When the flu season arrives, people are starting to get sick. Getting sick affects their everyday life and behavior. This change in behavior should propagate in their purchases in the supermarket. So they will buy products that will reflect the fact that they are sick. We also study human movements that are inherently massive, dynamical, and complex. But understanding the individual mobility patterns, could be of such a fundamental importance for so many different phenomena. We decided to exploit these patterns in order to study and predict the attraction of different socio-economic factors of human environment. In our first approach we study the distribution of the travelling sub-populations in Tuscany region in Italy, to the airports of the region and we built a dynamic model for the interplay of attraction of availability of air travel and an airport's popularity among the population. Based on this model, we forecast the future evolution of the airports in the region. In our second approach, we identifiy and categorize industrial clusters in Veneto region in Italy, by size and population dynamics and measured their attraction. We create a real-time system which help us to feel the pulse of a city, and predict the rise of new industrial clusters or the death of existing ones. Finally, we attempt prediction in social networks, introducing the interaction prediction problem, trying to predict intra-community interactions, interactions that may occur in the interior of the same community, and we applied the same approach to predict inter-community interactions, the weak links that keep together the modular structure composing complex networks.

Acknowledgements

I would like to thank all the people that have been there for me during these years of my Ph.D. It would been a rather long and hard journey without them. Thanks for holding my hand through everything I went through and for being there for me. Ph.D. is not a journey you can do alone.

I am very grateful to my advisors Dino Pedreschi and Salvatore Rinzivillo that believed in me and supported me. They guided me through the world of research and thanks to their enthusiasm I learned to love the world of research and my work. Many thanks to the head of the KDDLab Fosca Giannotti for her guidance, as well as to the the senior researchers Anna, Mirco, Roberto and Alina. Another big thank goes to the other KDDLab colleagues for sharing ideas and reciprocal help: Giulio, Luca, Farzad, Lorenzo, Paolo, Vittorio, Francesca, Letizia, Valerio, Daniele, Michela, Barbara, Chiara, Laura, Viola, and Roberto.

Many thanks to my friends that are here in Pisa or half the way across the world: Maria, Vicky, Sophia, Despoina, Panos, Vasilis, Michalis, Farzad, Sara, Dimitris, Khatia, Eleutheria, Stamatis, Marco, Thanassis, Sophia, Yiannis, Zoi, Surya. They kept me company, when most probably I was the worst possible company of the world. Thank you about the long discussions about life and its meaning. Hoping we will find it one day! Some of you may know what a Ph.D. is all about, but most of you, don't, and that's really a charm.

A big thank you goes to my family, and more specifically to my sisters: Athina, Olympia and Nicky. They have always been there for me and supported me through all my choices, crazy or not. Feeling grateful to have them in my life.

Finally, I would like to thank my partner in life during these years, Riccardo, because no matter what, he has always been there. No matter the fact that computer science means nothing to him, no matter the fact that the research world is so complex and unknown to him, he has always tried to support me, help me and be there for me. Thank you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

One of the most pressing, and fascinating challenges of our time is understanding the complexity of the global interconnected society we inhabit. This connectedness reveals in many phenomena: in the rapid growth of the Internet and Web, in the ease with which global communication and trade now takes place, and in the ability of news and information as well as epidemics, trends, financial crises and social unrest to spread around the world with surprising speed and intensity. Ours is also a time of opportunity to observe and measure how our society intimately works: the *Big Data* originating from the digital breadcrumbs of human activities promise to let us scrutinize the ground truth of individual and collective behaviour at an unprecedented detail in real time. Multiple dimensions of our social life have Big Data proxies nowadays. Each one of us produces an unthinkable amount of data while performing our daily activities.

Think of your everyday routine. Perhaps your alarm rings to wake you up on your mobile, you check your email or what's been posted on social media sites, you travel to work on public transport, or in a car along streets where traffic flow is monitored, and then you buy some bread on the way home with your supermarket card. What kind of data do your everyday activities generate, and what might such data be useful for?

Social relationships leave traces in the network of our phone or email contacts, in the friendship links of our favourite social networking site. Our shopping patterns leave traces in the transaction records of our purchases. Our movements leave traces in the records of our mobile phone calls, in the GPS tracks of our on-board navigation systems. Big Data means one very specific thing and this is how humans changed their behaviour by interacting with very large technological systems. Looking at ourselves we can try to capture new clues about how human complex society works [24].

The use of big data analytics for measuring and understanding social phenomena is a recent but very lively arena. Sensing Big Data at a societal scale has the potential of providing a powerful social microscope, which can help us understand many complex and hidden socio-economic phenomena. Such challenge clearly requires high-level analytics, modelling and reasoning across all the social dimensions above, an activity that it is often referred to as *data mining*: the task of making sense of Big Data by extracting meaningful information from large, messy and noisy data [139]. Big Data is something which in this day and age none of us can afford not to understand. It is a gigantic ocean of information which we can exploit in order to learn something about human behaviour.

Previously, the only way of measuring how humans behave was to put them in an experiment, or ask them to write down answers for a survey. Now, we increasingly rely on networked computer systems and smart cards to support our everyday activities, and everything we do generates data: buying bread at the supermarket or calling a friend for a chat. The difference to all other kinds of traditional official data is that digital generated data is available "real-time" and has

therefore the ability to give an insight into various aspects of collective human behaviour in a previously unimaginable way. The diffusion of *data science* constitutes a genuine opportunity to bring powerful new tools to complement the traditional survey data and official statistics, adding depth and different aspects of human behaviour and experience, narrowing both time and knowledge gaps thanks to the real-time character of the insights gathered through *data mining*. The complex connectedness and documentation of online activity has revolutionized social sciences and the traditional attempts of quantifying scientifically peoples' socio-economic behaviour. The biggest advantage of online-based big data in comparison with traditional data, is the opportunity for an improved and current understanding of human behaviour that can lead to a real-time awareness and generate a more appropriate and timely feedback.

Taking into account these large new data sets that are available nowadays, a question comes in mind; can we use some of these data sources, which are immediately accessible after their creation, to help here? Can we scientifically predict our future? Scientists and pseudo scientists have been pursuing this mystery for hundreds and perhaps thousands of years. But now, they are starting to use these new data sets to better measure what people are doing in the world right now and they are revealing patterns in human behaviour previously thought to be purely random. Precise, orderly, predictable patterns... Barabási, describes a revolutionary new theory showing how we can predict human behaviour in Bursts, an original investigation into human nature in the light of Big Data [27]. His approach relies on the digital reality of our world, from mobile phones to the Internet and email, because it has turned society into a huge research laboratory. All those electronic trails of time stamped texts, voicemails, and internet searches add up to a previously unavailable massive data set of statistics that track our movements, our decisions, our lives. Analysis of these trails is offering deep insights into the rhythm of how we do everything. His finding? We work and fight and play in short flourishes of activity followed by next to nothing. The pattern isn't random, it's "bursty." Randomness does not rule our lives in the way scientists have assumed up until now. Barabási's wide range of examples from seemingly unrelated areas include how dollar bills move around the U.S., the pattern everyone follows in writing email, the spread of epidemics, and even the flight patterns of albatross. Bursts reveals what this amazing new research is showing us about where individual spontaneity ends and predictability in human behaviour begins. Can we use these new large data sets to even possibly forecast what humans are going to do in the future?

The power of big data lies in the fact that history repeats itself. So what you did yesterday, you'll often do tomorrow as well. And if you can process these huge data sets, then you can start identifying these repeating patterns and use that information to make better predictions about what's going to happen in the future. Such predictions could help us better anticipate where crimes might occur, or how diseases might spread.

Big data might help us to make better decisions in an economic context as well. Making a decision can be quite complex, and this might be quite complex for a range of people – for us as humans in general in our everyday life, but also for commercial stakeholders or governmental stakeholders. They need to rely on key statistics like unemployment rate or GDP, and they rely on the latest and most up-to-date figures. This is a problem because most of these key statistics come with one major disadvantage – it's so difficult and time consuming to calculate these numbers that it can take two weeks, or a month, or even more than that, to get to the final number ready for decision makers.

So this is the problem of *forecasting*. Traditionally, you rely on what has happened in the past, and you want to forecast the next step. But here we have a problem: the latest number which is available comes with a significant time lag. We know right now, maybe, the unemployment rate a month ago or two months ago. In order to forecast a little bit closer to the time where we are right now, this becomes ultimately a *nowcasting* challenge – we want to forecast the present rather than the future. Focusing on the power to forecast near-time or so-called *nowcast* values

of activity in various sectors, such as the unemployment rate, automobile sales and epidemics, is becoming of particular interest to many organizations and researchers, trying to gain timelier forecasts of economic indicators, future sales and disease spreading.

Nowcasting is particularly relevant for those key macro economic variables which are collected at low frequency, typically on a quarterly basis, and released with a substantial lag. To obtain "early estimates" of such key economic indicators, nowcasters use the information from data which are related to the target variable but collected at higher frequency and released in a more timely manner. For example, euro area GDP, which is the key statistic describing the state of the economy, is only available at quarterly frequency and is released six weeks after the close of the quarter. However, there are several variables related to GDP, such e.g. industrial production or various surveys, available at monthly frequency and published with shorter delay. These can be used to construct early estimates of GDP. Heuristics are applied to triangulate and form stable estimates. With the advent of real time data and big data computing such estimates can be very precise, as it can be seen in weather forecasts.

So, in a *nowcasting* problem, there is a time-dependent variable $X(t)$ that cannot be measured instantaneously. We can only know the value of $X(t)$ at time $t+d$, where d is a delay. *Nowcasting* is the act of (probabilistically) approximating the value of $X(t)$ at time t, based on a set of n (possibly latent) variables $Y_1(..), Y_2(...), ...Y_n$, which can be measured at time t or time $t+d_{Y_{1,2,...n}}$ where $d_{1,2,...n} < d$.

Nowcasting is referring to a prediction of variable $X(t)$ based on information available here and now. As a blend of the words now and forecast (= a statement about what is likely to happen) a nowcast is a prediction informed by analysis of data currently available. In other words, it's a hypothetical model rooted in the most up-to-date set of circumstances, which could for instance differ significantly from what was anticipated the previous week, the day before, or even a number of hours earlier.

The speed at which nowcasting is done, is related to each specific phenomena and the availability of the variables $Y_1(..), Y_2(...), ...Y_n$. In real time, some data series have observations through the current period, whereas for others the most recent observations may be available only for a week or a month earlier. For example, there are phenomena that the value $X(t)$ arrives with a delay d of a week and the variables $Y_1(..), Y_2(...), ...Y_n$ are available within few hours or days. There are other that the value $X(t)$ arrives with a delay d of 10 years and the variables $Y_1(..), Y_2(...), ...Y_n$ are available within few days or weeks. In all these cases, the availability of these variables $Y_1(..), Y_2(...), ...Y_n$ in almost real-time, allows us to predict the "now" - *nowcast* the value of $X(t)$. When we use these variables to predict the future values $X(t)$ then in that case we are *forecasting* the evolution of the phenomena.

When facing a nowcasting problem, there are certain challenges arising. Each moment we need to have all the data available, so the acquisition of the data becomes of critical importance. Our systems have to be very efficient in order to provide the predictions on time. We cannot afford to calculate the requested values in 2 weeks time, in order to predict the values of next week. In a forecasting problem on the contrary, our predictions are only based on information about what has happened already. So the efficient acquisition of the required data as well as the efficiency of the systems is not of such an importance. On the special case of nowcasting, where the phenomenon being nowcasted occurs over long periods of time, the efficient data acquisition becomes less important, but the systems continue to have necessity of fast calculation time.

We can use all these large data sets, Big Data, as signals, as proxies for forecast and nowcast so many different phenomena, and even more social phenomena. We can manage to describe and predict how humans and society works. We can use geolocated data to observe and measure the behaviour of a population, to build better cities tailored to the movement of the population, with lower commuting times and lower pollution. We can exploit medical data to build classifiers able to help in diagnosing and curing diseases. We can use industrial data to improve the production

processes, and create smarter and more secure factories. We can do a lot of other incredible and useful things with the support of data and analytical tools able to extract useful knowledge from raw data. You can see many applications in Chapter 3.

Understanding how to predict the future and even more the present and understanding how nowcasting could be better achieved and used in order to face problems of the future is a really interesting problem of our time. And that's the problem that captured the interest and created the motivation for this thesis.

Of course, we cannot answer all these questions in a Ph.D thesis, but we can definitely try answering a few. We worked on different applications based on different sources of data, such as retail market data, human mobility data, social network data and collaboration data. In all the cases, we are interested on how to resolve problems in the context of nowcasting and forecasting independently of the nature of the phenomena under study or the type of the dataset. We are focusing on the development of the necessary models, algorithms and systems to nowcast and forecast as accurately as possible the different phenomena.

The first question we got interested into was predicting seasonal influenza. So we had this massive dataset of retail market data, and we wondered whether it would be useful in any way. We decided to use the retail market data as an external signal in order to predict the curve of an internal time series, the influenza one. Our reasoning was that when the flu season arrives, people are starting to get sick. Getting sick affects their everyday life and behaviour. This change in behaviour should propagate in their purchases in the supermarket. So they will buy products that will reflect the fact that they are sick. So we tried, and you can see the results in Chapter 5.

Our second thought, was quite different. We got interested in humans, and their mobility. Human movements are inherently massive, dynamical, and complex. But understanding the individual mobility patterns, could be of such a fundamental importance for so many different phenomena. We decided to exploit these patterns in order to study and predict the attraction of different socio-economic factors of human environment. Our first approach was to study the distribution of the travelling sub-populations in Tuscany region in Italy, to the airports of the region and for that reason we built a dynamic model for the interplay of attraction of availability of air travel and an airport's popularity among the population. Based on this model, we wanted to forecast the future evolution of the airports in the region. In our second approach, we identified and categorized industrial clusters in Veneto region in Italy, by size and population dynamics and measured their attraction. Our goal was to create a real-time system which will help us to feel the pulse of a city, and being able to predict the rise of new industrial clusters or the death of existing ones. More details, see in Chapter 6.

Finally, we got interested in social networks and we wanted to exploit the temporal information carried by the appearance and disappearance of edges in a fully dynamic context or to unveil hidden connections among existing nodes. We focused our prediction on intra-community interactions, interactions that may occur in the interior of the same community, and we applied the same approach to predict inter-community interactions, the weak links that keep together the modular structure composing complex networks. You can see the results in Chapter 7.

Of course, it was not so easy in the beginning. We came across several challenges on our way, research challenges that are rather normal in every data scientist's work. How can we actually find these sensors, these proxies to aid us in our predictions? These opportunistic sensors, external to the phenomena we want to study and predict are of major importance to these phenomena whose data are not available, or are only available with a significant lag. But even when we acquire the data from these sensors, we still have to worry about validation. We have to remain vigilant as these data can contain noise and biases, hard to detect and remove, and that could make any possible attempt to describe and nowcast these phenomena even harder. It's our responsibility as data scientists to extract useful information from the data, be ready to validate our models

and results, and provide correct conclusions and explanations for the results obtained. There may be limitations to the usefulness of big data analytics, which can identify correlations but not necessarily cause. Correlations can be extremely useful for making predictions or measuring previously unseen behaviour, if they occur reliably. However, they may also be misleading.

After the introduction of the research questions, Chapter 2 sets the stage for our work introducing the notions of Big Data, Data Science, Nowcasting and Forecasting. In Chapter 3 we present studies on nowcasting different phenomena from economic indicators to human and social indicators, while Chapter 4 presents the different datasets used in our work. In Chapter 5 we introduce a nowcasting and forecasting model for epidemic spreading, more specifically we use retail market data to predict the evolution of influenza in the next season. In Chapter 6 we introduce two forecasting models for socio-economic attractors, where using human mobility data we study the impact of airport investment on mobility in Tuscany region and the effect of industrial clusters on regional dynamics in Veneto region. In Chapter 7 we use social networks data to predict interactions inside and between the communities in the networks. Finally, Chapter 8 concludes the thesis and indicates some open issues.

Summing up, in this Ph.D. thesis we introduce data-driven as well as model-driven approaches to predict different phenomena, from epidemics to socio-economic attraction. We use Big Data deriving from our everyday life as external proxies to nowcast and forecast the evolution of phenomena whose study relies only on historical data or data that come only with a significant lag. And that's where the importance of our work lies.

Chapter 2

Setting the Stage

2.1 Big Data Revolution

Modern everyday life is threaded with countless interactions with massive technological systems that support our communication, our transport, our retail activities, and much more. Through these interactions, we are generating increasing volumes of Big Data, documenting our collective behaviour at an unprecedented scale.

We live life in the network. When we wake up in the morning, we check our e-mail, make a quick phone call, walk outside (our movements captured by a high definition video camera), get on the bus (swiping our RFID mass transit cards) or drive (using a transponder to zip through the tolls). We arrive at the airport, making sure to purchase a sandwich with a credit card before boarding the plane, and check our phones shortly before takeoff. Or we visit the doctor or the car mechanic, generating digital records of what our medical or automotive problems are. We post blog entries confiding to the world our thoughts and feelings, or maintain personal social network profiles revealing our friendships and our tastes. Each of these transactions leaves digital breadcrumbs which, when pulled together, offer increasingly comprehensive pictures of both individuals and groups, with the potential of transforming our understanding of our lives, organizations, and societies in a fashion that was barely conceivable just a few years ago.

The amount of data available is really impressive: digital technology has become ubiquitous and very much part of public and private organizations and individuals. People and things have become increasingly interconnected. Smartphones, buildings, cities, vehicles and other environments and devices have been filled with digital sensors, all of them creating evermore data. New high-throughput scientific instruments, telescopes, satellites, accelerators, supercomputers, sensor networks, and running simulations have generated and are generating massive amounts of data.

In a world more and more connected, Big Data gives us the opportunity to observe and measure how our society intimately works: the digital breadcrumbs of human activities carried the capacity to scrutinize the ground truth of individual and collective behaviour at an unprecedented detail. Multiple dimensions of our social life have been increasingly "proxied" by Big Data: automated payment systems record the tracks of our purchases; search engines record the logs of our queries on the web; wireless networks and mobile devices record the traces of our movements; social media record the traces of our opinions and emotions; social networks record the traces of our interactions. Thanks to the massive availability of this data, human behaviour can be observed at large scale. New powerful data-driven tools may be designed and developed to exploit this data for improving the world in many different ways. We can use GPS/GSM data to observe and measure the behaviour of a population, to build better cities tailored to the movement of the population, with lower commuting times and lower pollution. We can exploit medical data to build classifiers able to help in diagnosing and curing diseases. We can use in-

dustrial data to improve the production processes, and create smarter and more secure factories. We can do a lot of other incredible and useful things with the support of data and analytical tools able to extract useful knowledge from raw data. These data describing human activities are at the heart of the idea of a knowledge society, where the understanding of social phenomena is sustained by the knowledge extracted from the miners of Big Data across the various social dimensions by using data mining technologies.

Big Data have been blossoming together with the hope to harness the knowledge they hide to solve the key problems of society, business and science. However, turning an ocean of messy data into knowledge and wisdom is an extremely challenging task. Heterogeneity, scale, timeliness, complexity, and privacy problems with Big Data are the key issues to be addressed at all phases of the pipeline that can create value from data. Twenty-five years ago, most statisticians and computer scientists looked with scepticism at the novel community of KDD(Knowledge discovery in databases) scientists, trying to reformulate the analytical process as data driven discovery. Indeed, such visionary endeavour, combined with the advent of Big Data and spectacular advances in high performance computing, has brought what we call today *Data Science*: a disruptive paradigm shift impacting all disciplines that pushes towards novel scientific methods where "top down" modelling of phenomena coexists with "bottom up" discoveries from data.

Data abundance combined with powerful Data Science techniques has the potential to dramatically improve our lives by enabling new services and products, while improving their efficiency and quality. Many of today's scientific discoveries are already fueled by developments in statistics, data mining, machine learning, network science, databases, and visualization, and we can expect advances in any field related to the comprehension of complex phenomena as in medicine and health (network/personalized medicine), manufacturing (industry 4.0), social dynamics, urban planning, sustainable development.

2.2 Data Science

Starting from the early 90's, the availability of massive data has pushed various disciplines and technologies towards cooperation with the aim of devising ever better models, methods and algorithms for data analysis: database technology and data mining, machine learning and artificial intelligence, complex system theory and network science, statistics and statistical physics, information retrieval and text mining, natural language processing, applied mathematics. Since then, amazing advances in data-driven discovery of patterns, in learning classification and prediction models and in analysing complex networks, paved the road to modern *Data Science*.

Data mining algorithms for automated pattern discovery highlight the structure hidden in massive datasets, such as the clusters of consumers with similar behaviour emerging from large user bases, or the modules of proteins with similar functions emerging from the biological networks of protein-to-protein interactions. Data mining tasks can be divided into two main categories: descriptive and predictive tasks. Each category of tasks has different objectives of analysis and describes different types of possible data mining activities. Descriptive tasks have the goal of presenting the main features of the data: they essentially derive models that summarize the relationship in data, permitting in this way to study the most important aspects of the data. Predictive tasks have the specific objective of predicting the value of some target attribute of an object on the basis of observed values of other attributes of the object or of other objects.

Machine learning methods exploit large "training" datasets of examples to learn general rules and models to classify data and predict outcomes (e.g., classify a taxpayer as fraudulent, a consumer as loyal, a patient as affected by a specific disease, an image as representing a specific object, a post on a social media as expressing a positive emotion). These analytical models allow scientists to "produce reliable, repeatable decisions and results" and uncover "hidden insights" through learning from historical relationships and trends in the data. Machine learning tasks are

typically classified into two broad categories, depending on whether there is a learning "signal" or "feedback" available to a learning system: supervised and unsupervised learning. In supervised learning tasks the computer is presented with example inputs and their desired outputs, given by a "teacher", and the goal is to learn a general rule that maps inputs to outputs. In unsupervised learning tasks there are no labels given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end (feature learning).

Network science unveiled the magic of shifting from the statistics of populations to the statistics of interlinked entities, connected by the ties of their mutual interactions: this change of perspective reveals the universal patterns underlying complex social, economic, technological and biological systems, and is beginning to understand the dynamics of how opinions, epidemics, or innovations spread in our society, as well as the mechanisms behind complex systemic diseases, such as cancer and metabolic disorders and to reveal hidden relationships among them.

Data science emerged due to three concurring factors, unleashed by the digital transformation of society: i) the advent of Big Data; ii) the advances in data analysis and learning techniques; and iii) the advances in scalable high-performance computing infrastructures. The three factors together are an explosive mix: Big Data provide the critical mass of factual examples to learn from; analytics are able to produce predictive models and behavioural patterns from these data; scalable computing platforms make it possible to ingest data and perform analytics.

Data science is an interdisciplinary and pervasive paradigm aiming to turn data into knowledge and value. Data may be structured or unstructured, big or small, static or streaming. Knowledge and value may be provided in the form of predictions, automated decisions, models learned from data, or any type of data visualization delivering insights. As these better, richer and larger data become available, data science can step up from descriptive analytics ("What happens?"), to diagnostic ("Why did it happen?"), to predictive ("What will happen?"), to prescriptive ("How to make it happen?").

Data Science presents many exciting opportunities to improve modern society and boost social progress. It can support policy making, offer novel ways to produce high-quality and high-precision statistical information, empower citizens with self-awareness tools, promote ethical uses of Big Data. Data science may empower citizens, NGOs and policy makers with the means to gain a better understanding of complex socio-economic systems, methods for introspection of complex global processes, tools for assessing the implications of decisions beforehand, and hence to improve our capacity to sustainably manage our society on the basis of well-founded knowledge and inclusive participation.

Public interest around Data Science and Big Data is mounting as data-driven decision making becomes visible in everyday life. Society has shifted from being predominantly "analog" to "digital" in just a few years: society, organizations, and people are increasingly interconnected and "Always On". As mentioned above, as a consequence of digitization, data are collected about anything, at any time, and at any place. The spectacular growth of Big Data makes it possible to record, observe, and analyse the behaviour of people, machines, and organizations. The Internet of Things (IoT) is rapidly expanding: our homes, cars, factories, and cities are expected to become "smarter" by exploiting this wealth of collected data.

These developments are also changing the way scientific research is performed. Model-driven approaches are supplemented with data-driven approaches. A new paradigm emerged, where theories and models and the bottom up discovery of knowledge from data mutually support each other. Experiments and analyses over massive datasets are functional not only to the validation of existing theories and models, but also to the data-driven discovery of patterns emerging from data, which can help scientists design better theories and models, yielding deeper understanding of the complexity of social, economic, biological, technological, cultural and natural phenomena.

2.3 Measuring human behaviour with Big Data

How many people are unemployed right now? How many people currently have the flu?

Policymakers and business people rely on measurements of what is currently happening in society to make good decisions. However, measurements of key indicators such as unemployment rates or the spread of infections are often delayed, such that there are only measurements for the previous week, or the previous month, rather than the current week. This is because traditional methods of measuring these quantities can be slow.

Big Data offer the potential capability of creating a digital nervous system of our society, enabling the measurement, monitoring and prediction of relevant aspects of socio-economic phenomena in quasi real time. Analysing Big Data, we can observe and measure how our society works. This potential has fuelled, in the last few years, a growing interest around the usage of Big Data to support official statistics in the measurement of individual and collective economic well-being, as they provide new means to the statistic institutes.

Big Data might help us to make better decisions in an economic context. Making a decision can be quite complex, and this might be quite complex for a range of people – for us as humans in general in our everyday life, but also for commercial stakeholders or governmental stakeholders.

Economists, investors, and journalists avidly follow monthly data releases on economic conditions. This is a problem because most of these key statistics come with one major disadvantage – it's so difficult and time consuming to calculate these numbers that reports are only available with a lag: the data for a given month is generally released about halfway through the next month, and are typically revised several months later. Many important measurements are published on a periodic basis. For example, the United States government releases GDP figures every quarter, and unemployment figures every month. These data are published with a lag; the employment rate for March of 2018 was released in April of 2018. Even once published, many of these time series are still subject to later revisions as more information becomes known. Because of these issues, such data do not provide an up-to-date estimate of the statistic they are tracking. It would clearly be helpful to have more timely estimates of these measurements. Big Data, differently from official approaches, has the potential to produce continuous statistics.

Moreover, in official statistics, information are collected by means of survey or census; however, census are complex and expensive, so surveys represent a feasible alternative to collect data. In order to make reasonable inferences on the target population surveys should be drawn representative of the population. The official statistics is based mainly on surveys or administrative data and often these are not accessible, while most of Big Data are public or belong to private companies. In this scenario, Big Data represent a concrete opportunity for understanding social complexity.

Traditionally, you rely on what has happened in the past, and you want to forecast the next step. But here we have a problem: the latest number which is available comes with a significant time lag. We know right now, maybe, the unemployment rate a month ago or two months ago. The need of more timely forecast of these economic indicators pushed economists to consider several sources of data on real-time economic activity, such as transaction data from private sector companies with the aim to be able to "contemporaneously forecast" changes and events in nearly real-time. In order to forecast a little bit closer to the time where we are right now, this becomes ultimately a *nowcasting* challenge – we want to forecast the present rather than the future.

Taking into account the types of data sets that are available nowadays, a question comes in mind, 'can we use some of these data feeds, which are immediately accessible after their creation, to help here?' Can we analyse data from sites such as Google, Wikipedia, Twitter and Flickr to find out whether we can use data from the internet to measure and even predict what humans do in the real world?

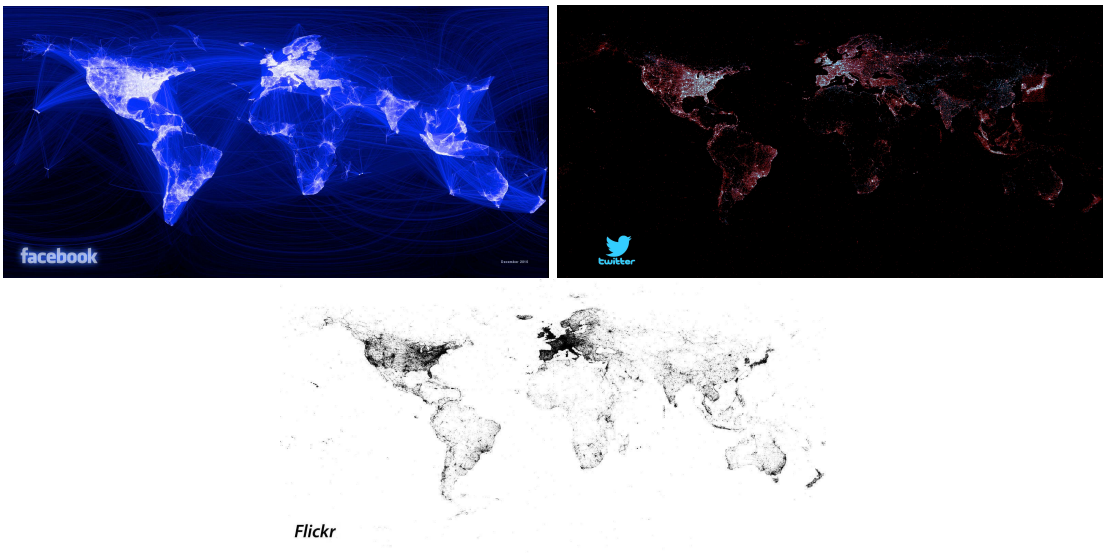


Figure 2.1: Map of the locations of Facebook users, tweets from Twitter and Flickr photos taken by users all over the world.

For example, Google searches – they are available for analysis more or less immediately after we have carried out searches on a global, worldwide scale. Google has made available a public web facility, called Google Trends [1], based on Google Search, that shows how often a particular search-term is entered relative to the total search-volume across various regions of the world, and in various languages. A lot of recent studies have considered the predictive capability of web search data made available by Google Trends [73, 292].

Another interesting source of data derives from social networks, such as Facebook, Twitter, Flickr, Instagram, etc. Vast data-streams from these social networks contain people's opinions, fears and dreams. But they can also tell where the users are at the moment they use them. Take for example Figure 2.1, where it shows the locations of users of Facebook, tweets from Twitter and photos uploaded by users of Flickr.

The shapes of the continent are emerging, visualising the global coverage of this new type of data set. At the same time, given that these data sets can be retrieved from an online platform, these data sets are extremely cheap to handle. They are naturally occurring because people are using these services where they are at the moment in time. They are interested in something.

This gives scientists the opportunity to study human behaviour on a worldwide scale in a very, very cheap way, and it allows them to get this data as soon as it is generated, so the process is very fast. If this is compared to traditional methods in the social sciences research area, then this might complement these methods, which are mainly consisting of laboratory based experiments and online or offline surveys. These have the advantage that there is control over the participants. They are asked specific questions, and they give their absolute attention to the scientist. But at the same time, the number of participants is very limited, so the global scale of these new data sources might complement these existing sources.

Research groups around the world, across business and across policy, have been looking at how they can use these new data sets to make better decisions about what humans do and make better decisions about how resources can be allocated better. Examples of how people use these data sets across areas as diverse as economics, health, happiness, and crime. This data is opening up a new era for understanding of human behaviour – and also for policy making and business processes which depend upon this understanding. Research has shown how data can give insight into the risk of an upcoming stock market crash; decrease delays in measuring the spread of

illness; or even allows to predict where crimes might occur.

How people are starting to use these online data, the kind of data that's generated through our usage of search engines, such as Google, or social network sites like Twitter or Facebook can offer us quicker measurements of important aspects of what's going on in the world at the moment and even possibly forecast what they're going to do in the future.

In the offline world, too, humans are generating huge amounts of data just by using credit cards, taking public transport, shopping in the supermarket, or making a phone call to the police. If scientists manage to identify repeating patterns of behaviour in these datasets, it can help predict where crimes might occur, or work out where people are going to move to, which is something really crucial in understanding where diseases are going to spread.

2.4 Nowcasting and Forecasting

Nowcasting As mentioned above in 2.3, Big Data can help in getting faster measurements of human behaviour; a process we came to call *nowcasting*. Nowcasting is a novel promising field of research and it has been successfully combined with the analysis of large datasets of human activities. It is generally a collective or global approach since it requires a data model which capture the behaviour of a mass of people.

The term is a contraction for now and forecasting and has been used for a long-time among meteorologists to forecast near-term weather conditions, particularly violent weather approaching a particular region, based on vast amounts of incoming data. Knowing that a tornado is minutes away from your neighbourhood can be the difference in whether your family survives or not. The goal of *nowcasting*, which is also called "predicting the present", is to estimate up-to-date values for phenomena whose actual observations are available only with a delay. Thanks to the pervasiveness of Big Data, nowcasting is useful beyond weather.

It has recently become popular in economics as standard measures used to assess the state of an economy, e.g., gross domestic product (GDP), are only determined after a long delay, and are even then subject to subsequent revisions. While weather forecasters know weather conditions today and only have to predict future weather, economists have to forecast the present and even the recent past.

Nowcasting is particularly relevant for those key macro economic variables which are collected at low frequency, typically on a quarterly basis, and released with a substantial lag. To obtain "early estimates" of such key economic indicators, nowcasters use the information from data which are related to the target variable but collected at higher frequency and released in a more timely manner. For example, euro area GDP, which is the key statistic describing the state of the economy, is only available at quarterly frequency and is released six weeks after the close of the quarter. However, there are several variables related to GDP, such e.g. industrial production or various surveys, available at monthly frequency and published with shorter delay. These can be used to construct early estimates of GDP. Heuristics are applied to triangulate and form stable estimates. With the advent of real time data and big data computing such estimates can be very precise, as it can be seen in weather forecasts.

So, in a *nowcasting* problem, there is a time-dependent variable $X(t)$ that cannot be measured instantaneously. We can only know the value of $X(t)$ at time $t+d$, where d is a delay. *Nowcasting* is the act of (probabilistically) approximating the value of $X(t)$ at time t, based on a set of n (possibly latent) variables $Y_1(..), Y_2(...), ...Y_n$, which can be measured at time t or time $t+d_{Y_{1,2,...n}}$ where $d_{1,2,...n} < d$. Often these variables are latent, meaning that they are the projections on a smaller-dimensional space, of a set of variables in a larger multi-dimensional space that can be readily measured.

Nowcasting models have been applied in many institutions, in particular central banks, and the technique is used routinely to monitor the state of the economy in real time. Monetary policy

decisions in real time are based on assessments of current and future economic conditions using incomplete data. Because most data are released with a lag and are subsequently revised, both forecasting and assessing current-quarter conditions, *nowcasting* are important tasks for central banks. Central banks (and markets) pay particular attention to selected data releases either because the data are released early relative to other variables or because they are directly tied to a variable the central banks want to forecast (e.g. employment or industrial production for nowcasting gross domestic product, GDP).

For example, the European Central Bank has published a paper [25] on how it uses nowcasting of huge volumes of economic data to get an early jump on GDP estimates, which are critical to establishing short-term and long-term fiscal policy. The authors argue that the nowcasting process goes beyond the simple production of an early estimate as it essentially requires the assessment of the impact of new data on the subsequent forecast revisions for the target variable. As mentioned before, nowcasting is particularly relevant for those key macro economic variables which are collected at low frequency, typically on a quarterly basis, and released with a substantial lag. To obtain "early estimates" of such key economic indicators, nowcasters use the information from data which are related to the target variable but collected at higher frequency, typically monthly, and released in a more timely manner. For example, euro area GDP, which is the key statistic describing the state of the economy, is only available at quarterly frequency and is released six weeks after the close of the quarter. However, there are several variables related to GDP, such e.g. industrial production or various surveys, available at monthly frequency and published with shorter delay. The authors used this information to construct early estimates of GDP. One of the key features of an effective nowcasting tool is to incorporate the most up-to-date information in an environment in which data are released in a non-synchronous manner and with varying publication lags.

Choi and Varian [73] used the term *nowcasting* as opposed to *forecasting*, to advocate the tendency of web searches to correlate often with various economic indicators, which may reveal helpful for short term prediction. The authors claim that Google Trends data may help predict the future or even better the present. They also suggest that nowcasting can be used to predict consumer behaviour in specific regions. As they write, by using Google Trends data for a given geography, there can be made accurate predictions for major economic activities, such as retail, automotive and house sales as well as travels. For example, the volume of queries on a particular brand of automobile during the second week in June may be helpful in predicting the June sales report for that brand, when it is released in July. Or by modelling search data for, say, "real estate agents" in Texas, they are able to more accurately estimate home sale activity in the region. Their research has been validated elsewhere. For example, economists in Hungary have followed Varian and Choi's nowcasting methodology, and created extremely refined estimates of consumer activity in that country [344].

Whether Google search data is able or not to show patterns in online search behaviour and can predict "early warning signs" for moves in the stock market is the topic of another recent study, which analysed a broad number of finance related search terms and concluded with the result that the combination of major data sets can be seen as a mirror of collective human financial behaviour [292]. Although the reliability of data from the Internet is still subject of criticism, all of the studies mentioned above, unveil strong links between online search behaviour and real world events.

Speaking about the search data and the potential of nowcasting in 2012, Hal Varian, Google chief economist and one of the first economists who used search engine data to forecast near-term values of economic indicators, said: "the hope is that as you take the economic pulse in real time, you will be able to respond to anomalies more quickly."

But nowcasting can also be used in the field of epidemiology. Traditional flu monitoring depends in part on national networks of physicians who report cases of patients with influenza-

like illness (ILI); a diffuse set of symptoms, including high fever, that is used as a proxy for flu. That estimate is then refined by testing a subset of people with these symptoms to determine how many have flu and not some other infection. But the near-global coverage of the Internet and burgeoning social-media platforms such as Twitter have raised hopes that these technologies could open the way to easier, faster estimates of ILI, spanning larger populations. The mother of these new systems was Google Flu Trends [2], launched in 2008. Based on research by Google and the Center for Disease Control (CDC), it relied on data mining records of flu-related search terms entered in Google's search engine, combined with computer modelling. Its estimates had almost exactly matched the CDC's own surveillance data over time — and it delivered them several days faster than the CDC can. The system had been rolled out to 29 countries worldwide, and had been extended to include surveillance for a second disease, dengue.

The authors in [286] and [143] showed that search data could help predict the incidence of influenza-like diseases. Intuitively, there is a close relationship between how many people search for flu-related topics and how many people actually have flu symptoms. Of course, not every person who searches for "flu" is actually sick, but a pattern emerges when all the flu-related search queries are aggregated. The authors compared the query counts with traditional flu surveillance systems and found that many search queries tend to be popular exactly when flu season is happening. By counting how often we see these search queries, it is possible to estimate how much flu is circulating in different countries and regions around the world. This finding is important because traditional flu surveillance agencies deliver their estimates with a delay of weeks, while the web search-based service delivers basically daily, and timeliness is crucial to enable public health officials and health professionals to better respond to seasonal epidemics and pandemics.

These examples very nicely highlight the power which lies in these naturally occurring data sets in order to improve our understanding and our forecast of what is going on in the world right now. Ultimately, this can inform decision making and allows us to make better decisions than actually it was possible before.

So in few words, nowcasting is the discipline of determining a trend or a trend reversal objectively in real time. It employs correlated data that are real-time, or more frequently updated than the desired statistic. Methods for this task leverage observations of correlated time series to estimate values of the target series. Nowcasting is fact-based, focuses on the known and knowable, and therefore avoids forecasting.

Forecasting *Forecasting* is the process of making predictions of the future based on past and present data and most commonly by analysis of trends. A commonplace example might be estimation of some variable of interest at some specified future date. An important distinction in forecasting, in comparison to nowcasting, is that the future is completely unavailable and must only be estimated from what has already happened.

Forecasting has applications in a wide range of fields where estimates of future conditions are useful. Not everything can be forecasted reliably, if the factors that relate to what is being forecast are known and well understood and there is a significant amount of data that can be used very reliable forecasts can often be obtained. If this is not the case or if the actual outcome is effected by the forecasts, the reliability of the forecasts can be significantly lower.

There are different categories of forecasting methods; qualitative forecasting techniques are subjective, based on the opinion and judgement of consumers, experts; they are appropriate when past data are not available. They are usually applied to intermediate- or long-range decisions. Instead, quantitative forecasting models are used to forecast future data as a function of past data. They are appropriate to use when past numerical data is available and when it is reasonable to assume that some of the patterns in the data are expected to continue into the future. These methods are usually applied to short- or intermediate-range decisions.

Time Series A most common source of these data, used both for nowcasting as well as forecasting, are *time series*, a collection of data points collected at constant time intervals. These are analysed to determine the short term trend so as to nowcast the present and long term trend so as to forecast the future. The purpose of time series analysis is generally twofold: to understand or model the stochastic mechanisms that gives rise to an observed series and to predict or forecast the future values of a series based on the history of that series.

A time series is a series of data points indexed (or listed or graphed) in time order. Most commonly, a time series is a sequence taken at successive equally spaced points in time. Thus it is a sequence of discrete-time data. Observations that unfold over time usually represent valuable information subject to analysis, classification, indexing, forecasting, or interpretation. Real-world examples include financial data (e.g., stock market fluctuations), medical data (e.g., electrocardiograms), computer data (e.g., log sequences), or mobility data (e.g., location of moving objects).

Time series analysis comprises methods for analysing time series data in order to extract meaningful statistics and other characteristics of the data. Time series forecasting is the use of a model to predict future values based on previously observed values.

A core issue when dealing with time series is determining their pairwise similarity, i.e., the degree to which a given time series resembles another. Since time series are ubiquitous, the assessment of their similarity is a core part of many computational systems. In particular, the similarity measure is the most essential ingredient of time series clustering and classification systems.

2.4.1 Problems in Nowcasting

As Castle et al. [62] point out, contemporaneous forecasting - nowcasting, is valuable in itself, but it also raises a number of interesting research questions involving topics such as variable selection, mixed frequency estimation, and incorporation of data revisions, to name just a few.

Goel in [145] provide a useful survey of work in this area and describe some of the limitations of web search data. As they point out, search data is easy to acquire and is often helpful in making forecasts, but may not provide dramatic increases in predictability. But, they concluded that in the absence of other data sources, or where small improvements in predictive performance are material, search queries provide a useful guide to the near future.

The most famous example of problematic nowcasting comes from Google itself with the nowcasting model of Google Flu Trends failing to provide accurate predictions in different moments in history.

The Google Flu Trends case As mentioned above, a lot of studies have considered the predictive capability of web search data, such as those made available by the Google Trends service [1]. When a small team of software engineers first started working on Google Flu Trends (GFT) in 2008 [2], they wanted to explore how real-world phenomena could be modelled using patterns in search queries. Since its launch, Google Flu Trends has provided useful insights and served as one of the early examples for *nowcasting* based on search trends, which is increasingly used in health, economics, and other fields. Over time, they' ve used search signals to create prediction models, updating and improving those models over time as they compared their prediction to real-world cases of flu.

This would be a success story, if the story stopped here. But it doesn't. What people recognised over time was that this Google-based estimate of flu cases, from time to time, wasn't so accurate as people might have hoped. In particular, as Lazer and al. notice in [207], in February 2013, Google Flu Trends made headlines but not for a reason that Google executives or the creators of the flu tracking system would have hoped. *Nature* reported that GFT was

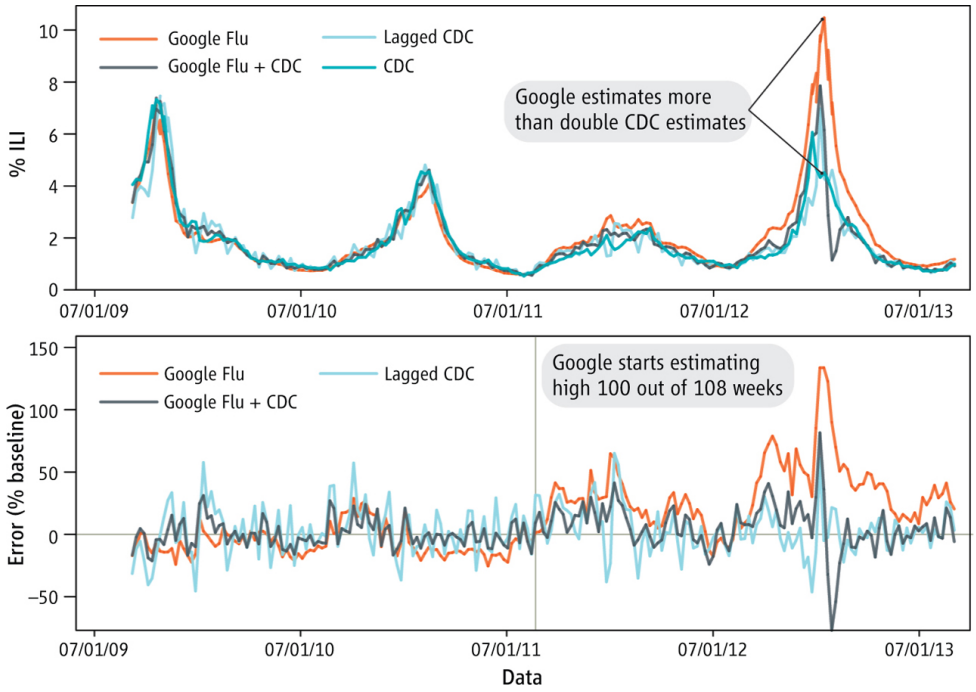


Figure 2.2: GFT overestimation. (Top) Estimates of doctor visits for ILI. "Lagged CDC" incorporates 52-week seasonality variables with lagged CDC data. "Google Flu + CDC" combines GFT, lagged CDC estimates, lagged error of GFT estimates, and 52-week seasonality variables. (Bottom) Error [as a percentage [Non-CDC estimate) (CDC estimate)]/(CDC) estimate). Both alternative models have much less error than GFT alone. Mean absolute error (MAE) during the out-of-sample period is 0.486 for GFT, 0.311 for lagged CDC, and 0.232 for combined GFT and CDC.

predicting more than double the proportion of doctor visits for influenza-like illness (ILI) that the Centre for Disease Control (CDC), which bases its estimates on surveillance reports from laboratories across the United States [266, 57]. Google had dramatically overestimated the number of people having the flu (see Figure 2.2), and people start to question why might this have happened and what the underlying reasons were.

And that would not be the first time that a flu season has tripped Google up. In 2009, GFT completely missed the nonseasonal influenza A-H1N1 [80]. According to the authors it was a glitch attributed to changes in people's search behaviour as a result of the exceptional nature of the pandemic. GFT engineers updated the algorithm in 2009, but its comparative value as a stand-alone flu monitor is questionable. A study in 2010 [145] demonstrated that GFT accuracy was not much better than a fairly simple projection forward using already available (typically on a 2-week lag) CDC data.

The head of the CDC Influenza Surveillance and Outbreak Response Team told Nature News that she monitors GFT (and other digital disease detection sentinels) "all the time", likely in the sense that some data are better than no data [57]. Moreover, some investigators are beginning to use GFT as ground truth for epidemiologic studies [263].

The point is that when new flu strains (in particular, H1N1) was subject to public discussion, then these Google-based estimates weren't so precise as they could have been. So a number of people came up with the idea that when a lot of people – triggered by maybe media coverage – are looking up online flu and flu-related symptoms, then this obviously doesn't any longer closely match with what is going on in the world right now in terms of actual disease cases. So we have to be careful about changes in the underlying behaviour of people, which might be triggered by

all sorts of external factors.

Lazer and al. [207] name it *Big Data Hubris*, that refers to the assumption that Big Data are a substitute for, rather that a supplement to, traditional data collection and analysis. Also is quite likely that GFT was an unstable reflection of the prevalence of the flu because of the changes the engineers themselves were making in Google's search algorithm in order to improve the commercial service, but also the changes the customers made in using that service, that affected GFT' s tracking. The Google search algorithm was not a static entity, as the company was constantly testing and improving search. But also, modifications to the search algorithm were presumed to be implemented so as to support Google's business model. And last but not least, GFT has never documented the 45 search terms used, and the examples that have been released appeared misleading [80]. Neither were core search terms identified nor larger search corpus provided. It wouldn't be ethically acceptable maybe to make the data available, but no restrictions should apply to the derivative, aggregated data.

In a brief working paper, Google described the need to revise GFT [82]. They acknowledged that a multivariable approach would enhance the accuracy of GFT, but that paper also shared many of the weaknesses inherent in the original and first revision to GFT. First, the methods lacked transparency, as the working paper did not identify the model they were implementing. Second, the predictive validity of the revised GFT remains unknown, because Google only included 5 weeks of data in the paper estimating the predictive accuracy. Last, their revision still relied on investigator opinion to select/ omit some queries and failed to incorporate automatic updating.

2.5 Solutions

As people increasingly turn to the Internet for news, information, and research purposes, it is tempting to view online activity at any moment in time as a snapshot of the collective consciousness, reflecting the instantaneous interests, concerns, and intentions of the global population. From this perspective, it is a short step to conclude that what people are searching for today is predictive of what they will do in the near future. By appropriately aggregating counts of search queries related to retail activity, moviegoing, or travel, one might be able to predict collective behaviour of economic, cultural, or political interest. Determining the nature of behaviour that can be predicted using search, the accuracy of such predictions, and the time scale over which predictions can be usefully made are therefore all questions of interest.

The coverage of "traditional data" generated by official statistics and surveys is often limited and the analysis expensive. It takes time to be collected, processed, verified and eventually published. The traditional sources will continue to generate useful and important information, but the revolution of digital data presents the opportunity to get richer and deeper insights into human experience, aiming to complete and enrich socio-economic indicators that already exist. There is a tendency for Big Data research and more traditional applied statistics to live in two different realms—aware of each other's existence but generally not very trusting of each other [207]. Big Data offer enormous possibilities for understanding human interactions at a societal scale, with rich spatial and temporal dynamics, and for detecting complex interactions and nonlinearities among variables. We contend that these are the most exciting frontiers in studying human behaviour. However, traditional "small data" often offer information that is not contained (or containable) in Big Data, and the very factors that have enabled Big Data are enabling more traditional data collection. The Internet has opened the way for improving standard surveys, experiments, and health reporting. Instead of focusing on a *Big Data revolution*, perhaps it is time we were focused on an *all data revolution*, where we recognize that the critical change in the world has been innovative analytics, using data from all traditional and new sources, and providing a deeper, clearer understanding of our world.

As mentioned above, examples of different areas engaging Big Data analytics have shown and underlined the growing importance and promise of real-time digital data as valuable source in reflecting and sometimes predicting collective human behaviour and could therefore also eventually forecast changes and increase prevention from unexpected crisis. The real-time awareness of the well-being of a country, the anomalous changes in how a society accesses or searches for goods and services and a real-time feedback on the effectiveness of policy changes may serve as indicators of changes and lead to a more adaptive approach, to a greater resilience and better outcomes. The term "real-time" though, is not to be misunderstood as "occurring immediately", it should rather be considered as information which is produced and made available in relative short and relevant period of time and which also can be analysed within a timeframe that allows action to be taken in response.

What we would like to achieve if it is possible, is to build more sophisticated forecasting models than those describe already in the existing bibliography in order to be able to achieve predicting the present in a more effective way. Of course, the existing models could serve as baselines to help us get started with our own modelling efforts in order to be able also to refine them for our specific applications. Nowcasting is becoming a valuable tool for anyone interested in economic and business conditions in the present. And knowing with better precision what's happening now will give us a better understanding of what might be in store for the future.

Chapter 3

Nowcasting at Work

3.1 History of Nowcasting

Nowcasting was first used in meteorology and is a technique for very short-range forecasting that maps the current weather, then uses an estimate of its speed and direction of movement to forecast the weather a short period ahead, assuming the weather will move without significant changes.

Nowcasting is a very old technique. When Admiral FitzRoy first produced forecasts at the Met Office [UK] in the 1860s, he did it by collecting reports of storms from around the coast, and then sharing these reports with coastal ports that may be downwind, so that they knew there was bad weather coming. This was a simple form of nowcasting.

The term *nowcasting* was actually coined in the 1980s by Met Office scientist Professor Keith Browning, to describe the process of extrapolating a sequence of radar images to produce a very short-range rainfall forecast.

Nowadays, the concept of nowcasting is also used in other scientific fields, such as economics, healthcare and social phenomena.

Something we've been particularly fascinated by is data from the internet. So data on what people are looking for on Google, data on what pages people are looking at on Wikipedia, or data on who's talking to who on Twitter for example. Researchers in this theme investigate how new Big Data, in particular data from online sources such as Google, Wikipedia, Flickr and Twitter, can be used to measure human behaviour and even anticipate what decisions people may make in the future. Work in this theme is designed to deliver both behavioural measurements and predictions of practical value to external commercial and governmental stakeholders, as well as empirical insights to support the construction of theories of human behaviour.

So far as we know, the first published paper that suggested that web search data was useful in forecasting economic statistics was Ettredge in [115], which examined the US unemployment rate. At about the same time Cooper in [81] described using Internet search volume for cancer-related topics. Since then there have been several papers that have examined web search data in various fields.

In the Table 3.1 you can see a classification of the studies in Nowcasting that are presented in this Chapter. The studies are classified based on their *application domain* as well as on the type of data used for the study, *structured* or *unstructured*.

	Type of Data	
Application Domain	Structured	Unstructured
Economy	[73, 292, 115, 295, 74, 293, 261, 198, 91, 46, 72, 194, 245, 349, 105, 35]	[197, 45, 11]
Unemployment	[115, 15, 109, 337, 159, 20, 240]	[222, 343]
GDP	[293, 261, 156, 132, 19, 312, 192, 136, 125, 212]	
Stock Market	[292, 295, 198, 91, 46, 245, 294, 247]	[45, 11]
Humans and Society	[143, 295, 293]	
Human Behaviour	[145, 301, 100, 107, 321, 211, 155, 352, 220, 47]	[16, 195, 205, 148]
Natural Disasters	[34]	[291, 150]
Civil Unrest and Crime	[298, 187, 48]	[63, 8, 186, 185]
Well-being	[134, 133, 32, 302, 204, 110, 270]	[176, 67, 68, 90, 44, 96, 217, 59, 230, 231, 43]
Happiness	[166]	[102, 243, 126, 146, 86, 196]
Mental Health	[44, 217, 59, 43, 265, 300, 206, 276, 355, 356, 179, 236, 154]	
Suicide	[370, 238, 199, 200]	[161, 38, 305]
Poverty	[254, 304, 309, 180, 42, 327, 129, 332, 335]	
Epidemiology	[286, 143, 81, 362, 173, 127, 376, 364, 277, 347, 119, 18, 69, 12, 106, 264, 34, 23, 114, 232, 184, 118, 188, 218, 339, 241]	[231, 265, 322, 75, 262, 329, 323, 373, 365, 153, 169, 325, 275, 239, 316, 269, 93, 328, 202, 203, 54, 83, 77, 55, 313, 336, 66, 111, 120, 163, 359]

Table 3.1: Classification of the studies in Nowcasting based on their *application domain* as well as on the type of data used for the study, *structured* or *unstructured*.

3.2 Nowcasting for Economy

Recent studies in the area of economics have started to focus on the analysis of data describing online behaviour, stemming from services such as the search engine Google [295, 74, 293, 261, 292, 198, 91], the search engine Yahoo! [46], the online encyclopaedia Wikipedia [245, 197], the microblogging platform Twitter [45], as well as investigating data from more traditional news sources such as the Financial Times [11].

Choi and Varian [73, 72] looked into the problem of nowcasting, and they investigated a range of problems where our online search behaviour might help to close the nowcasting problem gap in terms of time delay. In [74], an updated and streamlined version of those two initial papers, they described how to use Google Trends data to predict several economic metrics including initial claims for automobile demand, unemployment, and vacation destinations. First, they looked at automotive sales. They were able to correlate human searches on individual automotive brands, parts of vehicles and cars to the number of cars sold; so that's interesting. They found a correlation. They found a relationship between what people search for in a car, automotive-related domain, and how many cars are actually sold in this period. So if you take this into account, in addition to what you know has happened before then you have historic automotive sales, maybe one month or two months old. And you take into account online searches on this domain, on this topic, which is basically related to the current period, let's say the current

month. Then you are able to improve the nowcasts, so this means the forecasts of automotive sales right now. Choi and Varian successfully demonstrated that this is possible.

They looked also at a second example, and this relates to human travel behaviour. They got hold of data of Hong Kong. Hong Kong records how many international visitors are coming to the city, and they split these visitors up in terms of countries, so they know how many visitors, how many tourists, are coming from Germany, from Italy, from France, and all other countries all over the world. This is a number which gets aggregated, and this aggregation is also a very slow process, so this number is not available immediately for governmental stakeholders or other people who might be interested in the question, 'how many tourists are coming to the city?' because they can use this information to make a better decision for their business directions.

So Choi and Varian were able to show that there is a correlation between, let's say, inhabitants of the country of Germany googling for Hong Kong (and all related subjects like restaurants, hotels, and all other places you can imagine in the city of Hong Kong) and how many people were subsequently visiting Hong Kong coming from Germany. This was a relationship which wasn't there for one country only. They were able to show across a number of countries that there is a very significant relationship, which subsequently shows that also in this case you might be better by incorporating a much quicker and, yes, much more easily accessible data feed in your forecasting equations in terms of getting a better idea of what is going on in the world right now. Choi and Varian realised that there is a relationship between people in their home country searching for venues, for places, in Hong Kong and later visiting, indeed, the city of Hong Kong. Given that this information is much more quickly available for analysis, they were also able to show in this example that it is possible to improve forecasts of the present of people coming from a specific country visiting Hong Kong later on.

Koop and Onorante in [194] add to the nowcasting ideas of Choi and Varian, by using dynamic model selection (DMS) methods which allow for model switching between time-varying parameter regression models. They extended the DMS methodology by allowing for the model switching to be controlled by the Google variables through Google probabilities. That is, instead of using Google variables as regressors, they allow them to determine which nowcasting model should be used at each point in time.

The majority of economic indicators is published by official statistics. In official statistics, information are collected by means of survey or census; however, census are complex and expensive, so surveys represent a feasible alternative to collect data. But data from surveys or administrative data are often not accessible, while most of Big Data are public or belong to private companies. In this scenario, Big Data represent a concrete opportunity for understanding social complexity. For example, in 2010 Statistics Netherlands has introduced a new method for computing the Dutch Consumer Price Index (CPI) based on supermarket scanner data [349]. In a similar line of work, in [105] Dubey and Gennari presented the uses of big data in the domain of food prices, from producing official statistics to nowcasts for food security early warning while in [35] Bernardini et al. from the Italian National Institute of Statistics (ISTAT) experimented on price index evaluation, starting from the comparison among different formulas for the elementary indices combined with different ways of using supermarket scanner data through sampling selection schemes.

In early 2016, the European Statistical System (Eurostat) launched a competition[1] to develop nowcasting tools using big data sources on key statistics, such as unemployment, price indexes and tourism. Later that year, Mazzi from Eurostat published an operational step by step approach aiming to facilitate the use of Big Data in nowcasting cases [237]. By 2017, Eurostat issued a practical guide for processing supermarket scanner data to calculate the CPIs of EU countries to ensure the comparability of the values across Europe as well as to modernise the official

[1]BDCOMP: the Big Data for Official Statistics Competition (*http* : *//ec.europa.eu/eurostat/cros/content/bdcomp*$_e$*n*)

statistics[2].

3.2.1 Unemployment

The first to try nowcasting for unemployment, were Ettredge et al. in [115] where they found that counts of the top 300 search terms during 2001-2003 were correlated with US Bureau of Labor Statistics unemployment figures.

Latern on, the authors in [15, 109, 337] have confirmed the value of search data in forecasting unemployment in the US, Germany and Israel. Guzman in [159] has examined Google data as a predictor of inflation. Baker and Fradkin in [20] have used Google search data to examine how job search responded to extensions of unemployment payments. In more recent works, social media data has been used to nowcast employment status and shocks in [222, 343]. Finally, [240] summarise how online search data can be used for economic nowcasting by central banks, as they show that the volume of online searches can be used as indicators of economic activity, more specifically for unemployment and housing markets in the United Kingdom.

3.2.2 GDP

One of the most used measures of the economic health of a nation is the Gross Domestic Product (GDP): the market value of all officially recognized final goods and services produced within a country in a given period of time. GDP, prosperity and well-being of the citizens of a country have been shown to be highly correlated. However, GDP is an imperfect measure in many respects. GDP usually takes a lot of time to be estimated and arguably the well-being of the people is not quantifiable simply by the market value of the products available to them. In [156] Guidotti et al. use a quantification of the average sophistication of satisfied needs of a population as an alternative to GDP, as it can be calculated more easily than GDP and it proves to be a very promising predictor of the GDP value, anticipating its estimation by six months. The measure is arguably a more multifaceted evaluation of the well-being of the population, as it tells us more about how people are satisfying their needs. An alternative methodology uses electronic payment data to nowcast GDP [132]. However in this case the only issue addressed is the timing issue, but no attempt is made into making the measure more representative of the satisfaction of people's needs.

Monetary policy decisions in real time are based on assessments of current and future economic conditions using incomplete data. Because most data are released with a lag and are subsequently revised, both forecasting and assessing current-quarter conditions (nowcasting) are important tasks for central banks. Central banks (and markets) pay particular attention to selected data releases either because the data are released early relative to other variables or because they are directly tied to a variable the central banks want to forecast (e.g. employment or industrial production for nowcasting gross domestic product, GDP). In principle, however, any release, no matter at what frequency, may potentially affect current-quarter estimates and their precision. In nowcasting current-quarter GDP growth, qualitative judgement is typically combined with simple small-scale models that sometimes are called "bridge equations". The idea is to use small models to "bridge" the information contained in one or a few key monthly data with the quarterly growth rate of GDP, which is released after the monthly data [19, 312, 192, 136]. Other examples can be found on the Eurozone [125], or on different targets such as income distribution [212].

Another study has shown that Internet users from countries with a higher per capita GDP are more likely to search for information about years in the future than years in the past [293]. Preis et al. used these Google Trends logs to introduce a future orientation index to quantify

[2]https://circabc.europa.eu/sd/a/8e1333df-ca16-40fc-bc6a-1ce1be37247c/Practical-Guide-Supermarket-Scanner-Data-September-2017.pdf

the degree to which Internet users worldwide seek more information about years in the future than years in the past and they found a striking correlation between the country's GDP and the predisposition of its inhabitants to look forward. Building upon this study, Noguchi et al. in [261] used search engine query data to construct measures of the time-perspective of nations, and tested these measures against per-capita GDP. Their results indicate that nations with higher per-capita GDP are more focused on the future and less on the past, and that when these nations do focus on the past, it is more likely to be the distant past.

3.2.3 Stock Market

Financial markets are truly fascinating. If you have some idea what they are going to do in the near future, then you have the opportunity to become extremely rich. At the same time, if you are wrong with your forecast, then you can potentially lose a lot of money. From a behavioural point of view, financial markets are fascinating because we can study what people have decided to buy or sell. When stock markets became electronic trading platforms in the 1990s, these were one of the first big data generators on human behaviour, very detailed in a very detailed fashion, recording what people have decided to do.

Crises in financial markets affect humans worldwide. Detailed market data on trading decisions reflect some of the complex human behaviour that has led to these crises. At their core, financial trading data sets reflect the myriad of decisions taken by market participants. In today's world, information gathering often consists of searching online sources. All these massive new data sources resulting from human interaction with the Internet may offer a new perspective on the behaviour of market participants in periods of large market movements.

This kind of data provides insights into our economic life on different scales. A steadily increasing number of Internet users visit websites of search engines every day. Each query request can be seen as an individual vote: using search engines, we leave information about our interests codified as search terms. Thus, search engines can collect our interests on the smallest possible scale, the scale of individual requests. On larger time scales, our interest forms trends. Aggregated search volume data can be used for uncovering such trends that affect our economic life on large scales, such as the international financial crisis [295].

In 2010, T. Preis and his colleagues [295] provided evidence that search engine query data and stock market fluctuations are correlated. They used big data in combination with approaches from physics to help us understand infrequent but catastrophic stock market crises. Data on information flow via the Financial Times, Wikipedia and Google can be linked to trading patterns in financial markets. Changes in searches for financial and political information on Google and Wikipedia may have contained early warning signals of stock market moves.

Search engine query data deliver insight into the behaviour of individuals who are the smallest possible scale of our economic life. Individuals are submitting several hundred million search engine queries around the world each day. In order to investigate whether Internet search volume is correlated with financial market fluctuations, the largest possible scale of our economic they used search volume data provided by Google. They studied weekly search volume data for various search terms from 2004 to 2010 and they asked the question whether there is a link between search volume data and financial market fluctuations on a weekly time scale. Both collective 'swarm intelligence' of Internet users and the group of financial market participants can be regarded as a complex system of many interacting subunits that react quickly to external changes. They found clear evidence that weekly transaction volumes of Standard & Poor's 500 (S&P 500) companies are correlated with weekly search volume of corresponding company names.Increasing transaction volumes of stocks coincide with an increasing search volume and vice versa. Thus, one can conclude that search volume reflects the present attractiveness of trading a stock. But it seems that neither buying transactions nor selling transactions are preferred when one detects an

increased search volume. Thus, the commonly accepted reasons for financial market movements, news and volume, are clearly linked together because news should be the most likely reason for searching company names in Internet search engines. In addition, they discovered that present price movements seem to influence the search volume of the corresponding company name in the following weeks.

In addition, they applied a method for quantifying complex correlations in time series, which was previously introduced in [294], with which they found a clear tendency that search volume time series and transaction volume time series show recurring patterns. This fact supports their finding that there is a clear link between weekly transaction volumes and weekly search volumes. More important, there is not only a linear dependence but also complex dependencies, which raises hopes that search volume data can contribute to understand financial crises. Uncovering these mechanisms and dependencies, which are useful to understand the formation of financial crises, is of crucial importance as an effective crises observatory could contribute in protecting the stability of financial systems.

In [292] the authors investigated the intriguing possibility of analyzing search query data from Google Trends to provide new insights into the information gathering process that precedes the trading decisions recorded in the stock market data. They analysed changes in Google query volumes for search terms related to finance and they found patterns that may be interpreted as "early warning signs" of stock market moves. Their results illustrate the potential that combining extensive behavioural data sets offers for a better understanding of collective human behaviour. In [91] instead of choosing topics for which search data should be retrieved and investigating whether links exist between the search data and financial market moves, the authors present a method that allows to identify topics for which levels of online interest change before large movements of the Standard & Poor's 500 index (S&P 500). They found that search volumes from Google relating to politics and business can be linked to subsequent stock market moves. One possible explanation for the results is that increases in searches around these topics may constitute early signs of concern about the state of the economy—either of the investors themselves, or as society as a whole.

This demonstration of a link between stock market transaction volume and search volume has also been replicated using Yahoo! data [46], where Bordino et al. showed in particular that query volumes anticipate in many cases peaks of trading by one day or more. A further study analysed data from Twitter and considered the emotions of traders, rather than their information gathering processes, suggesting that changes in the calmness of Twitter messages could be linked to changes in stock market prices [45]. In [247] the authors have shown that the number of clicks on search results stemming from a given country correlates with the amount of investment in that country. Moat et al. in [245] showed that data on views of Wikipedia pages can also be related to market movements, providing evidence that increases in the number of views of financially related pages on Wikipedia could be detected before stock market falls. Evidence has also been provided that Google Trends data can be used to measure the risk of investment in a stock [198]. Finally, in [11] the authors quantify the relationship between movements in financial news and movements in financial markets by exploiting a corpus of six years of financial news from the *Financial Times*. Their results suggest that greater interest in a company in the news is related to the greater interest in the corresponding company in stock markets. The results provide quantitative support for the suggestion that movements in financial markets and movements in financial news are intrinsically interlinked.

3.3 Nowcasting for Humans and Society

Recent years have witnessed a revolution in the social sciences. Mammoth amounts of data are now being generated through society's extensive interactions with technological systems,

automatically documenting collective human behaviour in a previously unimaginable fashion [191, 208, 351]. Analysis of such Big Data opens up new opportunities for a more precise and extensive quantification of real world social phenomena that was difficult to attain using complicated and expensive surveys and laboratory experiments alone.

A particularly fruitful area of research has focused on the analysis of Internet user search queries, as logged by search engines such as Google. Strong links have been found between changes in the information users are seeking online and events in the real world, ranging from reports of flu infections across the USA[143] to the trading volume in the US stock markets[295].

Something particularly exciting about Google data is its global breadth. Never before had been possible to measure what information people are interested in all around the world. But it can be tricky to compare such data between countries, because people in different countries search in different languages. People in France might search in French, for example. People in Germany might search in German. The authors in [293] realised there is one thing which is almost universal between languages. And that's the year in Arabic numerals, so 2014, 2015, 2013, for example. So using data from 2010, they considered all countries which have more than 5 million internet users. And they measured how often they were searching for the next year, 2011, and how often they were searching for the previous year, 2009. They used these Google Trends logs to demonstrate that Internet users from countries with a higher per capita GDP are more likely to search for information about the future than information about the past. They introduced a future orientation index to quantify the degree to which Internet users worldwide seek more information about years in the future than years in the past and they found a striking correlation between the country's GDP and the predisposition of its inhabitants to look forward.

3.3.1 Human behaviour

Making good decisions requires us not only to make our best possible guess about what is going on in the world right now, but also to make our best estimate of what we think might happen in the future. As humans, we try to predict what other people are going to do all the time. We often do this by identifying repeating patterns in behaviour, for example, someone' s tendency to be late, or times at which traffic is particularly bad. Large data sources can help us make better predictions of future behaviour, by allowing computers to spot patterns in large datasets that humans alone might not have identified.

As people increasingly turn to the Internet for news, information, and research purposes, it is tempting to view online activity at the moment in time as a snapshot of the collective consciousness, reflecting the instantaneous interests, concerns, and intentions of the global population. From this perspective, it is a short step to conclude that what people are searching for today is predictive of what they will do in the near future. Consumers contemplating buying a new camera may search to compare models; moviegoers may search to determine the opening date of a new film, or to locate cinemas showing it; and individuals planning a vacation may search for places of interest, to find airline tickets, or to price hotel rooms. If so, it follows that by appropriately aggregating counts of search queries related to retail activity, moviegoing, or travel, one might be able to predict collective behaviour of economic, cultural, or political interest. Determining the nature of behaviour that can be predicted using search, the accuracy of such predictions, and the time scale over which predictions can be usefully made are therefore all questions of interest.

Google makes aggregated data on what people search for online available via its service Google Trends, offering unprecedented insight into people's interests and concerns [246]. Authors in [145] showed that what consumers are searching for online can predict their collective future behavior days or even weeks in advance. They looked at data from Yahoo on what films, songs, and games people had been looking for online to forecast the opening weekend box-office revenue for

feature films, first-month sales of video games, and the rank of songs on the Billboard Hot 100 chart. They showed that in all cases that search counts are highly predictive of future outcomes. Knowing the ranking of a song in the chart in the previous week, provides a very good prediction to what the ranking of that song is going to be in the chart this week. Using this data together with a predictive model, improves the predictions, and it's easier to find out where those films and songs and games are going to be in the next week. In [16] Asur et al. have shown how social media can be utilized to forecast future outcomes. Specifically, using the rate of chatter from almost 3 million tweets from the popular site Twitter, they constructed a linear regression model for predicting box-office revenues of movies in advance of their release.

In [155] Gruhl et al., based on an analysis of around half a million sales rank values for books over a period of four months, and correlating postings in blogs, media, and web pages, were able to draw several interesting conclusions. First, carefully hand-crafted queries produce matching postings whose volume predicts sales ranks. Second, these queries can be automatically generated in many cases. And third, even though sales rank motion might be difficult to predict in general, algorithmic predictors can use online postings to successfully predict spikes in sales rank.

Radinsky et al. [301] wanted to predict top terms that will prominently appear in the future news. They presented a methodology for using patterns of user queries from Google Trends to predict future events. In order to predict whether a term will appear in tomorrow's news, they examined if the terms in today's queries indicated this term in the past and they provided empirical support for the effectiveness of the method by showing its prediction power on news archives. Huang and Penna [100] examined the use of search data for measuring consumer sentiment while the authors in [352, 220] examined retail sales and consumption metrics.

Duncan and Elkan [107], in another interesting study, introduced a nowcasting technique called FDR (false discovery reduction) that combines tractable variable selection with a time-series model trained using a Kalman filter. They applied the method to sales figures provided by the United States census bureau, and to a consumer sentiment index. As side data, the experiments used timeseries from Google Trends of the volumes of search queries. The computational tractability of the FDR method allows variable selection from a large number of potential side variables, which reduces the need for choosing a small set of potentially predictive auxiliary time-series by hand, and thereby also allows for the discovery of unexpected correlations. Variable selection is a particularly important issue in the context of nowcasting because nowcasting relies on observations of correlated side data to make an up-to-date estimate, and in many cases the number of potential auxiliary timeseries to choose among is large. The FDR method outperformed baseline methods when nowcasting sales data from the United States census bureau and consumer sentiment, and has performance that is comparable with the state-of-the-art nowcasting method in [321], while allowing selection from over 250 times as many auxiliary timeseries.

In [195] the authors reported results which demonstrate that data on Facebook likes can be used to predict everything from race, religion and sexual orientation to intelligence and happiness, to a surprising degree of accuracy. The authors collected data at Cambridge University, generated by "MyPersonality", a Facebook app that has been running for over six years, where people were given the opportunity to take certain psychometric tests, like a big five personality test, or an IQ test, and so on. They were then asked if they would be willing to allow the app to take a snapshot of their Facebook profile, and so the researchers could have both the psychometric test results, for the participants, from the volunteered user profiles, they could also have their relationship status, age, gender, and other information. Their Facebook Likes, their friendship network, and so on and so forth. Whatever you can get from Facebook.

The results were surprisingly accurate estimates of race, age, political views, and even sexuality. Accuracy was lowest about 60% when it came to predicting whether users' parents were still together when they were 21. People whose parents divorced before they were 21 tended

to Like statements about relationships. Drug users were ID with about 65% accuracy, smokers with 73%, and drinkers with 70%. Sexual orientation was also easier to distinguish among men. 88%, right there. For women, it was about 75%. Gender, by the way, race, religion, and political views were predicted with high accuracy, as well. For instance, white versus black, 95%. Now here's the key– the more you Like something on Facebook, the easier you are to analyse. The correlation of someone's age to the model's prediction rises steadily as the number of Likes to analyse increases, as well. The same for gender.

Data-streams from social networks like Twitter and Facebook contain also people's opinions, fears and dreams. Scientists used these data from social media to nowcast certain aspects of society. In [205] the authors turned their attention to the issue of public mood, or sentiment - the mood of the nation. They associated each of the basic emotions (fear, joy, anger, sadness) with a list of words, generated by a combination of manual and automatic methods, and successively benchmarked on a test set. This is called citation-sentiment analysis. They found out each of the four key emotions changes over time, in a manner that is partly predictable (or at least interpretable). Joy rises in Christmas, fear in Halloween, and especially negative mood started in October 2010, where massive cuts where announced in UK. Finally, anger growth in the weeks leading up to the summer riots of August 2011, which seems to have started before the riots themselves.

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. Botta et al. in [47] explain how we can use data from mobile phones and Twitter to make surprisingly accurate estimates of the number of people in a given location at a given time. Using a football stadium and an airport as case studies, they presented 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. Their 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.

Another interesting study from Goncalves et al. [148], is looking at limits on how many friendships we can all maintain. There is a theory by a British anthropologist that states that if all of our cognitive capacity is in our brains somewhere, then somehow our brains must impose a limitation on this. The famous number of friendships is 150 and it's called Dunbar's number. The authors used data from Twitter to actually look for the signature and they found it. They looked at several tens of millions of conversations happening on Twitter and they found out that the signature that actually shows over an extended period of time is you do interact more strongly with some people than others and users can entertain a maximum of 100-200 stable relationships.

Finally, Letchford et al. in [211] investigate how the searches of Google users vary across U.S. states with different birth rates and infant mortality rates. They found that users in states with higher birth rates search for more information about pregnancy, while those in states with lower birth rates search for more information about cats. Similarly, they found that users in states with higher infant mortality rates search for more information about credit, loans and diseases.

3.3.2 Natural Disasters and Crises

Much has been said about the potential for big data to help avoid or mitigate the consequences of disasters, ranging from epidemics, to financial crises, to human reactions to natural disasters. Once again, many such hopes stem from perceived opportunities to gain quicker measurements of current human behaviour, and even predict future behaviour. Big data can be proved very useful to policymakers and public administration in order to support and evaluate public policies for natural disasters as well as crime. The city of Chicago, for example, started using text

analytics on Twitter and on 311 (the local emergency number) records to detect and prevent phenomena like rat infestations and to track gang and drug-related violence [63]. The Department of Transportation of Florida is using data collected from the Waze app to adjust traffic signalling in real-time and prevent traffic jams [8].

Natural disasters Recent disasters have drawn attention to the vulnerability of human populations and infrastructure, and the extremely high cost of recovering from the damage they have caused. Examples include the Hurricane Sandy in 2012, the Haiti earthquake of January 2010, the Wenchuan earthquake of May 2009, Hurricane Katrina in September 2005, the Indian Ocean Tsunami of December 2004. In all of these cases impacts were severe, in damage, injury, and loss of life, and were spread over large areas. In all of these cases modern technology has brought reports and images to the almost immediate attention of much of the world's population, and in the Katrina case it was possible for millions around the world to watch the events as they unfolded in near-real time. Images captured from satellites have been used to create damage assessments, and digital maps have been used to direct supplies and to guide the recovery effort, in an increasingly important application of Digital Earth.

These modern mega-events like natural disasters or public rallies are unlike their historic predecessors for one big reason that excites sociologists: Now these famous moments are insanely well-documented by the people who witness them. Take Hurricane Sandy, which unfortunately hit the Northeast coast of the United States in 2012. In addition to all that debris and damage, the storm left behind an extensive trail of tweets, Flickr photos, and even evolving Wikipedia edits. That data is now feeding the growing field of computation social science, more finely informing about how vast groups of people behave during extraordinary situations.

The authors in [291] were wondering whether Flickr can tell us something related to Hurricane Sandy and they retrieved from the photo sharing website Flickr via the API all the photos which were taken in a period of one month in which Hurricane Sandy occurred and they used only the photos which were entitled, or had the tag, relating to Hurricane Sandy. So specifically, they used the photos in which 'Hurricane' 'Sandy' or 'Hurricane Sandy' was used to describe the content of the photo. They used this stream of photos to calculate hourly counts on how many photos were taken by users around the globe relating to Hurricane Sandy. They used the atmospheric air pressure in order to study how this relates to the actual strength of the hurricane and they averaged the air pressure in the US state, New Jersey, where Hurricane Sandy had landed. They found out that the number of photos going up and atmospheric pressure going down is actually reaching the point when the hurricane made landfall where they actually register, in exactly the same hour, the most photos taken and subsequently uploaded to the photo sharing platform and the peak in atmospheric air pressure.

The implication? "We suggest," the researchers write, "that Flickr can be considered as a system of large scale real-time sensors documenting collective human attention." Of course, meteorologists and policy-makers don't need Flickr to tell them the extent and path of a catastrophic storm. But weather data can't necessarily tell us how people are responding to a disaster, or even whether most of them see it coming.

During the first weeks after the Haiti earthquake on January, 2010 there were reports of large population movements out of the severely affected capital, Port-au-Prince (PaP). In October 2010 a cholera outbreak also occurred in Haiti. Rumours about large population movements out of the outbreak area circulated and it was imperative to identify high-risk areas for the emergence of new outbreaks. The lack of reliable data on population movement made coordination and relief prioritization difficult, and it is within this context that the authors in [34] used data from mobile phones to estimate the magnitude and trends of population movements in Haiti following the 2010 earthquake and cholera outbreak and they showed that mobile phone use can successfully provide these estimates rapidly and with potentially high validity. Thus, this tracking method

could be useful in many parts of the world, including those particularly vulnerable to disasters.

In [150] the authors use volunteered geographic information (VGI), a rapidly evolving area of geospatial data and tools and subset of social networking and user-generated web content, to study a series of wildfire events in Santa Barbara area in 2007-2009. The short duration of the fires and the severity of the threat meant that approaches used to inform the public during the fire were no longer adequate and instead numerous postings of VGI, using services such as Flickr and Google maps provided an alternative to official sources.

Civil Unrest and Crime There is enormous interest in inferring features of human behaviour in the real world from potential digital footprints created online - particularly at the collective level, where the sheer volume of online activity may indicate some changing mood within the population regarding a particular topic. Civil unrest is a prime example, involving the spontaneous appearance of large crowds of otherwise unrelated people on the street on a certain day. The authors in [298] show that a simple low-level indicator of civil unrest can be obtained from online data at the aggregate level through Google Trends or similar tools. The study covers countries across Latin America during 2011-2014 in which diverse civil unrest events took place. In each case, they found that the combination of the volume and momentum of searches from Google Trends surrounding pairs of simple keywords, tailored for the specific cultural setting, provide good indicators of periods of civil unrest.

Bowers and Johnson [187, 48], realised that if you have data, not only on where the crimes have occurred, but also when the crimes have occurred, you can possibly make better predictions. Specifically, what they found was that if a house had been burgled, then the chance that another house in that street would be burgled over the next two weeks increased significantly. So if you know that a crime has occurred somewhere recently, then this is reason to believe there might be another crime there in the near future. Some scientists have recently built on this idea, as they realised that the maths that they use to model earthquakes and the aftershocks and to anticipate where they might occur and when, can also be used to better anticipate where and when crimes might occur. These maths are starting to be used in predictive policing software around the world, with a characteristic example of PredPol, a predictive analytics software serving law enforcement.

Johnson et al. [186] analysed data collected on ISIS-related websites involving 100K individual followers and they developed a statistical model aimed at identifying behavioural patterns among online supporters of ISIS and used this information to predict the onset of major violent events. Sudden escalation in the number of ISIS-supporting ad hoc web groups ("aggregates") preceded the onset of violence in a way that would not have been detected by looking at social media references to ISIS alone. The model suggests how the development and evolution of such aggregates can be blocked.

In military planning, it is important to be able to estimate not only the number of fatalities but how often attacks that result in fatalities will take place. The authors in [185] uncovered a simple dynamical pattern that may be used to estimate the escalation rate and timing of fatal attacks. They proved that the time difference between fatal attacks by insurgent groups within individual provinces in both Afghanistan and Iraq, and by terrorist groups operating worldwide, gives a potent indicator of the later pace of lethal activity.

3.3.3 Well-being

Numerous studies on well-being are published every year. In the US, Gallup and Healthways produce a yearly report on the well-being of different cities, states and congressional districts [134], and they maintain a well-being index based on continual polling and survey data [133]. Other countries are also beginning to produce measures of well-being: in 2012, surveys measuring

national well-being and how it relates to both health and where people live were conducted in both the United Kingdom by the Office of National Statistics [32, 302] and in Australia by Fairfax Media and Lateral Economics [204].

While these and other approaches to quantifying the sentiment of a city as a whole rely almost exclusively on survey data, there are now a range of complementary, remote-sensing methods available to researchers. The explosion in the amount and availability of data relating to social media in the past 10 years has driven a rapid increase in the application of data-driven techniques to the social sciences and sentiment analysis of large-scale populations.

A direction of research is that of using social network measures, derived from nation-wide data like phone call records, as proxy of socio-economic indicators of poverty, well-being and progress. The rationale is that social networks shape the fabric of society and form the backbone of social and economic life, but only recently we are having access to big data that expose a nation's network structure and allow to study its social impact. A reference study in this line is [110], where the authors combined a complete records of national communication networks with national census data, aimed at investigating the relation between the structure of social networks and the access to socio-economic opportunities. They found that the diversity of individuals' relationships is strongly correlated with the economic development of communities. Intuitively, the diversity of one's social contacts is believed to be proportional to the access to opportunities: when aggregated over a territory or local community, network diversity is expected to measure the opportunities that such community offers, hence its level of well-being.

Another direction would be that of taking into account mobility of people, again considering big data from either mobile phone call records or GPS trajectories from cars covering entire countries or large regions [270]. In this study the authors found a stronger correlation between well-being and diversity of movement: the greater the diversity of mobile behaviour of the inhabitants of a territory, the greater its prosperity. They found also that the mobility diversity is strongly correlated with several indicators of the socio-economic level of geographic units: a greater mobility entropy within a territory implies a lower value of various poverty indicators, and a larger per capita income. These findings open an interesting perspective to study human behaviour through mobile phone data, as new statistical indicators can be defined to describe and possibly predict (nowcasting) the economic health of a territory.

In [176] the authors proposed a set of well-being indicators, derived from a new supervised technique of web opinion analysis designed to capture several aspects of subjective well-being from on-line discussions and then tries to relate these information with health indicators. The several dimensions of subjective well-being extracted from web conversations are aggregated into a unique index called SWBI (Social Well-Being Index). The technique of opinion analysis used in [176] is the iSA (integrated Sentiment Analysis) algorithm [67, 68] that extracts the sentiment from texts posted on social networks and has already been used to capture instantaneous happiness from social media data [90].

Previous studies have demonstrated the association between the health and behavioural patterns of a person, and the possibility to predict health and well-being conditions using different sources of behavioural information from social media and mobile phones. Detection of emotional states, happiness levels and depressive disorders [44, 96, 217, 59], prediction of physical health conditions [230, 231] and stress levels [43].

Happiness The UN's 2012 World Happiness Report attempts to quantify happiness on a global scale with a 'Gross National Happiness' index which uses data on rural-urban residence and other factors [166].

A lot can be done with data on what information people are looking for online, for example on Google or on Wikipedia. However, there also lots of data on what information people are distributing online, on Twitter and Facebook, for example. A number of studies have tried to use

this data to better understand what affects our happiness, and how happiness spreads. Many of studies tend to base themselves on work which was previously done in linguistics. Now, in this linguistics work, researchers created large lists of words and they gave them to other people to judge, for example, how positive the words are or how negative the words are. Scientists can now take this big list of words and compare the words in the list to words that people post on Twitter and on Facebook. Twitter encourages its 200 million users worldwide to make their posts, commonly known as tweets, publicly available and tagged with the user's location. Tweets but also Facebook posts have another advantage: they tend to be of-the-moment, sent on impulse. They have immediacy; they reflect what the sender is feeling at the time, not what he or she feels looking back, a considered opinion from later.

Several authors have proposed to use tweets as a proxy for happiness, using several ways to map the 140 characters of text in each tweet into an emotional state, which is then aggregated using the geographical and temporal anchors. An early example is Twittermood[3] project, aimed at mapping the mood in the US throughout the day as inferred by hundreds of million tweets observed during several months. The content of each tweet is mapped into an emotional state using the Affective Norms for English Words (ANEW) method, providing a set of normative emotional ratings for a large number of words in the English language [49]. Emotional ratings are then grouped by US state and hour of the day, so that simple infographics reveal evident patterns.

A more comprehensive project is the Hedonometer[4] developed by Dodds and his colleagues at Univ. Vermont [102, 243]. The authors analysed over 46 billion words contained in nearly 4.6 billion expressions (tweets) posted over a 33 month span by over 63 million unique users, and constructed a tunable, real-time, remote-sensing, and non-invasive, text-based hedonometer, i.e., a thermometer of happiness. The project also investigates correlations between tweets across the US and a wide range of emotional, geographic, demographic, and health characteristics, annually surveyed in all 50 US states and around 400 urban populations. On this basis, the authors generate taxonomies of states and cities based on their similarities in word use; estimate the happiness levels of states and cities; correlate highly-resolved demographic characteristics with happiness levels; and connect word choice and message length with urban characteristics such as education levels and obesity rates, provided that better insight is gained on how to master the high self-selection bias inherent into the population of social media users, and how to compensate for it.

Christakis and Fowler studied the role of social networks in the spread of happiness [126], using data on 4739 individuals, collected from the Framingham Heart Study in the period 1983-2003 in Framingham, Massachusetts. The authors defined a suitable measure of happiness and after that, they implemented a social network analysis, finding that happy people tend to be connected to one another and the relationship of *ego* (the person whose behaviour is being analysed) and *alter* (a possible person who is potentially influencing the behaviour of the ego) happiness ranges over three degrees of separation. This means that not only if our friends are happy we are happy as well, but we are also felicitous if friends of our friends and friends of friends of our friends are happy. Authors established the impact of an ego on the happiness of others, depending on the nature of relationship: nearby mutual friends and next door neighbours influence happiness depression scale more than coresident spouses or siblings. These findings indicate the importance of physical proximity and dependence of happiness spreading more on frequent social contacts than on deep social connections.

In an early study, Golder and Macy [146] identified individual-level diurnal and seasonal mood rhythms in cultures across the globe, using data from millions of public Twitter messages. They found, as it would be easy to imagine, that people really like the weekend — people were

[3] http://ccs.neu.edu/home/amislove/twittermood/

[4] http://hedonometer.org

much happier on Saturdays and Sundays. The noticed, though, that this happiness drops off on a Sunday evening as people would prepare to go back to work, and equally, it starts to grow on a Friday. It seems, however, that people really don't like Tuesdays. They also found that individuals awaken in a good mood that deteriorates as the day progresses–which is consistent with the effects of sleep and circadian rhythm–and that seasonal change in baseline positive affect varies with change in daylength.

Another study used this same approach and looked at messages posted on Facebook. This study by Coviello et al. [86] compared what people were posting on Facebook to data that they had on the weather, specifically how much it was raining. They found in this study that, if it rains, people tend to post less happy messages on Facebook. So the rain seems to be affecting people's emotions. Perhaps most interestingly, though, this emotion seems to pass along their network. So for example, if a friend on Facebook is somewhere where it's raining and this affects the emotional content of what they post on Facebook, according to this study, then more likely her friends will also post a sadder message on Facebook as a result, even though it's not raining where they would be.

A further study, by Kramer et al. [196], took this a little bit further. The authors actually work at Facebook, so what they could do was analyse the emotional content of these messages with an algorithm without looking at the messages and then automatically, for a subset of Facebook users, remove the messages which were most strongly positive or which were most strongly negative. What they found was that Facebook users who saw fewer strongly positive messages were then more likely to post more negative messages and fewer positive messages. Equally, the other way around, users who' d seen fewer strongly negative messages would then post more positive messages and fewer negative messages.

Now web companies manipulate their websites to try and affect our behaviour all of the time. This is part of their business. It's just like salespeople try and affect our behaviour to encourage us to buy more products. However, this study caused a bit of an outcry, because on this occasion the researchers were trying to manipulate people' s behaviour in the name of science, without necessarily having got their consent to carry out this experiment. On the other hand, however, the results of this experiment were made publicly available. You didn' t need a university subscription or something to see the results, they were available to everybody — whereas normally, we don' t know what the web companies are doing, and we also don' t know what the outcomes of their investigations are.

These pieces of work seek to identify occurrence patterns for words with pre-defined affect scores at different levels of temporal granularity. Such approaches, with more sophisticated components for emotion recognition in social media content, can be alternatives to public surveys for mood and happiness indicators.

Mental Health The World Health Organisation describes mental health as "the foundation for well-being and effective functioning for an individual and for a community" and highlights the importance of selecting suitable indicators of mental health [167]. One can distinguish between macro-level indicators, which are meant to provide a picture of generic well-being across a large population, usually at national scale, and individual indicators of mental health.

Most of the macro measures typically use statistics from census, administrative and economic sources to measure the social and economic macro-environment as important determinants of mental health (e.g. Human Development Index, Gender Development Index, Human Poverty Indices). As stated in [108], "Measuring better lives has become even more important today, as many of our economies and societies have been stricken by the global financial crisis. Understanding how the lives of people have been affected and designing the best strategies to help those who have suffered the most requires looking well beyond the impact of the crisis on economic production and financial markets. It is thus important to have as accurate as possible

information on how both people' s economic and non-economic well-being have evolved during the crisis".

With the digital revolution of big data, such as mobile phone data and online search data, new sources of measurement of determinants of mental health are rising, acting as proxies to predict people's mental health conditions. Information on human mobility behaviour derived from mobile phones has been shown to be an invaluable source to leverage within the public health domain, both at an aggregated and individual level. In many cases, researchers were able to capture how massive population moves or the daily routines of individuals, and thus to study critical issues for public health like the spread of a disease or the detection of mental health problems such as depression [265].

We have several individual indicators of mental health, that include measures of positive mental health, such as coherence and meaning in life, self-esteem etc. as well as indicators of mental distress, such as negativity, anxiety, depression [167]. These measures can be used by experts or individuals for diagnostic and management purposes, but also in aggregation, for large scale surveys. However, the reliance on self-reporting required to obtain these measures is time consuming and expensive and can only produce sparse data on small populations. Moreover, self-reporting is likely to introduce bias into results. Recent work [300, 206, 59, 276] shows the potential of experience sampling using mobile devices for behavioural studies and clinical care, especially relating to mental health. A variety of longitudinal sensor data from a smart phone as well as location information, obtained passively from the user's phone, can be calibrated against the user's responses to behaviour or emotion related questions. The latter are usually harvested through regular prompts for input provided by a smart phone application.

Researchers have used mobile phones to assess student moods and stress by correlating data from phone sensors, daily probes on student states and termly behavioural surveys [355]. They have identified a strong correlation between automatic sensing data and a broad set of well-being scales. Their work focuses on the calibration of sensor data against self-reported mood without any indication of how these can be combined for prediction purposes. In a related study [356] employ mobile phone use data and survey data from students to predict their GPA score at the end of term. The temporal granularity here is rather coarse, while no textual data is considered and the predictive model does not consider raw data, but rather pre-built classifiers which feed into a regression model.

Jaques et al. in [179] applied a multi-task, multi-kernel approach for predicting students' well-being using survey, mobility, smartphone and physiology data over a one-month period; despite the ability of the prediction model to provide interpretable results by using one kernel per modality, the textual modality was not used while one of the most predictive modalities (survey data) demanded manual effort from the subjects, which is in contrast to our objective. Other studies focusing on stress detection [43] and happiness recognition [44] have also ignored the textual modality or require user input (e.g. personality traits) to be used by their model. Canzian and Musolesi employed well-established and novel metrics to associate human mobility characteristics and depressive states [59]. Their results show that they can identify depressive states by analysing the mobility routines of an individual and thus they can enable a continuous monitoring of her/his mental state by a therapist. In other lines of work, researchers employed mobile phone data in order to predict daily mood states [217] and to diagnose mood changes [236, 154].

Suicide Prevention of suicide has long been complicated by its multidimensional nature. Studies have identified many risk factors for suicide in the general population. Nevertheless, even with standardized measurement of suicide risks, recognition of suicidal individuals is still difficult and preventive intervention is often too late. Persons with suicidal intentions may not seek medical/psychiatric attention, but rather seek information regarding the means of suicide.

In recent years, the Internet has become an essential medium for people seeking health information, yet the Internet includes both helpful and harmful sources. One study demonstrates that suicide-risk individuals who went online for suicide information were likely to visit pro-suicide sites (i.e., websites providing a guide to suicide) [161]. Pro-suicide websites are easily accessed through searches and may be implicated in completed suicides [38, 305]. Unfortunately, how the public perceives and utilizes suicide information is still largely unknown. A better understanding of Internet search behaviour in relation to suicide may help design efficient suicide prevention programs.

Analysis of Internet searches has proven to be useful for predicting infectious disease outbreaks [143, 173, 286, 362], yet limited mental health studies have been reported using Internet search data. Researchers found that Internet searches for depression have seasonality, and the degree of seasonality is latitude-dependent [370]. A pilot study suggests that annually averaged Google search activity for "suicide" correlates to yearly suicide rate data in the United States [238]. Risk factors of suicide tend to interact and potentiate each other. Although Internet search volume data represent a collective phenomenon, the temporal relationship between Internet search trends and suicide data may shed light on sequential acts of suicidal behaviours. This study shows that searches for most medical, familial, and socioeconomic terms preceded suicide deaths and most searches for psychiatric-related terms coincided with suicide data.

The authors in [370] evaluated the association between suicide and Google searches trends for 37 suicide-related terms representing major known risks of suicide, in Taipei City, Taiwan. They used cross correlation analysis to estimate the temporal relationship between suicide and search trends and multiple linear regression analysis to identify significant factors associated with suicide from a pool of search trend data that either coincides or precedes the suicide death. Their results show that a set of suicide-related search terms, the trends of which either temporally coincided or preceded trends of suicide data, were associated with suicide death. Searches for "major depression" and "divorce", for example, accounted for at most, 30.2% of the variance in suicide data. Appropriate filtering and detection of potentially harmful source in keyword-driven search results by search engine providers may be a reasonable strategy to reduce suicide deaths.

Finally, Kristoufek et al. in [199, 200] examined whether online activity measured by Google searches could improve estimates of the number of suicide occurrences in England, and they found that estimates drawing on Google data are significantly better than estimates using previous suicide data alone.

Poverty In European Union (EU), three indicators are used for monitoring progress towards the Europe 2020 poverty and social exclusion reduction target: at-risk-of-poverty, very low work intensity and severe material deprivation. The timeliness of these indicators is crucial for monitoring the effectiveness of policies and the impact of macroeconomic conditions on poverty and income distribution. However, partly due to the complexity of the data collection process, estimates of the number of people at risk of poverty and social exclusion are released by Eurostat with a substantial time lag. Navicke et al. in [254] present a method that can be used to nowcast the current at-risk-of-poverty rate for the European Union (EU) countries based on microdata from a previous period. In [304] the authors present a microsimulation-based methodology for nowcasting changes in the distribution of income over a period for which EU statistics are not yet available, and assess the implications of these changes for the proportion of the population at risk of poverty.

While numerous high-resolution indicators of human welfare are routinely collected for populations in high-income countries, the geographical distribution of poverty in low- and middle-income countries is often uncertain. Additionally, the majority of the existing techniques rely on the availability of census data, which are typically collected every 10 years and often released with a delay of one or more years, making the updating of poverty estimates challenging.

Recently, there are promising signs that novel sources of high-resolution data can provide an accurate and up-to-date indication of living conditions. In particular, recent work illustrates the potential of features derived from remote sensing and geographic information system data [309, 180] and mobile operator call detail records (CDRs) [42, 327, 129, 332]. There has been an attempt to build predictive maps of poverty using both data sources such as aggregate data from mobile operators and geospatial data and the results show that the CDRs data produce accurate, high-resolution estimates in urban areas not possible using geospatial data alone [335].

3.3.4 Epidemiology

In recent years there has been an increasing interest in infectious diseases. Historical examples like the 1918 influenza pandemic, as well as recent threats like the 2002-2003 SARS epidemic, the 2009 A(H1N1) influenza pandemic, the 2012 MERS coronavirus outbreak, and the 2015 extensive Ebola virus (EBOV) in West Africa exemplify that national boundaries or oceans have never prevented infectious diseases to reach distant territories.

While the H1N1 pandemic was not as devastating as it was feared at the beginning of the outbreak in 2009, it gained a special role in the history of epidemics: it was the first pandemic whose course and time evolution was accurately predicted months before the pandemic reached its peak [23]. Balcan et al. were able to predict in advance, several months in advance, when the peak of the epidemic would occur in each country, and even sometimes for large countries in each region of each country and they were able to publish these predictions several months ahead of time.

Tracking the spread of an epidemic disease, like seasonal or pandemic influenza is an important task that can reduce its impact and help authorities plan their response. In particular, early detection and geolocation of an outbreak are important aspects of this monitoring activity. Various methods are routinely employed for this monitoring, such as counting the consultation rates of general practitioners or physicians. For seasonal or pandemic influenza, there are several traditional surveillance systems, including those employed by the U.S. Centres for Disease Control and Prevention (CDC) and the European Influenza Surveillance Scheme (EISS), that rely on both virologic and clinical data and include influenza-like illness (ILI) physician visits. CDC publishes national and regional data from these surveillance systems on a weekly basis, typically with a 1-2 week reporting lag.

Traditional measurements of the number of people who currently have the flu, rely on flu patients visiting their doctor, and their doctor reporting flu cases to a central health authority, such as the Centres for Disease Control and Prevention (CDC) in the US. This data collection process can take a while, and as a result, there is usually a delay of one or two weeks in making the data available. To alleviate this information gap, multiple methods combining climate, demographic, and epidemiological data with mathematical models have been proposed for real-time estimation of flu activity [322, 75, 262, 329, 323, 373].

Big data might be of use in improving health, as it can help us get quicker measurements of disease spreading, and even predict where diseases will spread to next. Rapidly identifying an infectious disease outbreak is critical, both for effective initiation of public health intervention measures and timely alerting of government agencies and the general public. Surveillance capacity for such detection can be costly, and many countries lack the public health infrastructure to identify outbreaks at their earliest stages. Furthermore, there may be economic incentives for countries to not fully disclose the nature and extent of an outbreak [365]. The Internet, however, is revolutionizing how epidemic intelligence is gathered, and it offers solutions to some of these challenges. Freely available Web-based sources of information may allow us to detect disease outbreaks earlier with reduced cost and increased reporting transparency.

In an attempt to provide faster detection, innovative surveillance systems have been created

to monitor indirect signals of influenza activity, such as call volume to telephone triage advice lines [114] and over-the-counter drug sales [232]. In 2006, about 90 million American adults were believed to search online for information about specific diseases or medical problems each year [127], making web search queries a uniquely valuable source of information about health trends. Previous attempts at using online activity for influenza surveillance have counted visitors to certain pages on a U.S. health website [184], and user clicks on a search keyword advertisement in Canada [118].

A vast amount of real-time information about infectious disease outbreaks is found in various forms of Web-based data streams. These range from official public health reporting to informal news coverage to individual accounts in chat rooms and blogs. Because Web-based data sources exist outside traditional reporting channels, they are invaluable to public health agencies that depend on timely information flow across national and subnational borders. These information sources, which can be identified through Internet-based tools, are often capable of detecting the first evidence of an outbreak, especially in areas with a limited capacity for public health surveillance. For example, the World Health Organization' s Global Outbreak Alert and Response Network relies on these data for day-to-day surveillance activities[153, 169]. Revised international health rules have authorized the World Health Organization to act on this information to issue recommendations to prevent the spread of diseases [363].

In order to improve early detection, researchers thought of monitoring health-seeking behaviour in the form of online web search queries, which are submitted by millions of users around the world each day. In particular, traditional measurements of disease spreading can be delayed, and online data such as data from Google or Yahoo! might help reduce such delays. Methods that harness Internet-based information have been proposed, such as Google [143], Yahoo [286, 81], and Baidu [376] Internet searches, Twitter posts [325, 275], Wikipedia article views [239], clinicians' queries [316], and crowdsourced self-reporting mobile apps such as Influenzanet (Europe) [269], Flutracking (Australia) [93], and Flu Near You (United States) [328].

In [81] the authors studied Yahoo! search activity related to the 23 most common cancers in the United States. They found out that the Yahoo! search activity associated with specific cancers correlated with their estimated cancer incidence, estimated cancer mortality, and volume of related news coverage, and that sharp increases in Yahoo! search activity scores from one day to the next appeared to be associated with increases in relevant news coverage. They concluded that media coverage appears to play a powerful role in prompting online searches for cancer information and that Internet search activity offers an innovative tool for passive surveillance of health information seeking behaviour.

Polgreen in [286] showed that search volume for handpicked influenza-related queries was correlated with subsequently reported caseloads over the period 2004-2008, and Hulth et al. in [173] found similar results in a study of search queries submitted on a Swedish medical Web site. An automated procedure for identifying informative queries is described in [143] from Ginsberg, and based on that methodology, *Google Flu Trends* (http://www.google.org/flutrends) was introduced by Google.org in 2008 to provide real-time estimates of flu incidence for more than 25 countries and to help predict outbreaks of flu.

Intuitively, there is a close relationship between how many people search for flu-related topics and how many people actually have flu symptoms. Of course, not every person who searches for "flu" is actually sick, but a pattern emerges when all the flu-related search queries are aggregated. The authors compared the query counts with traditional flu surveillance systems and found that many search queries tend to be popular exactly when flu season is happening. By counting how often we see these search queries, it is possible to estimate how much flu is circulating in different countries and regions around the world. This finding is important because traditional flu surveillance agencies deliver their estimates with a delay of weeks, while the web search-based service delivers basically daily, and timeliness is crucial to enable public health officials and health

professionals to better respond to seasonal epidemics and pandemics.

The authors in [364] used data from Google Flu Trends to study the spread of the pandemic H1N1 influenza in New Zealand during 2009. Influenza nowcasting was attempted also with data stemming from Wikipedia [239] and Twitter [325, 275] instead of search volume. In [202] the authors measured the prevalence of flu-like symptoms in the general UK population, based on the contents of Twitter. Their method is based on the analysis of hundreds of thousands of tweets per day, searching for symptom-related statements, and turning statistical information into a flu-score. They have tested it in the United Kingdom for several weeks during the H1N1 flu pandemic and they compared their flu-score with data from the Health Protection Agency, obtaining on average a statistically significant linear correlation which is greater than 95%. In a later work, the authors [203] used supervised computer-learning algorithms to map textual content to flu levels. Instead of choosing the key words and phrases themselves, they used machine learning algorithms to find out which words in the database of tweets occurred more often at times of elevated levels of flu. They ended up using a version of linear regression to map word frequencies to flu levels, and they obtained very positive results; flu epidemics, they claim, it turns out that can be detected based on Twitter content.

A number of studies have built on these findings related to different epidemics, including [54, 83, 277, 347, 362, 119, 18]. In [69, 12] the authors used Google queries to monitor Dengue epidemics, in [106] Dukic et al. used Google queries to predict hospitalizations for methicillin-resistant Staphylococcus aureus infections and in [264] Ocampo et al. for malaria surveillance. Chunara et al. in [77] used social and news media to validly estimate cholera, Wilson in [362] to monitor listerosis, while in [376] Yuan et al. monitored influenza epidemics in China with search queries from Baidu.

These Web-based data sources not only facilitate early outbreak detection, but also support increasing public awareness of disease outbreaks prior to their formal recognition. Through low-cost and real-time Internet data-mining, combined with openly available and user-friendly technologies, both participation in and access to global disease surveillance are no longer limited to the public health community. The availability of Web-based news media provides an alternative public health information source in under-resourced areas. However, the myriad diverse sources of infectious disease information across the Web are not structured or organized; public health officials, nongovernmental organizations, and concerned citizens must routinely search and synthesize a continually growing number of disparate sources in order to use this information. With the aim of creating an integrated global view of emerging infections based not only on traditional public health datasets but rather on all available information sources, the authors in [55] developed HealthMap, a freely accessible, automated electronic information system for organizing data on outbreaks according to geography, time, and infectious disease agent.

In another interesting line of work, researchers exploit the human mobility to understand epidemics spreading. The everyday movements of humans create the dynamic links that connect populations and enable geographic spread and sustained transmission of infectious diseases. Difficulties in measuring these types of human movements, traditionally estimated using travel surveys, road networks, or small-scale global positioning system (GPS) studies, have long hindered efforts to understand these dynamics. Mobility behaviours have been captured mainly by (i) Call Detail Records (CDRs) or Mobile Network Data generated by providers, and by (ii) smartphone applications. Mobile phone data in the form of call data records (containing information about the location of the mobile phone tower used during a call from a mobile phone) provide one of today's most exciting opportunities to study human mobility [149] and its influence on disease dynamics. Researchers are able to understand massive phenomena such as the spreading of epidemics [188, 218], mass-migration phenomena [10] or the exposure of a population to air-pollution [221]. It is worth noticing that both CDRs and Mobile Network Data are based on the cell towers of a provider, thus resulting in a coarser spatial granularity with

respect to the GPS data. In addition, CDRs suffer from low temporal resolution since they are event-driven (i.e. records are created by a call/SMS trigger), while the Mobile Network Data overcome this since they are network-driven (i.e. records are generated independently of the phone usage) [265].

Analogously, advances in wearable devices have radically improved our capability to track human contacts at high spatial and temporal resolution [64], affording a much more detailed characterization and understanding of social behaviours [282], complementing previous work based on large-scale surveys and self-reported information [251]. Objective measurements of social contact and mobility networks complement self-reported data and pave the way to a more accurate description of infectious disease dynamics. In particular, high quality data are needed to improve parametrization of large-scale computer simulation disease models. The introduction of these models has enabled us to broaden the traditional modeling perspective to encompass large numbers of individuals, rather than population aggregates. Mobile phone data have already been used to create realistic models of human mobility[149], predict the rate of spread of drug resistance [229], assess the prospects of malaria eradication [339], and monitor population movements during the Haiti cholera outbreak in near real-time [34]. Models based on recorded sequences of human contacts can inform the design of containment measures and of targeted immunization strategies [313] and marks an important departure from the static representation of contact networks [336]. Large-scale mobility data can be used to map the worldwide circulation of emerging infectious diseases such as the 2009 H1N1 pandemic [23, 241]. In other words, data are increasingly shaping the development of computer simulations that create in silico experiments hardly feasible in real systems with the goal of providing better scenario analysis for the policy making process and crisis management.

Mobile applications have also started being extensively used in health and well-being domains [231, 66, 265, 111]. Many applications rely on the longitudinal monitoring of an individual outside the clinical settings, leveraging on the multiple data sources provided by the current smartphones. The major advantage of this approach is that the collection of human behavioural routines is completely unobtrusive, fine-grained (e.g. GPS signal or calls/SMSs are collected directly from the user's device) and personalized at the individual level. In addition, the collection of potential symptoms (e.g. fever, cough, etc.) can be self-reported by using an ad-hoc mobile phone application. In this context, Fan et al. [120] proposed a hierarchical probabilistic model to simultaneously predict individuals' physical health by understanding how flu is spread within the proximity interaction networks dynamically captured by mobile phone Bluetooth data. They tested their model both on the MIT Social Evolution [231] data set as well as on the data collected within the iEpi Study [163], where 103 students reported their symptoms and shared their Bluetooth sensor data. In the former, they succeeded in predicting one step ahead the occurrence of the symptoms, while in the latter they revealed the underlying proximity interaction network features related with flu exposure and spreading. Wesolowski et al. in [359] outline the utility of mobile network data for understanding human mobility in the context of the Ebola, and highlight the need to develop protocols for rapid sharing of operator data in response to public health emergencies.

Flu continues to affect millions of people every year, and while it' s still early days for nowcasting and similar tools for understanding its spread, what definitely becomes clear is that is hard to think disease surveillance without data from already existing systems as well as big data deriving from the Internet or mobile devices. Accurate real-time tracking of influenza outbreaks is of major importance as it would help public health officials make timely and meaningful decisions that could save lives.

Chapter 4

Proxies of Human Behaviour

The term Big Data generally refers to any collection of data so large and complex that it becomes difficult to process using traditional data tools. We are under the Big Data microscope: as biologists observe micro-organisms under their microscopes, we can observe our personal actions under the powerful lenses of Big Data. We are in the Big Data era and almost everything we do nowadays requires the use of some digital device: from communications to travels, every human action is digitalized in some form.

As proxies of human life, in this thesis we focus on four types of data which are generated by the digitalization of some events or actions.

- *Mobility data*: car movements are stored in form of GPS points and trajectories.
- *Retail Market data*: retail purchases are stored in form of shopping sessions.
- *Social Netwok data*: data from a social media platform in which people share thoughts and opinions in form of social network.
- *Collaboration Data*: co-authorship data from several publications in form of social network.

Table 4.1: Information about the datasets used in the thesis.

Dataset	Type	Users	Events	Period	Used In
Octo	GPS traces	160,000	9.8 M	May 2011	Forecasting Attractors
Octo	GPS traces	250,000	18.9 M	26/1 - 16/3 2014	Forecasting Attractors
COOP	transactions	1,500,000	300.0 M	2010-2016	Predicting Influenza
FB-LIKE	network	1,899	59,835	23/3 - 26/10 2004	Interaction Prediction
DBLP	network	747,700	5,319,654	2001–2010	Interaction Prediction

Table 4.1 summarizes the characteristics of the datasets used for the analysis described in this thesis. In the following, we provide some details for each dataset.

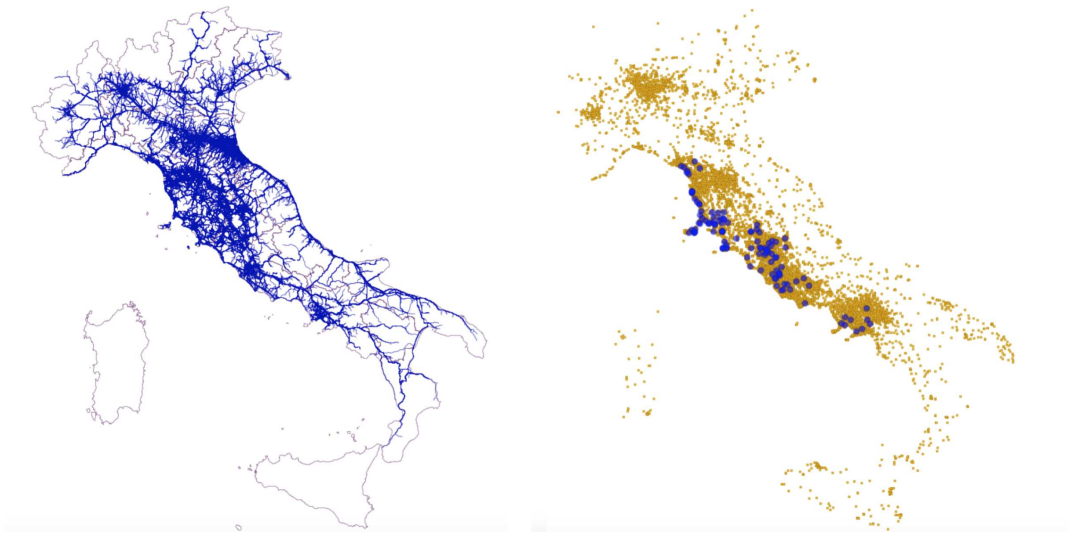
4.1 Human Mobility Data

The Global Positioning System (GPS) is a satellite navigation system that utilizes more than two dozen satellites. It broadcasts precise timing signals by radio to GPS receivers, allowing them to accurately determine their spatio-temporal location (longitude, latitude, timestamp). A GPS receiver calculates its position by precisely timing the signals sent by GPS satellites high above the Earth. Each satellite continually transmits messages which specify the precise positioning information, and the time the massage was transmitted. The receiver computes the distance to each satellite by determining the transit time of each message it receives. These

distances along with the satellites' locations are used to compute the position of the receiver, in form of latitude, longitude and other information like elevation, direction, and speed. GPS-enabled devices provide us with all the required information for trajectory tracking, giving access to accurate, time-stamped locations of each tracked moving point. Nowadays GPS receivers are embedded in many devices we use every day like smartphones and vehicles, allowing to easily track human mobility.

In this thesis we use a massive real-life GPS dataset, the Octo dataset, obtained from tens of thousands private vehicles with on-board GPS receivers. The owners of these cars are subscribers of a car insurance contract, under which the tracked trajectories of each vehicle are periodically sent to a central server for antifraud and anti-theft purposes. This dataset has been donated for research purposes by Octo Telematics Italia S.r.l. The market penetration of this service is variable on the territory, but in general covers around 3% of the total registered vehicles. The Octo dataset stores information of approximately 9.8 million different car travels from 160,000 cars tracked during May 2011 in a geographical area corresponding to Tuscany, central Italy (see Figure 4.1(left)) and 18.9 million different car travels from 250,000 cars tracked during February 2014.

The GPS device automatically turns on when the car starts, and the sequence of GPS points that the device transmits every 30 seconds to the server forms the historical movement of a vehicle. When the vehicle stops no points are logged nor sent. By employing an advanced version of [233], we exploit these stops to split the historical movement into several sub-movements named trajectories, that correspond to the travels performed by the vehicles. Clearly, the vehicle may have stops of different duration, corresponding to different activities. To ignore small stops like gas stations, traffic lights, bring and get activities and so on, we choose a stop duration threshold of 20 minutes: if the time interval between two consecutive GPS points is longer than 20 minutes, the first observation is considered as the end of a trip and the second observation is considered as the start of another trip. We also perform the extraction of the trips by using different stop duration thresholds 5, 10, 15, 20, 30, 40 minutes, without finding significant differences in the sample of short trips and in the statistical analysis we present in the current thesis.



(a) (b)

Figure 4.1: (*Left*) Octo dataset: GPS trajectories passed through central Italy in May 2011. (*Right*) Coop dataset: geographical distribution of shops (blue) and customers (yellow).

4.2 Retail Market Data

With respect to retail market data, i.e., transactional data, the dataset we employ is the Coop dataset. UniCoop is one of the largest Italian retail distribution company. The market chain serves several million customers covering an extensive part of the Italian territory. The 138 stores of the company sell about 570,000 different items. In particular, the stores of the company mainly cover the west coast of central Italy(see Fig. 4.1(right)). The shop distribution is not homogeneous: shops are located in a few Italian regions and therefore, the coverage of these regions is much more significant while customers from other regions usually shop only during vacation periods in these regions. The chain operates three different tiers of shops according to their size: Iper shops are the largest, the Italian equivalent of a US mall; Super are the middle level, a large supermarket; and Small is the smallest shop type, whose size is comparable to a dollar store. The dataset contains retail market data in a time window spanning from 2010 to 2016. The active and recognizable customers in that interval are about 1,500,000. A customer is active if she has purchased something during the data time window, while she is recognizable if the purchase has been made using a membership card. Through the card, customers can get a discount. The company is able to tie each shopping session to the card. In particular, for each shopping session, or basket, the company knows:

- which customer made the purchase;
- all single items composing the basket;
- the time and the day of the shopping session;
- in which shop the transaction happened.

4.3 Social Network Data

In this category fall all those datasets which are built over samples of online social network users (OSN), email exchange logs and call graphs. Due to their nature, in these networks two nodes (users, or actors) are connected by a link if among them has occurred an explicit interaction. Social networks can be differentiate by the reciprocity of the connections among users: in some online services (i.e. Facebook, Skype, Foursquare) explicit links have to be considered as a fully mutual relationship while in others (i.e. Twitter, Google+) users can specify one-way relations (a user A can follow a peer B even if it is not followed back). This differentiation leads to the adoption of two different models to represent OSNs: undirected and directed graphs.

Moreover, among the several type of information that a node/edge can carry a very relevant one is time. Social networks describe mutable realities: nodes and edges can appear and disappear. These changes in both local structures and more complex topologies can be captured and exploited during the analytical process. It is very important to understand that social networks can be seen as proxies for human sociality. The former builds up an overestimation of the real social connection peoples have: OSNs eliminate the costs that are needed to maintain and nurture a friendship (in terms of time, involvements, travel. . .) amplifying the perception of their users sociality.

The Facebook-like Social Network originate from an online community for students at University of California, Irvine. The dataset includes the users that sent or received at least one message (1,899 in total). A total number of 59,835 online messages were set over 20,296 directed ties among these users.

4.4 Collaboration Data

Collaboration networks record who works with whom in a specific setting. Co-authorship of scientific papers, co-working (in an office, movie, firm...), playing in the same team are all characteristic examples of activities of this type. Another example that has been extensively studied by sociologists is the graph on highly-placed people in the corporate world, with an edge joining two if they have served together on the board of directors of the same Fortune 500 company[244]. The on-line world provides new instances as well: the Wikipedia collaboration network (connecting two Wikipedia editors if they've ever edited the same article) [87, 193] and the World-of-Warcraft collaboration network (connecting two W-o-W users if they've ever taken part together in the same raid or other activity) [366].

Sometimes a collaboration network is studied to learn about the specific domain it comes from; for example, sociologists who study the business world have a substantive interest in the relationships among companies at the director level, as expressed via co-membership on boards. On the other hand, while there is a research community that studies the sociological context of scientific research, a broader community of people is interested in scientific co-authorship networks precisely because they form detailed, pre-digested snapshots of a rich form of social interaction that unfolds over a long period of time[258]. By using on-line bibliographic records, one can often track the patterns of collaboration within a field across a century or more, and thereby attempt to extrapolate how the social structure of collaboration may work across a range of harder-to-measure settings as well.

With respect to this type of data, the dataset we employ is deriving from the DBLP database. DBLP listed more than 3.66 million journal articles, conference papers, and other publications on computer science in July 2016, up from about 14,000 in 1995. We use an extraction of the database that contains data for a total of 10 years from 2001 to 2010 for 747,700 authors with 5.32 million papers co-authored.

Chapter 5

Nowcasting and Forecasting model for Epidemic Spreading

5.1 Influenza

With 3 to 5 million infected and 290,000–650,000 killed by influenza worldwide each year, influenza surveillance is of tremendous importance, providing necessary intelligence for hospitals facing staffing decisions, physicians facing active and accurate diagnoses, employers with workers at risk for infection, and public health officials making recommendations for protecting unvaccinated individuals. In addition to seasonal influenza, it exists a new strain of influenza virus against which no prior immunity and that demonstrates human-to-human transmission could result in a pandemic with millions of fatalities [177].

In the case of epidemic and infectious diseases, it is useful to know in advance how many people are going to be infected. Basically it's needed to know where they're going to be, so how many doctors are going to be needed, how many hospital beds, how many antivirals, etc. And it would be necessary to be able to do this ahead of time and to be able to prepare for it in time, in order in the end, to help reduce the size of the epidemic and reduce the consequences for public health. Early detection of disease activity, when followed by a rapid response, can reduce the impact of both seasonal and pandemic influenza [122, 223].

Traditional measurements of the number of people who currently have the flu, rely on flu patients visiting their doctor, and their doctor reporting flu cases to a central health authority, such as the Centres for Disease Control and Prevention (CDC) in the US. This data collection process can take a while, and as a result, there is usually a delay of one or two weeks in making the data available. To alleviate this information gap, multiple methods combining climate, demographic, and epidemiological data with mathematical models have been proposed for real-time estimation of flu activity [322, 75, 262, 329, 323, 373].

And that's where Big Data comes to help. Big data might be of use in improving health. In order to improve early detection, we could monitor health-seeking behaviour in the form of online web search queries, which are submitted by millions of users around the world each day. In particular, traditional measurements of disease spreading can be delayed, and online data such as data from Google might help reduce such delays. Methods that harness Internet-based information have been proposed, such as Google [143], Yahoo [286], and Baidu [376] Internet searches, Twitter posts [325, 275], Wikipedia article views [239], clinicians' queries [316], and crowdsourced self-reporting mobile apps such as Influenzanet (Europe) [269], Flutracking (Australia) [93], and Flu Near You (United States) [328].

5.1.1 The Google Flu Trends paradigm

One of the most fascinating examples of nowcasting using online data, is when the CDC teamed up with engineers at Google to investigate whether the number of flu infections could be estimated from data on how often Internet users had searched for flu-related keywords, such as flu symptoms. The CDC are tasked to come up with numbers for people in the US presenting themselves with influenza-like illness symptoms (ILI). These numbers are aggregated through several layers from a network of doctors based in the US. When the CDC get the final number, then they know how many influenza-like illness cases there were about two weeks ago. Now a question rises, how good can your decision be to maybe close an airport, to make another kind of intervention, to change the real-world system to minimise the impact of a flu wave; how good can this decision be, based on information which is two weeks old?

So when Google engineers and the CDC teamed up, they were looking into the possibility of finding relationships between online data – and in particular, what people search for online on Google – and the number of people presenting themselves to doctors with ILI symptoms. Data on how frequently Internet users are searching for keywords is available to Google with no delay, opening up an opportunity to get much quicker estimates of the number of people currently infected with the flu. What they did was a brute-force analysis, correlating all possible search terms people search for, with the number of flu cases in the US.

What they came up with, in the end, was a list of most-related search terms for this problem for the number of flu cases. So very interestingly, this list contained a lot of flu-related terms, symptoms you might have, and symptoms you might search for online, to find out what you should do if you have the flu. Using this information (which is very readily accessible right after its creation of typing in the search term) Google, together with the CDC, were able to provide improved forecasts of the present, nowcasts, of the number of ILI cases in the US [143]. Based on this finding, Google set up a service, called *Google Flu Trends (GFT)*, which provided figures and estimates of the current level of flu in a number of countries all over the world.

5.1.2 What went wrong?

This would be a success story, if the story stopped here. But it didn't. What people recognised over time was that this Google-based estimate of flu cases, from time to time, wasn't so accurate as people might have hoped. In particular, a couple of years ago, when Google forecasts were compared with the actual number of flu cases in the US, there was a mismatch. Google had dramatically overestimated the number of people having the flu, and people started to question why might this have happened and what the underlying reasons were.

People started criticising GFT [207, 317, 58] and they identified three limitations of the original GFT algorithm. First, it was shown that a static approach, which does not take advantage of newly available CDC's ILI activity reports as the flu season evolves, produced model drift, leading to inaccurate estimates. Second, the idea of aggregating the multiple query terms (the independent variables in the GFT model) into a single variable did not allow for changes in people's Internet search behaviour over time (and thus changes in query terms' abilities to track flu) to be appropriately captured. Third, GFT ignored the intrinsic time series properties, such as seasonality of the historical ILI activity, thus overlooking potentially crucial information that could help produce accurate real-time ILI activity estimates.

Practically when new flu strains (in particular, H1N1) were subject to public discussion, then these Google-based estimates weren't so precise as they could have been. So a number of people came up with the idea that when a lot of people – triggered by maybe media coverage – are looking up online flu and flu-related symptoms, then this obviously doesn't any longer closely match with what is going on in the world right now in terms of actual disease cases.

This is the second part of the Google Flu Trends story, where scientists need to be a little bit

careful about changes in the underlying behaviour of people, which might be triggered by all sorts of external factors, but also in the use of these sources in general. A study by Poletti et al. [285] introduces a model for influenza transmission, accounting for spontaneous behavioural changes on the perceived risk of infection in Italy. During the initial phases of the epidemic the Italian population has been exposed to a massive information campaign on the risks of an emerging influenza pandemic, that contributed to alter the perceived risk as well. Knowing in advance how to account for spontaneous behavioural changes could greatly improve the predictive power of epidemic transmission models and the effectiveness of control strategies.

5.1.3 Later Work

But also, this is not the end because questions have been raised as to whether equally good estimates of current flu levels could be obtained from forecasting models using historic ILI records alone, particularly if it was assumed that CDC measurements were only delayed by one week [145, 266].

Additionally, a number of people came up with the idea of actually using adaptive nowcasting models. Instead of training a model over the entire time period for which they have data available, and using this as a static method, they actually adapt their model. They train it on only short parts of the time series, which also incorporates online behaviour, during this time period only. What Google and others did, they trained one model in the first place and used this model without alterations and modifications, leading to mismatches when the behaviour of people changed. Scientists now by using shorter time windows in which they train their models and produce nowcasts for the next period (in the case of flu, for the next time step) and then retrain their model one week later, they are able to better incorporate search volume and its changing nature.

In [290] the authors build forecasting models which are dynamically retrained over time. Using these models, they wanted to quantify the extent to which relevant search queries aggregated in GFT could have been used to improve estimates of weekly influenza levels in the USA between 2010 and 2013, beyond the forecasts which can be made from historic ILI data. They found that when using GFT data in combination with historic flu levels, the error of in-sample nowcasts can be significantly reduced, compared with a baseline model that uses historic data on flu levels only. At [317] the authors introduced an alternative methodology capable of producing more accurate predictions of influenza activity than GFT, and doing so autonomously with dynamic updating of the model each week as new information about CDC-reported ILI become available, as developed in 2013.

In [372] the authors introduced a model, named ARGO (AutoRegression with Google search data), an autoregressive model with Google search queries as exogenous variables. ARGO utilizes Google search data to estimate current influenza-like illness activity level, and the authors claimed that it manages to outperform all available Google-search–based real-time tracking models for influenza epidemics at the national level of the United States, including GFT. In addition, they provided a theoretical framework that justifies the prevailing use of linear models in the digital disease detection literature by incorporating causality arguments through a hidden Markov model. Google data are also used in [315], where a machine learning-based methodology was introduced, capable of nowcasting and forecasting estimates of influenza activity in the US by leveraging data from multiple data sources including: Google searches, Twitter microblogs, nearly real-time hospital visit records, and data from a participatory surveillance system.

Another interesting line of research, includes data produced by the so-called participatory surveillance systems, who aim at capturing seasonal influenza activity directly from the general population through Internet-based surveys. Participatory surveillance systems for seasonal influenza are currently running in 13 countries worldwide and collect, aggregate and communicate

data in real time during the course of every influenza season. Specifically, the systems that are currently online are: Influenzanet, a network of Web platforms running in eleven European countries [3, 269], FluNearYou in the United States [76, 88, 328] and FluTracking in Australia [60, 61, 94, 95, 273]. Participatory surveillance systems have proven to be accurate and reliable for ILI surveillance, as the detected timing and relative intensities of influenza epidemics are consistent with those reported by sentinel doctors [88, 94, 269, 283, 350]. Furthermore, it has been shown that participatory surveillance systems can also provide relevant information to estimate age-specific influenza attack rates [78, 274, 283, 53] and influenza vaccine effectiveness [61, 112], to assess health care usage [340], and to estimate risk factors for ILI [7]. While the large majority of published studies have focused on forecasting the influenza activity in the United States only, few works have applied their approaches to other countries as well [377, 378] with particular notice to [284], the only influenza nowcasting and forecasting approach in Italy. While the advantages of participatory surveillance, compared to traditional surveillance, include its timeliness, lower costs, and broader reach, it is limited by a lack of control over the characteristics of the population sample. Modelling and simulation can help overcome this limitation as well as provide real-time and long-term forecasting of influenza activity in data-poor parts of the world [53].

Obviously, in an ideal world, however, we'd like also to where that epidemic is going to spread to and how bad an outbreak is going to be. Again, this is something which big data and big models can possibly help us with. As known, for a disease to spread between two people, they have to share some kind of proximity, or some element of the environment around them, like a door handle, for example. So a good model of how a disease is spread is going to have to take into account how many people are near a person that is infected. However, people don't tend to stay still. For example, we all tend to travel to work and back every day and many of us sometimes travel further afield to other countries, for example, for business reasons, perhaps to see family or perhaps simply because you want to go on holiday. Researchers, have looked at how they can build models that capture large scale data on how people travel between countries via air, for example, and also large sets of smaller scale data on how people commute in local areas. The results they have to date suggest that integrating this information on how people travel into models of how diseases spread can give us a better idea of where a disease is going to spread to and also how bad the outbreak is going to be [21].

The everyday movements of humans create the dynamic links that connect populations and enable geographic spread and sustained transmission of infectious diseases. Understanding and modelling human mobility is of paramount importance for public health and is they key to predict the spatial and temporal diffusion of human-transmitted infections. There have been several studies using mobility data for modelling and predicting influenza cases [29, 128, 341, 377]. In [29], the authors studied the impact of people's mobility behaviours for predicting the future presence of flu-like and cold symptoms, using mobility traces from mobile phones and the daily self-reported flu-like and cold symptoms of individuals, while in [128], the authors introduced an agent-based model of epidemic spread using mobility and social network data. In cite [377] the authors presented an online platform that integrates current and historical surveillance data, mobility data, social data mining and several forecast models for real-time seasonal influenza prediction.

In the majority of the works using big data sources, the data have been used as input for statistical models that do not take into account a detailed individual level description of the disease dynamics. In [378] Zhang et al. propose the first seasonal influenza forecast framework based on a stochastic, spatially structured mechanistic model (individual level microsimulation) initialized with geo-localized microblogging data from Twitter. The framework provides forecasts for the United States, Italy and Spain and generates an ensemble of forecasts for the main indicators of the epidemic season: peak time and intensity. A characteristic feature of the

mechanistic modelling approach is in the explicit estimate of key epidemiological parameters relevant for public health decision-making that cannot be achieved with statistical models not considering the disease dynamics.

5.1.4 Take away message

All these lines of work, demonstrate that big data can help us get quicker measurements of disease spreading. Data on how often people search for flu symptoms online can help us estimate how many people have the flu at the moment. Data on how people fly around the world and commute within cities can help create forecasts of where diseases might spread in the future. Although novel digital data streams may suffer from a number of limitations including signal drifts that might affect the reliability of their forecasts, they have increasingly large data volumes, are highly contextual, geo-localized, and allows an unprecedented real-time access to information that can improve forecasting methodologies.

It's important to notice that in order to make a decision in whatever area of society, and specifically in this area of epidemics, emergency measures need to be taken. For example, keeping hospital beds free for people who suffer from very severe symptoms of the flu. However, to make these decisions there is need for real-time information, and the worst case scenario would be to not have any information at all. If there is no information then basically the hospitals are randomly guessing how many people will come through the door in the country next week, or next month, or whenever. Any kind of information in whatever kind of a societal problem, whether this is unemployment rates, people needing to deal with unemployment, benefit schemes, or any kind of demand, is better than having no information.

Any improvement on something, which obviously can come in with an error in many cases, is an improvement. So despite the fact that there could be an error, a numerical error of 10%, this might be a very good thing because maybe the error before introducing the novel method was actually 20%, or maybe 50% and maybe before that, there was no information about this sort of problem at all. Of course it is very important to know where this error was and obviously, there is space for further improvement in the future and in many domain areas. However, to get a better grasp on what is going on at the moment in the society, having enough information and being able to build prediction models and methods, is very important.

Of course, there is still an improvement to be made when incorporating online data and changes in online behaviour must be carefully analysed and used on an ongoing basis. While it's still early days for nowcasting and similar tools for understanding flu spread, another thing that becomes clear is that is hard to think disease surveillance without using data from already existing systems. It's proven for example, that greater value could be obtained by combining GFT with other near real-time health data. By combining GFT and lagged CDC data, as well as dynamically recalibrating GFT, scientists were able to improve on the performance of the prediction.

5.2 Predicting Seasonal Influenza in Italy using Supermarket Retail Records

5.2.1 Introduction

At individual level, each person generates more than 5Gb of data per year. An avalanche of information that, for the most part, consists of transactions (or baskets), i.e., a special kind of categorical data in the form of sets of event data, such as the items purchased in a shopping cart, the web pages visited in a browsing session, the songs listened in a time period, the clinical events in a patient's history. Thanks to these data, human activities are becoming observable,

measurable, quantifiable and, predictable. For instance, this is the case of individual transactional data – retail sales, web sessions, credit card transactions, etc. – where each user produces historical data of his behaviour. There are several works in bibliography studying this type of data but we focus on retail data, describing the shopping behaviour of people.

In [157], Guidotti et al. propose a new clustering method that is designed and improved for finding clusters in the specific context of transactional data. Exploiting these individual clusters and representative transactions they build a personal cart assistant that suggests to the customer the items to put in her shopping list, extending an earlier approach [158] where they introduced a method of recommendation for the customer's set of most necessary items. An interesting aspect of studying retail data, is the adoption rate. The authors in [310] identified the new products, innovations, meant to be a success, as well as the early adopters, individuals that that consistently tend to adopt successful innovations before they reach success. Using these individuals as signals, they manage to achieve high accuracy in the early-stage prediction of successful innovations. Finally, in [279] the authors apply the logic of complex system theory to the retail market. Using large quantities of data extracted from the retail activity of the customer subset of an Italian supermarket chain, they discover that highly ranked customers, with more sophisticated needs, tend to buy niche products, i.e., low-ranked products; on the other hand, low-ranked, low purchase volume customers tend buy only high-ranked product, very popular products that everyone buys. Additionally, they propose a simple marketing application useful for targeted advertising as they manage to classify the likelihood of a customer-product connection, spotting with a high accuracy the smallest customer set that is likely to start buying a given product. Finally they discover that the mobility of customers is more predictable than previously thought, by predicting part of the variance in their movements just by knowing what types of products are sold in different supermarkets of an area.

In [156] Guidotti et al. use a quantification of the average sophistication of satisfied needs of a population as an alternative to GDP, as it can be calculated more easily than GDP and it proves to be a very promising predictor of the GDP value, anticipating its estimation by six months.

In this study, our aim is to use supermarket retail data as a proxy for predicting seasonal influenza. In other words, we will use an external signal in order to predict the curve of an internal time series. When the flu season arrives, people are starting to get sick. Getting sick affects their everyday life and behaviour. This change in behaviour propagates in their purchases in the supermarket. They will buy products that will reflect the fact that they are sick.

We collect this information studying the purchases of each week in order to detect the more correlated products with the influenza adoption trend. We identify the customers that buy them during the influenza peak, and through them we identify the *sentinels*, frequent baskets of such customers during the peak. We use the weekly values, time series, of the next season for these sentinels as a proxy for the next flu season. Finally, using a regression model we assume that this week's influenza depends on the influenza in previous weeks as well as the proxy in the current week and previous weeks. Our approach produces nowcasting and forecasting values for up to 4 weeks ahead.

5.2.2 Proposed Approach

As mentioned before, predicting influenza is not an easy task. We built a data-driven approach, exploiting information extracted from external timeseries, those of retail market, using data mining and machine learning techniques, in order to be able to succeed and provide valid nowcasts and forecasts. In this section we present the approach we designed for this study (graphically represented in Fig. 5.1) with the following steps:

Step 1: For each product p we construct the timeseries S_p of the volume of purchases in a weekly

level. We filter the products who are more correlated with the influenza adoption trend S_i calculating the Pearson correlation measure between $\{S_p, S_i\}$.

Step 2: For each of the filtered products we identify the customers c that bought them during the influenza peak $[T-2, T, T+2]$.

Step 3: For these users, we obtain all their purchases during the same period and we keep the baskets of products bought together $b_{c1_1}, b_{c1_2}, ..., b_{c1_n}, b_{c2_1}, ...$ We apply the Apriori algorithm to identify the frequent baskets and we reconstruct the composite timeseries for these baskets-sentinels for the next season S_b.

Step 4: We feed these timeseries into the regression model in order to predict the influenza adoption trend for the next season.

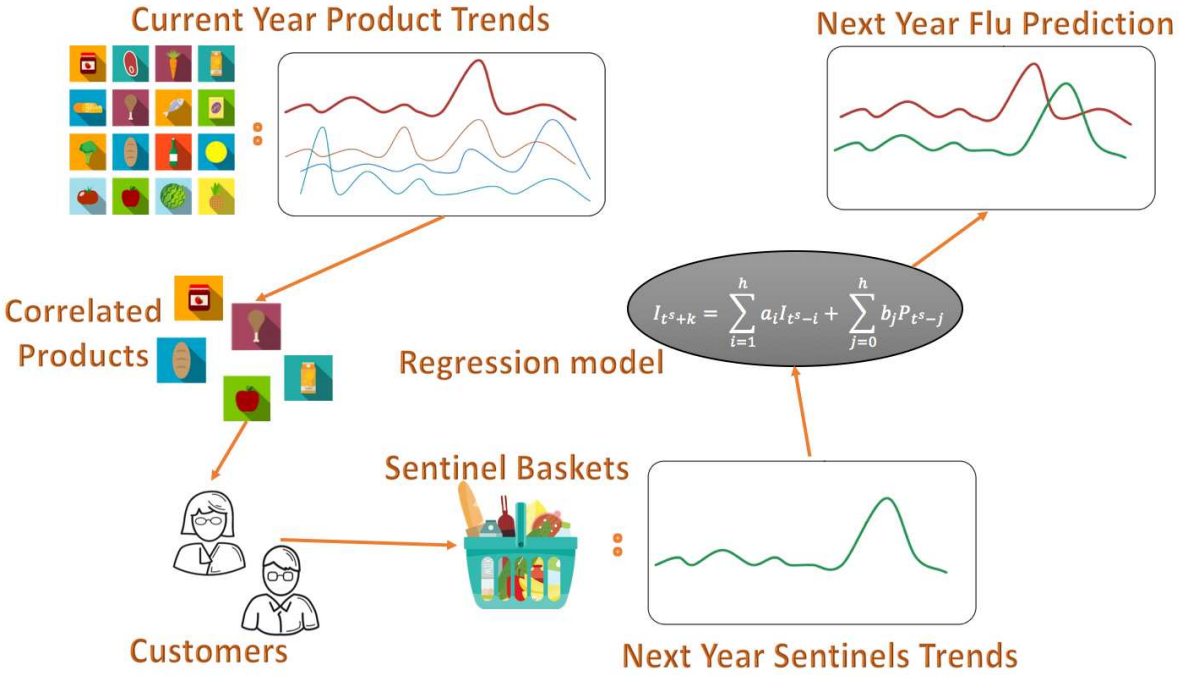


Figure 5.1: Proposed approach workflow. We consider the timeseries of all the products and we filter the most correlated ones *(Step 1)*. Then for each product we identify the customers that bought it during the influenza peak *(Step 2)*. Starting from these customers, we reconstruct all their baskets during the same period. We obtain the composite timeseries for these baskets for the values of the next season and these will be our sentinels *(Step 3)*. Finally, we feed these signals into the regression model and we produce the final predictions *(Step 4)*

Step 1: Correlated Products

We need to define the time granularity of our observation period for the retail data. We choose to use a weekly aggregation mainly because influenza reports are on a weekly base. We prepare the retail data in order to correspond into the weekly reports of influenza and we work on a 'Subcategory' level, referred from now on as *product p* (See Section 5.2.3). We report the weekly sales for each of the products for all the weeks of interest (42nd week of the year until the last week of April of the following year) producing the final retail time series S_p.

It is important to notice that even working in an aggregated level in the retail hierarchy, our timeseries are 2,665. So it is really important to filter out the products that are not correlated

with the influenza adoption trend S_i, so we can work mainly with products that have a similar adoption trend. We choose to use the Pearson Correlation, as it is one of the most commonly used correlation measures.

In statistics, the *Pearson correlation coefficient*, also referred to as Pearson's r, is a measure of the linear correlation between two variables X and Y. It has a value between +1 and -1, where 1 is total positive linear correlation, 0 is no linear correlation, and -1 is total negative linear correlation. It is widely used in the sciences and it was developed by Karl Pearson from a related idea introduced by Francis Galton in the 1880s.

Using Pearson correlation coefficient we calculate the correlation r between each product's time series S_p with the influenza time series S_i and we filter out the time series with a negative correlation in order to identify the products that have adoption trend similar to the influenza trend, the most *correlated products*.

Step 2: Sentinel Customers

We are interested in studying human behaviour mainly during the influenza peak. We identify the influenza peak week and we define the *period of interest* $[T - \delta, T, T + \delta]$ with $[\delta] = 2$, considering 2 weeks before and after the week of the peak.

Using the sentinel products we identified in the previous step, we trace their sales during the period of interest and we identify the customers that bought them, through the receipts matching each customer with her corresponding purchases. These customers become our *sentinel customers*. We are interested in the purchases of these specific customers, since those individuals would have higher possibility to be either infected or close to an infected individual. We have to notice that customers are using loyalty cards, linking them with their purchases throughout the whole period of interest, and that a loyalty card normally represents the whole household, with the probability of more than a person per household.

Step 3: Sentinel Composite Baskets

In this step, we are working backwards. Using the *sentinel customers*, we track all their purchases during the period of interest, through their receipts, and we obtain their corresponding baskets, products bought together under the same receipt. We obtain the baskets for each of these *sentinel customers* and we create a pool of baskets, losing the information of who bought what. We are only interested from now on, on the information contained in the products bought together, and in the patterns we can extract through customers behaving with a similar way and buying products at the supermarket in a similar way. For that reason, we use the *Apriori algorithm*.

Apriori algorithm, is an algorithm for frequent item set mining and association rule learning over transactional databases. The algorithm uses a bottom up approach as it proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. The algorithm terminates when no further successful extensions are found. It uses breadth-first search and a Hash tree structure to count candidate item sets efficiently. It generates candidate item sets of length k from item sets of length *k-1*. Then it prunes the candidates which have an infrequent sub pattern. According to the downward closure lemma, the candidate set contains all frequent k-length item sets. After that, it scans the transaction database to determine frequent item sets among the candidates.

Using Apriori algorithm, we extract the most frequent baskets in our pool. For every product composing the most frequent baskets we obtain the corresponding timeseries and then for each basket we create a cumulative value of all the products that belong to it and we create the corresponding composite timeseries. We use the measure presented in Step 1 and we calculate

the Pearson correlation between the influenza timeseries S_i and the timeseries for each of the baskets S_b and we keep the most correlated baskets, the *sentinel composite baskets*.

We move on to the next year and we extract the timeseries for each of these *sentinel composite baskets* that we had obtained from the past year. We achieve this by repeating the procedure mentioned before, as we create a cumulative value of all the products that belong to each basket-sentinel.

Step 4: Regression Model

Inspired by Yang, Shihao et al[372], we assume that the influenza depends both on the history of influenza itself, using the autoregression model, as well as the *sentinel composite baskets* from the retail data, the proxy.

We assume that the kth week ahead influenza depends on the influenza in h previous weeks as well as the proxy in the current week and h previous weeks:

$$I_{t^s+k} = \sum_{i=1}^{h} a_i I_{t^s-i} + \sum_{j=0}^{h} b_j P_{t^s-j} \tag{5.1}$$

where t^s denotes the week t in season s, k denotes the kth week ahead, I_t^s denotes the influenza in week t^s, P_t^s denotes the proxy in week t^s, and a_i and b_i are coefficients.

The coefficients are solved by Support Vector Regression (SVR) model with kernel 'rbf' (see section 5.2.2). For each forecasting week t^s and forecasting target k, the SVR model is trained by the data starting from the first week in the previous season $s-1$ to the last week t^s-1. The h in equation 5.1 as well as the parameters in SVR are selected by the LOSO cross-validation.

AutoRegression Model The AutoRegressive model (AR) assumes that the kth week ahead influenza depends only on the influenza in h previous weeks:

$$I_{t^s+k} = \sum_{i=1}^{h} a_i I_{t^s-i} \tag{5.2}$$

where t^s denotes the week t in season s, k denotes the kth week ahead, I_t^s denotes the influenza in week t^s, and a_i are coefficients.

Similar to the Regression model, the coefficients are solved by SVR with kernel 'rbf', with training data in the last and current seasons, with the same LOSO cross-validation.

Leave-one-season-out Cross Validation In all the regression models, the possible adjustable parameters are chosen by leave-one-season-out (LOSO) cross-validation with grid search introduced by EL Ray et al [117]. The LOSOCV use one season as the validation set and the remaining seasons as the training set, and repeat validating on each season.

Support Vector Regression Support Vector Regression (SVR) solves generalized linear relation between a vector input x and a scalar output $f(x)$:

$$f(x) = K(x)^T \beta \tag{5.3}$$

where K is the kernel function (discussed later), and β is the coefficient vector.

For example, in the case of equation 5.1, the input is $x_t^s = (I_{t^s-h}, I_{t^s-(h-1)}, ..., I_{t^s-1}, P_{t^s-h}, P_{t^s-(h-1)}, ..., P_{t^s})$, the output is $f(x_t^s) = I_{t^s+k}$, and the coefficient is $\beta = (a_h, a_{h-1}, ..., a_1, b_h, b_{h-1}, ..., b_0)$.

$$H(\beta, \beta_0) = C \sum_{t=1}^{N} V_\epsilon(y_t, f(x_t)) + \frac{1}{2}||\beta||^2 \tag{5.4}$$

$$V_\epsilon(r) = \begin{cases} 0, & \text{if } |r| < \epsilon, \\ |r| - \epsilon, & \text{otherwise.} \end{cases} \tag{5.5}$$

where t_0 and t_N are the first and last points in the training timeline, y_t is the observed value at week t, C and ϵ are adjustable parameters. $V_\epsilon(r)$ denotes the error from the fitting model to the observed values. As in Eq. 5.5, error of fitting is zero when $r < \epsilon$, indicating that SVR allows the existing of margins within which the training data points doesn't contribute to the loss function. While the remaining data points are called the support vectors.

SVR is parameters sensitive. Obviously, when ϵ is too large, the margin encloses all the data points, leading to no support vector in the model. While when ϵ is too small, almost every data point becomes support vector, causing the overfitting issue. As for the C, when too large, the loss function ignores the $||\beta||^2$ term, causing overfitting; when too small, the error from support vectors will be ignored, leading to no complexity captured by the model from the data.

Another key in SVR model is the kernel function $K(x)$. The kernel function could be in any form, which is the reason that SVR solves "generalized" linear relation. The most commonly used kernel, radial basis function ('rbf'), has a similar form with the Gaussian distribution [164]:

$$K(x, x_i) = exp(-\gamma||x - x_i||^2) \tag{5.6}$$

where γ is an adjustable parameter. When γ is too large, the influence of the support vector x_i decays too fast with the distance to x. In this case, each x will be "supported" by a small number of support vector, leading to a low accuracy of prediction. While γ too small, each support vector has equal influence on x, and no complexity will be captured by the model.

In implementation, C, ϵ and γ are determined by GridSearch and LOSO cross validation.

Sensitivity Test on the Number of Proxies In each season, the top proxies (sentinel composite baskets) are pre-calculated and pre-selected. Instead of using only one proxy as in equation 5.1, we extend the model to be:

$$I_{t^s+k} = \sum_{i=1}^{h} a_i I_{t^s-i} + \sum_{n=1}^{N_P} \sum_{j=0}^{h} b_j^n P_{t^s-j}^n \tag{5.7}$$

where the P^n is the top nth proxy, and N_P is the total number of proxies used in this model.

5.2.3 Data

Influenza Dataset

Since we want to investigate the correlation between retail market data and influenza data, we need a reliable data source for influenza. In developed and developing countries, national syndromic (i.e. based on observed symptoms) surveillance systems monitor levels of influenza-like illness (ILI) cases among the general population by gathering information from physicians, known as sentinel doctors, who record the number of people seeking medical attention and presenting ILI symptoms. Influenza activity in Italy is officially monitored by the Italian National Institute of Health, "Istituto Superiore di Sanita" (ISS) and the Interuniversity Research Centre on Influenza (Ciri), through a system called Influnet. The Influnet system collects data from a network of about 900 sentinel General Practitioners (GPs) and paediatricians and compiles a weekly report in which the national and regional incidence rates by age group are published during the winter season, generally from week 42 to the last week of April of the following year (around week 17). The system covers about 2% of the Italian population. Doctors who participate in the monitoring are required to identify and write down daily, on their own register, each new case of influenza.

Figure 5.2: Data Model of the Data Warehouse.

Each week, the aggregate number of cases seen by any physician (divided by age groups and by risk category) is transmitted to the relevant Reference Center. The ISS processes the data at national level and produces a weekly report. Data are published with at least one-week lag and typically new reports provide a first estimate of the weekly ILI incidence which is then updated in the following weeks as more data from sentinel GPs are recorded.

We collected the Influnet reports for five influenza seasons, from 2011-2012 to 2015-2016, from week 47 to week 17. The reports are available at the following URLs: `http://www.iss.it/flue/index.php?lang=1&anno=2016&tipo=13`.

Retail Dataset

Our analysis is based on real world data about customer behaviour. The dataset we used is the retail market data describing the purchases of the customers of COOP, one of the largest Italian retail distribution companies. The conceptual data model of the data warehouse is depicted in Figure 5.2. We analysed a dataset of 30M shopping sessions occurred in Leghorn province over 2010-2016, corresponding to about 150,000 active and recognizable customers. A customer is active if she has purchased something during the data time window, while she is recognizable if the purchase has been made using a loyalty card. Customers are provided with a loyalty card which allows to link different shopping sessions, and therefore reconstruct their personal shopping history. The customers of this supermarket with a card are very engaged in the shop itself: the supermarket is in fact a cooperative and whoever has the card is considered a member. This makes the data more valuable as the customers with a card have very high incentives to buy whatever they can in this supermarket, making it the primary (and sometimes only) source of the products they buy. In fact, a study by Bocconi showed that COOP is able to score among the highest in the metrics of customer fidelity [1]. The 138 stores of the company cover the whole west coast of Italy, selling 571,092 different items. For each customer, we have N$\sim$150 baskets, D$\sim$100 different items, and an average basket length T of $\sim$8 items.

An important dimension of the data warehouse is Marketing, representing the classification of products: it is organized as a tree and it represents a hierarchy built on the product typologies, designed by marketing experts of the company. The top level of this hierarchy is called 'Area' that splits the products into three fundamental categories: 'Food', 'No Food' and 'Other' that refers to medical products. The bottom level of the hierarchy, the one that contains the leaves of the tree, is called 'Segment' and it contains 7,656 values. Hence, for each item contained in

[1]The news of the study, in Italian, can be found at `http://www.viasarfatti25.unibocconi.it/notizia.php?idArt=6527`. The PI of the study can be reached at isabella.soscia@skema.edu.

the dataset, there is an entry assigning it to the right path of the hierarchy tree.

The marketing hierarchy goes like that: i) Area (3 values), ii) Macro sector (4 values), iii) Sector (13 values), iv) Department (76 values), v) Category (529 values), vi) Subcategory (2,665 values), vii) Segment (7,656 values), viii) Item (571,092 values).

Given that the dataset contains more that many customers and items, to build a matrix 'customers × items' would generate an enormous cells matrix, that is redundant for our purposes. Hence we need to reduce both dimensions (customers and items). There are two main criteria to select the customers: on the basis of their purchase behaviour (e.g. excluding from the analysis all the people that did not purchased at least a total number x items) or geographically (e.g. considering just the customers of an area). We decided to apply the latter filter, since we do not want to exclude any customer behaviour apriori. For that reason, we only use data from the Leghorn province, one of the best represented areas of Italy, with regards to number of shops in the area, as well as number of loyal users.

Another issue regards the cardinality of products. There is a conceptual problem in using the level of detail of 'Item': the granularity is too fine, making the analysis impractical as it would consider a very low detail level. The distinction between different packages of the same product, e.g. different sizes of bottles containing the same liquid, is not of interest in our study ('Item' level). Equally, the distinction between products of different brands, e.g. milk from company A or B, is not of interest of our study ('Segment' level). A natural way to solve this problem is to use the marketing hierarchy of the products, substituting the item with its marketing 'Subcategory' value. In this way, we reduce the cardinality of the dimension of the product (from 571,092 to 2,665), aggregating logically equivalent products. Throughout our study, we will refer to those subcategories as products.

Figure 5.3 shows that the observed customers cover the entire territory of Italy. However, the shop distribution is not homogeneous. Shops are located in a few Italian regions. Therefore, the coverage of these regions is much more significant, while customers from other regions usually shop only during vacation periods in these regions. Our analysis is performed on national influenza data, because regional influenza data are not reliable enough. However, the national data doesn't suffer from the same limitation. We have to assume though, that influenza spreads in an homogeneous way all over the country, which for a relatively small country as Italy could be true.

5.2.4 Experiments and Results

Baseline predictions As a reference, we produced influenza predictions using only historical reported influenza. We achieved this via the autoregressive model presented in section 5.2.2. These predictions were used to assess the added value provided by our digital disease detection systems information.

According to the value of k, a distinction can be made between nowcasting ($k = 0$), i.e. inferring the present influenza incidence value and forecasting ($k > 0$), i.e. predicting the influenza incidence value in k weeks. Influenza predictions generated with the baseline model are then contrasted with those produced by the regression model that integrate retail data, to assess their added value.

Evaluation metrics We report 3 evaluation metrics to compare the performance of the regression model: *Pearson correlation*, *maximum absolute percent error (MAPE)* and *mean percentage error (MPE)*. The definitions are given below. Our notation is as follows: y_i denotes the observed value of the influenza at time t_i, x_i denotes the predicted value by the model at time t_i, $\bar{y}$ denotes the mean or average of the values y_i and similarly $\bar{x}$ denotes the mean or average of the values x_i.

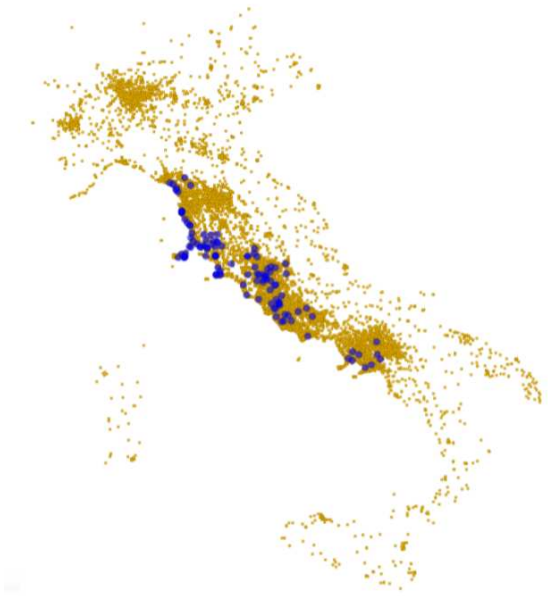


Figure 5.3: Coop Shop Distribution

Pearson Correlation, a measure of the linear dependence between two variables during a time period $[t_1, t_n]$, is defined as:

$$r = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})}} \tag{5.8}$$

Maximum Absolute Percent Error (MAPE), a measure of the magnitude of the maximum percent absolute difference between predicted and true values, is defined as:

$$MAPE = (\max_i \frac{|y_i - x_i|}{y_i}) \times 100 \tag{5.9}$$

Maximum Percent Error (MPE), a measure of the magnitude of the maximum percent difference between predicted and true values, is defined as:

$$MPE = (\max_i \frac{y_i - x_i}{y_i}) \times 100 \tag{5.10}$$

Predictions Predictions of influenza activity were produced using the regression model for the 2013/14 influenza season. The forecasts were made assuming we had access only to the historical influenza reports and retail data up to the previous week of estimation. We compared the model's predictions with the baseline: the reported influenza activity level, published typically with 1- or 2-wk delay, by calculating the values of the Pearson correlation coefficient and the mean absolute percentage error (MAPE), mentioned above. It's important to notice that we use the lower threshold of 0.02 for influenza in order to filter out the weeks with a very low signal. As a result, the predictions start with week 51 and end with week 11. We report the predictions for up to 5 most correlated sentinel baskets, as from further experimentation it's proven that the performance remains stable for a larger number of proxies.

Table 5.1 presents the performance of the regression model for four time horizons 1-, 2-, 3- and 4-wlp.

As expected, we notice that the values of Pearson correlation decrease and the values of MAPE increase as we make forecasts with a larger lead. Indeed, one-week nowcasts predict with

	Pearson correlation				MAPE			
	1-wlp	2-wlp	3-wlp	4-wlp	1-wlp	2-wlp	3-wlp	4-wlp
autoregr	*0.99*	*0.94*	*0.91*	*0.85*	*9.54*	*18.38*	*28.47*	*28.02*
top1	0.98	0.97	0.95	0.91	7.37	8.63	9.70	19.55
top2	0.99	0.98	0.95	0.86	9.42	9.99	13.94	23.34
top3	0.99	0.98	0.94	0.90	8.55	10.16	15.59	21.05
top4	0.98	0.97	0.94	0.88	11.23	9.66	15.19	18.06
top5	0.97	0.97	0.95	0.86	11.49	9.32	13.52	19.93

In all cases p-value < 0.01.

Table 5.1: Pearson correlations and MAPE from comparison of forecasts between the regression model and the baseline for season 2013/14

a percent error of about 7 to 11%, two-week forecasts show percent errors of about 10%, three-week forecasts show percent errors of about 10 to 15%, and finally the four-week forecasts show percent error of about 18 to 23%. Table 5.1 quantitatively shows the added value of using retail data over a simple historical autoregressive approach. Our forecast estimates in all time horizons (up to four weeks ahead) show at least comparable accuracy to "real-time" estimates obtained with a purely autoregressive model, and with two-week, three-week and four-week forecasts ahead the errors are almost cut in half.

Figure 5.4 displays the predictions against the reported influenza activity level for the four time horizons 1-, 2-, 3- and 4-wlp. The top panel of the figure graphically show the influenza activity level along with the predictions of the regression model with the sentinel baskets, as a function of time. The MAPE and MPE errors are displayed in the bottom panel of the figure.

Overall predictions track very accurately the influenza activity level, as shown in the top panel of the figure. Close inspection shows that, the one-week ahead predictions from the regression model with the sentinel baskets and the baseline are very similar, with small errors. For two-weeks, three-weeks and four-weeks ahead it's becoming more and more clear that the autoregression model predictions' are failing to track the influenza activity level either because of overshooting or shifting the peak. On the contrary our predictions seems to track rather closely the influenza activity level with a small overshooting in some cases.

5.2.5 Conclusions

We considered as our reference a baseline autoregressive model that considers only influenza data from the traditional surveillance Influnet. The results presented demonstrate the superiority of the approach, compared with the autoregression baseline. Specifically, we show that our methodology can produce predictions one week ahead of influenza with comparable accuracy. We also show that our forecasts (up to three weeks into the future) always improve predictions produced with the baseline autoregressive model, thus proving quantitatively the added value of incorporating retail data in our flu prediction model.

Our methodology allows us to use retail data, reflecting human's behaviour with regards to influenza and historical influenza information that complement one another. This fact suggests that combining information from multiple sources is advantageous and this is the case not only for real-time predictions but also for the one, two and three week forecasts ahead. The value of these results is even higher given the fact that they were produced with national level data of influenza, as the regional would be of low-quality. It is highly likely that our methodology would lead to even more accurate results if we could use results of a finer aggregation level.

While the results presented here are for influenza-like illnesses at the national level within Italy, our approach shows promise to be easily extended to accurately track not only influenza in other countries where similar data sources may be available but also other infectious diseases.

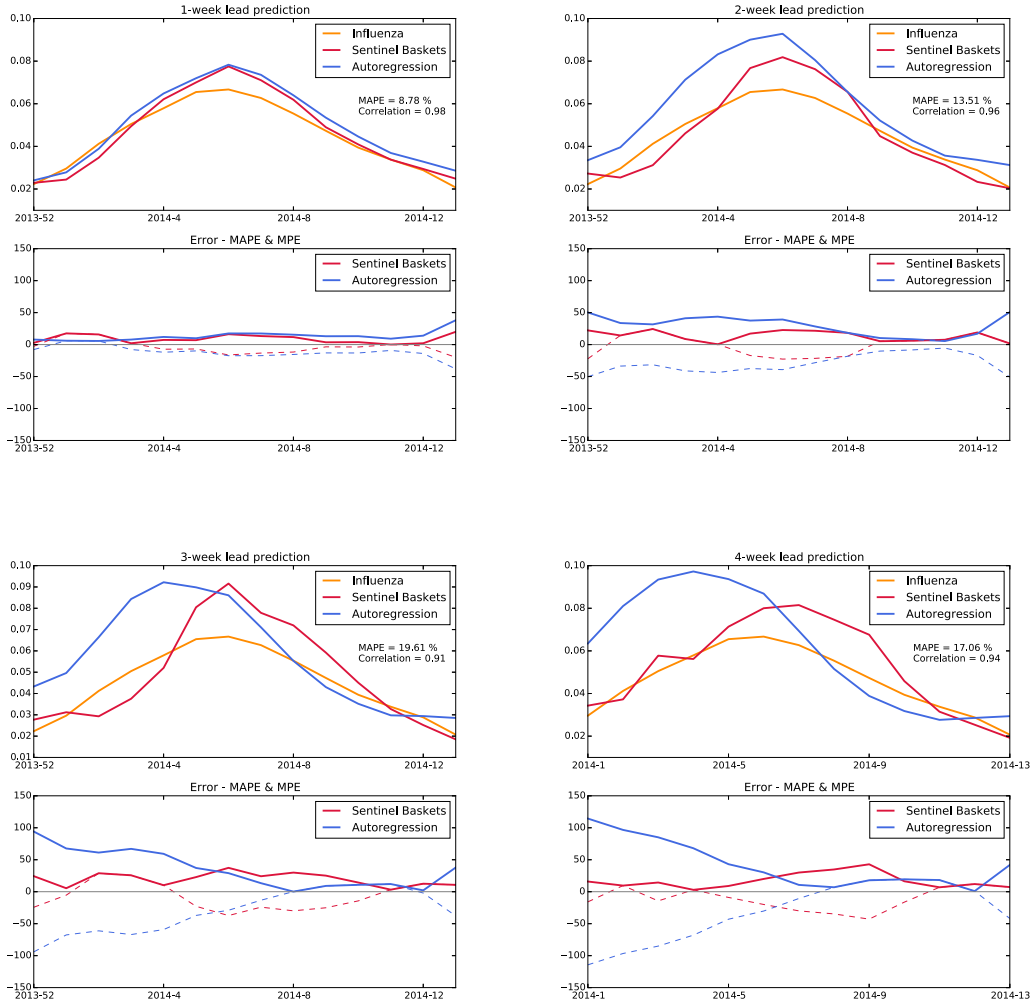


Figure 5.4: Predictions for 1 to 4 weeks ahead.

Chapter 6

Forecasting models for Socio-Economic Attractors

6.1 Human Mobility

We live in an era in which understanding individual mobility patterns is of fundamental importance for epidemic preventions and urban planning. Human movements are inherently massive, dynamical, and complex. Aided by modern transportation technologies, we can now travel to any place on the globe in just one day or two. On the other hand, while the mobility of our fellow species is mainly governed by mating needs and food resources, human mobility is fundamentally driven by ourselves, from job-imposed restrictions and family related programs to involvement in routine and social activities. Therefore, quantifying the regularities and singularities behind human movements is of major importance.

Even more, human mobility is a key driving force behind various spatio-temporal phenomena on geographical scales. In the past, approaches to human interactions and mobility have mostly relied on census and survey data, which were often incomplete and/or limited to a specific context. Despite advances in the study of human transport [113, 97], this lack of data has hindered the construction of a general framework of human mobility based on dynamical principles at the individual level with the ability to bridge spatial scales, from small communities to large urban areas and countries, in a bottom-up perspective. In the last decades, the wide-spread availability of geo-localization devices and the technologies to store and analyse the data they generate had a huge impact on the research areas dealing with spatial and spatio-temporal data. These large-scale datasets generated by various domains of modern technologies, ranging from registration of dollar bills to mobile phone services and GPS devices, from vehicles for instance, quickly became a focus for researchers from several disciplines, especially in the domain of data mining, as they represent society-wide proxies for human mobile activities. These mobility data provide a new powerful social microscope, which may help us understand human mobility, and discover the hidden patterns and models that characterize the trajectories humans follow during their daily activity.

Analysts reason about high-level concepts, such as systematic versus occasional movement behaviour, purpose of a trip, and commuter/resident/visitor patterns. Accordingly, the mainstream analytical tools of transportation engineering, such as origin/destination matrices, are based on semantically rich data collected by means of field surveys and interviews. It is therefore obvious that big, yet raw, mobility data can be used to overcome the limits of surveys, namely their high cost, infrequent periodicity, quick obsolescence, incompleteness, and inaccuracy. On the other hand, automatically sensed mobility data are ground truthed: real mobile activities are directly and continuously sampled as they occur in real time, but clearly they do not have any semantic annotation or context.

In 2006, Dirk Brockmann and his colleagues used the geographic circulation of bank notes in the United States as a proxy for human traffic, assuming that individuals transport money as they travel [51]. They showed that popular Web sites for currency tracking (such as http://en.eurobilltracker.com and www.wheresgeorge.com) collect a massive number of records on money dispersal that can be used as a proxy for human mobility. They found that most bills remain in the vicinity of their initial entry, yet a small but a significant number have traversed distances of the order of the size of USA, consistent with the intuitive notion that short trips occur more frequently than long ones.

This work opened the path to the general exploitation of proxy data for human interaction and mobility [52]. Analogously, modern mobile phones and personal digital assistants combine sophisticated technologies such as Bluetooth, Global Positioning System, and WiFi, constantly producing detailed traces on our daily activities [281, 138]. For instance, in a recent study, Gonzalez et al. [149] used mobile phone data to track the movements of 100,000 people over a 6-month time span. Contrary to bills, mobile phones are carried by the same individual during her daily routine, offering the best proxy to capture individual human trajectories. Each time an individual makes a call the mobile phone operator registers the coordinates of the cell towers communicating with the phones, effectively tracking her locations.

Furthermore, it is now possible to use sensors and tags that produce data at the microscale of one-to-one interactions [208, 281]. A large body of research has flourished in the last years, aimed at better understanding of the human mobility patterns, attracting scientists from diverse disciplines, being not only a major intellectual challenge, but also given its importance in domains such as urban planning, sustainable mobility, transportation engineering, public health and economic nowcasting and forecasting.

The records of mobile communications collected by telecommunications carriers provide extensive proxies of individual trajectories and social relationships, by keeping track of each phone call between any two parties and the localisation in space and time of the party that initiates that call. The high penetration of mobile phones implies that such data capture a large fraction of the population of an entire country. The availability of these massive CDRs (Call Detail Records) has attracted the interest of many researchers in many fields with a variety of interesting results. In [253] Nanni et al. demonstrated that using mobile phone data of all the inhabitants in the Ivory Coast in Africa, it is possible to estimate precisely mobility flows and support the creation of sustainable planning tools, even in a developing country without a sensor infrastructure on roads. The results indicate that a reliable, continuous estimate of mobility flows from GSM (Global System for Mobile communications) data is within the reach of the current state-of-the-art.

An example of how to use GSM data to continuously monitoring some demographic variables of interest is the sociometer of urban population proposed in [131] aimed at estimating the proportion of city users that fall into these categories: residents, commuters, and visitors. The authors used mobile phone call records to characterize the call profiles of the people: residents call essentially any time, commuters tend to call only during weekdays and working hours, visitors call sporadically.

In [84, 308], a social network analysis view is adopted to mobility data to reach a better understanding of human mobility patterns, leveraging the underlying, hidden connections that human mobility establishes among different places. The authors construct a network whose nodes are the territory zones and the weighted edges between any two zones represent the number of travels between them. The analysis phase consists in discovering densely connected subgraphs in this network using a community detection algorithm, to highlight group of zones that are highly connected by many travels compared to the lower connectivity among different modules/clusters. When mapped back to geography, these modules suggest very definite borders, delimiting the mobility basins dictated by the ground truth of human whereabouts.

There are different approaches to study human mobility, from global models forming universal

laws for human movements [149, 330, 331] to studies on individual mobility [137, 271, 307, 272].

In [149] Gonzalez et al. use radius of gyration to show how human mobility is more structured than previous models which use the concept of Lévy flight or random walk. In [330] the authors refine the model of [149] and introduce a revised model incorporating both exploration and preferential return, meaning that an individual has the possibility to move from an already known area or to explore a new one. In [331] Song studied the predictability of human mobility and succeeded in predicting human behaviour.

Studying the mobility of individuals, in [137] the authors develop a framework to analyse human mobility based on trajectory data, while in [271] they use the methods and tools of [137] to understand mobility data in an urban environment. Rinzivillo et al. in [307] synthesise the complexity of human mobility into a model called the *individual mobility network* and exploit this model to enrich spatio-temporal data with semantic information on the activity of the trip, to find the "purpose of motion" as they call it.

The analytical systems for mobility data mining and network analysis described in [137] has been used to create an urban mobility atlas[1], i.e., a comprehensive catalogue of the mobility behaviours in a city based on GPS data from private cars. Each city is portrayed by an infographic showing the key mobility variables, such as the radius of gyration of residents and its distribution.

Given the importance of human mobility, there have now been studies based in the different disciplines from public health to traffic monitoring and urban planning. Since human mobility is responsible for the increased geographical spread of human infectious diseases, models now include human movement as a major element. Belik et al. [33] study features of human mobility relevant for geographical epidemic spread, Longini et al. [224] use stochastic epidemic simulations to investigate the effectiveness of targeted antiviral prophylaxis to contain influenza and in [223] more specifically to contain influenza at its source. These ideas and more have been incorporated in online monitoring platforms, such as GLEAMviz [342]. Human mobility studies are useful also for traffic monitoring and prediction as in [209], where the authors describe a model chain in order to simulate air quality and dynamic exposure and in [297], where Puzis et al. suggest a new traffic assignment model that is based on the concept of a shortest path betweenness centrality measure to optimize the locations of traffic monitoring units, hence reducing the cost and increasing the effectiveness of traffic monitoring. Finally, urban planners are interested in human mobility in order to propose a new network of road usage such as in [354] or to study how different groups of citizens interact with places in metropolitan areas as in [183].

Another interesting line of research for human mobility is location and trajectory prediction. The approaches proposed in literature for location and trajectory prediction can be classified on the basis of the prediction strategy used. In the literature, a lot of works addressing the location prediction problem propose methods that base the prediction only on the movement history of the object itself [13, 65, 147, 182, 190, 219, 260, 320, 345, 357, 371]. We say that these approaches use the individual strategy for the prediction of user future positions. Some approaches of this category adopt time series analyses [65, 320] to forecast user behaviour in different locations. Time series analyses enable estimations as the time of the future visits and expected residence time in those locations [320]. In this kind of works, it is necessary to define the set of interesting locations to be considered in the analysis. In [65] these locations are areas statically defined, while [320] provides a method for extracting significant locations among which users move frequently.

Others prediction approaches are based on Markovian processes [260] and on machine learning techniques such as classification [13, 345]. In particular, in these two last works the location prediction problem is treated as a classification problem: in [13] the location information considered for classification refers to the history of user movements, that is represented by a vector of h time-ordered locations crossed by a user; while in [345] the classification tree is built based

[1] http://kdd.isti.cnr.it/uma/

on simple, intuitive features extracted from the user visit sequence data with associated a semantic meaning. In [219], in order to capture aspects of the individual's mobility behaviours, the authors propose a modified Brownian Bridge model that incorporates linear extrapolation. Other works such as [182, 190] provide methods for the prediction of the movement ahead of a moving object whose movement is constrained to a road network. In [190] the authors assume that the objects' destinations are known. Considering the road network most of the works in this category transform the trajectory into a path on the graph representing the road network. Finally, some works combine historical spatial and temporal data about the user with contextual data such as accelerometer, Bluetooth and call/sms log [147] or with social relationships with friends [371].

The main problem of approaches implementing the individual strategy is that they fail in predicting future locations of non-systematic users. In these cases applying a collective strategy could improve the prediction. Prediction approaches belonging to this category first extract mobility behaviour for each user considering only the user's movement history, like in the individual strategy, and then they merge all the individual models for the construction of the predictor [98, 201, 375]. These works typically use a grid for obtaining cells instead of points like in [201, 375], or extract semantic places from raw data by grouping different spatial coordinates that identify a stop [375].

Other approaches address the location prediction problem by using a global strategy, i.e., they extract movement behaviours from the movement history of all the users in the database and use this global knowledge to forecast the next location visited by a specific moving object. The basic assumption, in this case, is that people often follow the crowd, i.e., individuals tend to follow common paths. This strategy was followed in many papers; most of them extract frequent patterns and association rules from data [70, 181, 189, 210, 213, 225, 248, 249, 250, 374] using methods based on Apriori, PrefixSpan and FPGrowth techniques. Some recent works instead use probabilistic models and in particular Markovian models [56, 140, 299, 369]. Some of these approaches are suitable for predicting the next location by using GSM data [189, 213, 225, 374]; while others work well with GPS data [70, 181, 210, 248, 140, 249, 250, 299, 369]. Solutions based on GPS data typically apply a spatial discretization to make easier finding frequent or interesting locations. Two main types of discretization are applied: the first one extracts interesting places applying density based clustering techniques [181, 210, 213]; while the second one simply uses a grid on the space, determining for each trajectory the sequence of intersected cells [70, 210, 248, 140, 249, 250].

Another interesting way to exploit user mobility information for predicting the next user location is based on the idea to combine the global and individual strategies in order to obtain more accurate predictions. In particular, the idea is to have a global predictor constructed using all users' mobility data and for each user also producing a predictive model based only on her individual movements. Therefore, during the prediction the idea is to use one of these two predictors: when using the individual predictor is not possible to provide a valid and accurate prediction then the global predictor is used [14, 30, 71, 379]. The [14] is based on GPS data but applies a discretization based on clustering; while the others are based on GSM data. In [101] it is used a global model to improve the personalized model: the prediction score that is a combination of the global score and individual one.

6.2 The impact of airport investment on mobility in Tuscany

6.2.1 Introduction

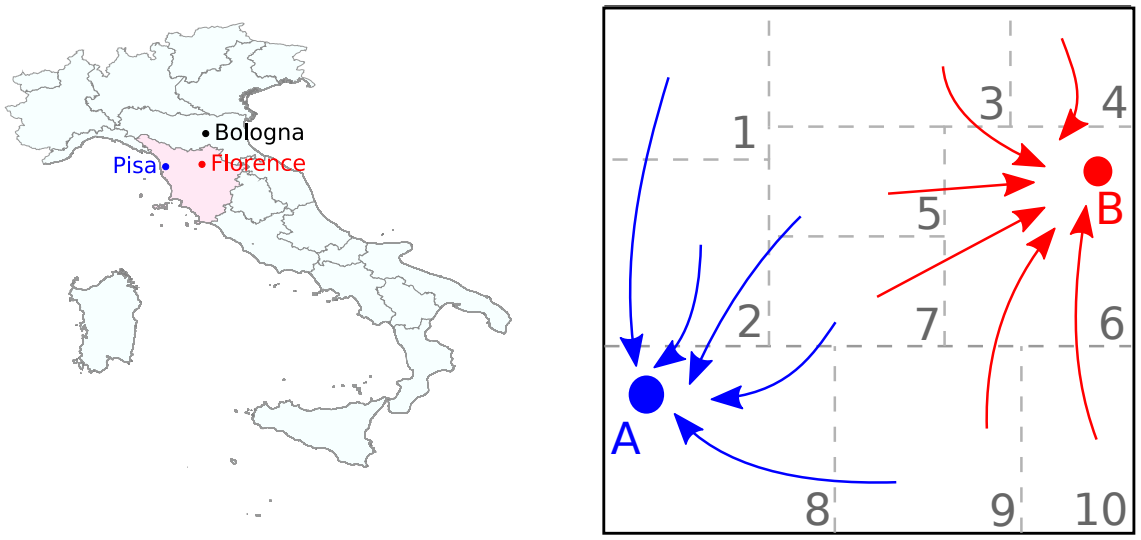
Airports are nodal centres in the air transport network, where researchers consider long range human mobility. In [142] the authors investigate the productivity or performance of airports,

and how changes in the industry may have affected them by studying a dataset of US airports while in [171] the concepts of total factor productivity (TFP) are applied in an empirical study of Australian airports. In [174] the authors examine the changing nature of the performance measurement of airports, while in [9] Adler and Berechman study the same problem but from the airlines' point of view, as they state that the performance of airports has a strong effect on the airlines' choice of hubs. In [278] the authors develop a statistical model for the passengers' choice of airport and airline, in order to study which variables influence the level of demand in a multiple airport region and how airports can use these insights in their efforts to attract more passengers. In [135] Gautreau et al. study the US airport network in order to introduce a model of dynamical networks, which reproduces the main empirical features, both for stationary and local dynamical properties. Balcan et al. in [22] present a model for digital epidemiology integrated with population mobility data from airline traffic to simulate the spread of epidemics at the worldwide scale.

The market of air transport is a competitive one with airports having to justify their existence by attracting and facilitating sufficient passengers to make a profit. In a multiple airport region, airports will also compete with each other for origin (destination) passengers, and perhaps even for transfer passengers. In this work we study the general concept of the impact of major attracting centres in the dynamics of competitive systems, but more specifically we address the question of what variables influence the passengers' airport choice, and we quantify the attractiveness of an airport. We study how airports and other places of interest are able to attract costumers and thereby influence individuals' mobility, by incorporating competition between airports for existing costumers as well as the potential to attract new costumers.

Starting from a retail model introduced in [361] that incorporates the distribution of spending power by the various sub-populations over different shopping centres, we develop our model for the distribution of the travelling sub-populations in Tuscany to the airports of the region. In [361] Wilson states that there are similarities of form between urban system models and models of ecosystems. He uses urban techniques to incorporate spatial competition effects into ecosystem models and he effectively interprets the complex dynamics. Managers of urban resources make use of Wilson's methods and tools to diagnose the farm and territorial sustainability of urban agricultural systems as in [17] and to propose a land use suitability strategy model to orient land uses of non-urbanised areas (NUAs) as in [152]. In [289] Portugali et al. highlight Wilson's contribution to develop complexity theories for cities with implications for urban planning and design and in [360] the authors build new models to represent urban structure, making use of Wilson's ecological models.

First we build a non-spatial model for the interplay of attraction of air travel and frequent flyer population, and we define and calculate the attractiveness of the airport as well as the population of frequent flyers. We introduce a spatially explicit variant of the model, with a number of different locations considered for the airports and the subpopulations. In order to validate such models, we need empirical observations, such as financial data and mobility data from vehicles. In order to indicate the usefulness of such an approach a particular application is considered to allow insight to be easily transferred to other domains. Here, we study the case of two Italian airports, Pisa and Florence airports, that operate in the same region in Tuscany, in Italy, and the subpopulations of the different provinces of the region. To do this, we construct a dynamic model for the interplay of availability of air travel and an airport's popularity among the population. In Figure 6.1 we illustrate our modelling, where a zonal approach is used for the population who travel to one of two airports.



(a) (b)

Figure 6.1: Illustration of the area of interest and the modelling approach. Panel (a) highlights the region of Tuscany in Italy with the airports of Pisa, Florence and Bologna indicated by dots. In panel (b) Tuscany is divided schematically into zones with the different regions indicated by numbers from 1 to 10 and the idealised locations of two airports indicated by dots labelled as 'A' and 'B'. An arrow pointing from a zone to an airport indicates that the airport attracts residents of that zone. Some zones have individuals who are attracted to both airports.

6.2.2 Development of a mathematical model

The non-spatial model

First, we formulate a non-spatial compartmental model to capture the basic dynamics of the system. Variable A is the measure of attractiveness of the airport, which is assumed to be proportional to the number of seats on flights served by the airport per unit time. N is the total population of the region (assumed fixed), and F is the population of travellers (so frequent fliers), $F \leq N$. Note that the number of passengers served by an airport does not equal the number of seats (maximum number of passengers). Model variables A and F are functions of time t, and the new model, which considers the rates of change of A and F, is given by

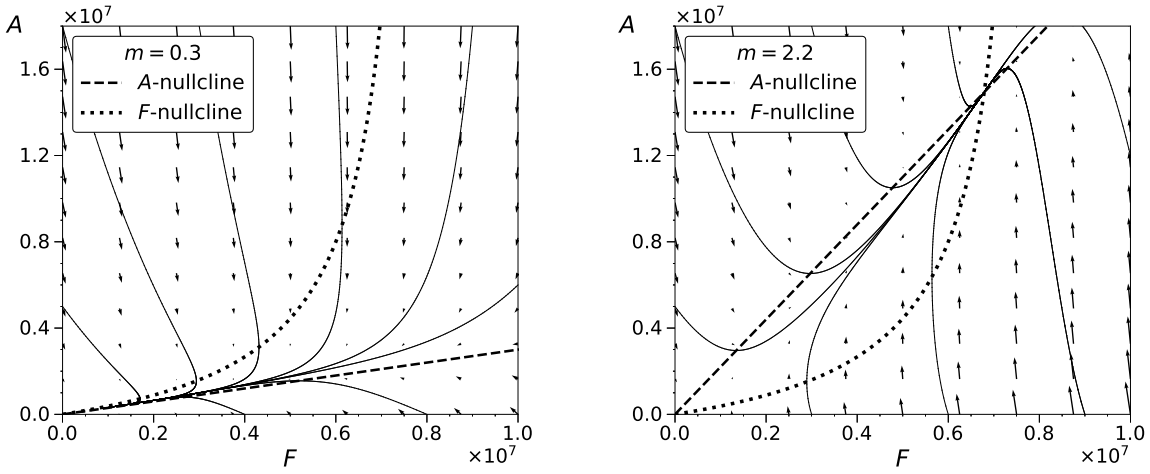
$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}A &= s(mF - (k+e)A), \\ \frac{\mathrm{d}}{\mathrm{d}t}F &= -rF + re\frac{bA}{1+bhA}. \end{aligned} \tag{6.1}$$

The parameter k represents the cost of operating the airport and m is the average amount of money a person spends on air travel over a fixed period of time. The parameter s determines the speed of response to these effects, and e measures the extra effort that the airport makes to increase the number of their customers, associated with an investment. The population of travellers declines by rate a r ($r > 0$) in the absence of other airports in its neighbourhood, and it is also affected by availability of flights. This dependence is modelled by the "uptake" term $re(bA)/(1 + bhA)$, which is proportional to e and also related to parameter b. Moreover the functional form $(bA)/(1 + bhA)$, often referred to as type II functional response term in biological sciences, describes the situation when the uptake of flight availability by the population is increasing but decelerating (for instance, used in predatory-prey systems of ecology [170] and in enzyme kinetics [242]). Parameter b describes the slope of the term at $A = 0$, and the parameter

h is chosen so that $1/h$ characterizes the level of saturation of the uptake. The uptake term approaches re/h if flight availability is unbounded.

Note that instead of interpreting A as the revenue of an airport, we can consider the number of seats on all flights operated by an airport per unit time. Data for the number of flights per airport is available, and multiplying the number of flights with 200 gives an estimate on the maximum number of passengers that an airport could accommodate. It is a reasonable assumption to make that the number of seats offered by an airport is proportional to the revenue of the airport. Note that the number of passengers served by an airport does not equal the number of seats (maximum possible number of passengers).

In the special case when there is no investment and $e = 0$, equations become uncoupled and therefore F decays exponentially. This uncoupled system is equivalent to the retail model described by [361] (albeit with different notation).



(a) 1 equilibrium.

(b) 2 equilibria.

Figure 6.2: Vectorplot and nullclines of the model (6.1) to showcase two qualitatively different types of behaviour for 1 equilibrium and 2 equilibria for two different values of the average spending of people. The dashed line indicates the A-nullcline (where $\frac{\mathrm{d}}{\mathrm{d}t}A = 0$) and the dotted curve indicates the F-nullcline (where $\frac{\mathrm{d}}{\mathrm{d}t}F = 0$). Intersections of the nullclines determine the equilibria. Parameters are set as $r = 0.001$, $s = 0.003$, $k = 0.9$, $e = 0.1$, $b = 30$, $h = 1.25 \times 10^{-8}$, $m = 0.6$ for panel (a) and $m = 2.2$ for panel (b).

We start our analysis of the model by noting that solutions with non-negative initial values for the model variables remain non-negative. Furthermore, considering a variation of the number of travellers, it is easy to see that for the evolution of F, the interval $[0, e/h)$ is invariant, as

$$\frac{\mathrm{d}}{\mathrm{d}t}F > 0 \quad \Leftrightarrow \quad \left(A > \frac{F}{bh\left(\frac{e}{h}-F\right)},\ F < \frac{e}{h}\right) \text{ or } \left(A < \frac{F}{bh\left(\frac{e}{h}-F\right)},\ F > \frac{e}{h}\right)$$

therefore the limit e/h cannot be breached by F. In what follows we assume that $0 \leq F < e/h$. Next, we determine nullclines and equilibria.

$$\begin{array}{rcl} \frac{\mathrm{d}}{\mathrm{d}t}A = 0 & \Leftrightarrow & F = \frac{e+k}{m}A, \\ \frac{\mathrm{d}}{\mathrm{d}t}F = 0 & \Leftrightarrow & F = \frac{ebA}{1+bhA}, \end{array}$$

and it follows that the trivial equilibrium $\boldsymbol{E}^0 = (F^0, A^0) = (0, 0)$ always exists. Non-trivial equilibria $\boldsymbol{E}^* = (F^*, A^*)$ satisfy

$$\frac{e+k}{m} = \frac{eb}{1+bhA},$$

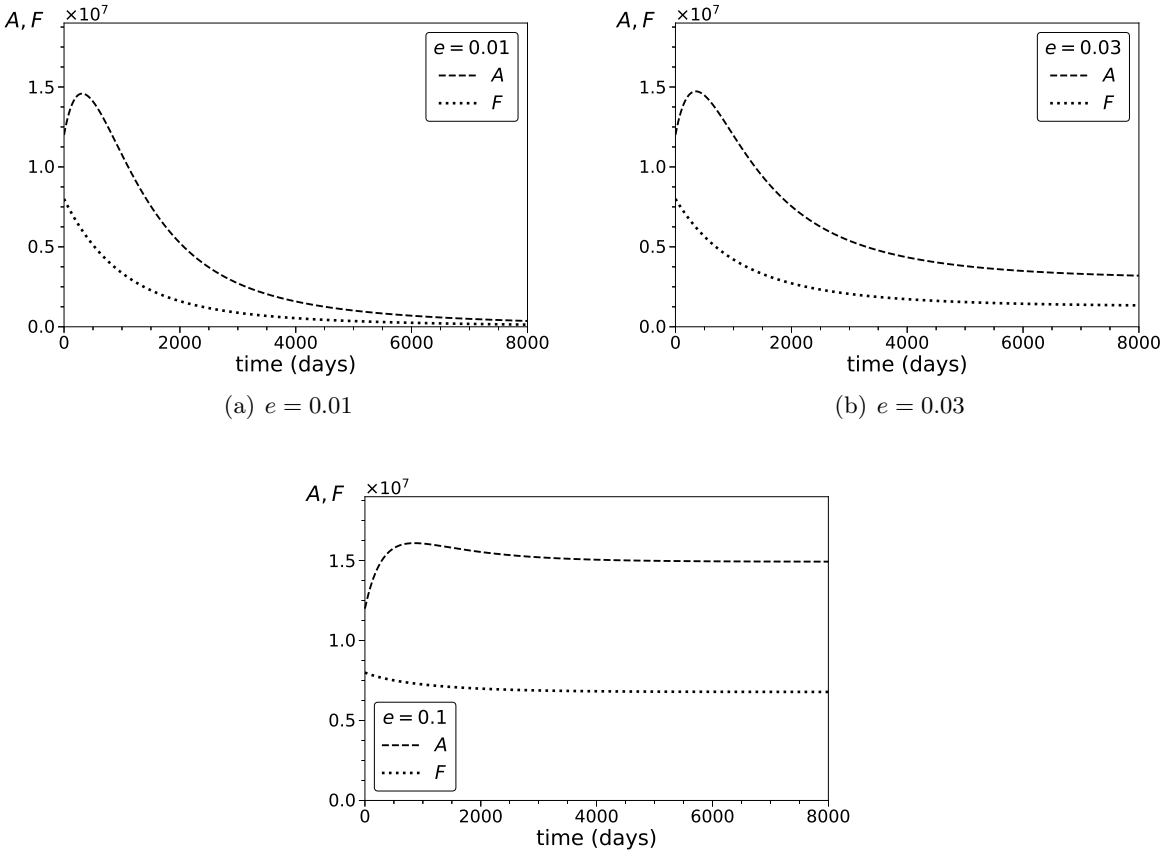
therefore

$$\begin{array}{lcll} (F^0, A^0) & = & (0,0) & \text{always exists,} \\ (F^*, A^*) & = & (\frac{e}{h} - \frac{k+e}{mbh}, \frac{em}{h(k+e)} - \frac{1}{bh}) & \text{exists if and only if } e(mb-1) > k. \end{array}$$

It is also useful to note that

$$\begin{array}{rcl} \frac{\mathrm{d}}{\mathrm{d}t}A > 0 & \Leftrightarrow & A < \frac{m}{k+e}F, \\ \frac{\mathrm{d}}{\mathrm{d}t}F > 0 & \Leftrightarrow & A > \frac{F}{bh\left(\frac{e}{h}-F\right)}, \end{array}$$

therefore by using a simple vector plot analysis it follows that when the non-trivial equilibrium $\boldsymbol{E}^*$ exists, it attracts all solutions, otherwise all solutions converge to the trivial equilibrium (see also Figure 6.2). In what follows it is assumed that $m > 1/b$, the condition necessary for the existence of the positive equilibrium.



(a) $e = 0.01$

(b) $e = 0.03$

(c) $e = 0.1$

Figure 6.3: Time evolutions of A and F for the non-spatial model (A: dashed line, F: dotted line) for three different values of e, representing increasing investment to secure passengers. Panel a) corresponds to $e = 0.01$. For this value only on equilibrium solution exists at $\boldsymbol{E}^0 = (0,0)$. Panel b) corresponds to $e = 0.03$, for which A and F converge to the non-trivial equilibrium point $\boldsymbol{E}^*$. Panel c) corresponds to $e = 0.1$, for which A and F also converge to the non-trivial equilibrium as it was also shown in panel b) of Figure 2. Parameters are given by: $A(0) = 1.2 \times 10^7$, $F(0) = 8 \times 10^6$, $r = 0.001$, $s = 0.003$, $m = 2.2$, $k = 0.9$, $b = 30$, $h = 1.25 \times 10^{-8}$.

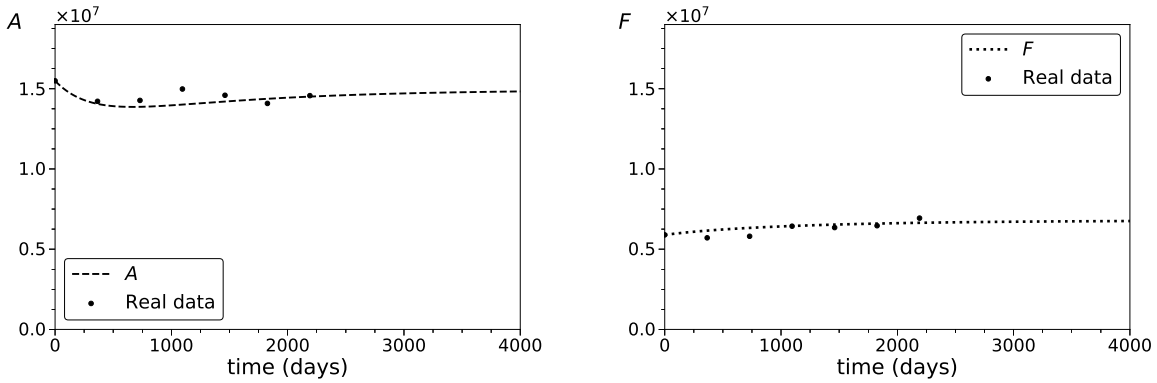
It is clear from the condition for the existence of the positive equilibrium that increasing e can rescue the airport from bankruptcy. There is a threshold $\bar{e} = k/(mb-1)$ such that only the

trivial equilibrium exists for $e < \bar{e}$ but as e increases through $\bar{e}$ the globally attractive positive equilibrium emerges. Furthermore, if the positive equilibrium $\boldsymbol{E}^*$ exists then the computations

$$\frac{\mathrm{d}}{\mathrm{d}e}F^* = \frac{1}{h} - \frac{1}{mbh} > 0,$$
$$\frac{\mathrm{d}}{\mathrm{d}e}A^* = \frac{m}{h} \times \frac{(k+e)-e}{(k+e)^2} > 0$$

show that increasing e –corresponding to increasing investments to bring in more customers– is always beneficial for an airport in the long run. In Figure 6.3 we illustrate dynamic behaviour of the model (6.1) for three different values of e. For the parameter values indicated in the caption of Figure 6.3, the threshold for the existence of the positive equilibrium is $\bar{e} \approx 0.0138$.

To provide insight we also fit this non-spatial model to real data. Data from Tables 6.1 and 6.2 contain information about the number of passengers at Pisa and Florence airports and about the total number of flights at each airport. The number of seats is assumed to be equal to the number of flights multiplied by 200. In the Tables, it can be seen that the total traveller population of both airports monotonically increases over time, however the revenue of airports changes in a non-monotonic way. The fit is presented in Figure 6.4, where in panel a) the dots are the sum of the total number of seats for both airports and the curve represents the time evolution of A according to the model. In panel b), dots represent the sum of the total number of passengers for both airports and the curve corresponds to the time evolution of variable F according to our model. Note that $m = 2.2$ means that individuals spend 2.2×365 Euro on average at airports per year, and the ratio $9:1$ of k and e was chosen based on an assumption that approximately 10 % of spendings of an airport is related to efforts of increasing customer numbers.



(a) (b)

Figure 6.4: Solution curves of the non-spatial model fitted to real data drawn from Tables 6.1 and 6.2 (dots, in the yearly breakdown between 2008–2014). In panels a) and b) the sum of the total number of seats and the sum of the total number of passengers for both Pisa and Florence airports are represented by dots. In panel a) the dashed line represents the time evolution of the variable A, in panel b), the dotted line represents the time evolution of the variable F. Time is expressed in days on the horizontal axis. Parameters for the calculation of the curves according to the model are $r = 0.001$, $s = 0.003$, $m = 2.2$, $k = 0.9$, $e = 0.1$, $b = 30$, $h = 1.25 \times 10^{-8}$.

The spatially explicit model

We now extend our model to incorporate spatial aspects. Motivated by the spatially explicit retail model in [361], we incorporate multiple airports and divide the population into a number

of groups based on geographic location. For each airport we assume that

$$\frac{\mathrm{d}}{\mathrm{d}t}A_j = s(D_j - (k_j + e_j)A_j), \qquad j = 1, \ldots, a, \tag{6.2}$$

where a is the number of airports and A_j is the measure of attractiveness of airport j, that is assumed to be proportional to the number of seats on flights operated by the airport j. Parameter k_j represents the cost of operating the airport j, e_j measures the effort of airport j to bring in new costumers, and D_j is the total revenue attracted into A_j. In the special case $e_j = 0$, the model (6.2) is equivalent to the retail model in [361] which incorporates the distribution of spending power by the various sub-populations over the shopping centres, with the population in each region assumed constant. In contrast, the novelty of our approach lies in the use of two important assumptions: we allow population sizes to vary and consider the general situation when the number of sub-populations differs from the number of airports.

We consider the total population distributed over z different zones, and denote by F_p the frequent fliers population in zone p, with $p = 1, 2, ..., z$. To compute D_j, the total revenue attracted to airport j, let m_p denote the average amount of spending on air travel per person from zone p over a fixed period of time, hence $m_p F_p$ is distributed over all airports. We let the proportion of $m_p F_p$ attracted by airport j be given by

$$m_p F_p \frac{A_j^\alpha \mathrm{e}^{-\beta d_{jp}}}{\sum_i A_i^\alpha \mathrm{e}^{-\beta d_{ip}}},$$

where e must not be confused with e, as the former is the base of the exponential function while the latter is one of the model parameters. Here, α and β are parameters and d_{jp} measures the "distance" between zone p and airport j, with $p \in \{1, \ldots, z\}$, $i, j \in \{1, \ldots, a\}$, (here, "distance" does not necessarily mean geographical distance but it could also be a measure of the accessibility of an airport by various means of transportation, and various other factors that can influence an individual's willingness to travel to that airport). Therefore D_j can be given by

$$D_j = \sum_{p=1}^{z} m_p F_p \frac{A_j^\alpha \mathrm{e}^{-\beta d_{jp}}}{\sum_i A_i^\alpha \mathrm{e}^{-\beta d_{ip}}}, \qquad j = 1, 2, \ldots, a.$$

If we choose $\alpha = 0$ then people's choice of airport depends solely on the distance, on the other hand by putting $\beta = 0$ we assume that the significance of distance as a factor is negligible and people make their decision based on the flight availability of the airport.

The parameter e_j in (6.2) measures the effort of the airport j to increase the number of their customers. H_p measures the impact of the attraction power of airport on the traveller population in zone p. In analogy with the non-spatial model (6.1), the total contribution of airports to the growth of the traveller population is $(b\sum_{j=1}^{a} e_j A_j)/(1 + bh\sum_{j=1}^{a} A_j)$, and to distribute this over the sub-populations we use the formula

$$H_p = \frac{b\sum_{j=1}^{a} e_j A_j}{1 + bh\sum_{j=1}^{a} A_j} \times \frac{F_p^\gamma}{\sum_q F_q^\gamma}, \qquad p = 1, 2, \ldots, z, \quad \gamma = 1, 2, \ldots, z,$$

where γ is another parameter.

We thus arrive at the spatially explicit model for a airports and z sub-populations:

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}t}A_j = s(D_j - (k_j + e_j)A_j), \\ &\frac{\mathrm{d}}{\mathrm{d}t}F_p = -rF_p + rH_p, \qquad\qquad j = 1, 2, \ldots, a, \;\; p = 1, 2, \ldots, z. \end{aligned} \tag{6.3}$$

We can use model (6.3) to derive equations for the dynamics of the traveller population $\sum_{p=1}^{z} F_p$ and total number of seats on flights from all airports, which we equate with $\sum_{j=1}^{a} A_j$. We note that if $e_j = e_i$ and $k_j = k_i$ for all $i, j = 1, \dots a$ and $m_p = m_q$ for all $p, q = 1, \dots z$ then the aggregation of the spatially explicit model (6.3) leads to the non-spatial model (6.1).

In what follows we let $\gamma = 1$ in our analysis, corresponding to a reasonable assumption that the attractive power of airports towards air travel is homogeneous for all individuals.

6.2.3 A special situation: 2 airports, 1 population

We consider the model (6.2) in the special situation where there are only two airports and one population. For brevity we denote by d_1 and d_2 the distance to airport 1 and airport 2 respectively. We perform our analysis for two particular cases, which differ in the way individuals choose between the airports.

Case 1 : $\alpha = 0$

When $\alpha = 0$ individuals choose between airports solely based on the distance to the airport which nonetheless takes into account accessibility and geographical distance. Other attributes of an airport (flight offers, size, etc.) have no impact on an individual's decision. System (6.3) becomes

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}A_1 &= s\left(mF\frac{\mathrm{e}^{-\beta d_1}}{\mathrm{e}^{-\beta d_1}+\mathrm{e}^{-\beta d_2}} - (k_1+e_1)A_1\right),\\
\frac{\mathrm{d}}{\mathrm{d}t}A_2 &= s\left(mF\frac{\mathrm{e}^{\beta d_2}}{\mathrm{e}^{-\beta d_1}+\mathrm{e}^{-\beta d_2}} - (k_2+e_2)A_2\right),\\
\frac{\mathrm{d}}{\mathrm{d}t}F &= -rF + r\frac{b(e_1A_1+e_2A_2)}{1+bh(A_1+A_2)}.
\end{aligned}$$

It can be shown that the equilibria of this system are the trivial equilibrium $\boldsymbol{E}^0 = \left(A_1^0, A_2^0, F^0\right) = (0,0,0)$, which always exists, and another equilibrium $\boldsymbol{E}^* = (A_1^*, A_2^*, F^*)$, which satisfies

$$\begin{aligned}
A_1^* &= \frac{mF^*}{k_1+e_1} \times \frac{\mathrm{e}^{-\beta d_1}}{\mathrm{e}^{-\beta d_1}+\mathrm{e}^{-\beta d_2}},\\
A_2^* &= \frac{mF^*}{k_2+e_2} \times \frac{\mathrm{e}^{-\beta d_2}}{\mathrm{e}^{-\beta d_1}+\mathrm{e}^{-\beta d_2}}.
\end{aligned}$$

Here F^* is given by $F^* = \frac{c_2}{c_1 c_3} - \frac{1}{c_3}$. The non-trivial equilibrium is only possible when $c_1 < c_2$, as we require it to be positive. The non-negative constants c_1, c_2 and c_3 are given by

$$\begin{aligned}
c_1 &= \frac{1}{bm},\\
c_2 &= \frac{e_1(k_2+e_2)\mathrm{e}^{-\beta d_1} + e_2(k_1+e_1)\mathrm{e}^{-\beta d_2}}{(k_1+e_1)(k_2+e_2)(\mathrm{e}^{-\beta d_1}+\mathrm{e}^{-\beta d_2})},\\
c_3 &= mbh\frac{(k_2+e_2)\mathrm{e}^{-\beta d_1} + (k_1+e_1)\mathrm{e}^{-\beta d_2}}{(k_1+e_1)(k_2+e_2)(\mathrm{e}^{-\beta d_1}+\mathrm{e}^{-\beta d_2})}.
\end{aligned} \tag{6.4}$$

Proposition 1. *In the case when $\alpha = 0$ the positive equilibrium $\mathbf{E}^*$ of the system is locally asymptotically stable when it exists. The trivial equilibrium $(0,0,0)$ is unstable if the positive equilibrium exists, and is locally asymptotically stable otherwise.*

Proof. First, we recall some definitions from [123]. For any square matrix B, we denote by $s(B)$ the maximum real part of all eigenvalues of B, and by $\rho(B)$ is the dominant eigenvalue of B. We say that a matrix is non-negative if all entries are non-negative. We denote by Z the set of all real square matrices whose off-diagonal entries are non-positive. We say that a square matrix B from the class Z is a non-singular M-matrix if there exists a matrix $C \geq 0$ and a number $c > \rho(C)$ such that $B = c \cdot \mathrm{Id} - C$, where Id denotes the identity matrix of the same type as B. [Theorem 5.1] on [123] establishes several alternative definitions for non-singular M-matrices: for instance, a matrix B from the class Z is a non-singular M-matrix if $B^{-1} \geq 0$. The following result is useful, and can be proved by similar arguments to those in [Theorem 2] in [348].

For a square matrix B, consider a splitting $B = F - V$ where F is a non-negative matrix and V is a non-singular M-matrix. Then it holds that $s(B) < 0$ if and only if $\rho\left(FV^{-1}\right) < 1$, $s(B) = 0$ if and only if $\rho\left(FV^{-1}\right) = 1$, and $s(B) > 0$ if and only if $\rho\left(FV^{-1}\right) > 1$.

In order to study stability of equilibria of the system in the case when $\alpha = 0$ we linearise the system and obtain the matrix of the linearised system, as

$$J = \begin{pmatrix} -s(k_1+e_1) & 0 & \frac{sm\exp(-\beta d_1)}{\exp(-\beta d_1)+\exp(-\beta d_2)} \\ 0 & -s(k_2+e_2) & \frac{sm\exp(-\beta d_2)}{\exp(-\beta d_1)+\exp(-\beta d_2)} \\ \frac{rb(e_1+bhA_2(e_1-e_2))}{(1+bh(A_1+A_2))^2} & \frac{rb(e_2+bhA_1(e_2-e_1))}{(1+bh(A_1+A_2))^2} & -r \end{pmatrix}.$$

Stability of an equilibrium is determined by the sign of $s(J)$ where J is the matrix of the linearised system evaluated at the equilibrium. The equilibrium is locally asymptotically stable if $s(J) < 0$ and it is unstable if $s(J) > 0$. We let

$$F = \begin{pmatrix} 0 & 0 & \frac{sm\exp(-\beta d_1)}{\exp(-\beta d_1)+\exp(-\beta d_2)} \\ 0 & 0 & \frac{sm\exp(-\beta d_2)}{\exp(-\beta d_1)+\exp(-\beta d_2)} \\ \frac{rb(e_1+bhA_2(e_1-e_2))}{(1+bh(A_1+A_2))^2} & \frac{rb(e_2+bhA_1(e_2-e_1))}{(1+bh(A_1+A_2))^2} & 0 \end{pmatrix},$$

$$V = \begin{pmatrix} s(k_1+e_1) & 0 & 0 \\ 0 & s(k_2+e_2) & 0 \\ 0 & 0 & r \end{pmatrix},$$

and obtain a splitting of J as $F - V$. The matrices F and V satisfy the conditions of the result cited above, therefore it remains to investigate for each equilibrium, if $\rho(FV^{-1}) < 1$ (local asymptotic stability) or $\rho(FV^{-1}) > 1$ (instability).

Eigenvalues of FV^{-1},

$$FV^{-1} = \begin{pmatrix} 0 & 0 & \frac{sm\exp(-\beta d_1)}{r(\exp(-\beta d_1)+\exp(-\beta d_2))} \\ 0 & 0 & \frac{sm\exp(-\beta d_2)}{r(\exp(-\beta d_1)+\exp(-\beta d_2))} \\ \frac{rb(e_1+bhA_2(e_1-e_2))}{s(k_1+e_1)(1+bh(A_1+A_2))^2} & \frac{rb(e_2+bhA_1(e_2-e_1))}{s(k_2+e_2)(1+bh(A_1+A_2))^2} & 0 \end{pmatrix},$$

are found by solving the characteristic equation $\lambda^3 - \lambda\sigma = 0$, where

$$\sigma = \frac{bm}{(\exp(-\beta d_1) + \exp(-\beta d_2))(1+bh(A_1+A_2))^2} \times \left(\frac{(e_1 + bhA_2(e_1-e_2))\exp(-\beta d_1)}{(k_1+e_1)} + \frac{(e_2+bhA_1(e_2-e_1))\exp(-\beta d_2)}{(k_2+e_2)} \right).$$

Roots are obtained as $\lambda_1 = \sqrt{\sigma}, \lambda_2 = -\sqrt{\sigma}, \lambda_3 = 0$, thus it is clear that it is necessary and sufficient to determine that if $\lambda_1 < 1 \Leftrightarrow \sigma < 1$ or $\lambda_1 > 1 \Leftrightarrow \sigma > 1$.

For the trivial equilibrium $\mathbf{E}^0 = (0, 0, 0)$, the formula for σ reduces to

$$\sigma = \frac{bm}{(\exp(-\beta d_1) + \exp(-\beta d_2))} \times \left(\frac{e_1 \exp(-\beta d_1)}{(k_1 + e_1)} + \frac{e_2 \exp(-\beta d_2)}{(k_2 + e_2)} \right),$$

and it is straightforward to see that if the condition for the existence of the non-trivial equilibrium does not hold (i.e., $c_1 \geq c_2$ with the definitions in (6.4)) then $\sigma < 1$, while if the condition holds (i.e., $c_1 < c_2$) then $\sigma > 1$.

Next we consider the positive equilibrium $\mathbf{E}^*$. Using the equalities derived from the steady state equations

$$\frac{(A_1)^*}{F^*} = \frac{m \exp(-\beta d_1)}{(\exp(-\beta d_1) + \exp(-\beta d_2))(k_1 + e_1)},$$
$$\frac{(A_2)^*}{F^*} = \frac{m \exp(-\beta d_2)}{(\exp(-\beta d_1) + \exp(-\beta d_2))(k_2 + e_2)},$$

the expression for σ reduces to

$$\begin{aligned}
\sigma &= b \frac{(e_1 + bh(A_2)^*(e_1 - e_2))(A_1)^* + (e_2 + bh(A_1)^*(e_2 - e_1))(A_2)^*}{F^*(1 + bh((A_1)^* + (A_2)^*))^2} \\
&= b \frac{e_1(A_1)^* + e_2(A_2)^*}{F^*(1 + bh((A_1)^* + (A_2)^*))^2} \\
&= \frac{1}{1 + bh((A_1)^* + (A_2)^*)},
\end{aligned}$$

where in the last step we have used the steady state equation $F^* = \frac{b(e_1(A_1)^* + e_2(A_2)^*)}{1 + bh((A_1)^* + (A_2)^*)}$. It is therefore trivial that $\sigma < 1$, that is, the positive equilibrium is locally asymptotically stable when it exists. □

Our numerical simulations and local stability analysis above suggest that the positive equilibrium is globally asymptotically stable whenever it exists, that is, all solutions converge to this equilibrium meaning that both airports persist. If the condition for the existence of the positive equilibrium does not hold, then all solutions converge to the trivial equilibrium, meaning that the traveller population and an airport's attractiveness declines over time.

It is useful to note that

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} A_1 > 0 &\Leftrightarrow A_1 < \frac{mF}{\mathrm{e}^{-\beta d_1} + \mathrm{e}^{-\beta d_2}} \times \frac{\mathrm{e}^{-\beta d_1}}{k_1 + e_1}, \\
\frac{\mathrm{d}}{\mathrm{d}t} A_2 > 0 &\Leftrightarrow A_2 < \frac{mF}{\mathrm{e}^{-\beta d_1} + \mathrm{e}^{-\beta d_2}} \times \frac{\mathrm{e}^{-\beta d_2}}{k_2 + e_2},
\end{aligned}$$

therefore if the condition $(k_1 + e_1)\mathrm{e}^{\beta d_1} = (k_2 + e_2)\mathrm{e}^{\beta d_2}$ holds, then one airport's attractiveness cannot grow or decline without that of the other's.

Case 2 : $\alpha = 1$

This corresponds to the case when individuals choose between airports based on both distance (accessibility, geographical distance) and other attributes of an airport (flight offers, size, etc.).

System (6.3) reduces to

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}A_1 &= s\left(mF\frac{A_1\mathrm{e}^{-\beta d_1}}{A_1\mathrm{e}^{-\beta d_1}+A_2\mathrm{e}^{-\beta d_2}}-(k_1+e_1)A_1\right),\\
\frac{\mathrm{d}}{\mathrm{d}t}A_2 &= s\left(mF\frac{A_2\mathrm{e}^{-\beta d_2}}{A_1\mathrm{e}^{-\beta d_1}+A_2\mathrm{e}^{-\beta d_2}}-(k_2+e_2)A_2\right),\\
\frac{\mathrm{d}}{\mathrm{d}t}F &= -rF+r\frac{b(e_1A_1+e_2A_2)}{1+bh(A_1+A_2)}.
\end{aligned}$$

We now look for equilibria of the system. It is easy to see that $F^*\neq 0$ holds for any equilibrium (A_1^*,A_2^*,F^*). We consider three possibilities.

a). $F^*\neq 0$, $A_1^*\neq 0$, $A_2^*=0$ (only airport 1 persists) An equilibrium of these characteristics satisfies $\boldsymbol{E}^1=(A_1^*,A_2^*,F^*)$, with

$$\begin{aligned}
A_1^* &= \frac{me_1}{(k_1+e_1)h}-\frac{1}{bh},\\
A_2^* &= 0,\\
F^* &= \frac{e_1}{h}-\frac{k_1+e_1}{mbh},
\end{aligned}$$

that is a non-negative equilibrium if and only if $k_1+e_1<me_1b$.

b). $F^*\neq 0$, $A_1^*=0$, $A_2^*\neq 0$ (only airport 2 persists)
We derive the equilibrium $\boldsymbol{E}^2=(A_1^*,A_2^*,F^*)$ with

$$\begin{aligned}
A_1^* &= 0,\\
A_2^* &= \frac{me_2}{(k_2+e_2)h}-\frac{1}{bh},\\
F^* &= \frac{e_2}{h}-\frac{k_2+e_2}{mbh}.
\end{aligned}$$

It is non-negative if and only if $k_2+e_2<me_2b$.

c). $F^*\neq 0$, $A_1^*\neq 0$, $A_2^*\neq 0$ (both airports persist)
It can be shown that an equilibrium of these characteristics exists if and only if $(k_1+e_1)e^{\beta d_1}=(k_2+e_2)e^{\beta d_2}$ holds. Airports (or perhaps local government) have control over the parameters e_1 and e_2 so they can be tuned to make sure the condition is met; otherwise one airport will dominate the other as discussed in cases 1 and 2. Assuming that the condition for coexistence holds, we derive

$$\frac{mb(e_1A_1+e_2A_2)}{1+bh(A_1+A_2)}=(k_1+e_1)A_1+(k_2+e_2)A_2.$$

This equation defines a curve on the two-dimensional (A_1,A_2) plane, where every point on the curve is an equilibrium. This curve goes through the origin and the 2 non-zero equilibrium points $\boldsymbol{E}^1$ and $\boldsymbol{E}^2$, determined in cases 1 and 2, therefore positive solutions exist if and only if at least one of $\boldsymbol{E}^1,\boldsymbol{E}^2$ exists (which means, are feasible outcomes for the model, i.e. non-negative). In particular, an equilibrium where $A_1^*=A_2^*=A^*$ satisfies

$$A^*=\frac{(e_1+e_2)m}{2h(k_1+e_1+k_2+e_2)}-\frac{1}{2bh},$$

and it is a positive equilibrium if and only if $k_1+e_1+k_2+e_2<mb(e_1+e_2)$ holds.

Proposition 2. *In the system for the case when $\alpha = 1$, the equilibrium $\mathbf{E}^1$ is locally asymptotically stable if $(k_1 + e_1)\exp(\beta d_1) < (k_2 + e_2)\exp(\beta d_2)$ holds, and unstable if the inequality is reversed. The equilibrium $\mathbf{E}^2$ is locally asymptotically stable if $(k_1 + e_1)\exp(\beta d_1) > (k_2 + e_2)\exp(\beta d_2)$ holds, and unstable if the inequality is reversed.*

Proof. First we study the stability of the equilibrium $\mathbf{E}^1$, where airport 2 does not persist ($(A_2)^* = 0$). The matrix for the linearised system about any point where $A_2 = 0$, is given by

$$\begin{pmatrix} -s(k_1+e_1) & -smF\exp(-\beta d_2)/(A_1\exp(-\beta d_1)) & sm \\ 0 & smF\exp(-\beta d_2)/(A_1\exp(-\beta d_1)) - s(k_2+e_2) & 0 \\ re_1b/(1+bhA_1)^2 & re_2b & -r \end{pmatrix}$$
$$= \begin{pmatrix} -s(k_1+e_1) & -s(k_1+e_1)\exp(-\beta d_2)/(\exp(-\beta d_1)) & sm \\ 0 & s(k_1+e_1)\exp(-\beta d_2)/(\exp(-\beta d_1)) - s(k_2+e_2) & 0 \\ re_1b/(1+bhA_1)^2 & re_2b & -r \end{pmatrix}.$$

Using the fact that the equalities $mF/A_1 = k_1 + e_1$ and $e_1b/(1 + bhA_1) = F/A_1 = (k_1 + e_1)/m$ hold for the equilibrium $\mathbf{E}^1$, the characteristic equation reads

$$\begin{aligned} &\Big(s(k_1+e_1)\exp(\beta(d_1-d_2)) - s(k_2+e_2) - \lambda\Big) \times \Big((s(k_1+e_1)+\lambda)(r+\lambda) - re_1bsm/(1+bhA_1)^2\Big) \\ &= \Big(s(k_1+e_1)\exp(\beta(d_1-d_2)) - s(k_2+e_2) - \lambda\Big) \\ &\quad \times \Big(\lambda^2 + \lambda(r + s(k_1+e_1)) + s(k_1+e_1)r - s(k_1+e_1)r/(1+bhA_1)\Big) \\ &= \Big(s(k_1+e_1)\exp(\beta(d_1-d_2)) - s(k_2+e_2) - \lambda\Big) \times (\lambda^2 + \lambda B + C) = 0, \end{aligned}$$

where we let $p = r + s(k_1 + e_1)$ and $q = sr(k_1 + e_1)bhA_1/(1 + bhA_1)$. Roots of the characteristic equation are obtained as

$$\lambda_1 = s\Big((k_1+e_1)\exp(\beta(d_1-d_2)) - (k_2+e_2)\Big), \qquad \lambda_{2,3} = \frac{-p \pm \sqrt{p^2-4q}}{2},$$

and it is easy to see that $p > 0$ and $0 < q < sr(k_1 + e_1)$ which implies that $p^2 - 4q > 0$ holds. Therefore, since $p > \sqrt{(p^2 - 4q)}$, we conclude that $\lambda_2, \lambda_3 < 0$ and the sign of λ_1 determines stability of the equilibrium. It follows that this equilibrium is locally asymptotically stable if $(k_1 + e_1)\exp(\beta d_1) < (k_2 + e_2)\exp(\beta d_2)$ holds, and unstable if the inequality is reversed.

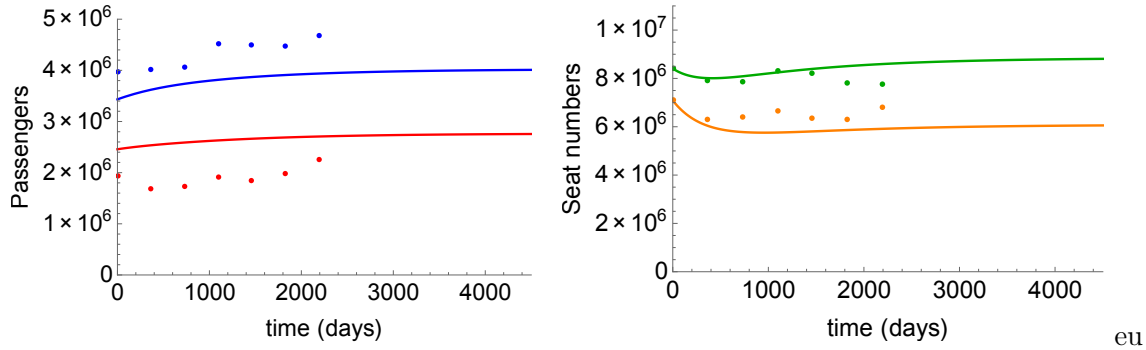
One can show in a similar way that the equilibrium $\mathbf{E}^2$ is locally asymptotically stable if $(k_1 + e_1)\exp(\beta d_1) > (k_2 + e_2)\exp(\beta d_2)$ holds, and unstable if the inequality is reversed. □

Summarizing, an equilibrium where both airports persist ($A_1^* > 0, A_2^* > 0$), can exist if and only if the condition $(k_1 + e_1)\mathrm{e}^{\beta d_1} = (k_2 + e_2)\mathrm{e}^{\beta d_2}$ holds, in which case there must exist at least one of the equilibria $\boldsymbol{E}^1$ and $\boldsymbol{E}^2$. However, neither of these equilibria can be locally stable if a positive equilibrium ($A_1^* > 0, A_2^* > 0$) exists. Our numerical simulations suggest that all solutions converge to a positive steady state.

If both equilibria $\boldsymbol{E}^1$ and $\boldsymbol{E}^2$ exist but the positive equilibrium does not (i.e., $(k_1+e_1)\mathrm{e}^{\beta d_1} \neq (k_2 + e_2)\mathrm{e}^{\beta d_2}$), then the condition for stability holds for exactly one of the equilibria $\boldsymbol{E}^1$ and $\boldsymbol{E}^2$. Numerical simulations confirm that in this case the locally stable equilibrium attracts all solutions.

If the positive equilibrium does not exist and only one of the equilibria $\boldsymbol{E}^1$ and $\boldsymbol{E}^2$ exists, then there are two cases: either its stability condition holds or it does not. Through numerical simulations we find that in the former case this equilibrium attracts all solutions while in the

latter all solutions converge to $\boldsymbol{E}^0$.



(a) (b)

eu

Figure 6.5: Solution curves of the model (6.3) for two airports and one population, fitted to real data drawn from Tables 6.1 and 6.2. On the vertical axis: blue = passengers who go to airport 1, red = passengers who go to airport 2, green = A_1 (number of seats at Pisa airport), orange = A_2 (number of seats at Florence airport), on the horizontal axis: time is days t. Continuous lines indicate solution curves of the model and dots of the same colour indicate real data obtained in the yearly breakdown between 2008–2014. Parameters are $r = 0.001$, $s = 0.003$, $m = 2.2$, $k_1 = k_2 = 0.9$, $e_1 = e_2 = 0.1$, $b = 30$, $h = 1.25 \times 10^{-8}$, $\alpha = 0.2$, $d_1 = 0$, $d_2 = 0.3$.

In Figure 6.5 we illustrate the dynamics of system (6.5) and fit our model to real data. The system setup allows us not only to model the dynamics of the traveller population F but also to calculate how it is distributed over the two airports. The number of individuals who choose airport 1 (Pisa airport) and airport 2 (Florence airport), respectively, are given by $\frac{FA_1^\alpha \mathrm{e}^{-\beta d_1}}{A_1^\alpha \mathrm{e}^{-\beta d_1} + A_2^\alpha \mathrm{e}^{-\beta d_2}}$ and $\frac{FA_2^\alpha \mathrm{e}^{-\beta d_2}}{A_1^\alpha \mathrm{e}^{-\beta d_1} + A_2^\alpha \mathrm{e}^{-\beta d_2}}$ respectively, and the sums of these numbers give the total population of travellers.

6.2.4 A special case: 2 airports, 2 populations

In the special case of two airports (as currently exists in Tuscany) and two populations, the model (6.1) reads

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}A_1 &= s\left(\frac{m_1 F_1 A_1^\alpha \mathrm{e}^{-\beta d_{11}}}{A_1^\alpha \mathrm{e}^{-\beta d_{11}} + A_2^\alpha \mathrm{e}^{-\beta d_{21}}} + \frac{m_2 F_2 A_1^\alpha \mathrm{e}^{-\beta d_{12}}}{A_1^\alpha \mathrm{e}^{-\beta d_{12}} + A_2^\alpha \mathrm{e}^{-\beta d_{22}}}\right) \\
&\quad - s(k_1 + e_1)A_1, \\
\frac{\mathrm{d}}{\mathrm{d}t}A_2 &= s\left(\frac{m_1 F_1 A_2^\alpha \mathrm{e}^{-\beta d_{21}}}{A_1^\alpha \mathrm{e}^{-\beta d_{11}} + A_2^\alpha \mathrm{e}^{-\beta d_{21}}} + \frac{m_2 F_2 A_2^\alpha \mathrm{e}^{-\beta d_{22}}}{A_1^\alpha \mathrm{e}^{-\beta d_{12}} + A_2^\alpha \mathrm{e}^{-\beta d_{22}}}\right) \\
&\quad - s(k_2 + e_2)A_2, \\
\frac{\mathrm{d}}{\mathrm{d}t}F_1 &= -rF_1 + r\frac{F_1}{F_1 + F_2} \times \frac{b(e_1 A_1 + e_2 A_2)}{1 + bh(A_1 + A_2)}, \\
\frac{\mathrm{d}}{\mathrm{d}t}F_2 &= -rF_2 + r\frac{F_2}{F_1 + F_2} \times \frac{b(e_1 A_1 + e_2 A_2)}{1 + bh(A_1 + A_2)}.
\end{aligned}$$

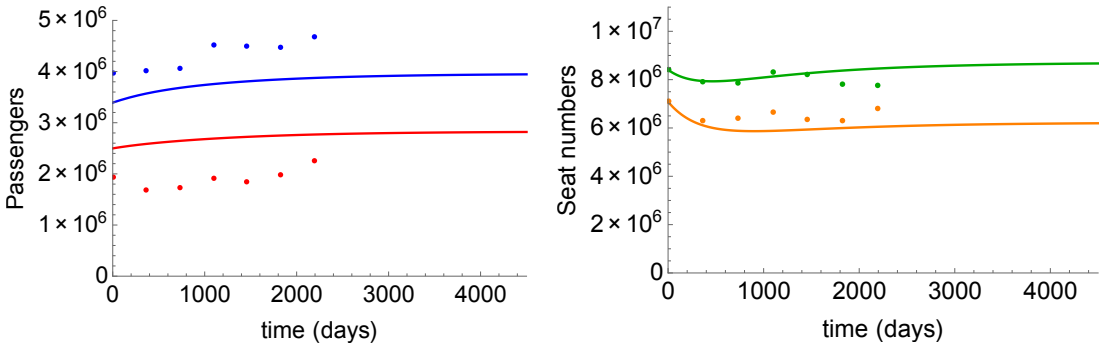
Note that the variables F_1 and F_2 stand for the traveller populations of zone 1 and 2, respectively. These variables do not give the number of passengers visiting airport 1 and airport

2, which are obtained as

$$V_1 = \frac{F_1 A_1^\alpha e^{-\beta d_{11}}}{A_1^\alpha e^{-\beta d_{11}} + A_2^\alpha e^{-\beta d_{21}}} + \frac{F_2 A_1^\alpha e^{-\beta d_{12}}}{A_1^\alpha e^{-\beta d_{12}} + A_2^\alpha e^{-\beta d_{22}}},$$
$$V_2 = \frac{F_1 A_2^\alpha e^{-\beta d_{21}}}{A_1^\alpha e^{-\beta d_{11}} + A_2^\alpha e^{-\beta d_{21}}} + \frac{F_2 A_2^\alpha e^{-\beta d_{22}}}{A_1^\alpha e^{-\beta d_{12}} + A_2^\alpha e^{-\beta d_{22}}},$$

respectively. Time evolution of the number of passengers served by each airport and the total number of airplane seats operated by each airport is illustrated in Figure 6.6. The two airports correspond to Pisa airport and Florence airport. We identify two zones within Tuscany for the two populations, as we divide the total population into two parts based on their distances from each airport. Zone 1 is identified using data in Table 6.5, as the collection of locations that are closer in distance to Pisa airport than to Florence airport. All other territories are marked as zone 2.

Data (as shown in Table 6.5) indicates that people living closer to Pisa airport prefer Pisa airport, to a large degree, over Florence airport. At the same time, people who live closer to Florence airport than to Pisa airport, do not have such strong preference, in fact they choose both airports in almost equal measure. For this reason, we let $d_{21} >> d_{11}$ but let $d_{12} \approx d_{22}$.



(a) (b)

Figure 6.6: Solution curves of the model (6.3) for two populations and two airports, fitted to real data drawn from Tables 6.1 and 6.2. On the vertical axis: blue = passengers who go to airport 1, red = passengers who go to airport 2, green = A_1 (number of seats at Pisa airport), orange = A_2 (number of seats at Florence airport), on the horizontal axis: time in days t. Continuous lines indicate solution curves of the model and dots of the same colour indicate real data obtained in the yearly breakdown between 2008–2014. Parameters are $r = 0.001$, $s = 0.003$, $m_1 = m_2 = 2.2$, $k_1 = k_2 = 0.9$, $e_1 = e_2 = 0.1$, $b = 30$, $h = 1.25 \times 10^{-8}$, $\alpha = 0.2$, $d_{11} = 0$, $d_{21} = 2$, $d_{22} = 0$, $d_{12} = 0$. The population of Tuscany is split into two zones based on data in Table 6.5.

6.2.5 Data

To validate our models and to make sure that all important aspects have been incorporated, we use data regarding the number of passengers and the number of flights on offer of each airport. Since we are interested in the behaviour of people in the provinces of Tuscany region, we also take under consideration the distance that each inhabitant lives from each airport, meaning the time that they need to spend in order to get to an airport either by car or train. Finally, we use mobility data, in order to study people's movements, so that we can know in reality how many people choose which airport and the location where they come from and in which province they live.

Airports statistical Data

In the Tuscany region of Italy there are two major airports operating, the Galileo Galilei airport of Pisa and the Amerigo Vespucci airport of Florence. The ENAC report about the development of Italian airports [4] tells that the two airports have different characteristics and history, which are the results of different policy choices along the time, and environmental constraints. A. Vespucci has some issues related to its inclusion in a problematic urban context, in addition to small size and flight infrastructure limitations. Situated northwest of Florence, 4 km far from the city centre, this airport is connected by highways, but it is not accessible by trains so the public transport consists of only urban and extra-urban buses. G. Galilei holds the role of access door of Tuscany region, and its importance grew with the introduction of low-cost flights. It is situated south of Pisa, very close to the urban area, and it is connected with highways and motorways, as well as by railway and urban and extra-urban public transports. We are also interested in the role played by the Guglielmo Marconi airport of Bologna, since this is a major competitor for this region.

We access information from the financial reports that are publicly available in the website of the "Toscana Aeroporti S.p.A" management company[2], which currently owns both these airports and has done so since June 1st, 2015. It was created through the merger of AdF - Aeroporto di Firenze S.p.A. (the management company of the Florence Amerigo Vespucci airport) and SAT - Società Aeroporto Toscano S.p.A. (the management company of the Galileo Galilei airport of Pisa). We access information for the G. Marconi airport of Bologna from the financial reports available in the website of "Aeroporto di Bologna"[3].

Galileo Galilei airport of Pisa In Pisa, we obtain data from 2008 until 2014 and we notice that during that period the number of passengers increased from almost 4 million to 4.7 million. We also notice a decrease in the total number of flights, with variations over the years of about 3-4 thousand flights (Table 6.1).

	Airport of Pisa						
	31/12/08	31/12/09	31/12/10	31/12/11	31/12/12	31/12/13	31/12/14
Total Passengers	3963717	4018662	4067012	4526723	4494915	4479690	4683811
Total Flights	42034	39461	39337	41676	41194	38961	38868

Table 6.1: Passengers and flights data from the airport of Pisa.

Amerigo Vespucci airport of Florence In Florence, we obtain data from 2008 until 2014 and we notice that the number of passengers increases from 1.9 million to 2.3 million over the years. We also note again a decrease in the total number of flights, with variations over the years of about 1-3 thousand flights (Table 6.2).

	Airport of Florence						
	31/12/08	31/12/09	31/12/10	31/12/11	31/12/12	31/12/13	31/12/14
Total Passengers	1928432	1688747	1737904	1906102	1852619	1983268	2251994
Total Flights	35429	31616	32018	33232	31769	31459	33976

Table 6.2: Passengers and flights data from the airport of Florence.

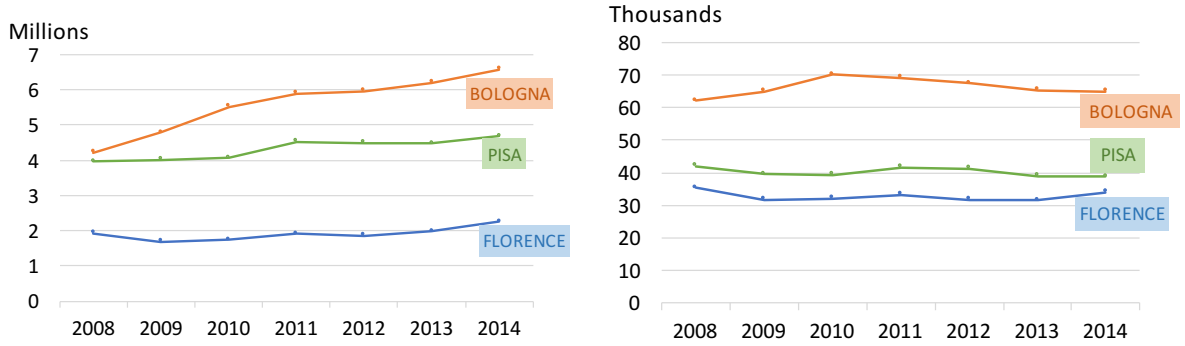
[2]http://www.toscana-aeroporti.com/en/home/investor-relations/financial-information/financial-reports.html
[3]http://www.bologna-airport.it/it/investor-relations/bilanci-e-relazioni.aspx?idC=62038&LN=it-IT

Guglielmo Marconi airport of Bologna In Bologna, we obtain data from 2008 until 2014 and we notice that during that period the number of passengers increased from 4.2 million to 6.6 million. We also notice an increase in the total number of flights, with variations over the years of about 3-8 thousand flights (Table 6.3).

	Airport of Bologna						
	31/12/08	31/12/09	31/12/10	31/12/11	31/12/12	31/12/13	31/12/14
Total Passengers	4225446	4782284	5511669	5885688	5958395	6193783	6580481
Total Flights	62042	64925	70270	69153	67527	65392	65058

Table 6.3: Passengers and flights data from the airport of Bologna.

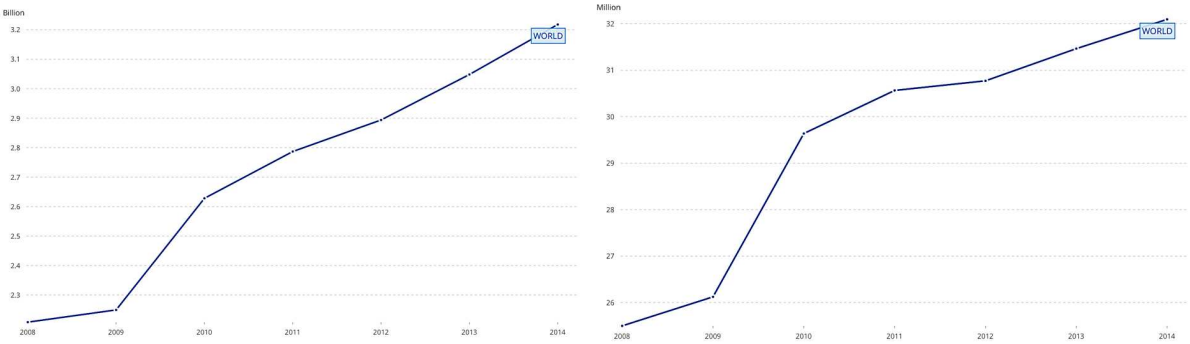
We observe in Figure 6.7 the data for passengers and flights for the three airports of Pisa, Florence and Bologna.



(a) (b)

Figure 6.7: Time evolution of the number of passengers in panel (a) and flights in panel (b) in the airports of Pisa, Florence and Bologna.

We compare the results with global characteristics and observe some similarities with the data of air transport worldwide (from the World Bank[4]), as seen in Figure 6.8. Air traffic, worldwide, keeps growing strongly, with 2.2 billion passengers carried in 2008 growing to almost 3.5 billion carried in 2014.



(a) (b)

Figure 6.8: Time evolution of the number of worldwide passengers in panel (a) and flights in panel (b).

[4]http://data.worldbank.org/

Population data

We obtain data for the official population of the municipalities and provinces of the Tuscany area in Italy from I.Stat[5], the statistical database currently produced by the Italian National Institute of Statistics. As can be seen in the Table 6.4, we have data from the last national census in Italy in 2011 as well as the data from the year 2014, as published on 01/01/2015, based on the calculations carried out on the population resident in each district of each province.

Province	Census of 9/10/2011	01/01/2015
Prato	245916	252987
Pistoia	287866	292509
Pisa	411190	421816
Arezzo	343676	346442
Siena	266621	270285
Grosseto	220564	224481
Lucca	388327	393478
Firenze	973145	1012180
Livorno	335247	339070
Massa-Carrara	199650	199406
Total	3672202	3752654

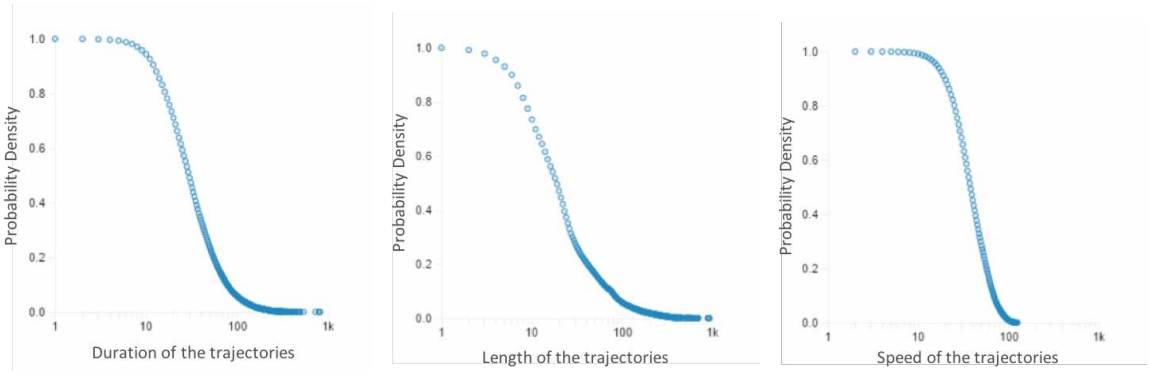
Table 6.4: Data regarding the official population of the provinces of Tuscany.

GPS Data

GPS data has been obtained from Octo Telematics Italia Srl, an insurance company who collect GPS trajectories for insurance purposes from on-board navigation systems in private cars, throughout the Tuscany region, in Italy.

We have datasets for vehicles for two distinct years, 2011 and 2014. The first dataset monitors almost 160,000 vehicles from 1/5/2011 to 31/5/2011, and the second monitors 250,000 vehicles for 7 weeks from 26/1/2014 to 16/3/2014. We break the raw trajectories following stops longer than 20 minutes and produce the final trajectories. We have 9.8 million trajectories for the dataset of 2011 and 18.9 million trajectories for the dataset of 2014. We assign each origin and destination point of the trajectories to the corresponding Italian census cell, using information provided by the Italian National Institute of Statistics (ISTAT). Based on the frequency of a person visiting a location, we define the most frequent location, L_1, the place where an individual has the highest probability to be found when stationary, as being their home.

[5]http://dati.istat.it/Index.aspx



(a) (b) (c)

Figure 6.9: Distribution of duration in panel (a), length in panel (b) and speed of the trajectories in panel (c).

Based on this mobility dataset, we can extract useful information, as displayed in Figure 6.9 where we see the distribution of duration, length and speed for the trajectories of the database of 2011. We can also calculate the population of each province based on the most frequent location (L_1) of each vehicle. We calculate the most frequent location of each vehicle based on the highest frequency of visit in our dataset, and we define it as the residence place of the user. As a result, we assign to each province a local population based on which we can calculate the number of people visiting the two airports in Tuscany from each of the provinces. In order to calculate the number of people that go to Pisa or Florence airport we split the population of Tuscany region into two parts using two different methods. First, we split the population based on their distance from each airport. If a passenger lives closer to Pisa airport (based on the corresponding value of L_1) then they are assigned to Pisa population, and if they live closer to Florence airport then they are assigned to Florence population (Table 6.5).

Population	People travelling to Pisa airport	People travelling to Florence airport
Pisa	1018497	36911
Florence	351624	365037

Table 6.5: Data regarding the split of the population based on their distance from each airport.

Secondly, we split the population based on their airport of preference. If a passenger prefers the Pisa airport then they are assigned to Pisa population, and if they prefer the Florence airport then they are assigned to Florence population (Table 6.6).

Population	People travelling to Pisa airport	People travelling to Florence airport
Pisa	1194320	72084
Florence	175801	329864

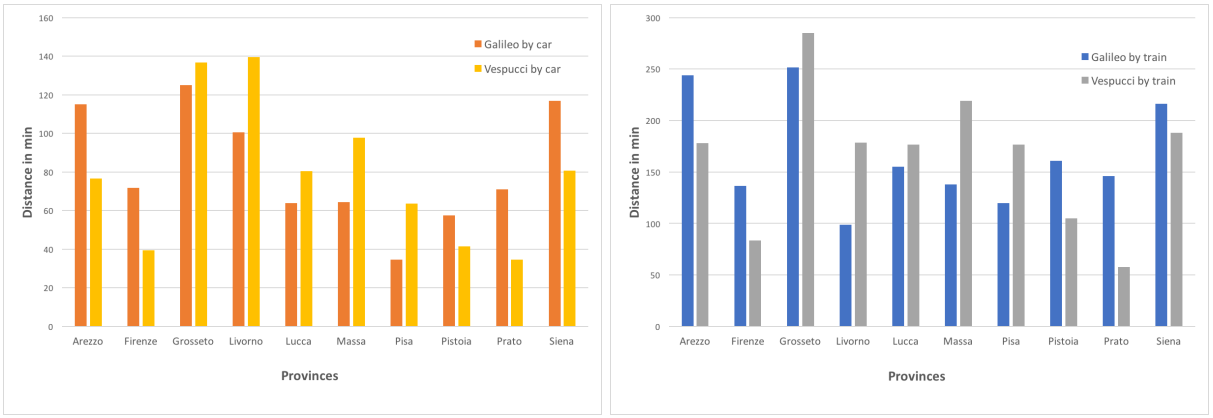
Table 6.6: Data regarding the split of the population based on their preference of airport.

Based on the data Table 6.6, the passengers who live closer (in geographical distance) to Pisa than to Florence, choose Pisa airport in approx 94% of the cases and choose Florence airport in approx 6% of the cases. Passengers who live closer (in geographical distance) to Florence than to Pisa, choose Pisa airport in approx 36% of the cases, and choose Florence airport in approx 64% of the cases. It is clear that most passengers choose Pisa, regardless of their distance from airports, so in the 3-dimensional model we let $d_1 < d_2$ to reflect this. In the 4-dimensional model $d_{11} < d_{21}$ and $d_{22} < d_{12}$. For example, $d_{11} = 0$, $d_{21} = 2$ (distance parameters for population that

lives closer to Pisa) and $d_{22} = 0$, $d_{12} = 0.9$ (distance parameters for population that lives closer to Florence) indicate that passengers prefer the airport that is closer to them, and the strength of this preference is stronger for people closer to Pisa than for people closer to Florence.

Distances between populations and airports

We calculate for each of the 10 provinces of Tuscany, their distance in time from the airports of Pisa and Florence with two distinct modes of transportation: by train and by car, as seen in Figure 6.10. We use the distances between populations and airports in our model in order to define whether the people choose an airport based on distance or attractiveness of the airport.



(a) Distance by car

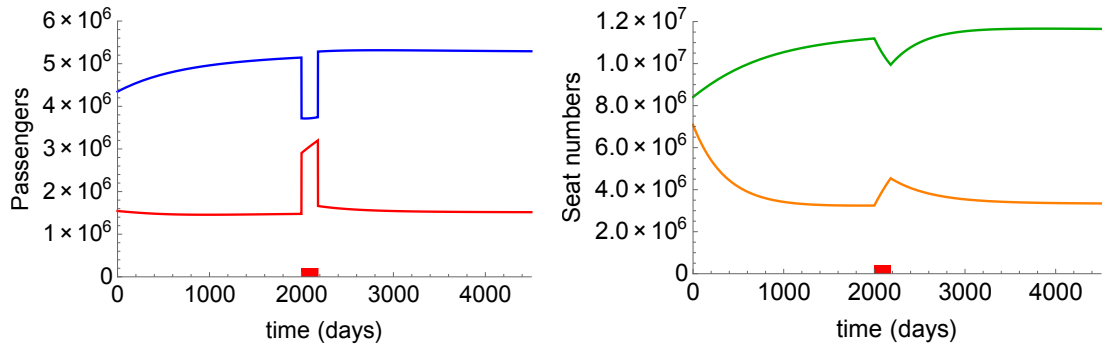
(b) Distance by train

Figure 6.10: Distance between provinces and airports of Tuscany region by car and train respectively.

6.2.6 Further modelling approaches

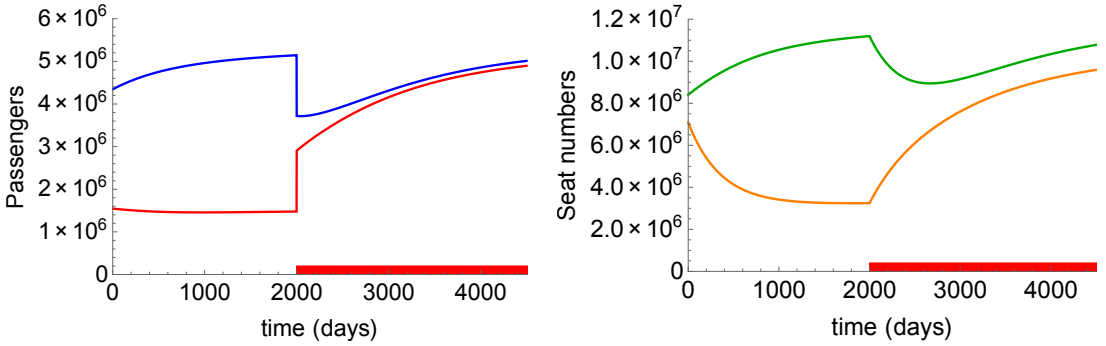
Interventions

We investigate the impact of interventions (e.g. investments or temporary closure) on the dynamics. We consider two airports and one population, with the assumption that airport 1 is more easily accessible than airport 2; to model this we utilize our model and let $d_2 = 1$, $d_1 = 0$. Airport 2 has the potential to bring in more customers by making an investment for a period of time (in our examples, for 6 months or for indefinite time) that involves doubling spending to accommodate their customers (that is, doubling e_2) and changing d_2 to 0 (that is, for the period of the investment they remove their disadvantage arising due to their larger distance). The impact of such interventions can be seen in Figure 6.11. Our results demonstrate that a temporary investment is capable of boosting customer numbers for a period of time (Figure 6.11 (a), (b)) but for long lasting improvement a permanent investment is needed, that proves beneficial in the long run (Figure 6.11 (c), (d)).



(a) investment for 6 months

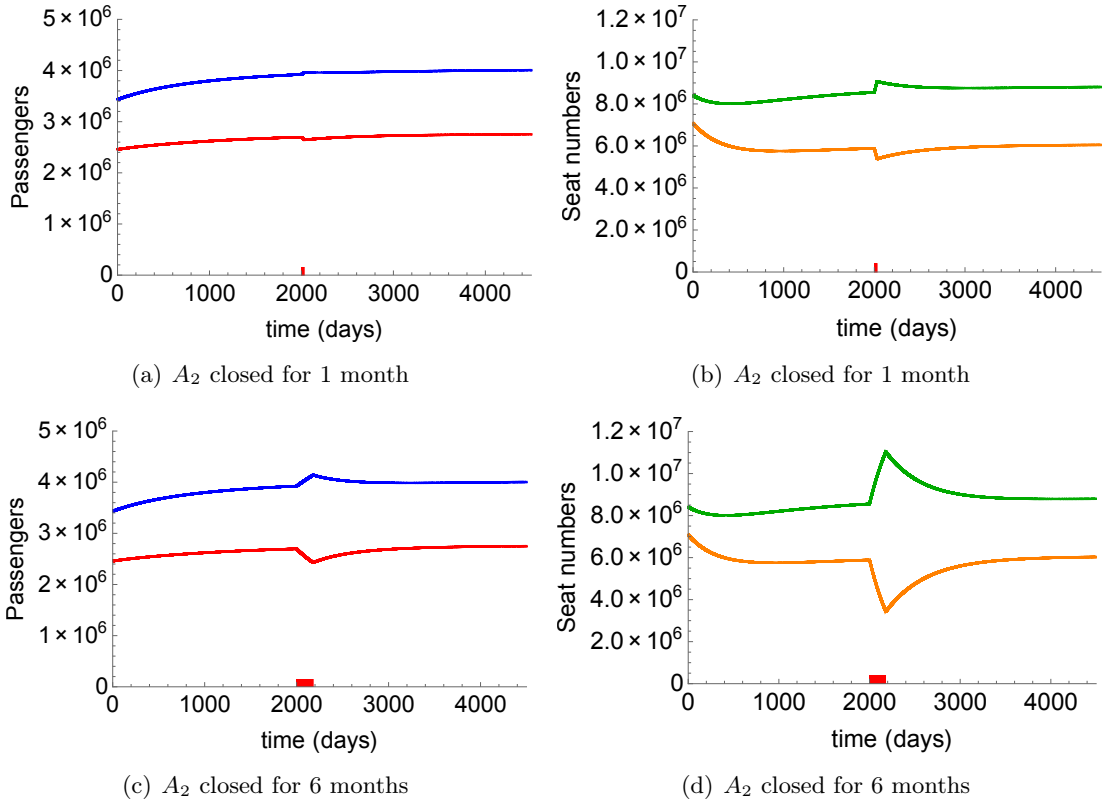
(b) investment for 6 months



(c) investment forever

(d) investment forever

Figure 6.11: Solution curves of the model produces time histories for one population and two airports, when airport 2 makes extra investment for a period of time of 6 months in panels (a) and (b) or forever in panels (c) and (d), starting at day 2000, as indicated on the horizontal axis. On the vertical axis: blue = passengers who go to airport 1, red = passengers who go to airport 2, green = A_1 (number of seats at Pisa airport), orange = A_2 (number of seats at Florence airport), on the horizontal axis: time in days t. Parameters are $r = 0.001$, $s = 0.003$, $m = 2.2$, $k_1 = k_2 = 0.9$, $e_1 = e_2 = 0.1$, $b = 30$, $h = 1.25 \times 10^{-8}$, $\alpha = 0.2$, $d_1 = 0$, $d_2 = 1$.



(a) A_2 closed for 1 month

(b) A_2 closed for 1 month

(c) A_2 closed for 6 months

(d) A_2 closed for 6 months

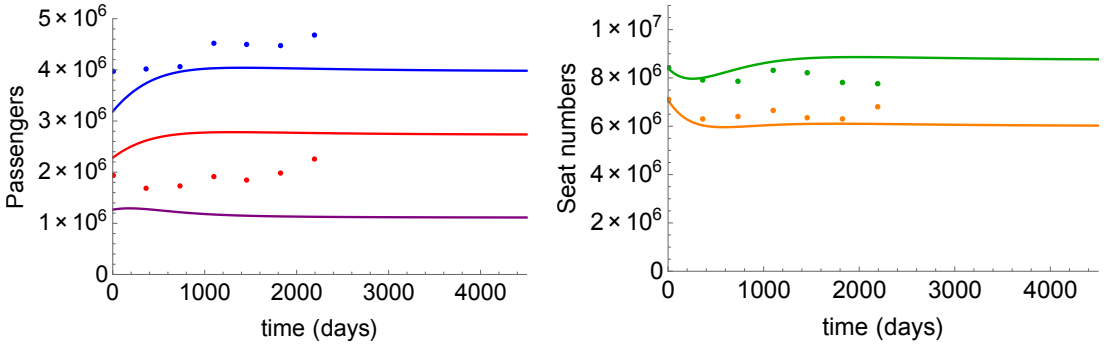
Figure 6.12: Solution curves of the model for one population and two airports, when airport 2 closes for a period of time of 1 month (first line of figures) or 6 months (second line of figures), starting at day 2000, as indicated on the horizontal axis. On the vertical axis: blue = passengers who go to airport 1, red = passengers who go to airport 2, green = A_1 (number of seats at Pisa airport), orange = A_2 (number of seats at Florence airport), on the horizontal axis: time in days t. Parameters are $r_1 = 0.001$, $s = 0.003$, $m = 2.2$, $k_1 = k_2 = 0.9$, $e_1 = e_2 = 0.1$, $b = 30$, $h = 1.25 \times 10^{-8}$, $\alpha = 0.2$, $d_1 = 0$, $d_2 = 0.3$.

If one of the airports needs to close down for a period of time then such action will clearly effect the dynamics of the growth of the other airports and the traveller population. We model such a scenario by assuming that airport 2 cannot serve customers for a period of time (in our examples, for 1 month and for 6 months). The impact of such action can be seen in Figure 6.12. For the period of time when airport 2 is closed, airport 1 takes all customers which greatly boosts its revenue (associated with seat numbers). After airport 2 reopens its revenue and customer numbers recover over time, however the longer the closure the longer recovery takes.

Incorporating a third airport: Bologna

Although Bologna is located outside of Tuscany, our data indicates that this airport is popular amongst some residents of Tuscany. We show that the fit presented in Figure 6.5 can be significantly improved when we incorporate a third airport. In the simulations presented in Figure 6.13 we made the assumption that approximately 20% of passengers of Bologna airport come from Tuscany, where passenger data and flight numbers for Bologna airport were drawn from Table 6.3. Comparing the fit in Figure 6.5 to that in Figure 6.13 indicates that the attracting power of airport(s) outside Tuscany maybe an important factor to incorporate in our analysis. Note that we do not aim to model the time evolution of passenger numbers in Bologna airport, the sole purpose of incorporating this airport is to examine whether this results in an improved

fit to data from Pisa and Florence.



(a) (b)

Figure 6.13: Solution curves of the model (6.3) for one population and three airports, fitted to real data drawn from Tables 6.1, 6.2 and 6.3. On the vertical axis: blue = passengers who go to airport 1, red = passengers who go to airport 2, purple = passengers who go to airport 3, green = A_1 (number of seats at Pisa airport), orange = A_2 (number of seats at Florence airport), on the horizontal axis: time in days t. Continuous lines indicate solution curves of the model and dots of the same colour indicate real data obtained in the yearly breakdown between 2008–2014. Parameters are $r = 0.001$, $s = 0.003$, $m = 2.2$, $k_1 = k_2 = k_3 = 0.9$, $e_1 = e_2 = e_3 = 0.1$, $b = 30$, $h = 1.25 \times 10^{-8}$, $\alpha = 0.2$, $d_1 = 0$, $d_2 = 0.3$, $d_3 = 1$.

6.2.7 Conclusions

In this work we have investigated how the number of customers of airports in Tuscany, Italy changes over time. A mathematical model reproduced real data of customer numbers and the mathematical analysis provided insight into what are the key factors that drive the dynamics.

Although this research was motivated by the need to understand how different airports are able to attract individuals and increase the number of their customers, the model formulation allows for this concept to be extended to more general problems. Airports in our model can be replaced by any retail centres or places of attraction. The model was constructed in a general way, such that it incorporates any number of sub-populations residing in different locations, and any number of airports (more generally, source of attraction). Analysis of simple cases have revealed the conditions for airports to be able to coexist, showing that cooperation of airports to meet this condition has been proved to be beneficial.

We also used our model to explore the effect of various interventions, such as a temporary closure of an airport, short- and long-term investment, and competition with other airports. Thus, this research has the potential to be incorporated into economic strategic planning, as the general framework and results allow for predictive analysis regarding the outcome of various interventions.

6.3 Using Mobility Data Analysis to Evaluate Effects of Industrial Clusters on Regional Dynamics - Ongoing Work

6.3.1 Introduction

What is a city? How does it evolve and can it die? These are central questions in a rapidly urbanising world. Cities are characterized by concentrating population, economic activity and services and they are quite fascinating. In 1950 only 30% of the word's population lived in cities. While developed countries are already highly urbanised with about 80% of the population living in cities, it is estimated that same will be true for the whole world by around the year 2050. 2 billion people will migrate to cities in next few decades, especially in China, India, Southeast Asia and Africa [160]. So this emphasises that it's really dramatic and crucially important to understand how cities work.

Over the past decades, the former president of the interdisciplinary Santa Fe Institute, Geoffrey West, together with a number of colleagues, investigated the growth of cities [36]. They looked at some low granularity data on crime, wealth and other aspects which describe cities. They were able to find that simple mathematical laws govern the properties of cities. Basically, given these laws, it is possible to describe properties like crime, average walking speed, health and many other quantities by one individual property. This is the population of a city. By investigating how this behaves across a number of cities of different size, they were able to find out to which extent these quantities scale with city growth.

In an urban world, it is of utmost importance to understand the processes that shape the growth of the cities and metropolitan regions. There is a prevailing paradigm of a mechanistic approach to the city development, where city is observed as a giant machine which needs to be controlled and directed – as obvious from Big Data and Smart Cities trend. This is especially true for fast developing nations such as China [37]. The mechanistic approach might be an urban engineer's dream, with a promising future of city management offices full of big screens showing huge amounts of data, but to truly develop cities is to allow them to evolve and self-organize. The mechanistic top-down approach emphasizes too much on control, what can be counter-productive in a case of many urban processes.

The opposite approach is the evolutionary approach, where terms such as ecosystem or nervous system are used to describe a city. Prominent field in this kind of approach to cities, and metropolitan regions in general, is Evolutionary Economic Geography. A city is an adaptive complex system whose dynamics depend on interactions of various heterogeneous agents, on all levels of the hierarchy. Therefore, a city is a result of bottom-up processes.

We are looking to reconcile these two approaches. Building upon ideas from Evolutionary Economic Geography, we observe urban systems as complex adaptive systems open to the evolutionary process. Economic system is population of rules, a structure of rules, and a process of rules [104], where rules are related to both physical technologies (metal machining, diesel engines, nano chips, etc.) and social technologies (the rule of law, money, private enterprise, etc.). Thus, the new rules emerge and evolve in the context of technological innovation and new ways of production of physical goods, as well as in the context of social innovation and the introduction of new rules of organisation and behaviour [255]. Cities, or more precisely, various clusters in and around urban areas are hotspots of organisational diversity and new rules creation.

Main goal of this study is to explore how to use mobility data to get further insights in the dynamics of industrial clusters and cities as a whole. Populations of firms exist in some space and time, and they are usually clustered in urban space. Firms create jobs and new firms, attracting labour and capital to its centres. During systematic shocks, like economic crises, it is expected that some clusters will shrink, while others will prove to be resilient. Their collapse, i.e. bankruptcy can lead to failure of the economy of the city - therefore inherent systemic risk. The

most famous case being Detroit, as a city focused on one big industry. So when that industry went downhill, so did the city. Instead of waiting for census data to understand processes that occurred during various stages of economic cycles, we can reach to Big Data and maybe even develop a real time system which will help us to feel the pulse of a city. We study the industrial clusters for the Veneto region in Italy from pre-crisis year (i.e., 2008) and one year post-crisis (i.e., 2010). We would also like to explore if the presence and typology of industrial clusters could affect the mobility of people.

6.3.2 Cluster Definition and Veneto Region

Clusters have been a hot topic of discussion for the last 25 years, ever since Porter introduced the term [287]. But even long before, the phenomenon of companies sticking close to each other and benefiting from it had been observed and analysed [235]. Clusters can provide a special environment for businesses. Knowledge seems to be flowing in the air, highly qualified labour pools, cooperation occurs not only along the value chain but also between competitors [235, 288]. The presence of clusters seems to be a primary driver of growth in employment and in other performance dimensions across essentially all regions and clusters [99]. But what exactly causes those positive effects? What other effects are there? Can the knowledge about clusters be utilized to positively effect the development a region? All those questions are highly relevant, not only in the academic discussion, but especially for decision makers in companies and governments.

There are several definitions of what an industrial cluster is. Enright in 1996 gave his definition "A regional cluster is an industrial cluster in which firms are in close proximity to each other.". Around the same time, Swann and Prevezer gave theirs "Clusters are defined as groups of firms within one industry based in one geographical area.". In 1997, Rosenfeld gave a slightly more detailed definition "A cluster is very simple used to represent concentrations of firms that are able to produce synergy because of their geographical proximity and interdependence, even though their scale of employment may not be pronounced or prominent." while in 1998 Feser says "Economic clusters are not just related and supporting industries and institutions, but rather related and supporting institutions that are more competitive by virtue of their relationships.". Roelandt and den Hertag in 1999 "Clusters can be characterized as networks of producers of strongly interdependent firms (including specialized suppliers) linked each other in a value-adding production chain." while Crouch and Farrell in 2001 speak about the concept of a cluster, "The more general concept of cluster suggests something looser: a tendency for firms in similar types of businesses to locate close together, though without having a particular important presence in an area.". But despite having so many definitions, the term cluster is anything but well defined. There seems to be a rough consensus on interconnected (linked / interdependent) entities within one industry (in a particular field / in similar types of business) which need to be in a geographic proximity (geographic concentration / geographic area).

The Veneto region in Northern Italy has been in the focus of research from the scratch, as it has a long history of economic strength. It has been the trade capital of the world from the 13th to the 15th century, and since the 1970s Veneto has risen to one of the richest and most industrialized regions of Italy. The unique business environment of northern Italy, namely a great quantity of small and medium-sized companies simultaneously competing and cooperating, spurred such a remarkable economic success around the 1980s that it became a classic example.

The Italian clusters, or *distretti industriali* (industrial districts) as they are called in Italian, are of such relevance that in classifications they even get a category to themselves (see [234] on Italian Clusters) from time to time. But the boom passed, and the economy was hit hard by several crises. The introduction of the Euro in 2000 and the financial crisis put the concept to a hard test [5]. The crisis of 2009 affected Veneto's economy strongly as the industrial production in the second quarter of 2009 fell by 19.5% from its level at the same time of the previous year

(Padova' s production even decreased by 27.9%).

Despite the high amount of data available on the Italian industrial districts, heretofore the information is very fragmented, concentrating more on the single cluster than the overall development, and incomparable due to non-standardized cluster definitions. The Italian census of 2001 adduces 34 industrial districts in Veneto for 1991 and 22 for 2001[6]. In the census of 2011 28 districts are stated[7].

Following-up to the new Regional Law on Clusters approved in May 2014, Veneto Region has re-organised its ecosystem and has officially recognised 17 clusters[8].

According to the new Law:

- A cluster is a local productive system, located in a specific area of the region and characterised by an elevated concentration of manufacturing and industrial companies, especially SMEs, operating in specific or connected production chains, relevant for the regional economy. In order to identify the clusters, the Region took into account elements like the presence of a documented history, the competitiveness in the international markets, the integration of vocational training and research centres and the presence of a cluster brand.

- An innovation network is a system of companies and public/private bodies located in the region but not necessarily next to each others, which can operate in different sectors and which are able to develop a coherent set of initiatives and projects relevant for the regional economy.

- An aggregation of companies is a group of at least 3 companies which cooperate with the aim of developing a common strategic project.

The decision has been taken by the Regional Government, who has also defined the geographical and sectoral scope of each cluster. The recognised clusters have been reduced from 40 to 17, according to criteria taking into account their history, the geographical localisation of an elevated number of industrial and handicraft companies operating in the same production chain and the competitiveness (also potential) on national and international markets. The traditional clusters deserve to be valorised for what they have represented in the regional and European scene, and because they have characterized the history of the productive system in Veneto.

The 17 industrial clusters recognised by the Region are the following[9]:

- Calzatura della Riviera del Brenta (Footware)
- Concia di Arzignano (Tanning)
- Meccanica dell'Alto Vicentino (Mechanical)
- Mobile del Livenza (Furniture)
- Occhialeria Bellunese (Eyewear)
- Orafo Vicentino (Goldsmith)
- Calzatura tecnica ed articoli sportivi di Asolo e Montebelluna (Sporting Goods)
- Ceramica artistica di Nove e Bassano del Grappa (Artistic Pottery)
- Elettrodomestici ed inox di Conegliano e del Trevigiano (Household Electrical Appliances)

[6] http://statistica.regione.veneto.it/strumenti_codifiche_classificazioni.jsp

[7] http://www.osservatoriodistretti.org/category/regione/veneto

[8] https://web.archive.org/web/20161107005835/http://www.alpclusters2020.eu/authors/Chiara-Beltrame/Simplification-cluster-ecosystem-Veneto-Region

[9] http://venetoclusters.it

- Condizionamento e refrigerazione del Padovano (Air-conditioning and Refrigeration)
- Giostra del Polesine (Carousel)
- Ittico del Polesine e del Basso Veneziano (Fishing)
- Marmo e pietra del Veronese (Marble and Stones)
- Mobile classico della Bassa Veronese (Furniture)
- Prosecco di Conegliano e Valdobbiadene (Wine)
- Vetro artistico di Murano e vetro del Veneziano (Artistic Glass)
- Vino della Valpolicella e Soave (Wine)

In this study, we identified and examined the data sources on the 17 clusters in the region Veneto. The available data was aggregated and visualized. The influential factors on cluster development *concentration*, *employment*, *specialization*, *productivity*, *export-orientation* and *homogeneity* were deduced from basic economic data and were used for a comparison of the clusters.

Despite the fact that the clusters of Veneto are famous and frequently used as an example in academic literature, the amount and quality of the data freely available is not sufficient to adequately describe the clusters and evaluate their development. One reason behind this is the vague and frequently changing definition of clusters. Thus clusters in different publications even with the same name differ in the geographic extend and the included companies. Furthermore, qualitative data has a large share in the definition and description of the clusters, making comparisons even more difficult. Another circumstance that makes the task of analysing the clusters of Veneto more challenging is that the data and its description usually only are provided in Italian language.

The Veneto region has 7 provinces and each of them has several municipalities (comuni), Venezia with 44 municipalities, Belluno with 67 municipalities, Padova with 104 municipalities, Rovigo with 50 municipalities, Treviso with 95 municipalities, Verona with 98 municipalities and Vicenza with 121 municipalities, for a total of 579 municipalities in the whole region.

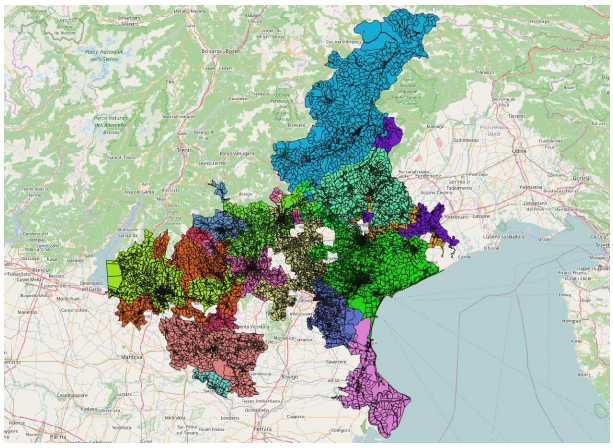


Figure 6.14: The official industrial clusters of Veneto region.

6.3.3 Proposed Approaches

In the first stage of our research we use ecosystemic approach to identify and categorize industrial clusters in Veneto region in Italy, by size and population dynamics based on official clusters data. Following that, we collect census data and analyse industrial dynamics before and

after the economic crises. While this approach has dynamics as its central topic, it completely depends on the census data, so our understanding of the ecosystems will always have certain time lag. To solve that problem we look at Big Data and its possibilities in understanding the processes that shape our cities [31]. We build upon ideas from social physics and from Bettencourt's observation [37] that cities are new kind of complex systems, part stars, part networks.

Therefore, in the second phase of our research we use mobility data to develop a novel method for identifying industrial clusters and measure their attraction. In particular, we are interested in mobility of individuals that derive from GPS trajectories. GPS traces contain geographical and temporal information regarding human mobility and they show great potential as a base for empirically investigations on human dynamics on a society wide scale. More specifically, we can study human movements in the territory in order to detect the poles of attraction during work hours, that could be a very good indicator about the industrial clusters.

In the last phase, we combine GPS traces with official industrial data, that could help us identify these poles of attractions, as industrial clusters, but also more specifically, which industrial cluster and whether it's a new or already known cluster.

It's important to mention that the work is not yet finished. For that reason, we present three different approaches. The first one makes use of networks, the second of territory annotation and the last one of the classification definition.

Networks Approach

Step 1: Mobility Networks

First, we construct different mobility networks where each of the municipalities is represented as a node. Directed, weighted edges have been set between each node A and B for the following cases leading to six networks:

- the no. of trajectories from A to B
- no. of users travelling at least once from A to B
- $\frac{\text{no. of trajectories from A to B}}{\text{no. of total outgoing trajectories from A}}$
- $\frac{\text{no. of trajectories from A to B}}{\text{no. of total ingoing trajectories to B}}$
- $\frac{\text{no. of users traveling at least once from A to B}}{\text{no. of total users starting at least once from A}}$
- $\frac{\text{no. of users traveling at least once from A to B}}{\text{no. of total users ending at least once at B}}$

In case there is eg. no trajectory, no edge was created. Hence, the weight of edges is always $w > 0$.

Step 2: Industry Network

In the second step, we construct the industrial network where each of the municipalities is represented as a node. Directed edges with weights $P(A, B)$ have been set between each node A and B creating one network for each sector (eg. manufacturing, fishing,...). The weights have been set using the Radiation Model, describing the flow of people between different locations [326]. In order for the radiation model to calculate the probability of a single trip to happen between two municipalities - nodes, we calculate the geographical coordinates of the centre of each municipality as well as its population, interpreted as number of people working in this municipality.

$$P(A,B) = \frac{W(A) \times W(B)}{(W(A) + W(R)) \times (W(A) + W(B) + W(R))}$$

$$W(A) = \text{no. of workers in this sector in municipality A;}$$

$$W(R) = \text{no. of workers in all municipalities without A and B}$$

Step 3: Topological Features

We calculate the betweenness centrality for each of the two networks, as well as the weighted in-degree.

- *Betweenness centrality* $c_B(e)$ of an edge e is the sum of the fraction of all-pairs shortest paths that pass through e:

$$c_B(e) = \sum_{(s,t) \in V} \frac{\sigma(s,t|e)}{\sigma(s,t)}, \tag{6.5}$$

  where V is the set of nodes,$\sigma(s,t)$ is the number of shortest (s,t)-paths, and $\sigma(s,t|e)$ is the number of those paths passing through edge e.

- The *weighted in-degree* $\Gamma(u,w)$ of a node is like the degree. It's based on the number of edges for a node, but ponderated by the weight of each edge. The degree is the sum of the edge weights adjacent to the node.

Step 4: Correlation Measures

To compare the mobility network with the industrial network we calculate the correlation between the two networks using Pearson product-moment correlation coefficient and Spearman's rank correlation coefficient.

In statistics, the *Pearson correlation coefficient*, also referred to as Pearson's r, is a measure of the linear correlation between two variables X and Y. It has a value between $+1$ and -1, where 1 is total positive linear correlation, 0 is no linear correlation, and -1 is total negative linear correlation. It is widely used in the sciences and it was developed by Karl Pearson from a related idea introduced by Francis Galton in the 1880s.

The *Spearman correlation* between two variables is equal to the Pearson correlation between the rank values of those two variables; while Pearson's correlation assesses linear relationships, Spearman's correlation assesses monotonic relationships (whether linear or not). If there are no repeated data values, a perfect Spearman correlation of $+1$ or -1 occurs when each of the variables is a perfect monotone function of the other. Intuitively, the Spearman correlation between two variables will be high when observations have a similar (or identical for a correlation of 1) rank (i.e. relative position label of the observations within the variable: 1st, 2nd, 3rd, etc.) between the two variables, and low when observations have a dissimilar (or fully opposed for a correlation of -1) rank between the two variables.

Step 5: Validation

We generate a null model with the mobility network's values constant and the industrial networks values randomly shuffled. We calculate the correlation again as in the previous step, and we repeat for 100 times. We want to verify that the networks are actually correlated and that the correlation has an explanation and it's not a random event.

Work-home Approach

In this approach, we are interested in identifying the nature of every municipality. Based on the number of incoming edges annotated as *tripstowork*, *tripstohome* or *tripstoother*, every municipality will be characterised as *working* or *living* area.

Step 1: Work-home Annotation

With the help of the GPS trajectories (see section 6.3.4), we can characterise the population of each municipality. We calculate the most frequent location (L_1) of each vehicle based on the highest frequency of visit in our dataset, and we define it as the home place of the user, and the second most frequent location (L_2) of each vehicle and we define it as the work place of the user. We calculate the entropy for each municipality, with entropy close to 1 meaning a balanced area of working and living people, and entropy close to 1 meaning that either work or home is prevalent. As a result, few municipalities will rise as 'working' or 'living' areas.

Step 2: Industrial Sectors

With the help of the industrial statistical data (see section 6.3.4) we apply a similar approach, where based on the number of active enterprises as well as the number of persons employed in each industrial sector, we identify concentration of enterprises of the same field in specific areas and we annotate them with a specific industrial sector label.

Using the annotation of Step 1, each 'working' area will be assigned to one or more sectors.

Step 3: Evaluation

For each industrial sector, we compare the annotated working areas with the official clusters. We have to point out that each industrial sector can contain one to many working municipalities. So it's possible to have a partial match between the official clusters and an industrial working area of ours.

There are two possible outcomes:

- Either we agree with the official clusters, and we obtain an almost real-time estimation of their evolution
- Either we disagree, so we establish a new definition/model based on these alternative sources of data

Classifier Approach

Step 1: Classifier

We use the official 17-cluster definition as ground truth and we train binary classifiers $C1..C17$ to determine if municipality M belongs to Cx. We test the following classifiers:

- *Logistic Regression* The binary logistic model is used to estimate the probability of a binary response based on one or more predictor (or independent) features. It allows one to say that the presence of a risk factor increases the odds of a given outcome by a specific factor. The model itself simply models probability of output in terms of input, and does not perform statistical classification (it is not a classifier), though it can be used to make a classifier, for instance by choosing a cut-off value and classifying inputs with probability greater than the cut-off as one class, below the cut-off as the other.

- *K-NN* The k-nearest neighbours algorithm (k-NN) is a non-parametric method used for classification. The input consists of the k closest training examples in the feature space. In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbours, with the object being assigned to the class most common among its k nearest neighbours (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbour.

- *Support Vector Machine (SVM)* Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting). An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall. SVM algorithms use a set of mathematical functions that are defined as the kernel. The function of kernel is to take data as input and transform it into the required form. We will use kernels that are linear, poly and sigmoid.

- *Naive-byes* Naive Bayes classifiers are a family of simple 'probabilistic classifiers' based on applying Bayes' theorem with strong (naive) independence assumptions between the features. Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. When dealing with continuous data, a typical assumption is that the continuous values associated with each class are distributed according to a Gaussian distribution, and that's the one we will use.

- *Decision Tree Classifier* Decision tree learning uses a decision tree (as a predictive model) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). Tree models where the target variable can take a discrete set of values are called classification trees; in these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels.

- *Random Forest Classifier* Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.

We use as features:

- in municipality M:
  - Official population
  - No of active enterprises in each of the 36 industrial sectors
  - No people employed in each of the 36 industrial sectors
  - No of people living (from mobility data) - x_h
  - No of people working (from mobility data) - x_w
  - home-work classification: $x_w - x_h$ (see Work-home Approach)
  - No of trajectories ending in M

  - No of trajectories starting from M
  - No of people working (from mobility data) weighted by distance travelled from home, where distance is:
    * Distance between municipality M marked as work place and home place (real distance)
    * Distance between municipality M marked as home place and work place (real distance)
    * Distance between municipality M and any other municipality (real distance)

- for each of the municipalities $N \neq M$
  - Official population
  - No of active enterprises in each of the 36 industrial sectors
  - No people employed in each of the 36 industrial sectors
  - distance between N and M
  - co-workers index between M and N on top of the mobility network (calculated by radiation model, see Networks Approach)
  - No of trajectories from N to M
  - No of trajectories from M to N
  - No of people travelling from N to M
  - No of people travelling from M to N

Step 2: Evaluation

First, we apply a simple *cross-validation* technique to evaluate each classifier. Cross-validation is when you reserve part of your data to use in evaluating your classifier. We split the data and we take 75% to be used for training, and then the remaining 25% of the data to evaluate each classifier's performance. The reason we need different data for training and evaluating the model is to protect against *overfitting*, one of the most fundamental problems in machine learning. If we optimize the model for the training data, then our model will score very well on the training set, but will not be able to generalize to new data, such as in a test set. When a model performs highly on the training set but poorly on the test set, this is known as *overfitting*, or essentially creating a model that knows the training set very well but cannot be applied to new problems. Therefore, the standard procedure for hyperparameter optimization accounts for overfitting through cross-validation.

Then we combine the multiple classifiers using *vote based classifier ensemble technique*, with soft voting. In order to determine the optimal settings we could try many different combinations to evaluate the performance of the voting classifier. However, evaluating it only on the training set can lead to *overfitting*. For that reason we evaluate the classifier using *k-fold cross validation*, normally with 10 folds and if data points are less than 15 then 5 folds.

We use *mean* and *standard deviation* as accuracy metrics in all the cases.

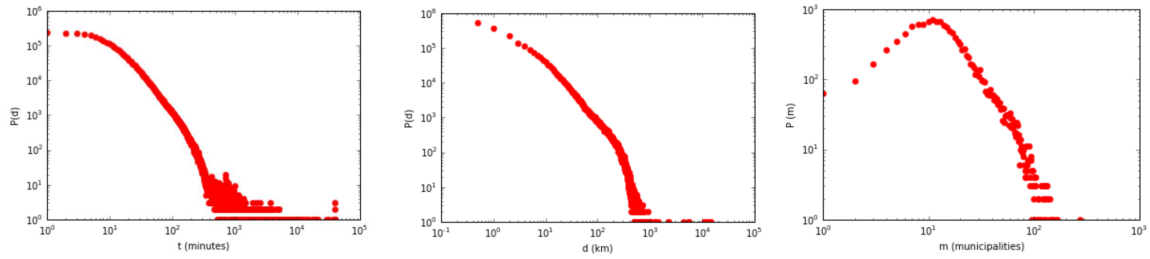
6.3.4 Data

GPS Data

We are studying mobility data - GPS trajectories, collected by Octo Telematics Italia Srl, an insurance company who collect GPS trajectories for insurance purposes from on-board navigation systems in private cars, in the region of Veneto in Italy. The dataset is divided into two parts of

two different periods of time. The first is from February 2008 and includes 13,000 vehicles and the second is from February 2010 and includes 19,000 vehicles. We break the raw trajectories following stops longer than 20 minutes and produce the final trajectories, 2.5 million trajectories for the dataset of 2008 and 3.5 million trajectories for the dataset of 2010.

The trajectories are structured like this : User ID, Trajectory ID, Time of start, Time of end, Location of start, Location of end (e.g. 180, 1, 2010-02-01 07:51:12, 2010-02-01 10:03:03, 25012_7 , 25016_41). The spatial resolution is on cell-census section level (Sezioni di censimento in Italian), that is the minimum unit of a municipality based on which is organized the census survey. We assign each origin and destination point of the trajectories to the corresponding Italian census cell, using information provided by the Italian National Institute of Statistics (ISTAT). Based on the frequency of a person visiting a location, we define the most frequent location, L_1, the place where an individual has the highest probability to be found when stationary, as being their *home*, and L_2, the place where an individual has the highest probability to be found during working hours, as being their *work*.



(a) Duration (b) Length (c) Number of municipalities

Figure 6.15: Distribution of duration in panel (a), length in panel (b) and number of municipalities visited in panel (c).

Based on these mobility data, we can extract much useful information, such as the distribution of duration, length and number of municipalities visited for the trajectories (see Fig. 6.15).

Industrial Data

Our second dataset consists of industry data from Business Register ASIA, a register that is yearly updated through a process of integration of administrative and statistical sources, for statistical analysis of the business population and its demography. We have data for each municipality of Veneto region, with respect to 36 different industrial sectors regarding number of businesses per municipality as well as number of people working in these businesses[10]. We have industrial sectors that cover from agriculture, manufacturing to construction, transportation and IT services. More specifically, we are interested in the number of active enterprises and the number of persons employed in these active enterprises, for each of the municipalities of Veneto. In Table 6.7, you can see the numbers aggregated per province.

[10]http://dati-censimentoindustriaeservizi.istat.it/Index.aspx?lang=en

Territory	Number of active enterprises	Number of persons employed in active enterprises
Veneto	**403,169**	**1,642,359**
Verona	75,408	320,553
Vicenza	70,983	315,527
Belluno	14,971	63,523
Treviso	71,734	291,967
Venezia	66,127	257,504
Padova	84,031	329,401
Rovigo	19,915	63,884

Table 6.7: Number of active enterprises and persons employed in these active enterprises for the provinces of Veneto region.

6.3.5 Experiments and Preliminary Results

We calculate the betweenness centrality and the weighted in-degree between each of the mobility networks and the industrial one. In Fig. 6.16, you can see the scatterplot of the values for the nodes of the first mobility network (no of trajectories) and the industrial one.

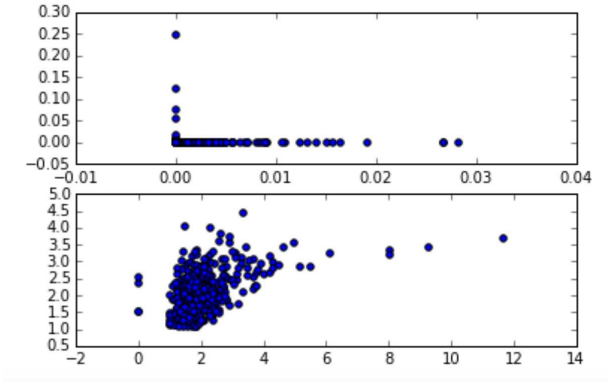


Figure 6.16: Betweenness centrality and weighted in-degree for the two networks

At this point, we calculate the correlation between the two networks using Pearson product-moment correlation coefficient and Spearman's rank correlation coefficient. We mention the correlations for the first mobility network (no of trajectories): $r_p = 0.653$ and $r_s = 0.637e^{-16}$ and in Fig. 6.17, you can see the scatterplot of the values for the nodes of the two networks.

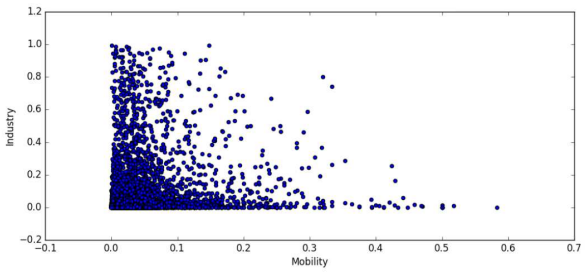


Figure 6.17: Pearson Correlation between the mobility network and the industry network.

We generate a null model with the mobility network's values constant and the industrial network's values randomly shuffled. We calculate the correlation again as in the previous step, and we repeat for 100 times.You can see the distribution of the values of the null model compared with the actual value, in the Fig. 6.18 (for two different industrial sectors).

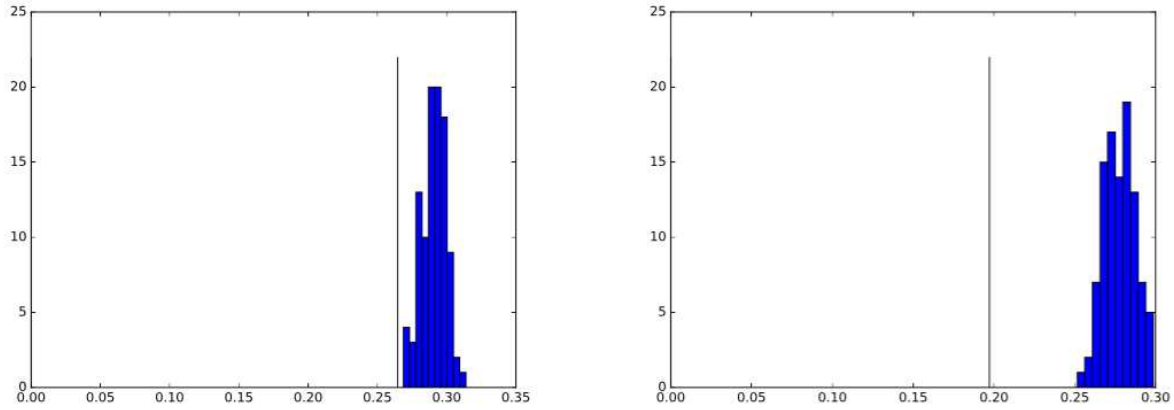


Figure 6.18: Distribution of the values of the null model compared with the actual value

It's clear that this approach doesn't work. That's the reason we considered other solutions.

Work-home Approach

In Fig. 6.19 we can see the annotation of the territory where the blue areas are the 'working' areas and the red are the 'living' areas, with stronger colour indicating the high prevalence of work or home places. In the rest of the territory, there is no high prevalence of either characterisation.

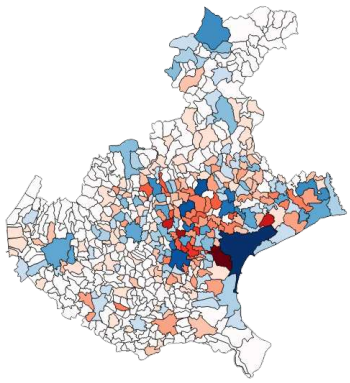
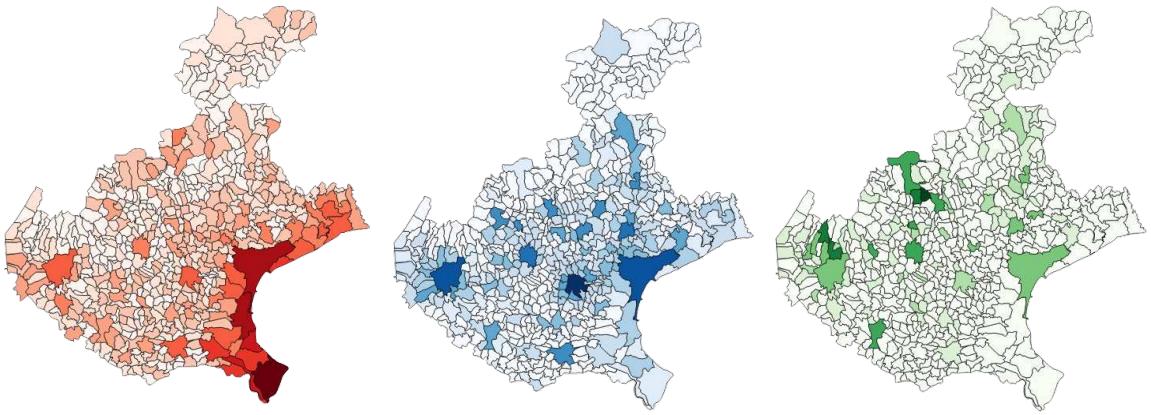


Figure 6.19: 'Working' and 'living' municipalities.

In Fig. 6.20 we can see few examples of the annotation of the territory based on the industrial sectors that are prevalent in each municipality.



(a) Agriculture (b) IT Services (c) Mining

Figure 6.20: Industrial sectors into the territory.

Classifier Approach

We have good results regarding each single classifier but not if we put all the 17 binary classifiers all together. In Tables 6.8 and 6.9 we can see the mean and standard deviation values for all the classifiers mentioned above, while in FIgure 6.21 we can see the mean and standard values for the voting classifier ensemble technique.

Our main limitation is the data size, as for so many features, we got maximum 120 data points for cluster 4 and minimum 12 for cluster 11.

Cluster	logreg	K-NN	Svm-poly	Svm-sig	Svm	Naive-bayes	Dec-tree	RFC
1	0.825	0.692	0.667	0.767	0.7	0.625	0.742	0.658
2	0.738	0.633	0.553	0.475	0.763	0.518	0.575	0.598
3	0.772	0.697	0.485	0.722	0.763	0.49	0.663	0.572
4	0.661	0.666	0.486	0.548	0.641	0.464	0.608	0.644
5	0.747	0.678	0.528	0.726	0.722	0.664	0.769	0.771
6	0.682	0.527	0.53	0.582	0.602	0.432	0.457	0.523
7	0.792	0.608	0.533	0.533	0.817	0.558	0.683	0.483
8	0.58	0.437	0.464	0.557	0.53	0.414	0.552	0.532
9	0.738	0.685	0.538	0.655	0.733	0.58	0.608	0.518
10	0.726	0.612	0.499	0.694	0.656	0.448	0.628	0.59
11	0.7	0.7	0.6	0.9	0.6	0.4	0.6	0.9
12	0.333	0.667	0.633	0.533	0.533	0.567	0.433	0.533
13	0.676	0.546	0.572	0.648	0.656	0.609	0.559	0.612
14	0.7	0.517	0.576	0.595	0.733	0.505	0.7	0.662
15	0.659	0.628	0.48	0.53	0.551	0.522	0.617	0.796
16	0.703	0.517	0.56	0.742	0.643	0.568	0.545	0.552
17	0.58	0.553	0.53	0.617	0.497	0.617	0.623	0.58

Table 6.8: Mean values for the classifiers.

Cluster	logreg	K-NN	Svm-poly	Svm-sig	Svm	Naive-bayes	Dec-tree	RFC
1	0.225	0.245	0.274	0.238	0.245	0.202	0.195	0.267
2	0.163	0.165	0.069	0.115	0.142	0.237	0.279	0.157
3	0.123	0.21	0.084	0.177	0.165	0.03	0.18	0.215
4	0.126	0.194	0.124	0.078	0.175	0.11	0.199	0.157
5	0.111	0.104	0.037	0.123	0.133	0.09	0.179	0.085
6	0.121	0.111	0.082	0.099	0.146	0.068	0.188	0.144
7	0.272	0.35	0.067	0.067	0.229	0.325	0.263	0.32
8	0.163	0.186	0.059	0.208	0.187	0.076	0.135	0.13
9	0.167	0.196	0.211	0.238	0.2	0.22	0.239	0.158
10	0.229	0.198	0.126	0.23	0.261	0.115	0.126	0.289
11	0.245	0.245	0.2	0.2	0.374	0.2	0.2	0.2
12	0.279	0.422	0.194	0.323	0.323	0.226	0.226	0.323
13	0.144	0.1	0.038	0.107	0.15	0.095	0.125	0.076
14	0.204	0.194	0.077	0.148	0.174	0.141	0.187	0.188
15	0.193	0.183	0.041	0.145	0.169	0.066	0.181	0.166
16	0.235	0.211	0.092	0.141	0.235	0.173	0.234	0.254
17	0.196	0.203	0.046	0.14	0.14	0.14	0.176	0.136

Table 6.9: Standard deviation values for the classifiers.

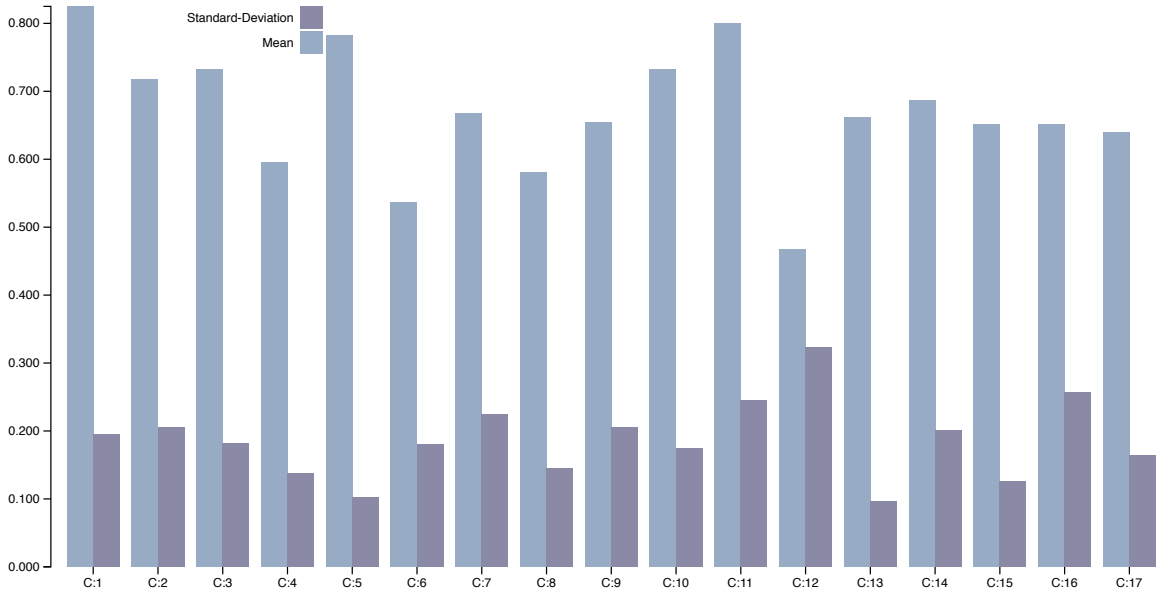


Figure 6.21: Accuracy metrics for the voting classifier.

In our next step we are interested in finding the best classifier for each cluster, with the appropriate parameter tuning. Additionally, we are interested in using *grid search* for the classifiers. Grid search means you have a set of classifiers, which differ from each other in their parameter values, which lie on a grid. What you do is you then train each of the classifiers and evaluate it using cross-validation. You then select the one that performed best.

Chapter 7

Studying Social Network Dynamics

7.1 Network Science

Networks are everywhere. The concept of networks is broad and general, and it describes how things are connected to each other. Networks are present in every aspect of life. From the Internet and its close cousin the World Wide Web to networks in economics, networks of disease transmission, and even terrorist networks, the imagery of the network pervades modern culture. What exactly do we mean by a network? What different kinds of networks are there? And how does their presence affect the way that events play out? In the past few years, a diverse group of scientists, including mathematicians, physicists, computer scientists, sociologists, and biologists, have been actively pursuing these questions and building in the process the new research field of network theory, or the **science of networks** [256].

Network science is a scientific discipline that examines the interconnections among diverse complex networks such as physical, engineered, information, biological, cognitive, semantic and social networks. A network has a very flexible definition: it is a set of nodes connected by links. In mathematical terms a network is represented by a *graph*, a pair of sets $G = (V, E)$, where V is a set of *nodes*, also called *vertices*, and E is a set of *edges*, also called *links*, that connects elements of V. According to the definition, any system with coupled elements can be represented as a network, so that our world is full of networks. In the scientific literature the terms network and graph are used interchangeably. Yet, there is a subtle distinction between the two terminologies; the *network, node*, and *link* combination often refers to real systems: the WWW is a network of web pages connected by URLs; society is a network of individuals connected by family, friendship or professional ties; the metabolic network is the sum of all chemical reactions that take place in a cell. In contrast, we use the terms *graph, vertex*, and *edge* when we talk about the mathematical representation of these networks: we talk about the web graph, the social graph (a term made popular by Facebook), or the metabolic graph.

The scientific study of networks has its roots in eighteenth century when the legendary mathematician Leonhard Euler solved the famous *Seven Bridges of Königsberg* problem [116]. In his paper Euler shows how relationships between simple graph measures, such as node degree and node cardinality, could be exploited in order to tackle everyday problems. Euler's mathematical description of vertices and edges was the foundation of *graph theory*, a branch of mathematics that studies the properties of pairwise relations in a network structure. The field of graph theory continued to develop, and specific networks like regular and disordered lattices were main objects of study in physics and natural sciences up to the end of the 20th century. Example of well-known problems that are formulated as graph tasks range from operational research (i.e. shortest path, minimum spanning tree, maximum flow) to the complexity theory (i.e. travelling salesman problem, graph isomorphism).

More recently other network science efforts have focused on mathematically describing differ-

ent network topologies. Duncan Watts reconciled empirical data on networks with mathematical representation, describing the small-world network. Albert-László Barabási and Reka Albert developed the scale-free network which is a loosely defined network topology that contains hub vertices with many connections, that grow in a way to maintain a constant ratio in the number of the connections versus all other nodes. Although many networks, such as the Internet, appear to maintain this aspect, other networks have long tailed distributions of nodes that only approximate scale free ratios.

Complex networks are nowadays used to describe a wide range of real-world phenomena: social and biological interactions, economic systems as well as optimization problems are examples of how broad is becoming the range of topics which are studied using network science approaches. This breadth of applicative scenarios is one of the main reasons for the renewed interest in network analysis that, in recent years, is emerged in the scientific community. Indeed, a wide class of network problems have been analysed and applied to several branches of research: community discovery, link prediction, node ranking and classification are only few of the several tasks extensively investigated. Among all those tasks, the most challenging and interesting ones aim to describe how networks evolve through time.

7.1.1 Link Prediction

Networks are rarely used to model static entities: i.e., in social contexts we can observe that as time goes by users appear and disappear, new interactions take place, and existing ones fell apart disrupting existing paths. Thinking about the evolution of a network we often wonder if there is any rule that regulate the rising of new edges. Or if couples of nodes exist, that are more likely to establish a connection than the others. Understanding these dynamics is the first step to obtain insights into the real nature of the phenomenon modelled by the observed network. Moreover, almost all the network problems can be reformulated in order to take into account the temporal dimension: communities can be tracked through all their life cycle to unveil their history; incremental ranking can be computed in order to optimize execution costs; links can be predicted using information obtained by the analysis of topology changes in the local surroundings of nodes. Networks taking into account the temporal dimension are called dynamic. The topology of these networks evolves over time as new links and nodes may appear or disappear according to the interactions among their users.

In order to analyse dynamic networks in a reliable way, the social features affecting their structure and behaviour must be considered. Indeed, temporal changes are sometimes independent from the network topology itself and result from external factors. The problem of predicting the existence of hidden links or the creation of new ones in social networks is commonly referred to as the *link prediction problem*. A formal definition describe **Link Prediction** as the problem of identify, given a snapshot of a network G at a time t_0, the top-k edges that are most likely to appear among its set of nodes, at a time t_1. The prediction is restricted to those nodes that are not connected by edges during the first observation. In the context of a bibliographic collaboration network, link prediction estimates the likelihood that two authors will collaborate in the future. In the context of a social network such as Flickr or Facebook, it estimates the likelihood that two users will become friends.

Working with complex networks the idea of model their edges as static entities appears an oversimplification: a friendship established long time ago that was not renewed by successive interactions has, intuitively, lesser value when used as input for a prediction algorithm than one that is more recent or "active". For this reason link prediction approaches have to be revised and extended. Networks should be represented by a graph whose edges are annotated with temporal informations (i.e. time series of weighted interactions, strength or other measures) and all the possible edges of the network should be considered as candidate for prediction, because multiple

interactions could occur over time between the same couple of nodes.

Predicting a new link correctly is certainly a hard task to accomplish: for this reason several approaches were proposed in order to study this evolutive aspect of complex networks, trying both supervised and unsupervised approaches [227]. Unsupervised approaches are based only on local (neighbourhood-based), or global (path-based), topological aspects relative to the couples of nodes for which its needed a prediction. In particular, link prediction strategies may be broadly categorized into four groups: (1) similarity-based strategies, (2) maximum likelihood algorithms, (3) probabilistic models and (4) supervised learning algorithms [227].

The first group defines measures of similarity as a score between each pair of nodes. All non-observed links are ranked according to their scores, and the links connecting more similar nodes are supposed to be of higher existence likelihoods. Despite its simplicity, the definition of node similarity is a non-trivial challenge. A similarity index can be very simple or very complicated, and it may work well for some networks while fail for some others. For example, in [103] the authors introduce a unsupervised method based on ranking factors using the assumption that people make friends in different networks following similar principles.

The second set of methods is based on maximum likelihood estimation [79, 380, 306, 175]. Empirical studies suggest that many real-world networks exhibit hierarchical organization. Indeed, these algorithms presuppose some organizing principles of the network structure, with the detailed rules and specific parameters obtained by maximizing the likelihood of the observed structure. From the viewpoint of practical applications, an obvious drawback of the maximum likelihood methods is that it is very time consuming. In addition, the maximum likelihood methods are probably not among the most accurate ones. In [172] the authors use continuous-time stochastic process for predicting aggregate social activities, that is different activities between users in the same social network.

The third group of algorithms is based on probabilistic Bayesian estimation [130, 338, 165]. Probabilistic models aim at abstracting the underlying structure from the observed network, and then predicting the missing links by using the learned model. Given a target network, the probabilistic model will optimize a built target function to establish a model based on a group of parameters, which can best fit the observed data of the target network. Then the probability that a non-existent link will appear is estimated by the conditional probability. In [381] is proposed a way to develop non-parametric latent feature relational models to minimize an objective function for a normalized link likelihood model.

Among those approaches, the authors in [280] presented a solution based on the preferential attachment principle. The authors in [6] and [257] introduced models based on the quantitative characteristics of common neighbours. In [6] Adamic and Andar demonstrated a means of leveraging text, mailing list, in and out-link information to predict link structure. They have also characterized specific types of items from each of these categories which turn out to be good or bad predictors. Furthermore, because predictors vary between communities, they were able to infer characteristics of the communities themselves. The authors in [257] showed that the probability of scientists collaborating increases with the number of other collaborators they have in common, and that the probability of a particular scientist acquiring new collaborators increases with the number of his or her past collaborators.

Those methodologies, given their simple nature, shown performances that led to, at most, 10% of correct predictions: given the complexity of the problem, this value that could seem very low is actually a very good result. To improve this result, supervised approaches that exploit not only the topology of the network but even semantic informations attached to the nodes (and edges) were proposed.

Link prediction through supervised learning algorithms was introduced in [214]. The authors studied the usefulness of graph topological features by testing them on co-authorship networks. They used a classifier that was trained according to the knowledge that a link will be present

or not in future. Then the classifier was used to predict new links. After [214], a wide range of models exploiting several different strategies have been proposed. Indeed, there has been proved that supervised methods reach better performances than unsupervised ones, in terms of both AUC and precision. Two supervised approaches are the ones in [50, 39], where the first one allows also for the prediction of new nodes. In order to extend the semantic information provided by the network in [228] it was presented a link prediction framework that uses multiple data sources, while [252] proposed an analysis through the use of some graph proximity measure and weight of the existing links.

In order to build an efficient classifier, many works focused on finding an efficient set of features. In [178] is shown that only a small set of features are essential for predicting new edges and that contacts between nodes with high centrality are more predictable than nodes with low centrality. Following these principles, in [26] principal component analysis is used to determine the weights of the features. According to these weights is reduced the number of features taken in input by the regression algorithm used for prediction. A rank aggregation approach is proposed in [296]. The authors rank the list of unlinked nodes according to some topological measures, then at the new instant time each measure is weighted according to its performance in predicting new links. The learned weights are used in a reinforcing way for the final prediction. Finally, in [334] tensor factorization is used to select the more predictive attributes, while in [215] important features for link prediction are examined and it is provided a general, high-performance framework for the prediction task.

Like we did with community features, many works reinforce the classifier with other kind of knowledge. The authors of [324] used textual features besides the topological ones and applied SVM as supervised learning method. In [353], spatial and mobility information are used to help the classifier.

Despite the good performances achieved, all the works reported until now do not solve the interaction prediction problem. Some works which consider dynamic networks are [50] and [40]. In [50], association rules and frequent-pattern mining are used to search for typical patterns of structural changes in dynamic networks. The authors developed the Graph Evolution Rule Miner to extract such rules and applied these rules to predict future network evolution. In [40], the prediction is optimized through weights which are used in a linear combination of sixteen neighbourhoods and node similarity features by applying the covariance matrix adaptation evolution strategy. However, in this second work the authors tried to predict only new interactions and not re-occurring ones. Finally, other works like [92] and [318] show how an approach based on time series modelling the evolution of continue univariate features describing node characteristics substantially helps in solving the link prediction task.

As shown in [227], despite the high precision, supervised approaches can be prohibitively time-consuming for a large networks having over 10,000 nodes. Moreover, supervised methods are proved to reach better performances in terms of both accuracy and precision than unsupervised methods. Thus, given our interest in large, sparse networks, and given that all the works cited highlight the importance of using features outside the links' dimension, our focus on local information gathered from communities and time series features to train the classifier is justified. In order to reduce the computational complexity, several approaches such as [333] make use of clustering and community information. These analyses suggest that clustering information, no matter the algorithm used, improves link prediction accuracy.

In order to build an efficient classifier for link prediction, it is crucial to define and calculate a set of graph structural features. As stated by the papers mentioned previously, when dealing with large-scale graphs that may include millions of vertexes and links, one of the challenges is the computationally intensive extraction of such features. Using our approach, we dramatically reduce the features computation because the calculus is performed considering separately the links present in network's communities. Several studies related to link prediction such as [121,

124, 178, 216, 368] try to suggest which are optimal topological structures of a network and the best features to be used with. For example, in [121] it is analysed the relation between network structure and the performance of link prediction algorithm, while in [178] it is shown that only a small set of features are essential for predicting new edges and that contacts between nodes with high centrality are more predictable than nodes with low centrality. The authors finally claim that on networks with low clustering coefficient, link prediction methods perform poorly, while, as the clustering coefficient grows, the accuracy is drastically improved. [124] investigate the effectiveness of link prediction by gradually reducing the number of visible links in the studied networks. They demonstrate that classification quality degrades with the number of visible links and that a small fraction of visible links helps in solving the problem with chances significantly higher than random. The authors of [368] propose a feature selection framework based on ranking, weighting, correlation and redundancy. In particular, they focus on preserving the maximum accuracy by finding the minimum redundancy in the feature space by using a greedy scheme.

Finally, in the literature there are only few works treating the problem of weak ties in link prediction. Some studies show how and why weak ties can be useful in link prediction. In particular in [226] is shown how the accuracy in link prediction can be improved by exploiting the contribution of weak ties. The Weak Ties Theory [151] states that people usually obtain useful information or opportunities through the acquaintances often not the close friends, i.e., the weak links in their friendship network play a significant role. Recently, the authors of [267] demonstrated that the weak ties mainly maintain the connectivity in mobile communication networks, and in [89] is explained how weak ties maintain the stability of biological systems. In [367] is developed an unsupervised model to estimate relationship strength from interaction activity and user similarity, while in [141] is presented a predictive model that maps social media data to tie strength.

Social networks are highly dynamic objects; they grow and change quickly over time through the addition of new edges, signifying the appearance of new interactions in the underlying social structure. Various applications could use link prediction in order to provide their service. For example, suggest new friendships on a social network, co-autorships on a professional network or interesting products in an online-market are certainly facilities that online services need to offer to their users. Understanding the mechanisms by which networks evolve is a fundamental question that is still not well understood, and it forms the motivation for our work here.

7.2 Supervised Intra-/Inter-Community Interaction Prediction

7.2.1 Introduction

As mentioned above, the problem of predicting the existence of hidden links or the creation of new ones in social networks is commonly referred to as the *link prediction problem*. In this work, we propose an analytic process which, exploiting well-known state-of-the-art techniques, is able to tackle this challenging task in dynamic networks.

In order to capture how topological features evolve— knowledge needed to perform prediction in dynamic contexts—we made use of time series. Specifically, considering a dynamic social network, we built a time series for each social feature of each couple of nodes, that is a sequence of measures at successive points in time, spaced at uniform time intervals. In our approach, we used such structure to forecast future values of each feature: time series forecasts are then used to solve the link prediction problem.

Several works highlight that, when addressing link prediction through supervised learning, it does not appear to exist a set of features or a similarity index that is outperforming in all settings: depending on the network analysed, various measures could be particularly promising or

not [214]. This suggests that the predictors which work best for a given network may be related to the structure within the network rather than a universal best set of predictors. Topological similarity indexes encode information about the relative overlap between nodes' neighbourhoods. We expect that the more similar two nodes' neighbourhoods are (e.g., the more overlap in shared friends), the more likely they may be to exhibit a future link. Moreover, we exploit well-known social network characteristics such as power law degree distribution [28], the small-world phenomenon [358], and community structure [144].

In this study, a valuable topological information that we leverage regards the modular structure of social networks: indeed, social networks can be partitioned into densely and internally connected vertex sets and it has been extensively observed that such topologies provide bounds to the sociality of the users within them. Furthermore, in a dynamic scenario, more than in a static one, the evolution of such boundaries describes changes in people' social behaviours. Starting from such observation, we decided to divide the original problem into two disjoint tasks:

- intra-community interaction prediction;
- inter-community interaction prediction.

Following the hypothesis that friends of friends are more likely to become friends than individuals who have no friends in common [151, 303], in the former task we restrict our attention to the prediction of new links at time $t + 1$ which occur between individuals who are in the same community at least once in $[0, t]$. This strategy has the computationally not negligible advantage of calculating only the features among nodes belonging to the same community. The latter task, on the other hand, focuses on the forecast of future bridges across network modules: such interactions represent the weak ties that keep together the overall network structure.

We propose a data mining process able to provide a solution to both tasks: moreover, we formalize the link prediction problem for dynamic networks, the *Interaction Prediction*. Our approach predicts future interactions by combining dynamic social networks analysis, time series forecast, feature selection and network community structure.

7.2.2 Interaction prediction problem

The classic formulation of link prediction involves the use of the observed network status to predict new edges that are likely to appear in the future or to unveil hidden connections among existing nodes. To satisfy this definition, a wide set of approaches were proposed and tested on several different domains both in supervised and in unsupervised fashion. However, graph structures are often used to describe rapid-scale human dynamics: social interactions, call graphs, buyer-seller scenarios and scientific collaborations are only few examples. This is exactly the reason why link prediction has become the principal instrument used to address the need of dealing with networks that evolve through time.

In this work, our aim is to exploit the temporal information carried by the appearance and disappearance of edges in a fully dynamic context: doing so, we plan to overcome the limitations imposed by the analysis of a static scenario when making predictions. To model rapid-scale dynamics, we will adopt the *interaction network* model:

Definition 1 (**Interaction Network**). *An interaction network $G = (V, E, T)$, is defined by a set of nodes V and a set of time-stamped edges $E \subseteq V \times V \times T$ describing the interactions among them. An edge $e \in E$ is thus described by the triple (u, v, t) where $u, v \in V$ and $t \in T$.Each edge e represents an interaction between nodes u and v that took place at time t.*

To easily analyse an interaction network G, we discretize it into τ consecutive snapshots of the same duration, thus obtaining a set of graphs $\mathcal{G} = \{G_0..., G_\tau\}$. We assume that the interactions belonging to G_t are only the ones that appear in the interval $(t, t + 1)$. Such modelling choice

allows us to make predictions not only for interactions that will take place among previously unconnected nodes, but also for predicting edges that have already appeared in the past. This decision is made in order to better simulate the dynamics that real interaction networks exhibit allowing nodes and edges both to rise and to fall. In real interaction networks, this model is a good proxy for structural dynamics since it allows to implicitly assign a time to leave to links (i.e., in a call graph, it enables to weight more recent interactions w.r.t. older ones when predicting future contacts among a pair of nodes). Due to the adoption of this more complex graph model, hereafter we will refer to this peculiar formulation of the LP problem as *Interaction Prediction problem*:

Definition 2 (**Interaction Prediction**). *Given a set* $\mathcal{G} = \{G_0, ...G_t, ...G_\tau\}$ *of ordered network observations, with* $t \in T = \{0...\tau\}$*, the interaction prediction problem aims to predict new interactions that will took place at time* $\tau + 1$ *thus composing* $G_\tau + 1$.

In the following section, we introduce our analytical workflow, built upon a supervised learning strategy, designed to solve the Interaction Prediction problem.

7.2.3 Proposed approach

The Interaction Prediction problem introduces new challenges to an already complex task. Due to the evolutionary behaviour of the networks subject of our investigation, a particular effort is needed in order to find a reasonable way to take care of structural dynamics during the prediction phase. To this extent, we make use of time-stamped network observations and community knowledge besides classical features in order to learn a robust machine learning model able to forecast new interactions. We design our approach to follow four steps (graphically represented in Fig. 7.1):

Step 1: Given an interaction network G as input, for each temporal snapshot $t \in T$ we compute a partition $\mathcal{C}_t = \{C_{t,0}, ..., C_{t,k}\}$ of G_t using a community discovery algorithm. Then we define, for each t and C, $G_{C_t} = (V_{t,C}, E_{t,C})$ as the subgraph induced on G_t by the nodes in C_t, such that $V_{t,C} \subseteq V_t$ and $E_{t,C} \subseteq E_t$.

Step 2: For each $t \in T$, we consider the interaction communities $\mathcal{C}_t$ of G_t and compute a set of measures F for each pair of nodes pair $(u, v) \in W_{t,\mathcal{C}}$ such that $W_{t,\mathcal{C}} = \{(u, v) : u, v \in V_{t,C} \wedge C_t \in \mathcal{C}_t\}$ that is (u, v) belong to the same community at time t. Thus, we obtain values $f_t^{u,v}$ describing *structural* features, *topological* features and *community* features of the node pairs (u, v) at time t.

Step 3: With these values, for each couple of nodes $(u, v) \in W_{t,\mathcal{C}}$ and feature $f \in F$ we build a time series $S_f^{u,v}$ using the sequence of measures $f_0^{u,v}, f_1^{u,v}, ..., f_\tau^{u,v}$. Then, we apply well-known forecasting techniques in order to obtain its future expected value $f_{\tau+1}^{u,v}$.

Step 4: Finally, we use the set of expected values $f_{\tau+1}^{u,v}$ for each feature $f \in F$ to build a classifier that will be able to predict future intra-community interactions.

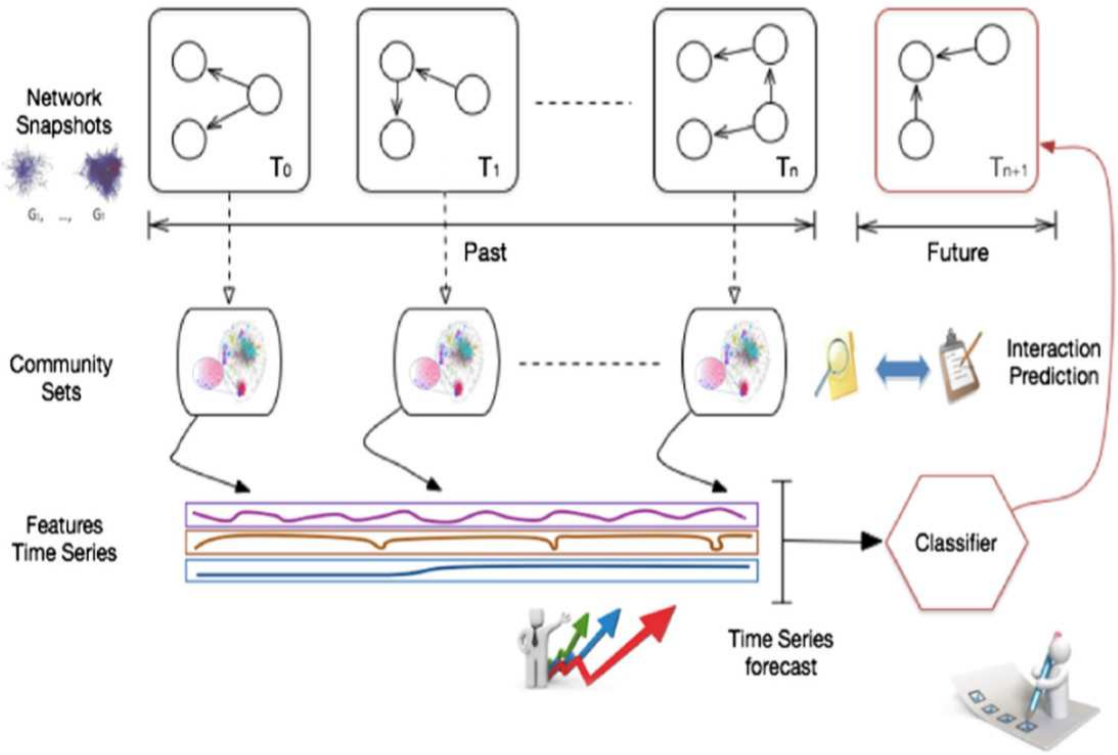


Figure 7.1: Proposed approach workflow. The interaction network is split into network snapshots and each snapshot is partitioned using a community discovery algorithm *(Step 1)*. Then for each community, a large set of features describing nodes and links are calculated *(Step 2)*. Using these values, different time series are built and a forecast of their future values is provided for the time of the prediction *(Step 3)*. Finally, these expected values are used to train a classifier able to predict new interactions *(Step 4)*

In the following, we discuss each step by itself, proposing solutions that can be used to instantiate the described analytical process making use of well-known methodologies.

Step 1: Community Discovery

Partitioning a network into communities is a complex task: for this reason, several approaches were introduced during the last decade, each one of them tailored to extract communities carrying specific traits. Due to the absence of an universally shared community definition, in order to evaluate the impact of community structure on the predictive power of the proposed supervised learning strategy, we tested three different CD algorithms, namely *Louvain, Infohiermap* and *DEMON*. Here we provide a short description of their major characteristics, while in the experimental section we will discuss how they affect the predictive power of the described analytical process. We remind that we adopted community discovery algorithms to split interaction networks into communities, and then we used these communities to calculate the features that will be illustrated in the following and to perform the predictions of new interactions.

Louvain is an heuristic method based on modularity optimization[41]. It is fast and scalable on very large networks and reaches high accuracy on ad hoc modular networks. The optimization is performed in two steps. First, it looks for "small" communities by optimizing modularity locally. Second, it aggregates nodes belonging to the same community and builds a new network whose nodes are the communities. These steps are repeated iteratively until a maximum of

modularity is attained and a hierarchy of communities is produced. Louvain produces a complete non-overlapping partitioning of the graph. As most of the approaches based on modularity optimization, it suffers from a "scale" problem that causes the extraction of few big communities and a high number of very small ones.

Infohiermap is one of the most accurate and best performing hierarchical non-overlapping clustering algorithms for community discovery[311] studied to optimize community conductance. The graph structure is explored with a number of random walks of a given length and with a given probability of jumping into a random node. Intuitively, the random walkers are trapped in a community and exit from it very rarely. Each walk is described as a sequence of steps inside a community followed by a jump. By using unique names for communities and reusing a short code for nodes inside the community, the walk description can be highly compressed, in the same way as reusing street names (nodes) inside different cities (communities). The renaming is done by assigning a Huffman coding to the nodes of the network. The best network partition will result in the shortest description for all the walks.

DEMON is an incremental and limited time complexity algorithm for community discovery[85]. It extracts ego networks, i.e., the set of nodes connected to an ego node u, and identifies the real communities by adopting a democratic bottom-up merging approach of such structures. Following this approach, each node, through its ego network (i.e., the induced graph on his one-hop neighbourhood), gives the perspective of the communities surrounding it: all the different nodes perspectives are then merged together leading to an overlapping partition. To each ego network is applied a label propagation algorithm which ignores the presence of the ego itself in order to identify local micro-communities, and then, with equity, such individual micro-level is combined with the ones obtained by the rest of the nodes ego networks. The result of this combination is a set of overlapping modules, the guess of the real communities in the global system, made not by an external observer, but by the actors of the network itself.

We chose to use the aforementioned algorithms since, due to their formulations, they cover three different kinds of community definitions: modularity-, conductance- and density-based ones. Since in our test we vary the structural properties of the communities used to extract the classification features, in the experimental analysis we will be able to discuss which network partitioning approach is able to provide more useful insights into future interactions.

Step 2: Features Design

In order to efficiently approach the Interaction Prediction task using a supervised learning strategy, it is crucial to identify and calculate a valuable set of features to train the classifier. When dealing with large-scale graphs that may include millions of vertices and links, one of the challenges is the computationally intensive extraction of such features. Several studies related to link prediction such as [121, 124, 178, 216, 368] have tried to suggest which are the optimal topological structure of a network and the best features to be used. Moving from the results of such analysis, we decided to use information belonging to three different families: *pairwise structural features*, *global topological features* and *community features*. We recall that all the features were com- puted before the community extraction phase on node pairs sharing the same social context.

a) *Pairwise structural features*

In this class fall all the measures used in the literature to score the likelihood of new links in unsupervised scenarios. Given a graph G, we will use the following notation:$\Gamma(u)$ identifies the set of neighbours of a node u in G; $|\bullet|$ represents the cardinality of the set $\bullet$. Starting from the measures proposed in [214], we restricted our set to the following ones:

- *Common Neighbor (CN)* assigns as likelihood score of a new link the number of neighbours shared by endpoints [259]. More formally,

$$CN(u, v) = |\Gamma(u) \cap \Gamma(v)| \tag{7.1}$$

- *Jaccard Coefficient (JC)* measures the likelihood of two nodes to establish a new connection as the ratio among their shared neighbours and the total number of their distinct neighbours [314]. It is defined as,

$$JC(u, v) = \frac{|\Gamma(u) \cap \Gamma(v)|}{|\Gamma(u) \cup \Gamma(v)|} \tag{7.2}$$

- *Adamic Adar (AA)* refines CN by increasing the importance of nodes which possess less connections [6]. Its formal definition is:

$$AA(u, v) = \sum_{w \in \Gamma(u) \cap \Gamma(v)} \frac{1}{log|\Gamma(w)|} \tag{7.3}$$

- *Preferential Attachment (PA)* assumes that the probability of a future link between two nodes is proportional to their degree [28]. Hence, it is defined as:

$$PA(u, v) = |\Gamma(u)| \times |\Gamma(v)| \tag{7.4}$$

As a direct consequence to their formulation, CN, JC and AA share the same result set composed by all the pair of nodes at most two-hops in G. However, the values obtained by the three measures for the same edge do not correlate (i.e., having a high CN does not imply having high JC or AA). Conversely, PA generates scores for all the possible node pairs: we restrict its computation to nodes at most at distance two to uniform its result set to the ones of the other measures. We remind that in our calculus of the features G corresponds to G_{C_t}, that is the subgraphs induced on G_t for each time stamp t.

b) Global topological features

The features discussed so far look at the nodes immediate surroundings. However, also the position of a node within the network carries valuable information that can be exploited in order to predict which kind of nodes are attracted by it.

In the literature, a wide set of measures were proposed to estimate the centrality of nodes and edges as well as their rank within a network. These scores are, often, computationally expensive to calculate: for this reason we have decided to make use only of two of them, specifically:

- *Degree Centrality (DC)* relates the centrality of a node to its degree. More formally,

$$DC(u) = |\Gamma(u)| \tag{7.5}$$

- *PageRank (PR)* is a link analysis algorithm introduced by [268] and used by the Google Web search engine. It assigns a numerical score to each element of a hyperlinked set of documents with the purpose of measuring its relative importance within the set.

$$PR(u) = \frac{1 - d}{N} + d \sum_{(u,v) \in E} \frac{PR(v)}{|\Gamma(v)|} \tag{7.6}$$

where $PR(u)$ is the page rank score of node u, N is the total number of nodes and d is the dumping factor. In our experimentation we used the default value for d (0.85).

DC and PR scores were computed for both the endpoints of possible edges pairs: the underlying idea is to understand if there is some correlation among the centrality of two nodes and the likelihood of the appearance of a new interaction between them. This choice can be seen as a way to generalize the PA measure where the operator defining the combination of the individual scores is not fixed.

c) Community features

One of the most pressing issues related to LP regards the reduction of *false-positive* forecasts. To this extent, as briefly mentioned before, we exploit community discovery as a way to reduce the number of predictions provided by the chosen pairwise structural features.

Communities group together nodes that are tightly connected within each other than with the rest of the network. Making predictions only between nodes belonging to the same community allows the predictive process to focus on connections that are more likely to appear, thus discarding the ones connecting different graph substructures. However, following the general intuition behind the idea of community, we can take advantage of more specifically designed measures. Indeed, all the information we can gather from the topological analysis of the communities can be used as features describing the extended surroundings of nodes. With this aim we have decided to introduce the following features:

- *Community Size (CS)* number of nodes belonging to the community C. Defined $G_C = (V_C, E_C)$ as the graph induced on G by the nodes in C, we have,
$$CS(G_C) = |V_C| \tag{7.7}$$
- *Community Edges (CE)* number of edges within nodes in C. Formally,
$$CE(G_C) = |E_C| \tag{7.8}$$
- *Shared Communities (SC)*, given two nodes $u, v \in V$ and a set of communities $\mathcal{C} = \{C_0...C_n\}$, identifies the number of communities shared by a couple of nodes u and v. When dealing with network partitions, SC takes value in $0, 1$, while in case of overlapping communities its domain is $[0, |\mathcal{C}|]$.
$$CS(u, v, C) = |\{C | u \in V_C \wedge v \in V_C \forall C \in \mathcal{C}\}| \tag{7.9}$$
- *Community Density (D)* ratio of edges belonging to the community over the number of possible edges among all the nodes within it. Formally,
$$D(C) = \frac{|E_C|}{|V_C| \times (|V_C| - 1)} \tag{7.10}$$
- *Transitivity (T)* identifies the ratio of triangles with respect to open "triads" (two edges with a shared vertex) in G_C.
$$T = 3\frac{|triangles(G_C)|}{|triads(G_C)|} \tag{7.11}$$
- *Max Degree (MD)* identifies the degree (w.r.t. the community subgraph) of the principal hub for the community C.
$$MD(C) = max\{|\Gamma(u) : u \in V_C\} \tag{7.12}$$
- *Average Degree (AD)* identifies the average degree (w.r.t. the community subgraph) of the nodes within the community C.
$$AD(C) = \frac{\sum_{u \in V_C} |\Gamma(u)|}{|V_C|} \tag{7.13}$$

Step 3: Forecasting Models

The third step of our approach involves the adoption of time series forecasting models to obtain, given subsequent observation of the same feature for the same pair of nodes, an estimation of its future value. Since the behaviour of the observed time series is not known in advance, we adopt several forecasting models based on different underlying assumptions. This choice allows us to identify which one best describes the evolution of the network analysed later on. Since the time series we are analysing are not large, we have decided to not employ complex models that are known to be very efficient on extended observation periods. In fact, we tested four computationally efficient models that have shown to achieve good performances on short time series.

In the following definitions we use, we identify with $Z_t = (t = 1...\tau)$ time series with τ observations and with Θ_t its forecast at time t.

- *Last Value (Lv)* considers as forecast the last observed value of the time series. The forecast is defined as:

$$\Theta_t = Z_{t-1} \tag{7.14}$$

- *Average (Av)* is the average of all the observations in Z_t:

$$\Theta_t = \frac{\sum_{i=1}^{T} Z_i}{\tau} \tag{7.15}$$

- *Moving Average (Ma)* predicts the next value by taking the mean of the n most recent observed values of a series Z_t. In our experiments, we have ranged n in the interval $[1, \tau]$.

$$\Theta_t = \frac{\sum_{i=1}^{\tau} Z_i}{n} \tag{7.16}$$

- *Linear Regression (LR)* fits the time series to a straight line. The level α and the trend β parameter (used to estimate the slope of the line) were defined by minimizing the sum of squared errors between the observed values of the series and the expected ones estimated by the model. This forecast is defined as:

$$\Theta_{t+h} = \alpha_t + h\beta_t \tag{7.17}$$

Step 4: Classifier Models

Predicting correctly new interactions is not an easy task. The complexity is mainly due to the highly unbalanced class distribution that characterizes the solution space: real-world networks are generally sparse; thus, the number of new interactions over the total possible ones tends to be small. We have discussed how it is possible, at least to some extent, to mitigate this problem by restricting the prediction set (i.e., predicting only new edges among nodes that, during the network history, were involved at least in a common community).

However, even adopting such precautions we can expect a substantial unevenness between the positive and the negative classes. This translates into a very high, hard-to-improve, threshold for the baseline model (i.e., in case of a network having density 0.1, which identifies the presence of "only" 1 / 10 of the possible edges, the majority classifier is capable of reaching more than 0.9 of accuracy by simply predicting the absence of new interactions) even though no interactions will be actually predicted since every possible future links will be marked as not present). In order to better characterize our approach, we instantiated it in two different scenarios (both for inter- and for intra-community predictions):

- *Balanced class distribution* we adopted class balancing through downsampling as performed in previous works [215], thus obtaining balanced classes and a baseline model having 0.5 accuracy.

- *Unbalanced class distribution* in order to provide an estimate of the real predictive power expressed by our methodology, we tested it against the unbalanced class distribution as expressed by the original data.

Moreover, since the main focus of this work is to describe a data mining approach that can be used to solve the Interaction Prediction problem and not to discuss a specific classification model, we evaluated our strategy independently from a hosted classifier: for this reason, in the following section we will discuss results achieved by an ensemble of classifiers showing the scores only for the best performing ones. In detail, our supervised learning model set is composed by: decision tree (C4.5, C&R, CHAID, QUEST, random forest), neural network, SVM and logistic regression.

7.2.4 Data

We tested our approach on two networks: an interaction network obtained from a Facebook-like[1] *Social* network and a co-authorship graph extracted from *DBLP*[2]. These datasets allow us to test our procedure on two different grounds: a "virtual" context, in which people share thoughts and opinions via a social media platform, and a "professional" one. The general statistics of the datasets are shown in Table 7.1, while a brief resume is in the following:

Social The Facebook-like social network originates from an online community for students at University of California, Irvine. The dataset includes the users that sent or received at least one message during 6 months. We discretize the network in 6 monthly snapshot and use the first 5 to compute the features needed to predict the edges present in the last one.

DBLP We extract author-author relationships if two authors collaborated at least in one paper. The co-authorship relations fall in temporal window of 10 years (2001–2010). The network is discretized on yearly basis: we use the first 9 years to compute the features and set as target for the prediction the edges belonging to the last one.

In Table 7.1 we can observe the low average density μ_D of the studied networks across the various snapshots. We notice immediately how the low standard deviation σ_D and σ_{CC} guarantee the good approximation of the average density and clustering coefficient as statistic. For this reason, it is remarkable the fact that Social is more dense than DBLP even though its clustering coefficient is considerably lower than DBLP. This means that, due to its nature, when a new interaction appears in DBLP, more than a couple of users is involved, creating automatically a complete clique, while, in Social, a new interaction just expresses the exchange of a direct message between the two users.

Table 7.1: Networks statistics: average density μ_D, average clustering coefficient $\mu_C C$ and their standard deviations, σ_D and $\sigma_C C$ reported as representative aggregate among the various snapshot.

Network	Nodes	Interactions	#Snapshots	μ_{CC}	σ_{CC}	μ_D	σ_D
DBLP	747,700	5,319,654	10(years)	0.665	0.018	3.113e−05	9.602e−06
Social	1899	113,145	6(months)	0.105	0.015	8.600e−03	1.400e−03

We can observe how DBLP is more "partitioning prone" due to the high clustering coefficient. On the other hand, Social has denser snapshots.

[1] http://toreopsahl.com/datasets/
[2] http://dblp.org

7.2.5 Experiments and Results

In this section, we report the results obtained by applying our approach to two real-world interaction networks. In Sect. 7.2.5 are discussed the results obtained focusing the prediction on intra-community interactions: in such context both balanced and unbalanced class scenarios are proposed and used to evaluate our approach. Finally, in Sect. 7.2.5 the same approach is applied to the forecast of inter-community interactions, the weak links that keep together the modular structure composing complex networks.

Intra-community interaction prediction

The Interaction Prediction problem is computationally expensive to address since, in theory, a prediction should be outputted for each pair of nodes in the network analysed. However, social network are known to be sparse and easily to be partitioned in internally dense substructures. Leveraging this observation, our approach is designed to reduce the node pairs for which compute a prediction to the ones whose endpoints share at least one community membership. Operating this choice, we focus on analysing strong ties—the links inter-communities—and discard the bridges that connects different communities.

Balanced scenario It happens frequently, in the LP problem, that the two classes to be predicted, i.e., there will be a link or not, are highly unbalanced. In our case, we have highly unbalanced dataset with a proportion of unlinked–linked of 95.95– 4.055 % for Social, and of 98.13–1.87 for DBLP. Unfortunately, the classifiers used in our experiments need a balanced test set in order to build the predictive model in the proper way. Following what is generally done in the literature, we balanced every snapshot G_t for Social and DBLP. To evaluate the performances of the classifiers, we used the accuracy and AUC which are defined in terms of the confusion matrix of a binary classifier (see Table 7.2):

- *Accuracy* defined as $ACC = \frac{TP+TN}{TP+FN+TN+FP}$, measures the ratio of correct prediction over the total.
- AUC identifies the area under the receiver operating characteristic (ROC). It illustrates the performances of binary classifiers relating the true-positive rate $TPR = \frac{TP}{TP+FN}$ to the false-positive rate $FPR = \frac{FP}{FP+TN}$ and providing a visual interpretation useful to compare different models.

Table 7.2: Confusion matrix of a binary classifier.

	Predicted	
	Class 0	**Class 1**
Actual		
Class 0	TN (true neg.)	FP (false pos.)
Class 1	FN (false neg.)	TP (true pos.)

To better highlight how the proposed approach performs on real-world networks, we need to compare the outcome of its instantiations varying the combination of community discovery algorithms and time series forecast models used.

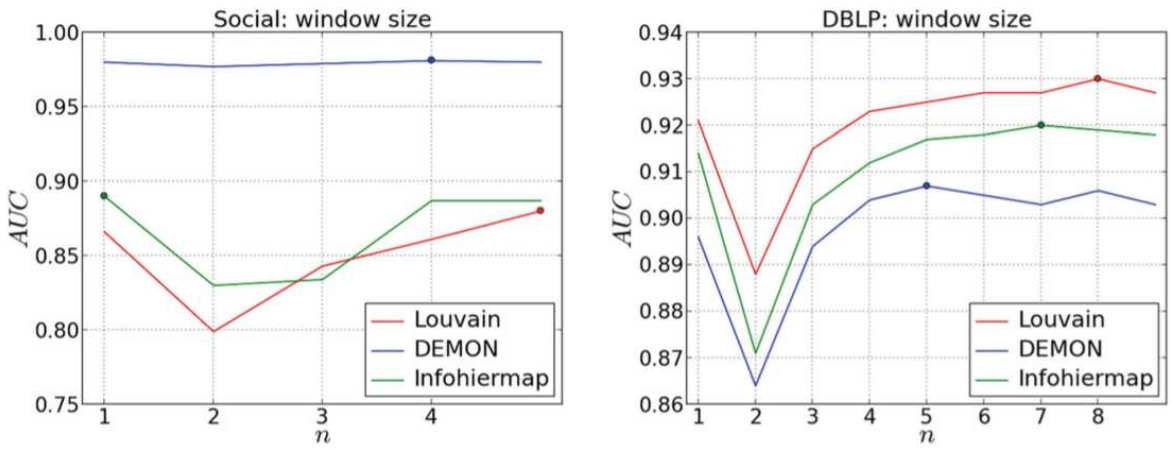


Figure 7.2: Balanced scenario. Accuracy AUC behaviour varying the observation window $n \in [0, tau]$ using the Moving Average Ma. Dots highlight highest values.

We carried out a preliminary study aimed at identifying the optimal window size n for the moving average (Ma) forecast having fixed the community discovery algorithm. By definition, the Lv and Av are special cases of the more general Ma: particularly, the former is equivalent to Ma when $n = 1$, while the latter when $n = \tau$. In Fig. 7.2 is shown, for the three community discovery algorithms, how the classification accuracy behaves varying the observation window n. We can observe different trends for Social and DBLP networks. In the former, the AUC is maximized by the classifier built upon DEMON communities, while in the latter the same approach is the one with worst performances. This is probably due to the particular definition of ego-network-based overlapping communities provided by this approach which is tailored explicitly for social contexts. Furthermore, by observing these plots we can conclude that, in order to obtain higher performances using Ma, two strategies are consistent: (1) minimize n using as forecast the last value (Lv) in order to make inference approximating the future with the actual network status, or (2) use $n \simeq t$ in order to have a better estimation of the whole historical trends. Hereafter, we make use of the best scoring classifiers in Fig. 7.2 to detail our analysis. We will refer to them as the Ma models for each specific network and community definition.

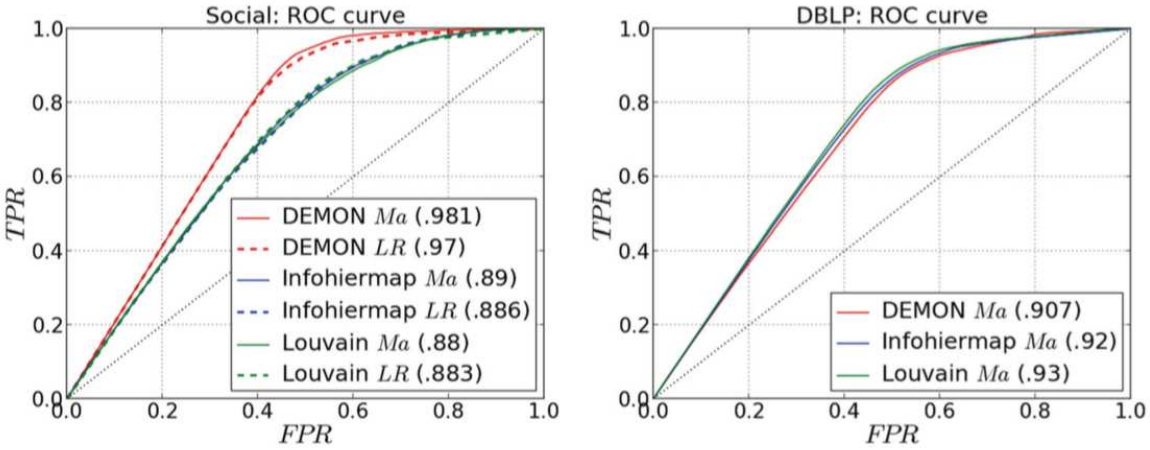


Figure 7.3: Balanced scenario. ROC curves of the various proposed workflow executed with different community discovery algorithms and forecasting methods. In Social the best performer is DEMON with Moving Average, while in DBLP there is not a combination considerably better than the others.

As second step, we compare the outcomes of the classifiers built using the LR forecast models with the Ma ones. In Fig. 7.3 are shown the ROC curves for both Social and DBLP datasets. In the former network, we can observe how LR and Ma provide very similar results even if the moving average is always capable of obtaining slightly better performances. DBLP shows the same trend with a small gap between the two approaches (for this reason, we omit the LR curve). We report in Table 7.3 the AUC and the ACC for all the comparisons.

Table 7.3: Balanced scenario.

Network	*DBLP*		*Social*	
Algorithm	AUC	ACC (%)	AUC	ACC (%)
DEMON Ma	0.907	85.58	**0.981**	93.55
DEMON LR	0.901	84.35	0.970	91.87
Louvain Ma	**0.930**	87.72	0.880	80.27
Louvain LR	0.926	87.48	0.883	81.37
Infohiermap Ma	0.920	86.69	0.890	81.34
Infohiermap LR	0.917	86.18	0.886	80.89

Compared performances varying community discovery and forecasting methods. In bold are the best performers. We can observe how the prediction is more method independent in DBLP than in Social.

Once identified the two best performers for Social (DEMON Ma and Infohiermap Ma) and for DBLP (Louvain Ma and Infohiermap Ma) w.r.t. AUC and ACC, we investigated which are the key features that contribute to their performances. We report in Fig. 7.4 the relative importance of the features used by the classifiers for each method. We can see how in Social the classifier built upon DEMON (a), as well as the one using Infohiermap communities (c), gives high importance to degree centrality and community measures (in particular to density, size and average degree) and tends to make less discriminating decision using pairwise structural features (with the exception of PA). Conversely, in DBLP (b, d, e) the community features set seems to show small predictive power for both the analysed algorithms. This discrepancy is probably due to the different nature of the studied networks: Social naturally models real social interactions

in a short period, while DBLP is inferred from connections (working collaborations) that are developed through years.

Table 7.4: Balanced scenario (Social)

Algorithm	AUC	ACC (%)
SF Ma	0.901	82.88
SF LR	0.895	82.18
FSF Ma	**0.956**	90.10
FSF LR	0.937	88.09

Baselines on structural features using only Structural Forecast (SF) features calculated in the whole network and Filtered Structural Forecast (FSF) calculated following the proposed approach.

In order to understand the boost provided to the classifier by the adoption of the right community discovery algorithm, we designed two different baselines: Structural Forecast (SF) and Filtered Structural Forecast (FSF). The SF model trains the classifier using only the forecasts for the pairwise structural features (CN, AA, PA and JC) computed on all the couple of nodes at distance at most 3 hops present in the whole network, not taking into account the presence/absence of shared communities among them. On the other hand, the FSF model restricts the computation to the pair of nodes belonging to the same community as the proposed approach does. As case study we report in Table 7.4 AUC and ACC of the best Ma and LR baselines for the Social dataset.

Since in Social our best performing approach is the one built upon DEMON communities, the structural features for the FSF baseline were computed using such partition of the network. The obtained results show that, using features extracted from the communities, we are able to gain 0.025 in AUC and 3.45% in ACC with respect to the FSF Ma baseline, and 0.08 in AUC and 10.67% in ACC with respect to the FS Ma one. These results highlight the importance of communities for the interaction prediction task, not only in providing features for pair of nodes, but also in filtering the dataset in order to determine a more accurate selection of nodes for the prediction. Without loss of generality, in the rest of this section, in order to reduce the number of comparisons, we will report a full analysis only for the Social dataset. Furthermore, the results obtained for the DBLP scenario do not differ significantly from the ones discussed with the exception, as seen previously, of the best community discovery algorithm (Louvain instead of DEMON). This divergence is due to the different nature and topology of the networks analysed.

Feature Class Prevalence Since our models are built upon three different classes of features (structural, topological and community related), it is mandatory to test their results against the classifiers using them separately.

Such analysis allows us to assess the predictive power of each class of features, giving an idea of their overall importance for the complete model. We built a classifier for each community discovery algorithm and each feature class by using together all the forecasted versions of the features belonging to it. As shown in Table 7.5, regardless of the community discovery algorithm used, the most predictive features are the ones belonging to the topology class, followed by structural and community ones. However, we can observe how the AUC and ACC are always higher for the model based on the DEMON approach: this trend suggests that this algorithm is the one that better bounds, at least for this network, the nodes that are more likely to establish future interactions.

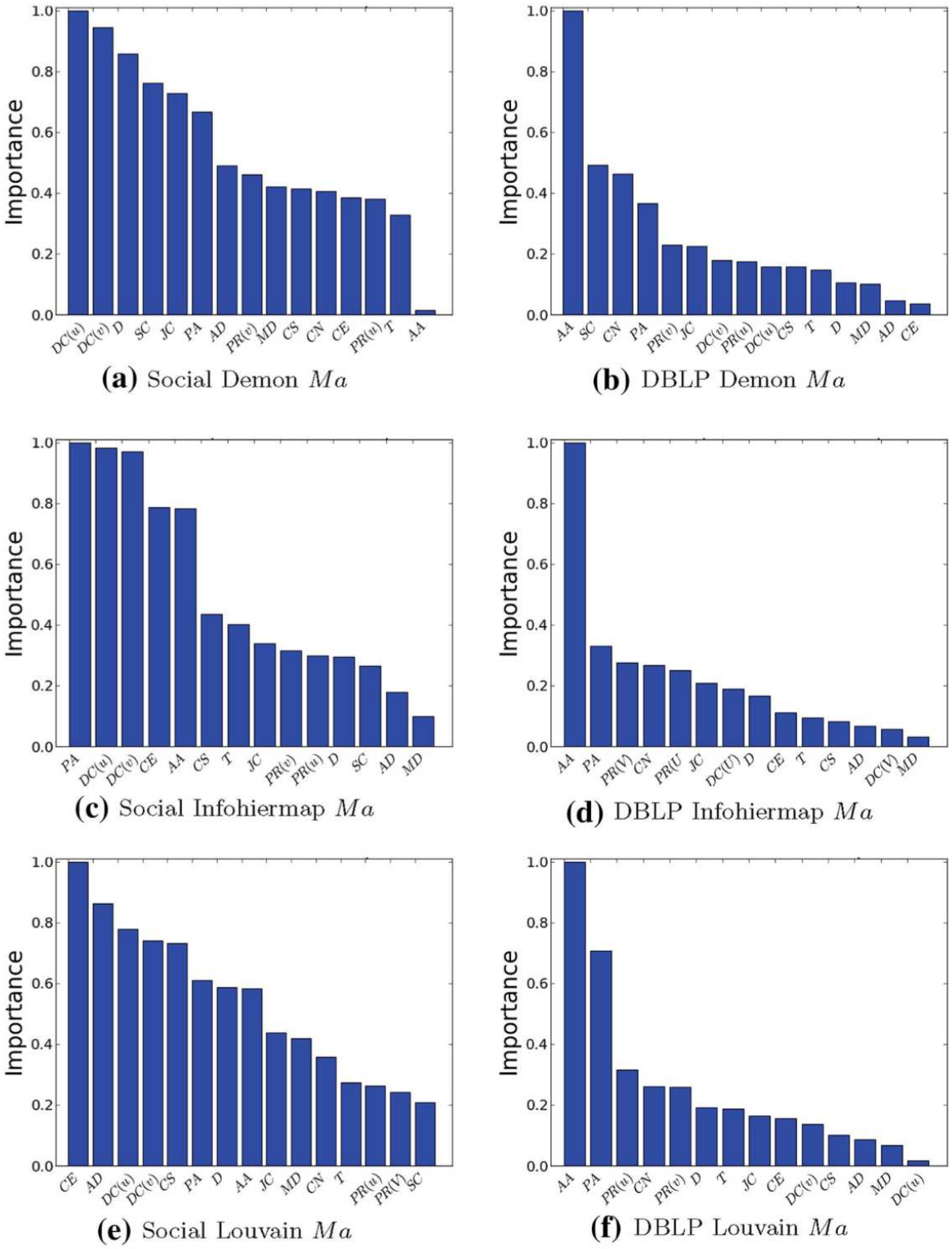


Figure 7.4: Balanced scenario. Features importance: the classifiers built for Social (in particular **a** and **c**) give high importance to community average degree DC, density D and size SC. On the other hand, for DBLP the most important features are the Adamic Adar AA and preferential attachment PA.

Table 7.5: Balanced scenario (Social)

Algorithm	AUC	ACC (%)
DEMON Structural	0.957	90.59
DEMON Topology	0.962	91.44
DEMON Community	0.903	83.53
Louvain Structural	0.850	78.63
Louvain Topology	0.875	79.38
Louvain Community	0.724	66.64
Infohiermap Structural	0.876	79.85
Infohiermap Topology	0.887	80.81
Infohiermap Community	0.667	62.11

Compared classifier performances for different classes of features. We can notice how independently from the community discovery algorithm the topological features always provide the highest performances.

Complete Classifier We investigated if the performances of the analysed classifiers can be improved by combining all the features obtained at the end of the forecasting stage (i.e., all the time series forecasts computed with Ma and LR). As we can see in Table 7.6, the performance boost is negligible with respect to DEMON Ma; in fact, we are able to gain only 0.35% in ACC maintaining the same AUC w.r.t. the results shown in Table 7.3. This means that the feature set used by our best classifier is "stable": its extension does not produce advantages that justify an increase of model complexity. Conversely, for Louvain and Infohiermap the gain in AUC and ACC is more evident: this is due to the different degree of approximation introduced for each feature in the forecasting stage.

Table 7.6: Balanced scenario (Social)

Algorithm	AUC	ACC (%)
DEMON All	**0.981**	93.90
Louvain All	0.901	83.05
Infohiermap All	0.894	81.91
FS All	0.959	90.44

Compared classifier performances using all the features. DEMON reaches the highest performances in terms of accuracy and area under the curve. In bold the AUC of the best performing approach.

Features forecast correlation As a consequence to the minor deviations in performances for different forecasting methods, we investigated which are the correlations among the forecasted values calculated by LR and Ma with $n \in [0, \tau]$. We analysed each feature separately observing the correlation average, median and variance. In Table 7.7, we report the average of the variances of these values aggregated for different classes of features. From this table emerges that, regarding structural features, Louvain has the lowest average of variances of correlations, while, for topological and community related features, it is DEMON with the lowest correlations.

Table 7.7: Balanced scenario (Social)

Algorithm	Structural	Topology	Community
DEMON	0.023	0.001	0.003
Louvain	0.009	0.017	0.018
Infohiermap	0.042	0.015	0.081

Mean of the variance of the correlations among the values forecasted with LR and Mv. The higher this value the most careful must be the choice in selecting the forecasting method.

As a result, we can say that, if we use Infohiermap (that has the highest average of the variances) to extract the communities from the interaction network, we should focus on the choice of the different forecasting methods applied. On the other hand, if we calculate the communities with DEMON, it does not matter very much which kind of forecast technique (LR or Ma) we use to calculate the expected values. This statement holds less strongly for Louvain which has a low correlation variance only for structural features.

Features forecast deviation We estimated how good is the proposed approach by analysing the deviation of the values calculated with the forecasting methods with the real values of the features at $\tau + 1$. The models built using the real features at $\tau + 1$ reach good performances (see Table 7.8).

Table 7.8: Balanced scenario (Social)

Algorithm	AUC	ACC (%)
DEMON	0.987	95.76
Louvain	0.888	81.16
Infohiermap	0.846	75.95

The high performances reached by the classifiers built using the real values at time $\tau + 1$ indicate that a good approximation of forecasting methods to these values is fundamental to build reliable classifiers.

This indicates that a good approximation of the real values is important to build a reliable classifier. As a consequence of these good performances, an analysis of the deviation of the expected values obtained with time series forecast with the real values is needed to understand which measures can be predicted better than others with a certain community discovery algorithm or a certain forecasting technique. Thus, we analysed the deviations $(f^{u,v}_{\tau+1} - \hat{f}^{u,v}_{\tau+1})^2$ of the expected values of the different forecasting methods with the real ones.

We analysed the sum of squared error (SSE) for each forecasting method of each feature in Fig. 7.5, and we observed that: (1) DEMON and Infohiermap perform better with Ma, (2) Louvain is generally worse than the others for every feature, (3) Infohiermap works better for structural and topological (4), and DEMON minimizes the error for the community features. However, independently from the community discovery algorithm or the forecasting method, the deviation is always very low justifying the good performances previously exposed.

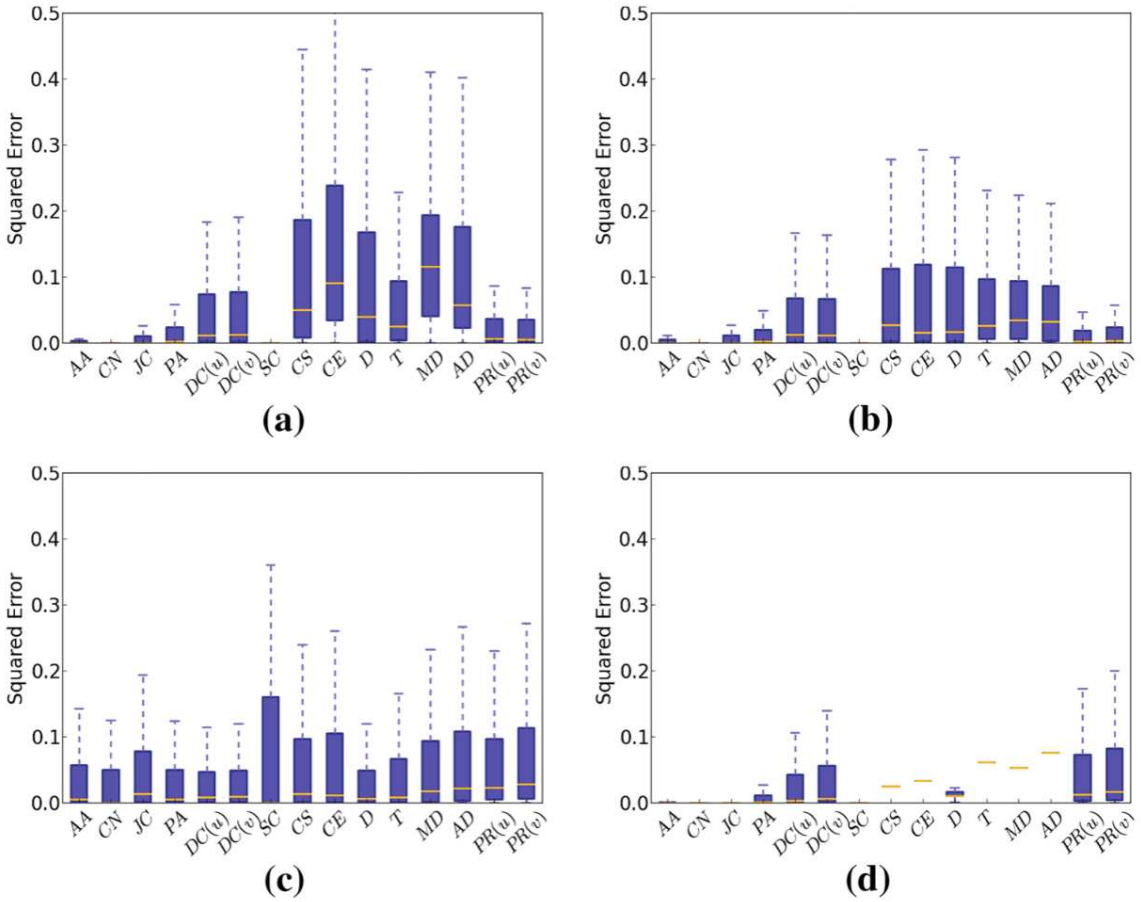


Figure 7.5: Balanced scenario (Social). The *boxplots* of squared errors per feature show how independently from the community discovery algorithm or the forecasting method the deviation is always very low especially for the most important features **a** Social Louvain Ma, **b** Social Louvain LR, **c** Social DEMON Ma, **d** Social Infohiermap Ma.

In particular, we found that, with respect to the other combinations, Infohiermap with LR has the highest SSE for each attribute. On the other hand, the best approximations are achieved by Infohiermap and DEMON with Ma with $n \in 3, 4$. Indeed, with the exception of AA, Louvain never has the lowest SSE among the features used. At the same time, by ranking the SSE among the different community discovery algorithms and forecasting techniques, it emerges that with Louvain the lowest SSE belongs to AA while the highest to SC. On the contrary, with DEMON the lowest SSE belongs to SC, while the highest changes with respect to the forecasting method. Finally, as far as Infohiermap is concerned, we cannot derive nothing interesting. Thus, probably, due to its nature related to ego networks, DEMON gives better results than the other community discovery algorithms for community features, while AA works really well with the communities extracted by Louvain.

Unbalanced scenario We have shown how the described analytic workflow is able to obtain good results when dealing with datasets having a balanced class distribution. Unfortunately, this scenario is not common when addressing the Interaction Prediction problem. Furthermore, making predictions on new interactions that will appear in a network involves, potentially, computing scores for all the $|V| \times (|V| - 1)$ pair of nodes of a network. Social networks are generally sparse, and this led to a high rate of false-positive predictions (in case of unsupervised approaches) or

to models that maintain high accuracy just predicting the absence of new links (the majority classifier in case of supervised learning). Indeed, predicting every object as belonging to the most frequent class guarantees high performances, but in general it leads to useless classification results. For this reason, evaluating the performances of classifiers in highly unbalanced scenarios is not an easy task, but is definitely a very important one.

Since we want to predict correctly new links, our primary purpose is to reach high precision avoiding the generation of false-positive predictions. This is the reason why in the unbalanced scenario we will discuss, besides AUC and ACC, the *Lift Chart* and *precision* of the tested classifiers.

Precision is defined as $PPV = \frac{TP}{TP+FP}$. It represents the ratio of correct predictions for a specific class (in our case the one representing the presence of the edge in the test set) with respect to the total predictions provided.

Lift Chart graphically represents the improvement that a mining model provides when compared against a random guess, and measures the change in terms of lift score. By comparing the lift scores for various portions of a dataset and for different models, it is possible to determine which model is the best and which percentage of the cases in the dataset would benefit from applying the model's predictions.

We report the precision instead of the accuracy because, unlike the balanced scenario (where starting from a ratio of 50–50 the accuracy has a strong significance), in the unbalanced one it is very easy to get a high, but meaningless, accuracy. This is due to the fact that, as a consequence to the sparsity of the interaction network, the majority classifier can predict always "no edge" with no effort reaching very high performances. Besides this we report the Lift Chart because, conversely from AUC and PPV (with which shares, describing isomorphic spaces, the conveyed information), it is able, even in unbalanced scenarios, to graphically emphasize the improvements provided by the tested classifier against a baseline model.

We preserved the original ratio between the node pairs with and without a future interaction in Social and DBLP datasets. For both networks, we used the DEMON algorithm to extract communities. This choice is due to the following reasons: (1) Social DEMON reaches the best performances in the balanced scenario; thus, we expect that it will behave well even in unbalanced scenario; (2) DBLP using Louvain (i.e., the best performer in the balanced scenario) in the unbalanced scenario, all the classification models output the majority classifier.

In Social, the ratio of negative class to the total amount of possible pairs is 95.947 %, that means that a majority classifiers predicting no edge for all the pairs would have an accuracy of almost 96 %. As output from the classification phase with Ma, we have a model which reaches an AUC of 0.966 with a prediction accuracy of 98.75 % and a precision w.r.t. the positive class of 95.61 %. These two are very significant results: on the one hand, we have an accuracy improvement of 2.803 % in an ideal window of 4.053 % (100–95.947 %) with respect to the majority classifier while, on the other, we have a very high precision on the positive class, considering that a classifier predicting always an edge would have a precision of 4.053 %. In addition to the Ma model, we also built three classifiers each one of them considering all the forecasts for a single category of features: topological, structural and community.

In Fig. 7.6-left, we show the Lift Chart of the four models for Social. From the chart emerges that after the Ma model, the most promising is the one built upon the topological features followed by structural and community ones.

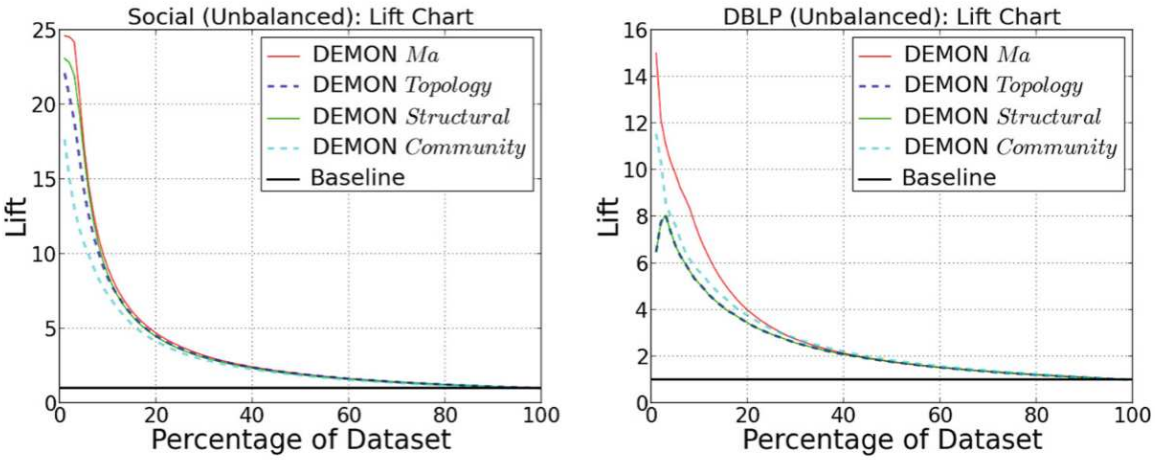


Figure 7.6: Unbalanced scenario. The lift charts of the compared methods show how in both networks DEMON with Moving Average is the combination able to reach the best performances.

Also in this unbalanced scenario, we want to "measure" how much the community approach provides efficiency just filtering in the "promising pairs". By building the dataset with all possible pairs without leveraging community information, we get a majority class, i.e., the absence of a link, with a ratio of 98.96 % over the total number of entries. In order to better compare the two cases, we filter out randomly some pairs with no edge, bringing the accuracy of the majority classifier at 95.947 % (like in the case with community discovery). Again we compare the performances for the SF and FSF, which are reported in Table 7.9, but now considering the precision instead of the accuracy. We can see that we gain almost a 10 % of precision just filtering out, in any time slot, all the pairs not belonging to the same community. These results are very significant. On the one hand, we have an accuracy improvement of 2.803 % in an ideal window of 4.053 % (100–95.947 %) with respect to the majority classifier. On the other hand, we have a very high precision on the positive class, considering that a classifier predicting always an edge would have a precision of 4.053 %. In addition to the Ma model, we try to build also classifiers considering all the forecast methods but grouped for "kind of measure": topological, structural and community. It emerges that after the Ma model, the most promising is the one built upon the topological features followed by structural and community ones.

Table 7.9: Unbalanced scenario (Social)

Algorithm	AUC	PPV (%)
SF Ma	0.897	64.06
SF LR	0.893	62.62
FSF Ma	0.918	**74.71**
FSF LR	**0.932**	72.45

Baselines on structural features using only Structural Forecast (SF) features calculated in the whole network and Filtered Structural Forecast (FSF) calculated following the proposed approach. In bold the AUC and PPV of the best performing approaches.

In DBLP case study, the resulting classifier has an AUC of 0.86, an ACC of 98.135 % and a precision with respect to the positive class of 44.78 %. The majority class (no link) has a ratio of 98.13 % over all the instances of the dataset. A possible reason for the lower performances obtained on DBLP w.r.t. Social is that in the latter an interaction represents a real social action between two different actors, while in DBLP an interaction models a relation of co-authorship in a paper, and the co-authorship is not, in our opinion, a strong representative of social interaction. However, we can notice that the performances are not completely bad: we have a precision of 44.78 %, starting from a ratio of positive class of 1.865 % (100–98.135 %), that is 24 times better than predicting for any pair the presence of the edge. Finally, we can observe from the Lift Chart in Fig. 7.6-right how, differently from the Social case, the most predictive set of features are the community ones, over the structural and topological.

Inter-community interaction prediction

So far, we have focused our attention on the task of predicting interaction within a community. We have shown that our approach is able to achieve good performances in case of both balanced and unbalanced class distributions and discussed the features that better predict the presence (or absence) of a new interaction. Here we address the complementary problem: prediction of *inter-community* interactions. Since the direct prediction of the network weak ties is a very complex problem prevalently due to the low stability of such links through time, we shift our interest to a related problem. We do not want to predict the specific endpoint of the interaction (i.e., user u of community C_j and user v of community C_z), but the presence of at least one interaction among users of two different communities, say C_j and C_z. To do so, we slightly modified our method:

- instead of using the original interaction network, we preprocess our data and build, for each snapshot, an induced graph using the previously extracted communities. In particular, for each snapshot graph G_i and related set of communities C_i we perform the transformation described by Algorithm 1;
- we compute the structural and topological features on the community-node pairs of each new induced graph;
- we apply the time series forecast and, on the forecasted feature values, we build the prediction model.

The main difference w.r.t. the original approach lies in the use of the communities as network nodes and not as filters (i.e., no community features are used to build the final model).

A crucial aspect is the process used to build each community-graph. As shown in Algorithm 1, for each community are identified the core nodes (lines 3-6): then, a new edge is created in the induced graph among the community C_j and C_z if there exists at least one edge in the original graph connecting two of their core nodes (lines 7-15). There are several ways to implement the IdentifyCommunityCores function: in our experiments, we use the top-k% high-degree nodes within each community (we fixed k to 5). After the construction of the community-network, we apply a reconciliation phase across consecutive snapshots in order to align the community ids. To build the evolutive chain of each community (i.e., to find the correspondence of a given community across time), we employed a well-established set matching procedure often used by dynamic community discovery approaches [162], namely the Jaccard matching:

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

Algorithm 1 BuildInducedGraph(G_i, C_i)

Require: G_i: network snapshot, C_i: community set.

```
CoreNodes = {}
IG = NewGraph
for c ∈ C_i do
    c_cores = IdentifyCommunityCores(C_i)
    CoreNodes[c] = c_cores
end for
for c_j ∈ C_i do
    for c_z ∈ C_i do
        if c_j ≠ c_z then
            if ∃(u, v) ∈ G_i, u ∈ CoreNodes[c_j], v ∈ CoreNodes[c_z] then
                IG.AddEdge(c_j, c_z)
            end if
        end if
    end for
end for
return IG
```

Given a community C at time t (C_t in the equation), we identify as its future expression in $t+1$ the community which maximizes the Jaccard function upon their node sets. We decided to evaluate the introduced methodology on a very specific case study: inter-community interaction prediction on the DBLP community-graph built upon the Infohiermap partition. The reasons behind such choice are the following:

- Among the previously analysed datasets, DBLP is the bigger one and it is always decomposed in a higher number of communities (ensuring community-graphs of meaningful size);
- DEMON generates overlapping communities; thus, the community-graph extraction loses some effectiveness (shared nodes generate a densely connected graph);
- Louvain as all modularity-based approaches suffers from the scale problem: this causes very sparse star-like community-graphs composed by few focal nodes (i.e., the bigger communities) linked to many satellites (i.e., very small communities that are rarely connected by interactions).

Balanced scenario In the intra-community scenario, w.r.t. the DBLP dataset and Infohiermap communities, we were able to produce predictions for, approximately, the 91 % of the interactions actually present in the test set. The filter produced by the application of Infohiermap was then able to discriminate weak ties across different network partitions and guarantee high AUC and Accuracy. Due to the community-graph construction defined in Algorithm 1 we now group together the remaining 9 % of the interactions in meta-links connecting different Infohiermap communities. Obviously, due to the IdentifyCommunityCores strategy, we will not able to make prediction for all the weak ties: however, the filtering introduced groups together 97 % of them producing a very reliable sample.

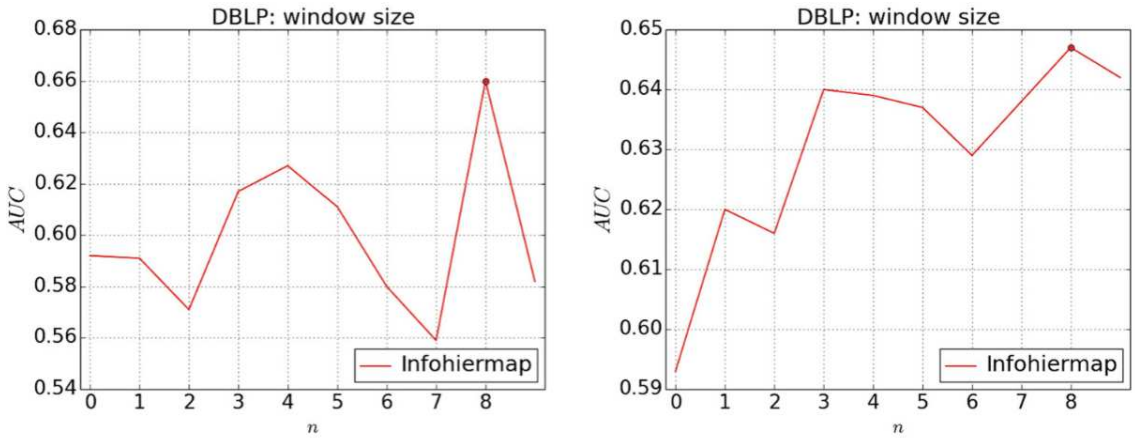


Figure 7.7: Inter-community prediction: *left* balanced and *right* unbalanced scenarios. AUC values varying $n \in [0, \tau]$ using the Moving Average Ma. Dots highlight highest values. In both scenarios, the optimal window size is 8.

Following the method designed for inter-community interaction prediction, we tested all the different time series forecasting strategies discussed in Section 7.2.3 and defined as Ma the one having high score (as shown in Fig. 7.7 for both the balanced and unbalanced scenarios). On the balanced class scenario, we obtained the results reported in Table 7.10. Our results are, as expected, not as good as the one obtained for the intra-community interaction problem. Here the best predictive power is expressed by the Ma time series forecast able to reach 66 % of accuracy w.r.t. the 50 % of the majority classifier. In order to better understand the impact of the time variable on such very volatile network structure, we also trained a classifier on the same feature set computed on the flattened community-graph (i.e., the graph built by keeping together nodes and edges of all the temporal snapshots). The obtained results suggest us that conversely from the intra-community settings here the adoption of time series does not play a crucial role even though it allows with Ma and Avg forecasting to slightly increase the prediction accuracy.

Table 7.10: Balanced scenario (DBLP)

Algorithm	AUC	ACC (%)
Lv	0.580	56.01
Avg	0.650	65.10
Ma	**0.660**	66.00
LR	0.581	58.10
Flat Graph	0.610	59.12
Baseline	0.500	50.00

Infohiermap performances for the inter-community prediction. The Moving Average Ma forecasted features allow for the best classification models. In bold the AUC of the best performing approach.

Unbalanced scenario To complete our analysis, we evaluated the effectiveness of our approach even in the unbalanced inter-communities setting. This scenario represents the most complex one we can design: we are targeting weak ties (i.e., the 9 % of the interactions not

covered by the intra-community predictions) when the majority class—no interaction—is approximately 98 %.

The results in Table 7.11 show a relatively high precision w.r.t. the minority class: while the baseline (the minority classifier) reaches 4.01 % precision, our approach is able to reach PPV = 50 % (even though the recall on the minority class drops from 100 % to "only" 65 %). Even in this scenario, the Ma time series forecast strategy is the one that offers higher quality models. Conversely from the balanced scenario, we can observe how the classifier built upon the flattened community-graph does not produce interesting results: even though it guarantees higher precision (PPV = 57.2 %) the overall model quality is lower (Flat graph AUC = .316 vs. Ma AUC = .647). The predictions made on the flattened networks are more precise, but the recall is low ($\sim$ 9 %). In an unbalanced scenario, the low stability of *inter-community* interactions amplifies the complexity of the predictive task: flattening the temporal dimension causes an increase of the false-negative predictions, which leads to performance degradation.

Table 7.11: Unbalanced scenario (DBLP)

Algorithm	AUC	PPV (%)
Lv	0.594	33.33
Avg	0.632	07.02
Ma	**0.647**	50.00
LR	0.596	50.00
Flat Graph	0.316	57.20
Baseline	0.504	4.01

Infohiermap performances for the inter-community prediction. Like in the balanced scenario, the Moving Average Ma forecasted features allow for the best classification models. In bold the AUC of the best performing approach.

7.2.6 Conclusions

In this work, we have tackled the Link Prediction problem in a dynamic network scenario. Since networks often model highly evolving realities that cannot easily be "frozen" in time without loss of information, a time-aware approach to link prediction is mandatory to achieve valuable results. Moreover, due to the intrinsic high computational cost of the approaches that solve this problem, it is important to reduce the list of possible candidates for which to compute a prediction (preferably avoiding the generation of false positives). To this extent, we have exploited the community structure of social networks to both bound the result set and design features whose analysis through time have allowed the description of a high-performance supervised learning strategy. Anyhow, using network partitions as filters make the proposed approach focus only on the prediction of intra-community interactions: to overcome this issue, we propose an experimental setting specifically designed to address inter-community interaction prediction. Using community-induced graphs, we show that the proposed analytical workflow can be applied to this complex problem and discuss the quality of the obtained results.

The results obtained with the proposed methodology open the way to several future lines of analysis. Indeed, more accurate time series forecast techniques can be evaluated in order to reduce the forecast error and evolutionary community discovery approaches can be used in order to incorporate communities life cycle features within the predictive process. Moreover, with respect to the type of dataset used, it could be possible to consider other types of features such

as mobility knowledge and spatial co-location. All these improvements will lead to more narrow and sophisticated classifiers that, taking into account more and more aspects, will be able to better predict future human interactions.

Chapter 8

Conclusions

Multiple dimensions of our social life have Big Data proxies nowadays. Each one of us produces a lot of data while performing our daily activities. Thanks to these data, human activities are becoming observable, measurable, quantifiable and, predictable. Big Data sources can provide data with better timeliness, reduced costs, improved robustness and applicability in cases where a developed statistical system is still not established. This may support researchers and policy makers by providing a more sophisticated depiction of the human environment, and enabling a more accurate evaluation of needed actions and expected results. Nowcasting can be used to monitor key indicators, allowing real-time adjustment and constant refinement.

In this thesis, we examined several different phenomena where nowcasting and forecasting could be achieved with various Big Data sources, available in quasi real-time, such as mobility data, retail data and social network data.

First, we recalled the basic notions of big data and data science, and we have presented the concept of nowcasting and forecasting, presenting relevant literature. Then, we have accurately discussed the current state of Nowcasting, which are the available applications and what they have to offer. Nowcasting is applied to so many different fields, from economy to well-being, from crime to epidemics. There are several interesting approaches, but of course several limitations and problems.

As first step, we use retail market data, reflecting human's behaviour with regards to influenza to predict the evolution of the seasonal influenza in Italy. We consider as our reference a baseline autoregressive model that considers only historical influenza data and our results demonstrate the superiority of our approach. Specifically, we show that our methodology can produce predictions one week ahead of influenza with comparable accuracy. We also show that our forecasts (up to three weeks into the future) always improve predictions produced with the baseline autoregressive model, thus proving quantitatively the added value of incorporating retail data in our flu prediction model. As second step, we use mobility data to study the attractiveness of airports over a territory. We investigate how the number of customers of airports in Tuscany, Italy changes over time. A mathematical model reproduced real data of customer numbers and the mathematical analysis provided insight into what are the key factors that drive the dynamics. We also used our model to explore the effect of various interventions, such as a temporary closure of an airport, short- and long-term investment, and competition with other airports. Thus, this research has the potential to be incorporated into economic strategic planning, as the general framework and results allow for predictive analysis regarding the outcome of various interventions. We also use mobility data to evaluate the effects of industrial clusters on regional dynamics. This is still an ongoing work, but our aim is to introduce a real-time system able to monitor industrial clusters, and predict the future evolution, such as birth of a new cluster or death of an existing one. As third step, in the last part of the thesis, we have tackled the Link Prediction problem in a dynamic network scenario. Since networks often model highly evolving

realities that cannot easily be "frozen" in time without loss of information, a time-aware approach to link prediction is mandatory to achieve valuable results. We exploited the community structure of social networks to both bound the result set and design features whose analysis through time have allowed the description of a high-performance supervised learning strategy. Anyhow, using network partitions as filters make the proposed approach focus only on the prediction of intra-community interactions: to overcome this issue, we proposed an experimental setting specifically designed to address inter-community interaction prediction and we obtained high quality results.

The results that we have obtained in this thesis represent useful examples of how nowcasting can be achieved in any setting, if the data is available. Many times the data for a specific phenomena is missing, or is only available with a significant lag. And that's where the importance of these approaches lies, as we use big data as external proxies to predict the evolution of these phenomena. Clearly, we don't intend to conclude our study on Nowcasting and Forecasting with the contents of this thesis, since many interesting research problems are still open. There are a few future research directions for the analytical approaches proposed in this thesis.

The first track of research is related to the influenza nowcasting paradigm. It would be rather useful to include human mobility behaviours to our approaches to predict the evolution of seasonal influenza. Understanding the patterns of human movements could be of great use into understanding the way epidemics spread and affect new individuals. This new source of data could lead to better predictions and as a result timelier response to a possible epidemic spread.

Additionally, the problem of nowcasting influenza using retail scanner data could be tackled using deep learning techniques, such as *recurrent neural networks (RNN)*. More specifically, the influenza time series could be the main input of a *long short-term memory (LSTM)* unit and the features extracted from the supermarket scanner data could be fed in an additional dimension in the LSTM. Deep learning techniques are often applicable to the problem of nowcasting, and could potentially lead to higher accuracy, earliness and stability.

As consequence of the model we introduced in the airports study, another track of research would include studying other places of attraction, such as retail centres or hospitals. We could make use of the vast dataset that we have in our possession regarding supermarkets all over Italy to study their attractiveness. In a similar way, this study could be applied to hospitals, in order to study the way they satisfy human needs and predict eventual needs that may rise in the future.

Finally, the last track of research derives from the Interaction Prediction problem and it involves the use of more accurate time series forecast techniques to reduce the forecast error, and evolutionary community discovery approaches in order to incorporate communities life cycle features within the predictive process. Moreover, with respect to the type of dataset used, it could be possible to consider other types of features such as mobility knowledge and spatial co-location, since mobility data is already available to us. All these improvements will lead to more narrow and sophisticated classifiers that, taking into account more and more aspects, will be able to better predict future human interactions.

Considering the relative novelty of the methodologies used to deal with these data, extra carefulness needs to be used to acknowledge possible shortcomings in terms of quality, accessibility, applicability, relevance, privacy policy and ownership of the data, all of which may affect the quality of policy evaluation and appraisal [319]. Nonetheless, we believe that big data sources can be successfully used to predict the future.

But one, cannot but wonder. Can we actually predict everything? An obvious answer, would be, yes you can; as soon as you have enough data and the capacity to learn from them. And what is enough? How much data do we actually need to be able to rely on our results and predictions? That still remains an open question, as the answer would rely on so many different factors, such as the nature of the questions you rise, the nature of the data available and their reliability. We

have to remain vigilant as these data can contain noise and biases, hard to detect and remove, and that could make any possible attempt to describe and nowcast these phenomena even harder. It's our responsibility as data scientists to extract useful information from the data, be ready to validate our models and results, and provide correct conclusions and explanations for the results obtained. Data can unintentionally be misused, and then, they can produce results which appear to be significant; but which do not actually predict future behaviour and cannot be reproduced on a new sample of data and bear little use.

Another important issue is interpretation. Because for any kind of analysis to be useful, interpretation as well as communication is needed. Interpreting big data needs to take context into account, such as how the data were collected, their quality and any assumptions made. As mentioned before, big data can contain noise and biases, but that's not the only reason that interpretation requires extra care. There may be limitations to the usefulness of big data analytics, which can identify correlations (consistent patterns between variables) but not necessarily cause. So what's the point of a study if you can' t claim causality? There is a difference between correlation and causation. If we find a correlation between one thing and another, it doesn't mean that we know that one thing has caused the other.

As scientists, we spend a lot of time trying to pin down these causal explanations, but it's possibly worth understanding that, although correlation is not the same as causation, if we want to demonstrate causality, we do really have to find correlations anyway. So basically, if we want to demonstrate causality, it's about finding a correlation and ruling out all of the other possible explanations. So this sort of result, which is indeed only a correlation at this point in time, is a building block that we can use to work towards what is, in many ways, the holy grail for us as scientists, these causal explanations and models of the world we live in.

Correlations can be extremely useful for making predictions or measuring previously unseen behaviour, if they occur reliably. However, they may also be misleading. For example, knowing that lots of people in an area are searching online for information on flu might be useful for targeting sales of flu remedies, but may not be a reliable predictor of a new flu epidemic. Also, techniques can be reductionist and not appropriate for all contexts. Some researchers have argued that big data is disconnected from social context and unable to capture the subjective experiences of individuals [168]. Opinions also differ on how well modelling can be used to predict dynamic social systems [346, 246].

Bibliography

[1] https://www.google.com/trends/.

[2] https://www.google.org/flutrends/.

[3] https://www.influenzanet.eu.

[4] Enac: Atlante degli aeroporti italiani.

[5] Oecd territorial reviews. Technical report oecd, Venice, Italy, 2010.

[6] Lada A. Adamic and Eytan Adar. Friends and neighbors on the web. *Social Networks*, 25:211–230, 2003.

[7] Alma J Adler, Ken TD Eames, Sebastian Funk, and W John Edmunds. Incidence and risk factors for influenza-like-illness in the uk: online surveillance using flusurvey. *BMC infectious diseases*, 14(1):232, 2014.

[8] Jeffrey Adler, John Horner, Jeanette Dyer, Alan Toppen, Lisa Burgess, Greg Hatcher, et al. Estimate benefits of crowdsourced data from social media. Technical report, United States. Dept. of Transportation. ITS Joint Program Office, 2014.

[9] Nicole Adler and Joseph Berechman. Measuring airport quality from the airlines' viewpoint: an application of data envelopment analysis. *Transport Policy*, 8(3):171 – 181, 2001.

[10] Mohammed N Ahmed, Gianni Barlacchi, Stefano Braghin, Francesco Calabrese, Michele Ferretti, Vincent Lonij, Rahul Nair, Rana Novack, Jurij Paraszczak, and Andeep S Toor. A multi-scale approach to data-driven mass migration analysis. In *SoGood@ ECML-PKDD*, 2016.

[11] Merve Alanyali, Helen Susannah Moat, and Tobias Preis. Quantifying the relationship between financial news and the stock market. *Scientific reports*, 3:3578, 2013.

[12] Benjamin M Althouse, Yih Yng Ng, and Derek AT Cummings. Prediction of dengue incidence using search query surveillance. *PLoS neglected tropical diseases*, 5(8):e1258, 2011.

[13] Theodoros Anagnostopoulos, Christos Anagnostopoulos, and Stathes Hadjiefthymiades. Mobility prediction based on machine learning. In *Mobile Data Management (MDM), 2011 12th IEEE International Conference on*, volume 2, pages 27–30. IEEE, 2011.

[14] Daniel Ashbrook and Thad Starner. Using gps to learn significant locations and predict movement across multiple users. *Personal and Ubiquitous computing*, 7(5):275–286, 2003.

[15] Nikolaos Askitas and Klaus F Zimmermann. Google econometrics and unemployment forecasting. *Applied Economics Quarterly*, 55(2):107–120, 2009.

[16] Sitaram Asur and Bernardo A Huberman. Predicting the future with social media. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01*, pages 492–499. IEEE Computer Society, 2010.

[17] C. Aubry, J. Ramamonjisoa, M.-H. Dabat, J. Rakotoarisoa, J. Rakotondraibe, and L. Rabeharisoa. Urban agriculture and land use in cities: An approach with the multifunctionality and sustainability concepts in the case of Antananarivo (Madagascar). *Land Use Policy*, 29(2):429 – 439, 2012.

[18] John W Ayers, Benjamin M Althouse, and Mark Dredze. Could behavioral medicine lead the web data revolution? *Jama*, 311(14):1399–1400, 2014.

[19] Alberto Baffigi, Roberto Golinelli, and Giuseppe Parigi. Bridge models to forecast the euro area gdp. *International Journal of forecasting*, 20(3):447–460, 2004.

[20] Scott Baker, Andry Fradkin, et al. What drives job search? evidence from google search data. *Discussion Papers*, pages 10–020, 2011.

[21] Duygu Balcan, Bruno Gonçalves, Hao Hu, José J Ramasco, Vittoria Colizza, and Alessandro Vespignani. Modeling the spatial spread of infectious diseases: The global epidemic and mobility computational model. *Journal of computational science*, 1(3):132–145, 2010.

[22] Duygu Balcan, Bruno Gonçalves, Hao Hu, José J. Ramasco, Vittoria Colizza, and Alessandro Vespignani. Modeling the spatial spread of infectious diseases: The global epidemic and mobility computational model. *Journal of Computational Science*, 1(3):132 – 145, 2010.

[23] Duygu Balcan, Hao Hu, Bruno Goncalves, Paolo Bajardi, Chiara Poletto, José J Ramasco, Daniela Paolotti, Nicola Perra, Michele Tizzoni, Wouter Van den Broeck, Vittoria Colizza, and Alessandro Vespignani. Seasonal transmission potential and activity peaks of the new influenza a(h1n1): a monte carlo likelihood analysis based on human mobility. *BMC Medicine*, 7(45), 2009.

[24] Philip Ball. *Why Society is a Complex Matter*. Springer, 2012.

[25] Marta Bańbura, Domenico Giannone, and Lucrezia Reichlin. Nowcasting. *European Central Bank Working Paper Series*, 1275, 2010.

[26] Zhifeng Bao, Yong Zeng, and YC Tay. sonlp: social network link prediction by principal component regression. In *Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on*, pages 364–371. IEEE, 2013.

[27] Albert-László Barabási. *Bursts: the hidden patterns behind everything we do, from your e-mail to bloody crusades*. Penguin, 2010.

[28] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[29] Gianni Barlacchi, Christos Perentis, Abhinav Mehrotra, Mirco Musolesi, and Bruno Lepri. Are you getting sick? Predicting influenza-like symptoms using human mobility behaviors. *EPJ Data Science*, 6(1), oct 2017.

[30] Dominique Barth, Samir Bellahsene, and Leila Kloul. Combining local and global profiles for mobility prediction in lte femtocells. In *Proceedings of the 15th ACM international conference on Modeling, analysis and simulation of wireless and mobile systems*, pages 333–342. ACM, 2012.

[31] Michael Batty. Big data, smart cities and city planning. *Dialogues in Human Geography*, 3(3):274–279, 2013.

[32] J. Beaumont and J. Thomas. Measuring national well-being - health. Technical report, UK Office for National Statistics, 2012.

[33] Vitaly Belik, Theo Geisel, and Dirk Brockmann. Natural human mobility patterns and spatial spread of infectious diseases. *Phys. Rev. X*, 1:011001, Aug 2011.

[34] Linus Bengtsson, Xin Lu, Anna Thorson, Richard Garfield, and Johan Von Schreeb. Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: a post-earthquake geospatial study in haiti. *PLoS medicine*, 8(8):e1001083, 2011.

[35] A Bernardini, C De Vitiis, A Guandalini, F Inglese, and MD Terribili. Measuring inflation through different sampling designs implemented on scanner data.

[36] Luís M. A. Bettencourt, José Lobo, Dirk Helbing, Christian Kühnert, and Geoffrey B. West. Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the National Academy of Sciences*, 104(17):7301–7306, 2007.

[37] Luís MA Bettencourt. The origins of scaling in cities. *science*, 340(6139):1438–1441, 2013.

[38] Lucy Biddle, Jenny Donovan, Keith Hawton, Navneet Kapur, and David Gunnell. Public health: Suicide and the internet. *BMJ: British Medical Journal*, 336(7648):800, 2008.

[39] Mustafa Bilgic, Galileo Mark Namata, and Lise Getoor. Combining collective classification and link prediction. *ICDMW*, pages 381–386, 2007.

[40] Catherine A. Bliss, Morgan R. Frank, Christopher M. Danforth, and Peter Sheridan Dodds. An evolutionary algorithm approach to link prediction in dynamic social networks. *Journal of Computational Science*, 5(5):750–764, 2013.

[41] 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(10):P10008, 2008.

[42] Joshua Blumenstock, Gabriel Cadamuro, and Robert On. Predicting poverty and wealth from mobile phone metadata. *Science*, 350(6264):1073–1076, 2015.

[43] Andrey Bogomolov, Bruno Lepri, Michela Ferron, Fabio Pianesi, and Alex Sandy Pentland. Daily stress recognition from mobile phone data, weather conditions and individual traits. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 477–486. ACM, 2014.

[44] Andrey Bogomolov, Bruno Lepri, and Fabio Pianesi. Happiness recognition from mobile phone data. In *Social Computing (SocialCom), 2013 International Conference on*, pages 790–795. IEEE, 2013.

[45] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of computational science*, 2(1):1–8, 2011.

[46] Ilaria Bordino, Stefano Battiston, Guido Caldarelli, Matthieu Cristelli, Antti Ukkonen, and Ingmar Weber. Web search queries can predict stock market volumes. *PloS one*, 7(7):e40014, 2012.

[47] Federico Botta, Helen Susannah Moat, and Tobias Preis. Quantifying crowd size with mobile phone and twitter data. *Royal Society open science*, 2(5):150162, 2015.

[48] Kate J Bowers, Shane D Johnson, and Ken Pease. Prospective hot-spotting: The future of crime mapping? *British Journal of Criminology*, 44(5):641–658, 2004.

[49] Margaret M Bradley and Peter J Lang. Affective norms for english words (anew): Instruction manual and affective ratings. Technical report, Citeseer, 1999.

[50] Björn Bringmann, Michele Berlingerio, Francesco Bonchi, and Aristides Gionis. Learning and predicting the evolution of social networks. *IEEE Intelligent Systems*, 25(4):26–35, 2010.

[51] D. Brockmann, L. Hufnagel, and T. Geisel. The scaling laws of human travel. *Nature*, 439(7075):462–465, 01 2006.

[52] Dirk Brockmann and Fabian Theis. Money circulation, trackable items, and the emergence of universal human mobility patterns. *IEEE Pervasive Computing*, 7(4), 2008.

[53] John S Brownstein, Shuyu Chu, Achla Marathe, Madhav V Marathe, Andre T Nguyen, Daniela Paolotti, Nicola Perra, Daniela Perrotta, Mauricio Santillana, Samarth Swarup, Michele Tizzoni, Alessandro Vespignani, Anil Kumar S Vullikanti, Mandy L Wilson, and Qian Zhang. Combining Participatory Influenza Surveillance with Modeling and Forecasting: Three Alternative Approaches. *JMIR Public Health and Surveillance*, 3(4):e83, nov 2017.

[54] 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(21):2153–2157, 2009.

[55] John S Brownstein, Clark C Freifeld, Ben Y Reis, and Kenneth D Mandl. Surveillance sans frontieres: Internet-based emerging infectious disease intelligence and the healthmap project. *PLoS medicine*, 5(7):e151, 2008.

[56] Ingrid Burbey. *Predicting future locations and arrival times of individuals.* PhD thesis, Virginia Tech, 2011.

[57] Declan Butler. When google got flu wrong. *Nature*, 494:155–156, 2013.

[58] Declan Butler. When google got flu wrong. *Nature*, 494(7436):155, 2013.

[59] Luca Canzian and Mirco Musolesi. Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*, pages 1293–1304. ACM, 2015.

[60] Sandra J Carlson, Craig B Dalton, David N Durrheim, and John Fejsa. Online flutracking survey of influenza-like illness during pandemic (h1n1) 2009, australia. *Emerging infectious diseases*, 16(12):1960, 2010.

[61] Sandra J Carlson, David N Durrheim, and Craig B Dalton. Flutracking provides a measure of field influenza vaccine effectiveness, australia, 2007–2009. *Vaccine*, 28(42):6809–6810, 2010.

[62] Jennifer L Castle, Nicholas WP Fawcett, and David F Hendry. Nowcasting is not just contemporaneous forecasting. *National Institute Economic Review*, 210(1):71–89, 2009.

[63] Charlie Catlett, Tanu Malik, Brett Goldstein, Jonathan Giuffrida, Yetong Shao, Alessandro Panella, Derek Eder, Eric van Zanten, Robert Mitchum, Severin Thaler, et al. Plenario: An open data discovery and exploration platform for urban science. *IEEE Data Eng. Bull.*, 37(4):27–42, 2014.

[64] Ciro Cattuto, Wouter Van den Broeck, Alain Barrat, Vittoria Colizza, Jean-François Pinton, and Alessandro Vespignani. Dynamics of person-to-person interactions from distributed rfid sensor networks. *PloS one*, 5(7):e11596, 2010.

[65] Michelangelo Ceci, Annalisa Appice, and Donato Malerba. Time-slice density estimation for semantic-based tourist destination suggestion. In *ECAI*, pages 1107–1108, 2010.

[66] Simone Centellegher, Marco De Nadai, Michele Caraviello, Chiara Leonardi, Michele Vescovi, Yusi Ramadian, Nuria Oliver, Fabio Pianesi, Alex Pentland, Fabrizio Antonelli, et al. The mobile territorial lab: a multilayered and dynamic view on parents' daily lives. *EPJ Data Science*, 5(1):3, 2016.

[67] Andrea Ceron, Luigi Curini, and Stefano Maria Iacus. *Social Media e Sentiment Analysis: L'evoluzione dei fenomeni sociali attraverso la Rete*, volume 9. Springer Science & Business Media, 2014.

[68] Andrea Ceron, Luigi Curini, and Stefano Maria Iacus. isa: A fast, scalable and accurate algorithm for sentiment analysis of social media content. *Information Sciences*, 367:105–124, 2016.

[69] Emily H Chan, Vikram Sahai, Corrie Conrad, and John S Brownstein. Using web search query data to monitor dengue epidemics: a new model for neglected tropical disease surveillance. *PLoS neglected tropical diseases*, 5(5):e1206, 2011.

[70] Ling Chen, Mingqi Lv, and Gencai Chen. A system for destination and future route prediction based on trajectory mining. *Pervasive and Mobile Computing*, 6(6):657–676, 2010.

[71] Meng Chen, Yang Liu, and Xiaohui Yu. Nlpmm: A next location predictor with markov modeling. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 186–197. Springer, 2014.

[72] Hyunyoung Choi and Hal Varian. Predicting initial claims for unemployment benefits. *Google Inc*, pages 1–5, 2009.

[73] Hyunyoung Choi and Hal Varian. Predicting the present with google trends. *Technical Report*, 2009.

[74] Hyunyoung Choi and Hal Varian. Predicting the present with google trends. *Economic Record*, 88(s1):2–9, 2012.

[75] Jean-Paul Chretien, Dylan George, Jeffrey Shaman, Rohit A Chitale, and F Ellis McKenzie. Influenza forecasting in human populations: a scoping review. *PloS one*, 9(4):e94130, 2014.

[76] Rumi Chunara, Susan Aman, Mark Smolinski, and John S Brownstein. Flu near you: an online self-reported influenza surveillance system in the usa. *Online Journal of Public Health Informatics*, 5(1), 2013.

[77] Rumi Chunara, Jason R Andrews, and John S Brownstein. Social and news media enable estimation of epidemiological patterns early in the 2010 haitian cholera outbreak. *The American journal of tropical medicine and hygiene*, 86(1):39–45, 2012.

[78] Rumi Chunara, Edward Goldstein, Oscar Patterson-Lomba, and John S Brownstein. Estimating influenza attack rates in the united states using a participatory cohort. *Scientific reports*, 5:9540, 2015.

[79] Aaron Clauset, Cristopher Moore, and Mark EJ Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453(7191):98, 2008.

[80] Samantha Cook, Corrie Conrad, Ashley L. Fowlkes, and Matthew H. Mohebbi. Assessing google flu trends performance in the united states during the 2009 influenza virus a (h1n1) pandemic. *PLOS One*, 6, 2011.

[81] Crystale Purvis Cooper, Kenneth P Mallon, Steven D. Leadbetter, Lori A. Pollack, Lucy A. Peipins, and J. K. Jansen. Cancer internet search activity on a major search engine, united states 2001-2003. In *Journal of medical Internet research*, 2005.

[82] Patrick Copeland, Raquel Romano, Tom Zhang, Greg Hecht, Dan Zigmond, and Christian Stefansen. Google disease trends: an update. *Nature*, 457:1012–1014, 2013.

[83] Courtney Corley, Armin R Mikler, Karan P Singh, and Diane J Cook. Monitoring influenza trends through mining social media. In *BIOCOMP*, pages 340–346, 2009.

[84] Michele Coscia, Salvatore Rinzivillo, Fosca Giannotti, and Dino Pedreschi. Optimal spatial resolution for the analysis of human mobility. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, pages 248–252. IEEE Computer Society, 2012.

[85] Michele Coscia, Giulio Rossetti, Fosca Giannotti, and Dino Pedreschi. Demon: a local-first discovery method for overlapping communities. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 615–623. ACM, 2012.

[86] Lorenzo Coviello, Yunkyu Sohn, Adam DI Kramer, Cameron Marlow, Massimo Franceschetti, Nicholas A Christakis, and James H Fowler. Detecting emotional contagion in massive social networks. *PloS one*, 9(3):e90315, 2014.

[87] David Crandall, Dan Cosley, Daniel Huttenlocher, Jon Kleinberg, and Siddharth Suri. Feedback effects between similarity and social influence in online communities. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 160–168. ACM, 2008.

[88] AW Crawley, O Wojcik, J Olsen, J Brownstein, and M Smolinski. Flu near you: Comparing crowd-sourced reports of influenza-like illness to the cdc outpatient influenza-like illness surveillance network, october 2012 to march 2014. In *2014 CSTE Annual Conference. Cste*, page 1, 2014.

[89] Peter Csermely. Strong links are important, but weak links stabilize them. *Trends in biochemical sciences*, 29(7):331–334, 2004.

[90] Luigi Curini, Stefano Iacus, and Luciano Canova. Measuring idiosyncratic happiness through the analysis of twitter: An application to the italian case. *Social Indicators Research*, 121(2):525–542, 2015.

[91] 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(32):11600–11605, 2014.

[92] Paulo Ricardo da Silva Soares and Ricardo Bastos Cavalcante Prudêncio. Time series based link prediction. In *Neural Networks (IJCNN), The 2012 International Joint Conference on*, pages 1–7. IEEE, 2012.

[93] Craig Dalton, David Durrheim, John Fejsa, Lynn Francis, Sandra Carlson, Edouard Tursan d'Espaignet, Frank Tuyl, et al. Flutracking: a weekly australian community online survey of influenza-like illness in 2006, 2007 and 2008. *Communicable diseases intelligence quarterly report*, 33(3):316, 2009.

[94] Craig B Dalton, Sandra J Carlson, Michelle T Butler, Elissa Elvidge, and David N Durrheim. Building influenza surveillance pyramids in near real time, australia. *Emerging infectious diseases*, 19(11):1863, 2013.

[95] Craig B Dalton, Sandra J Carlson, Lisa McCallum, Michelle T Butler, John Fejsa, Elissa Elvidge, and David N Durrheim. Flutracking weekly online community survey of influenza-like illness: 2013 and 2014. *Commun Dis Intell Q Rep*, 39(3):E361–E368, 2015.

[96] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. Predicting depression via social media. *ICWSM*, 13:1–10, 2013.

[97] Juan de Dios OrtÃozar and Luis G Willumsen. *Modelling transport*. John Wiley & Sons, 2011.

[98] Manlio De Domenico, Antonio Lima, and Mirco Musolesi. Interdependence and predictability of human mobility and social interactions. *Pervasive and Mobile Computing*, 9(6):798–807, 2013.

[99] Mercedes Delgado, Michael E Porter, and Scott Stern. Clusters, convergence, and economic performance. *Research policy*, 43(10):1785–1799, 2014.

[100] Nicolás Della Penna, Haifang Huang, et al. Constructing consumer sentiment index for us using google searches. Technical report, 2010.

[101] Trinh Minh Tri Do and Daniel Gatica-Perez. Where and what: Using smartphones to predict next locations and applications in daily life. *Pervasive and Mobile Computing*, 12:79–91, 2014.

[102] Peter Sheridan Dodds, Kameron Decker Harris, Isabel M Kloumann, Catherine A Bliss, and Christopher M Danforth. Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *PloS one*, 6(12):e26752, 2011.

[103] Yuxiao Dong, Jie Tang, Sen Wu, Jilei Tian, Nitesh V Chawla, Jinghai Rao, and Huanhuan Cao. Link prediction and recommendation across heterogeneous social networks. In *Data mining (ICDM), 2012 IEEE 12th international conference on*, pages 181–190. IEEE, 2012.

[104] Kurt Dopfer, John Foster, and Jason Potts. Micro-meso-macro. *Journal of evolutionary economics*, 14(3):263–279, 2004.

[105] Sangita Dubey and Pietro Gennari. Now-casting food consumer price indexes with big data: Public-private complementarities. *Food and Agriculture Organization of the United Nations Statistics Division*, 2014.

[106] Vanja M Dukic, Michael Z David, and Diane S Lauderdale. Internet queries and methicillin-resistant staphylococcus aureus surveillance. *Emerging infectious diseases*, 17(6):1068, 2011.

[107] Brendan Duncan and Charles Elkan. Nowcasting with numerous candidate predictors. *ECML PKDD*, pages 370–385, 2014.

[108] Martine Durand. The oecd better life initiative: How's life? and the measurement of well-being. *Review of Income and Wealth*, 61(1):4–17, 2015.

[109] Francesco D'Amuri and Juri Marcucci. 'google it!'forecasting the us unemployment rate with a google job search index. 2010.

[110] Nathan Eagle, Michael Macy, and Rob Claxton. Network diversity and economic development. *Science*, 328:1029–1031, 2010.

[111] Nathan Eagle and Alex Sandy Pentland. Reality mining: sensing complex social systems. *Personal and ubiquitous computing*, 10(4):255–268, 2006.

[112] W John Edmunds and Sebastian Funk. Using the internet to estimate influenza vaccine effectiveness. *Expert review of vaccines*, 11(9):1027–1029, 2012.

[113] Sven Erlander and Neil F Stewart. *The gravity model in transportation analysis: theory and extensions*, volume 3. Vsp, 1990.

[114] Jeremy U Espino, William R Hogan, and Michael M Wagner. Telephone triage: a timely data source for surveillance of influenza-like diseases. In *AMIA Annual Symposium Proceedings*, volume 2003, page 215. American Medical Informatics Association, 2003.

[115] Michael Ettredge, John Gerdes, and Gilbert Karuga. Using web-based search data to predict macroeconomic statistics. *Communications of the ACM*, 48(11):87–92, 2005.

[116] Leonhard Euler. Solutio problematis ad geometriam situs pertinentis. *Commentarii Academiae Scientiarum Imperialis Petropolitanae*, 1736.

[117] Nicholas G. Reich Evan L. Ray. Prediction of infectious disease epidemics via weighted density ensembles. *arXiv*, 1703.10936, 2017.

[118] Gunther Eysenbach. Infodemiology: tracking flu-related searches on the web for syndromic surveillance. In *AMIA Annual Symposium Proceedings*, volume 2006, page 244. American Medical Informatics Association, 2006.

[119] Gunther Eysenbach. Infodemiology and infoveillance: tracking online health information and cyberbehavior for public health. *American journal of preventive medicine*, 40(5):S154–S158, 2011.

[120] Kai Fan, Marisa Eisenberg, Alison Walsh, Allison Aiello, and Katherine Heller. Hierarchical graph-coupled hmms for heterogeneous personalized health data. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 239–248. ACM, 2015.

[121] Xu Feng, JC Zhao, and Ke Xu. Link prediction in complex networks: a clustering perspective. *The European Physical Journal B*, 85(1):3, 2012.

[122] Neil M Ferguson, Derek AT Cummings, Simon Cauchemez, Christophe Fraser, Steven Riley, Aronrag Meeyai, Sopon Iamsirithaworn, and Donald S Burke. Strategies for containing an emerging influenza pandemic in southeast asia. *Nature*, 437(7056):209, 2005.

[123] M. Fiedler. *Special matrices and their applications in numerical mathematics*. Martinus Nijhoff Publishers, Dodrecht, The Netherlands, 1986.

[124] Michael Fire, Rami Puzis, and Yuval Elovici. Link prediction in highly fractional data sets. In *Handbook of computational approaches to counterterrorism*, pages 283–300. Springer, 2013.

[125] Claudia Foroni and Massimiliano Marcellino. A comparison of mixed frequency approaches for nowcasting euro area macroeconomic aggregates. *International Journal of Forecasting*, 30(3):554–568, 2014.

[126] J. H. Fowler and N. A. Christakis. Dynamic spread of happiness in a large social network: Longitudinal analysis over 20 years in the framingham heart study. *BMJ*, 2008.

[127] Susannah Fox. *Online health search 2006*. Pew Internet & American Life Project, 2006.

[128] Enrique Frias-Martinez, Graham Williamson, and Vanessa Frias-Martinez. An agent-based model of epidemic spread using human mobility and social network information. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pages 57–64. IEEE, 2011.

[129] Vanessa Frias-Martinez and Jesus Virseda. On the relationship between socio-economic factors and cell phone usage. In *Proceedings of the Fifth International Conference on Information and Communication Technologies and Development*, ICTD '12, pages 76–84, New York, NY, USA, 2012. ACM.

[130] Nir Friedman, Lise Getoor, Daphne Koller, and Avi Pfeffer. Learning probabilistic relational models. In *IJCAI*, volume 99, pages 1300–1309, 1999.

[131] Barbara Furletti, Lorenzo Gabrielli, Chiara Renso, and Salvatore Rinzivillo. Analysis of gsm calls data for understanding user mobility behavior. In *Big Data, 2013 IEEE International Conference on*, pages 550–555. IEEE, 2013.

[132] John W Galbraith and Greg Tkacz. Nowcasting gdp with electronic payments data. Technical report, ECB Statistics Paper, 2015.

[133] Gallup-Healthways. Well-being index.

[134] Gallup-Healthways. State of well-being 2011: City, state and congressional district wellbeing reports. Technical report, Gallup Inc., 2012. Available: http://www.well-beingindex.com/files/2011CompositeReport.pdf.

[135] Aurelien Gautreau, Alain Barrat, and Marc Barthélemy. Microdynamics in stationary complex networks. *Proceedings of the National Academy of Sciences*, 106(22):8847–8852, 2009.

[136] Domenico Giannone, Lucrezia Reichlin, and David Small. Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics*, 55(4):665–676, 2008.

[137] Fosca Giannotti, Mirco Nanni, Dino Pedreschi, Fabio Pinelli, Chiara Renso, Salvatore Rinzivillo, and Roberto Trasarti. Unveiling the complexity of human mobility by querying and mining massive trajectory data. *The VLDB Journal*, 20(5):695, 2011.

[138] Fosca Giannotti and Dino Pedreschi. *Mobility, data mining and privacy: Geographic knowledge discovery*. Springer Science & Business Media, 2008.

[139] 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. *The European Physical Journal Special Topics*, 214(1):49–75, 2012.

[140] Győző Gidófalvi and Fang Dong. When and where next: individual mobility prediction. In *Proceedings of the First ACM SIGSPATIAL International Workshop on Mobile Geographic Information Systems*, pages 57–64. ACM, 2012.

[141] Eric Gilbert and Karrie Karahalios. Predicting tie strength with social media. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 211–220. ACM, 2009.

[142] David Gillen and Ashish Lall. Developing measures of airport productivity and performance: an application of data envelopment analysis. *Transportation Research Part E: Logistics and Transportation Review*, 33(4):261 – 273, 1997.

[143] 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, 2009.

[144] Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826, 2002.

[145] Sharad Goel, Jake M. Hofman, Sébastien Lahaie, David M. Pennock, and Duncan J. Watts. Predicting consumer behavior with web search. *PNAS*, 107(41):17486–17490, 2010.

[146] Scott A Golder and Michael W Macy. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science*, 333(6051):1878–1881, 2011.

[147] Joao Bártolo Gomes, Clifton Phua, and Shonali Krishnaswamy. Where will you go? mobile data mining for next place prediction. In *International Conference on Data Warehousing and Knowledge Discovery*, pages 146–158. Springer, 2013.

[148] Bruno Goncalves, Nicola Perra, and Alessandro Vespignani. Validation of dunbar's number in twitter conversations. *arXiv preprint arXiv:1105.5170*, 2011.

[149] Marta C. Gonzalez, Cesar A. Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 06 2008.

[150] Michael F Goodchild and J Alan Glennon. Crowdsourcing geographic information for disaster response: a research frontier. *International Journal of Digital Earth*, 3(3):231–241, 2010.

[151] Mark S. Granovetter. The strength of weak ties. *American Journal of Sociology*, 78(6):1360–1380, 1973.

[152] Paolo La Greca, Daniele La Rosa, Francesco Martinico, and Riccardo Privitera. Agricultural and green infrastructures: The role of non-urbanised areas for eco-sustainable planning in a metropolitan region. *Environmental Pollution*, 159(8):2193 – 2202, 2011. Selected papers from the conference Urban Environmental Pollution: Overcoming Obstacles to Sustainability and Quality of Life (UEP2010), 20-23 June 2010, Boston, USA.

[153] Thomas W Grein, KB Kamara, Guénaël Rodier, Aileen J Plant, Patrick Bovier, Michael J Ryan, Takaaki Ohyama, and David L Heymann. Rumors of disease in the global village: outbreak verification. *Emerging infectious diseases*, 6(2):97, 2000.

[154] Agnes Gruenerbl, Venet Osmani, Gernot Bahle, Jose C Carrasco, Stefan Oehler, Oscar Mayora, Christian Haring, and Paul Lukowicz. Using smart phone mobility traces for the diagnosis of depressive and manic episodes in bipolar patients. In *Proceedings of the 5th Augmented Human International Conference*, page 38. ACM, 2014.

[155] Daniel Gruhl, Ramanathan Guha, Ravi Kumar, Jasmine Novak, and Andrew Tomkins. The predictive power of online chatter. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 78–87. ACM, 2005.

[156] Riccardo Guidotti, Michele Coscia, Dino Pedreschi, and Diego Pennacchioli. Going Beyond GDP to Nowcast Well-Being Using Retail Market Data. In *Advances in Network Science*, pages 29–42. Springer International Publishing, 2016.

[157] Riccardo Guidotti, Anna Monreale, Mirco Nanni, Fosca Giannotti, and Dino Pedreschi. Clustering Individual Transactional Data for Masses of Users. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '17*. ACM Press, 2017.

[158] Riccardo Guidotti, Giulio Rossetti, Luca Pappalardo, Fosca Giannotti, and Dino Pedreschi. Market Basket Prediction Using User-Centric Temporal Annotated Recurring Sequences. In *2017 IEEE International Conference on Data Mining (ICDM)*. IEEE, nov 2017.

[159] Giselle Guzman. Internet search behavior as an economic forecasting tool: The case of inflation expectations. *Journal of economic and social measurement*, 36(3):119–167, 2011.

[160] UN Habitat. State of the world's cities 2010/2011: bridging the urban divide. *Earthscan, London*, 2010.

[161] Keith M Harris, John P McLean, and Jeanie Sheffield. Examining suicide-risk individuals who go online for suicide-related purposes. *Archives of Suicide Research*, 13(3):264–276, 2009.

[162] Tanja Hartmann, Andrea Kappes, and Dorothea Wagner. Clustering evolving networks. *arXiv preprint arXiv:1401.3516*, 2014.

[163] Mohammad Hashemian, Dylan Knowles, Jonathan Calver, Weicheng Qian, Michael C Bullock, Scott Bell, Regan L Mandryk, Nathaniel Osgood, and Kevin G Stanley. iepi: an end to end solution for collecting, conditioning and utilizing epidemiologically relevant data. In *Proceedings of the 2nd ACM international workshop on Pervasive Wireless Healthcare*, pages 3–8. ACM, 2012.

[164] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer New York, 2009.

[165] David Heckerman, Chris Meek, and Daphne Koller. Probabilistic entity-relationship models, prms, and plate models. *Introduction to statistical relational learning*, pages 201–238, 2007.

[166] John Helliwell, Richard Layard, and Jeffrey Sachs. World happiness report. 2012.

[167] Helen Herrman, Shekhar Saxena, Rob Moodie, World Health Organization, et al. Promoting mental health: concepts, emerging evidence, practice: a report of the world health organization, department of mental health and substance abuse in collaboration with the victorian health promotion foundation and the university of melbourne. 2005.

[168] Sarah Hetherington and Christian Madsbjerg. The thin data revolution. 2014.

[169] David L Heymann, Guénaël R Rodier, et al. Hot spots in a wired world: Who surveillance of emerging and re-emerging infectious diseases. *The Lancet infectious diseases*, 1(5):345–353, 2001.

[170] C. S. Holling. The components of predation as revealed by a study of small-mammal predation of the european pine sawfly. *The Canadian Entomologist*, 91(5):293–320, 1959.

[171] P.G Hooper and D.A Hensher. Measuring total factor productivity of airports— an index number approach. *Transportation Research Part E: Logistics and Transportation Review*, 33(4):249 – 259, 1997.

[172] Shu Huang, Min Chen, Bo Luo, and Dongwon Lee. Predicting aggregate social activities using continuous-time stochastic process. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 982–991. ACM, 2012.

[173] Anette Hulth, Gustaf Rydevik, and Annika Linde. Web queries as a source for syndromic surveillance. *PloS one*, 4(2):e4378, 2009.

[174] Ian Humphreys and Graham Francis. Performance measurement: a review of airports. *International Journal of Transport Management*, 1(2):79 – 85, 2002.

[175] Mikael Huss and Petter Holme. Currency and commodity metabolites: their identification and relation to the modularity of metabolic networks. *IET systems biology*, 1(5):280–285, 2007.

[176] Stefano Maria Iacus, Giuseppe Porro, Silvia Salini, and Elena Siletti. Social networks, happiness and health: from sentiment analysis to a multidimensional indicator of subjective well-being. *arXiv preprint arXiv:1512.01569*, 2015.

[177] WHO Influenza. Fact sheet no. 211, march 2014. *URL: http://www. who. int/mediacentre/factsheets/fs211/en*, 2014.

[178] Kazem Jahanbakhsh, Valerie King, and Gholamali C Shoja. Predicting human contacts in mobile social networks using supervised learning. In *Proceedings of the Fourth Annual Workshop on Simplifying Complex Networks for Practitioners*, pages 37–42. ACM, 2012.

[179] Natasha Jaques, Sara Taylor, Akane Sano, and Rosalind Picard. Multi-task, multi-kernel learning for estimating individual wellbeing. In *Proc. NIPS Workshop on Multimodal Machine Learning, Montreal, Quebec*, volume 898, 2015.

[180] Neal Jean, Marshall Burke, Michael Xie, W. Matthew Davis, David B. Lobell, and Stefano Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794, 2016.

[181] Hoyoung Jeung, Qing Liu, Heng Tao Shen, and Xiaofang Zhou. A hybrid prediction model for moving objects. In *Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on*, pages 70–79. Ieee, 2008.

[182] Hoyoung Jeung, Man Lung Yiu, Xiaofang Zhou, and Christian S Jensen. Path prediction and predictive range querying in road network databases. *The VLDB Journal*, 19(4):585–602, 2010.

[183] Shan Jiang, Joseph Ferreira, and Marta C. González. Clustering daily patterns of human activities in the city. *Data Mining and Knowledge Discovery*, 25(3):478–510, 2012.

[184] Heather A Johnson, Michael M Wagner, William R Hogan, Wendy W Chapman, Robert T Olszewski, John N Dowling, Gary Barnas, et al. Analysis of web access logs for surveillance of influenza. In *Medinfo*, pages 1202–1206, 2004.

[185] Neil Johnson, Spencer Carran, Joel Botner, Kyle Fontaine, Nathan Laxague, Philip Nuetzel, Jessica Turnley, and Brian Tivnan. Pattern in escalations in insurgent and terrorist activity. *Science*, 333(6038):81–84, 2011.

[186] NF Johnson, M Zheng, Y Vorobyeva, A Gabriel, H Qi, N Velasquez, P Manrique, D Johnson, E Restrepo, C Song, et al. New online ecology of adversarial aggregates: Isis and beyond. *Science*, 352(6292):1459–1463, 2016.

[187] Shane D Johnson and Kate J Bowers. The burglary as clue to the future: The beginnings of prospective hot-spotting. *European Journal of Criminology*, 1(2):237–255, 2004.

[188] Mohamed Kafsi, Ehsan Kazemi, Lucas Maystre, Lyudmila Yartseva, Matthias Grossglauser, and Patrick Thiran. Mitigating epidemics through mobile micro-measures. *arXiv preprint arXiv:1307.2084*, 2013.

[189] Juyoung Kang and Hwan-Seung Yong. A frequent pattern based prediction model for moving objects. *Int. J. Comput. Sci. Netw. Secur*, 10(3):200–205, 2010.

[190] Sang-Wook Kim, Jung-Im Won, Jong-Dae Kim, Miyoung Shin, Junghoon Lee, and Hanil Kim. Path prediction of moving objects on road networks through analyzing past trajectories. In *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, pages 379–389. Springer, 2007.

[191] Gary King. Ensuring the data-rich future of the social sciences. *science*, 331(6018):719–721, 2011.

[192] John Kitchen and Ralph Monaco. Real-time forecasting in practice: The us treasury staff's real-time gdp forecast system. 2003.

[193] Aniket Kittur and Robert E Kraut. Harnessing the wisdom of crowds in wikipedia: quality through coordination. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, pages 37–46. ACM, 2008.

[194] Gary Koop and Luca Onorante. Macroeconomic nowcasting using google probabilities. *University of Strathclyde*, 2013.

[195] Michal Kosinski, David Stillwell, and Thore Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805, 2013.

[196] 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(24):8788–8790, 2014.

[197] Ladislav Kristoufek. Bitcoin meets google trends and wikipedia: Quantifying the relationship between phenomena of the internet era. *Scientific reports*, 3:3415, 2013.

[198] Ladislav Kristoufek. Can google trends search queries contribute to risk diversification? *Scientific reports*, 3:2713, 2013.

[199] Ladislav Kristoufek, Helen Susannah Moat, and Tobias Preis. Estimating suicide occurrence statistics using google trends. *EPJ data science*, 5(1):32, 2016.

[200] Ladislav Kristoufek, Helen Susannah Moat, and Tobias Preis. Estimating current suicide rates using google trends. *World Academy of Science, Engineering and Technology, International Journal of Humanities and Social Sciences*, 4(5), 2017.

[201] John Krumm and Eric Horvitz. Predestination: Inferring destinations from partial trajectories. In *International Conference on Ubiquitous Computing*, pages 243–260. Springer, 2006.

[202] Vasileios Lampos and Nello Cristianini. Tracking the flu pandemic by monitoring the social web. In *Cognitive Information Processing (CIP), 2010 2nd International Workshop on*, pages 411–416. IEEE, 2010.

[203] Vasileios Lampos and Nello Cristianini. Nowcasting events from the social web with statistical learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(4):72, 2012.

[204] Annette Lancy and Nicholas Gruen. Constructing the herald/age–lateral economics index of australia's wellbeing. *Australian Economic Review*, 46(1):92–102, 2013.

[205] Thomas Lansdall-Welfare, Vasileios Lampos, and Nello Cristianini. Nowcasting the mood of the nation. *Significance*, 9(4):26–28, 2012.

[206] Neal Lathia, Daniele Quercia, and Jon Crowcroft. The hidden image of the city: sensing community well-being from urban mobility. In *International Conference on Pervasive Computing*, pages 91–98. Springer, 2012.

[207] David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. The parable of google flu: Traps in big data analysis. *Science Magazine*, 343(6176):1203–1205, 2014.

[208] David Lazer, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. Life in the network: the coming age of computational social science. *Science (New York, NY)*, 323(5915):721, 2009.

[209] W. Lefebvre, B. Degrawe, C. Beckx, M. Vanhulsel, B. Kochan, T. Bellemans, D. Janssens, G. Wets, S. Janssen, I. de Vlieger, L. Int Panis, and S. Dhondt. Presentation and evaluation of an integrated model chain to respond to traffic- and health-related policy questions. *Environmental Modelling and Software*, 40:160 – 170, 2013.

[210] Po-Ruey Lei, Tsu-Jou Shen, Wen-Chih Peng, and Jiunn Su. Exploring spatial-temporal trajectory model for location prediction. In *Mobile Data Management (MDM), 2011 12th IEEE International Conference on*, volume 1, pages 58–67. IEEE, 2011.

[211] Adrian Letchford, Tobias Preis, and Helen Susannah Moat. Quantifying the search behaviour of different demographics using google correlate. *PloS one*, 11(2):e0149025, 2016.

[212] Chrysa Leventi, Jekaterina Navicke, Olga Rastrigina, and Holly Sutherland. Nowcasting the income distribution in europe. 2014.

[213] Hongjun Li, Changjie Tang, Shaojie Qiao, Yue Wang, Ning Yang, and Chuan Li. Hotspot district trajectory prediction. In *International Conference on Web-Age Information Management*, pages 74–84. Springer, 2010.

[214] David Liben-Nowell and Jon Kleinberg. The link prediction problem for social networks. *CIKM*, pages 556–559, November 2007.

[215] Ryan N Lichtenwalter, Jake T Lussier, and Nitesh V Chawla. New perspectives and methods in link prediction. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 243–252. ACM, 2010.

[216] Ryan Lichtnwalter and Nitesh V Chawla. Link prediction: fair and effective evaluation. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, pages 376–383. IEEE Computer Society, 2012.

[217] Robert LiKamWa, Yunxin Liu, Nicholas D Lane, and Lin Zhong. Moodscope: Building a mood sensor from smartphone usage patterns. In *Proceeding of the 11th annual international conference on Mobile systems, applications, and services*, pages 389–402. ACM, 2013.

[218] A Lima, Manlio De Domenico, V Pejovic, and M Musolesi. Disease containment strategies based on mobility and information dissemination. *Scientific reports*, 5:10650, 2015.

[219] Miao Lin and Wen-Jing Hsu. Brownian bridge model for high resolution location predictions. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 210–221. Springer, 2014.

[220] Fredrik Lindberg. Nowcasting swedish retail sales with google search query data. *Stockholm University*, 2011.

[221] Hai-Ying Liu, Erik Skjetne, and Mike Kobernus. Mobile phone tracking: in support of modelling traffic-related air pollution contribution to individual exposure and its implications for public health impact assessment. *Environmental Health*, 12(1):93, 2013.

[222] Alejandro Llorente, Manuel Garcia-Herranz, Manuel Cebrian, and Esteban Moro. Social media fingerprints of unemployment. *PloS one*, 10(5):e0128692, 2015.

[223] Ira M. Longini, Azhar Nizam, Shufu Xu, Kumnuan Ungchusak, Wanna Hanshaoworakul, Derek A. T. Cummings, and M. Elizabeth Halloran. Containing pandemic influenza at the source. *Science*, 309(5737):1083–1087, 2005.

[224] Ira M. Longini, Jr., M. Elizabeth Halloran, Azhar Nizam, and Yang Yang. Containing pandemic influenza with antiviral agents. *American Journal of Epidemiology*, 159(7):623, 2004.

[225] Eric Hsueh-Chan Lu, Vincent S Tseng, and S Yu Philip. Mining cluster-based temporal mobile sequential patterns in location-based service environments. *IEEE transactions on knowledge and data engineering*, 23(6):914–927, 2011.

[226] Linyuan Lü and Tao Zhou. Role of weak ties in link prediction of complex networks. In *Proceedings of the 1st ACM international workshop on Complex networks meet information & knowledge management*, pages 55–58. ACM, 2009.

[227] Linyuan Lu and Tao Zhou. Link prediction in complex networks: A survey. *Physica A*, 390(6):1150–1170, 2011.

[228] Zhengdong Lu, Berkant Savas, Wei Tang, and Inderjit Dhillon. Supervised link prediction using multiple sources. *ICDM*, pages 923–928, 2010.

[229] Caroline Lynch and Cally Roper. The transit phase of migration: circulation of malaria and its multidrug-resistant forms in africa. *PLoS medicine*, 8(5):e1001040, 2011.

[230] Anmol Madan, Manuel Cebrian, David Lazer, and Alex Pentland. Social sensing for epidemiological behavior change. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, pages 291–300. ACM, 2010.

[231] Anmol Madan, Manuel Cebrian, Sai Moturu, Katayoun Farrahi, et al. Sensing the" health state" of a community. *IEEE Pervasive Computing*, 11(4):36–45, 2012.

[232] S Magruder. Evaluation of over-the-counter pharmaceutical sales as a possible early warning indicator of human disease. *Johns Hopkins APL technical digest*, 24(4):349–53, 2003.

[233] Gerasimos Marketos, Elias Frentzos, Irene Ntoutsi, Nikos Pelekis, Alessandra Raffaetà, and Yannis Theodoridis. Building real-world trajectory warehouses. In *Proceedings of the seventh ACM international workshop on data engineering for wireless and mobile access*, pages 8–15. ACM, 2008.

[234] Ann Markusen. Sticky places in slippery space. *The new industrial geography*, pages 98–126, 1999.

[235] Alfred Marshall. Principles of political economy. *Maxmillan, New York*, 1890.

[236] Aleksandar Matic, Venet Osmani, Andrei Popleteev, and Oscar Mayora-Ibarra. Smart phone sensing to examine effects of social interactions and non-sedentary work time on mood changes. In *International and interdisciplinary conference on modeling and using context*, pages 200–213. Springer, 2011.

[237] Gian Luigi Mazzi. Some guidance for the use of big data in macroeconomic nowcasting. In *CARMA 2016: 1st International Conference on Advanced Research Methods in Analytics*, pages 7–14. Editorial Universitat Politècnica de València, 2016.

[238] Michael J McCarthy. Internet monitoring of suicide risk in the population. *Journal of affective disorders*, 122(3):277–279, 2010.

[239] David J McIver and John S Brownstein. Wikipedia usage estimates prevalence of influenza-like illness in the united states in near real-time. *PLoS computational biology*, 10(4):e1003581, 2014.

[240] Nick McLaren and Rachana Shanbhogue. Using internet search data as economic indicators. 2011.

[241] Stefano Merler, Marco Ajelli, Andrea Pugliese, and Neil M Ferguson. Determinants of the spatiotemporal dynamics of the 2009 h1n1 pandemic in europe: implications for real-time modelling. *PLoS computational biology*, 7(9):e1002205, 2011.

[242] L. Michaelis and M. L. Menten. Die kinetik der invertinwirkung. *Biochem. Z*, 49(333-369):352, 1913.

[243] Lewis Mitchell, Morgan R Frank, Kameron Decker Harris, Peter Sheridan Dodds, and Christopher M Danforth. The geography of happiness: Connecting twitter sentiment and expression, demographics, and objective characteristics of place. *PloS one*, 8(5):e64417, 2013.

[244] Mark S Mizruchi. What do interlocks do? an analysis, critique, and assessment of research on interlocking directorates. *Annual review of sociology*, 22(1):271–298, 1996.

[245] 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.

[246] Helen Susannah Moat, Tobias Preis, Christopher Y Olivola, Chengwei Liu, and Nick Chater. Using big data to predict collective behavior in the real world 1. *Behavioral and Brain Sciences*, 37(1):92–93, 2014.

[247] Jordi Mondria, Thomas Wu, and Yi Zhang. The determinants of international investment and attention allocation: Using internet search query data. *Journal of International Economics*, 82(1):85–95, 2010.

[248] Anna Monreale, Fabio Pinelli, Roberto Trasarti, and Fosca Giannotti. Wherenext: a location predictor on trajectory pattern mining. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 637–646. ACM, 2009.

[249] Mikołaj Morzy. Prediction of moving object location based on frequent trajectories. In *International Symposium on Computer and Information Sciences*, pages 583–592. Springer, 2006.

[250] Mikołaj Morzy. Mining frequent trajectories of moving objects for location prediction. In *International Workshop on Machine Learning and Data Mining in Pattern Recognition*, pages 667–680. Springer, 2007.

[251] Joël Mossong, Niel Hens, Mark Jit, Philippe Beutels, Kari Auranen, Rafael Mikolajczyk, Marco Massari, Stefania Salmaso, Gianpaolo Scalia Tomba, Jacco Wallinga, et al. Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS medicine*, 5(3):e74, 2008.

[252] Tsuyoshi Murata and Sakiko Moriyasu. Link prediction of social networks based on weighted proximity measures. *Web Intelligence*, pages 85–88, 2007.

[253] Mirco Nanni, Roberto Trasarti, Barbara Furletti, Lorenzo Gabrielli, Peter Van Der Mede, Joost De Bruijn, Erik De Romph, and Gerard Bruil. Transportation planning based on gsm traces: a case study on ivory coast. In *Citizen in Sensor Networks*, pages 15–25. Springer, 2014.

[254] Jekaterina Navicke, Olga Rastrigina, and Holly Sutherland. Nowcasting indicators of poverty risk in the european union: a microsimulation approach. *Social Indicators Research*, 119(1):101–119, 2014.

[255] Mark W Nelson. Behavioral evidence on the effects of principles-and rules-based standards. *Accounting Horizons*, 17(1):91–104, 2003.

[256] Mark Newman, Albert-László Barabási, and Duncan J. Watts. *The Structure and Dynamics of Networks*. Princeton University Press, 2006.

[257] Mark E. J. Newman. Clustering and preferential attachment in growing networks. *Phys. Rev. E*, 64, 2001.

[258] Mark E. J. Newman. The structure of scientific collaboration networks. *PNAS*, 98(2):404–409, 2001.

[259] Mark EJ Newman. Clustering and preferential attachment in growing networks. *Physical review E*, 64(2):025102, 2001.

[260] Masaaki Nishino, Yukihiro Nakamura, Takashi Yagi, Shinyo Muto, and Masanobu Abe. A location predictor based on dependencies between multiple lifelog data. In *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks*, pages 11–17. ACM, 2010.

[261] 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(4):e95209, 2014.

[262] Elaine Nsoesie, Madhav Mararthe, and John Brownstein. Forecasting peaks of seasonal influenza epidemics. *PLoS currents*, 5, 2013.

[263] Elaine O Nsoesie, David L Buckeridge, and John S Brownstein. Guess who's not coming to dinner? evaluating online restaurant reservations for disease surveillance. *Journal of medical Internet research*, 16(1), 2014.

[264] Alex J Ocampo, Rumi Chunara, and John S Brownstein. Using search queries for malaria surveillance, thailand. *Malaria journal*, 12(1):390, 2013.

[265] Nuria Oliver, Aleksandar Matic, and Enrique Frias-Martinez. Mobile network data for public health: opportunities and challenges. *Frontiers in public health*, 3:189, 2015.

[266] Donald R. Olson, Kevin J. Konty, Marc Paladini, Cecile Viboud, and Lone Simonsen. Reassessing google flu trends data for detection of seasonal and pandemic influenza: A comparative epidemiological study at three geographic scales. *PLOS Computational Biology*, 2013.

[267] J-P Onnela, Jari Saramäki, Jorkki Hyvönen, György Szabó, David Lazer, Kimmo Kaski, János Kertész, and A-L Barabási. Structure and tie strengths in mobile communication networks. *Proceedings of the national academy of sciences*, 104(18):7332–7336, 2007.

[268] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

[269] D Paolotti, A Carnahan, V Colizza, K Eames, J Edmunds, G Gomes, C Koppeschaar, M Rehn, R Smallenburg, C Turbelin, et al. Web-based participatory surveillance of infectious diseases: the influenzanet participatory surveillance experience. *Clinical Microbiology and Infection*, 20(1):17–21, 2014.

[270] Luca Pappalardo, Dino Pedreschi, Zbigniew Smoreda, and Fosca Giannotti. Using big data to study the link between human mobility and socio-economic development. In *Big Data (Big Data), 2015 IEEE International Conference on*, pages 871–878. IEEE, 2015.

[271] Luca Pappalardo, Salvatore Rinzivillo, Zehui Qu, Dino Pedreschi, and Fosca Giannotti. Understanding the patterns of car travel. *The European Physical Journal Special Topics*, 215(1):61–73, 2013.

[272] Luca Pappalardo, Filippo Simini, Salvatore Rinzivillo, Dino Pedreschi, Fosca Giannotti, and Albert-László Barabási. Returners and explorers dichotomy in human mobility. *Nature*, 6:8166, 09 2015.

[273] Adriana Parrella, Craig B Dalton, Rodney Pearce, John CB Litt, and Nigel Stocks. Aspren surveillance system for influenza-like illness: A comparison with flutracking and the national notifiable diseases surveillance system. *Australian family physician*, 38(11):932, 2009.

[274] Oscar Patterson-Lomba, Sander Van Noort, Benjamin J Cowling, Jacco Wallinga, M Gabriela M Gomes, Marc Lipsitch, and Edward Goldstein. Utilizing syndromic surveillance data for estimating levels of influenza circulation. *American journal of epidemiology*, 179(11):1394–1401, 2014.

[275] Michael J Paul, Mark Dredze, and David Broniatowski. Twitter improves influenza forecasting. *PLoS currents*, 6, 2014.

[276] Veljko Pejovic, Neal Lathia, Cecilia Mascolo, and Mirco Musolesi. Mobile-based experience sampling for behaviour research. In *Emotions and Personality in Personalized Services*, pages 141–161. Springer, 2016.

[277] Camille Pelat, Clement Turbelin, Avner Bar-Hen, Antoine Flahault, and Alain-Jacques Valleron. More diseases tracked by using google trends. *Emerging infectious diseases*, 15(8):1327, 2009.

[278] Eric Pels, Peter Nijkamp, and Piet Rietveld. Airport and airline choice in a multiple airport region: An empirical analysis for the san francisco bay area. *Regional Studies*, 35(1):1–9, 2001.

[279] Diego Pennacchioli, Michele Coscia, Salvatore Rinzivillo, Fosca Giannotti, and Dino Pedreschi. The retail market as a complex system. *EPJ Data Science*, 3(1), dec 2014.

[280] David M. Pennock, Gary W. Flake, Steve Lawrence, Eric J. Glover, and C. Lee Giles. Winners don't take all: Characterizing the competition for links on the web. *PNAS*, 99(8):5207–5211, 2002.

[281] Alex Pentland. Reality mining of mobile communications: Toward a new deal on data. *The Global Information Technology Report 2008–2009*, 1981, 2009.

[282] Alex Pentland and Tracy Heibeck. *Honest signals: how they shape our world*. MIT press, 2010.

[283] Daniela Perrotta, Antonino Bella, Caterina Rizzo, and Daniela Paolotti. Participatory online surveillance as a supplementary tool to sentinel doctors for influenza-like illness surveillance in italy. *PloS one*, 12(1):e0169801, 2017.

[284] Daniela Perrotta, Michele Tizzoni, and Daniela Paolotti. Using Participatory Web-based Surveillance Data to Improve Seasonal Influenza Forecasting in Italy. In *Proceedings of the 26th International Conference on World Wide Web - WWW '17*. ACM Press, 2017.

[285] Piero Poletti, Marco Ajelli, and Stefano Merler. The effect of risk perception on the 2009 h1n1 pandemic influenza dynamics. *PloS one*, 6(2):e16460, 2011.

[286] Philip M. Polgreen, Yiling Chen, David M. Pennock, Forrest D. Nelson, and Robert A. Weinstein. Using internet searches for influenza surveillance. *Clinical Infectious Diseases*, 47(11):1443–1448, 2008.

[287] Michael E Porter. The competitive advantage of nations. *Competitive Intelligence Review*, 1(1):14–14, 1990.

[288] Michael E Porter. The economic performance of regions: measuring the role of clusters. In *Gothenburg: TCI Conference*, volume 19, page 2003, 2003.

[289] J. Portugali, H. Meyer, E. Stolk, and E Tan. *Complexity Theories of Cities Have Come of Age: An Overview with Implications to Urban Planning and Design*. Springer-Verlag Berlin Heidelberg, 2012.

[290] Tobias Preis and Helen Susannah Moat. Adaptive nowcasting of influenza outbreaks using google searches. *Royal Society open science*, 1(2):140095, 2014.

[291] Tobias Preis, Helen Susannah Moat, Steven R Bishop, Philip Treleaven, and H Eugene Stanley. Quantifying the digital traces of hurricane sandy on flickr. *Scientific reports*, 3:3141, 2013.

[292] Tobias Preis, Helen Susannah Moat, and H. Eugene Stanley. Quantifying trading behavior in financial markets using google trends. *Scientific Reports*, 2013.

[293] Tobias Preis, Helen Susannah Moat, H Eugene Stanley, and Steven R Bishop. Quantifying the advantage of looking forward. *Scientific reports*, 2:350, 2012.

[294] Tobias Preis, Wolfgang Paul, and Johannes J Schneider. Fluctuation patterns in high-frequency financial asset returns. *EPL (Europhysics Letters)*, 82(6):68005, 2008.

[295] 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 of London A: Mathematical, Physical and Engineering Sciences*, 368(1933):5707–5719, 2010.

[296] Manisha Pujari and Rushed Kanawati. Supervised rank aggregation approach for link prediction in complex networks. *WWW*, pages Pages 1189–1196, 2012.

[297] Rami Puzis, Yaniv Altshuler, Yuval Elovici, Shlomo Bekhor, Yoram Shiftan, and Alex (Sandy) Pentland. Augmented betweenness centrality for environmentally aware traffic monitoring in transportation networks. *Journal of Intelligent Transportation Systems*, 17(1):91–105, 2013.

[298] Hong Qi, Pedro Manrique, Daniela Johnson, Elvira Restrepo, and Neil F Johnson. Open source data reveals connection between online and on-street protest activity. *EPJ data science*, 5(1):18, 2016.

[299] Disheng Qiu, Paolo Papotti, and Lorenzo Blanco. Future locations prediction with uncertain data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 417–432. Springer, 2013.

[300] Kiran K Rachuri, Mirco Musolesi, Cecilia Mascolo, Peter J Rentfrow, Chris Longworth, and Andrius Aucinas. Emotionsense: a mobile phones based adaptive platform for experimental social psychology research. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, pages 281–290. ACM, 2010.

[301] Kira Radinsky, Sagie Davidovich, and Shaul Markovitch. Predicting the news of tomorrow using patterns in web search queries. In *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01*, pages 363–367. IEEE Computer Society, 2008.

[302] C. Randall. Measuring national well-being - where we live - 2012. Technical report, UK Office for National Statistics, 2012.

[303] Anatol Rapoport. Mathematical models of social interaction. 1963.

[304] Olga Rastrigina, Chrysa Leventi, and Holly Sutherland. Nowcasting: estimating developments in the risk of poverty and income distribution in 2013 and 2014. Technical report, EUROMOD Working Paper, 2015.

[305] Patricia R Recupero, Samara E Harms, and Jeffrey M Noble. Googling suicide: surfing for suicide information on the internet. *The Journal of clinical psychiatry*, 2008.

[306] Sid Redner. Networks: teasing out the missing links. *Nature*, 453(7191):47, 2008.

[307] S. Rinzivillo, L. Gabrielli, M. Nanni, L. Pappalardo, D. Pedreschi, and F. Giannotti. The purpose of motion: Learning activities from individual mobility networks. In *2014 International Conference on Data Science and Advanced Analytics (DSAA)*, pages 312–318, Oct 2014.

[308] Salvatore Rinzivillo, Simone Mainardi, Fabio Pezzoni, Michele Coscia, Dino Pedreschi, and Fosca Giannotti. Discovering the geographical borders of human mobility. *KI-Künstliche Intelligenz*, 26(3):253–260, 2012.

[309] David Rogers, Thomas Emwanu, and Timothy Robinson. Poverty mapping in uganda: an analysis using remotely sensed and other environmental data. 2006.

[310] Giulio Rossetti, Letizia Milli, Fosca Giannotti, and Dino Pedreschi. Forecasting success via early adoptions analysis: A data-driven study. *PLOS ONE*, 12(12):e0189096, dec 2017.

[311] Martin Rosvall and Carl T Bergstrom. Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems. *PloS one*, 6(4):e18209, 2011.

[312] Gerhard Rünstler and Franck Sédillot. Short-term estimates of euro area real gdp by means of monthly data. Technical report, ECB working paper, 2003.

[313] Marcel Salathé, Maria Kazandjieva, Jung Woo Lee, Philip Levis, Marcus W Feldman, and James H Jones. A high-resolution human contact network for infectious disease transmission. *Proceedings of the National Academy of Sciences*, 107(51):22020–22025, 2010.

[314] G Salton and MJ McGill. Introduction to modern information philadelphia, pa. american association for artificial intelligence retrieval. 1983.

[315] Mauricio Santillana, André T. Nguyen, Mark Dredze, Michael J. Paul, Elaine O. Nsoesie, and John S. Brownstein. Combining Search Social Media, and Traditional Data Sources to Improve Influenza Surveillance. *PLOS Computational Biology*, 11(10):e1004513, oct 2015.

[316] Mauricio Santillana, Elaine O Nsoesie, Sumiko R Mekaru, David Scales, and John S Brownstein. Using clinicians' search query data to monitor influenza epidemics. *Clinical infectious diseases: an official publication of the Infectious Diseases Society of America*, 59(10):1446, 2014.

[317] Mauricio Santillana, D Wendong Zhang, Benjamin M Althouse, and John W Ayers. What can digital disease detection learn from (an external revision to) google flu trends? *American journal of preventive medicine*, 47(3):341–347, 2014.

[318] Purnamrita Sarkar, Deepayan Chakrabarti, and Michael Jordan. Nonparametric link prediction in dynamic networks. *arXiv preprint arXiv:1206.6394*, 2012.

[319] Monica Scannapieco, Antonino Virgillito, and Diego Zardetto. Placing big data in official statistics: a big challenge. In *Conference on New Techniques and Technologies for Statistics, Brussels*, 2013.

[320] Salvatore Scellato, Mirco Musolesi, Cecilia Mascolo, Vito Latora, and Andrew T Campbell. Nextplace: a spatio-temporal prediction framework for pervasive systems. In *International Conference on Pervasive Computing*, pages 152–169. Springer, 2011.

[321] Steven L. Scott and Hal R. Varian. Bayesian variable selection for nowcasting economic time series. *Economic Analysis of the Digital Economy*, 2014.

[322] Jeffrey Shaman and Alicia Karspeck. Forecasting seasonal outbreaks of influenza. *Proceedings of the National Academy of Sciences*, 109(50):20425–20430, 2012.

[323] Jeffrey Shaman, Alicia Karspeck, Wan Yang, James Tamerius, and Marc Lipsitch. Real-time influenza forecasts during the 2012–2013 season. *Nature communications*, 4:2837, 2013.

[324] Naoki Shibata, Yuya Kajikawa, and Ichiro Sakata. Link prediction in citation networks. *Journal of the American Society for Information Science and Technology*, 63(1):78–85, 2012.

[325] Alessio Signorini, Alberto Maria Segre, and Philip M Polgreen. The use of twitter to track levels of disease activity and public concern in the us during the influenza a h1n1 pandemic. *PloS one*, 6(5):e19467, 2011.

[326] Filippo Simini, Marta C González, Amos Maritan, and Albert-László Barabási. A universal model for mobility and migration patterns. *Nature*, 484(7392):96, 2012.

[327] Christopher Smith-Clarke, Afra Mashhadi, and Licia Capra. Poverty on the cheap: Estimating poverty maps using aggregated mobile communication networks. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 511–520. ACM, 2014.

[328] Mark S Smolinski, Adam W Crawley, Kristin Baltrusaitis, Rumi Chunara, Jennifer M Olsen, Oktawia Wójcik, Mauricio Santillana, Andre Nguyen, and John S Brownstein. Flu near you: crowdsourced symptom reporting spanning 2 influenza seasons. *American journal of public health*, 105(10):2124–2130, 2015.

[329] Radina P Soebiyanto, Farida Adimi, and Richard K Kiang. Modeling and predicting seasonal influenza transmission in warm regions using climatological parameters. *PloS one*, 5(3):e9450, 2010.

[330] Chaoming Song, Tal Koren, Pu Wang, and Albert-Laszlo Barabasi. Modelling the scaling properties of human mobility. *Nat Phys*, 6(10):818–823, 10 2010.

[331] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

[332] Victor Soto, Vanessa Frias-Martinez, Jesus Virseda, and Enrique Frias-Martinez. Prediction of socioeconomic levels using cell phone records. In *International Conference on User Modeling, Adaptation, and Personalization*, pages 377–388. Springer, 2011.

[333] Sucheta Soundarajan and John Hopcroft. Using community information to improve the precision of link prediction methods. In *Proceedings of the 21st International Conference on World Wide Web*, pages 607–608. ACM, 2012.

[334] Stephan Spiegel, Jan Clausen, Sahin Albayrak, and Jérôme Kunegis. Link prediction on evolving data using tensor factorization. *New Frontiers in Applied Data Mining*, pages 100–110, 2011.

[335] Jessica E. Steele, Pål Roe Sundsøy, Carla Pezzulo, Victor A. Alegana, Tomas J. Bird, Joshua Blumenstock, Johannes Bjelland, Kenth Engø-Monsen, Yves-Alexandre de Montjoye, Asif M. Iqbal, Khandakar N. Hadiuzzaman, Xin Lu, Erik Wetter, Andrew J. Tatem, and Linus Bengtsson. Mapping poverty using mobile phone and satellite data. *Journal of The Royal Society Interface*, 14(127), 2017.

[336] Juliette Stehlé, Nicolas Voirin, Alain Barrat, Ciro Cattuto, Vittoria Colizza, Lorenzo Isella, Corinne Régis, Jean-François Pinton, Nagham Khanafer, Wouter Van den Broeck, et al. Simulation of an seir infectious disease model on the dynamic contact network of conference attendees. *BMC medicine*, 9(1):87, 2011.

[337] Tanya Suhoy. *Query indices and a 2008 downturn: Israeli data*. Bank of Israel, 2009.

[338] Ben Taskar, Ming-Fai Wong, Pieter Abbeel, and Daphne Koller. Link prediction in relational data. In *Advances in neural information processing systems*, pages 659–666, 2004.

[339] Andrew J Tatem, Youliang Qiu, David L Smith, Oliver Sabot, Abdullah S Ali, and Bruno Moonen. The use of mobile phone data for the estimation of the travel patterns and imported plasmodium falciparum rates among zanzibar residents. *Malaria journal*, 8(1):287, 2009.

[340] Natasha L Tilston, Ken TD Eames, Daniela Paolotti, Toby Ealden, and W John Edmunds. Internet-based surveillance of influenza-like-illness in the uk during the 2009 h1n1 influenza pandemic. *BMC public health*, 10(1):650, 2010.

[341] Michele Tizzoni, Paolo Bajardi, Adeline Decuyper, Guillaume Kon Kam King, Christian M Schneider, Vincent Blondel, Zbigniew Smoreda, Marta C González, and Vittoria Colizza. On the use of human mobility proxies for modeling epidemics. *PLoS computational biology*, 10(7):e1003716, 2014.

[342] Michele Tizzoni, Paolo Bajardi, Chiara Poletto, José J. Ramasco, Duygu Balcan, Bruno Gonçalves, Nicola Perra, Vittoria Colizza, and Alessandro Vespignani. Real-time numerical forecast of global epidemic spreading: case study of 2009 a/h1n1pdm. *BMC Medicine*, 10(1):165, Dec 2012.

[343] Jameson L Toole, Yu-Ru Lin, Erich Muehlegger, Daniel Shoag, Marta C González, and David Lazer. Tracking employment shocks using mobile phone data. *Journal of The Royal Society Interface*, 12(107):20150185, 2015.

[344] Istvan Janos Toth and Miklós Hajdu. Google as a tool for nowcasting household consumption: Estimations on hungarian data. *CIRET*, 2012.

[345] Le Hung Tran, Michele Catasta, Lucas Kelsey McDowell, and Karl Aberer. Next place prediction using mobile data. In *Proceedings of the Mobile Data Challenge Workshop (MDC 2012)*, number EPFL-CONF-182131, 2012.

[346] Emma Uprichard. Focus: Big data, little questions? *Focus*, 28:28, 2013.

[347] Antonio Valdivia and Susana Monge-Corella. Diseases tracked by using google trends, spain. *Emerging infectious diseases*, 16(1):168, 2010.

[348] P. van den Driessche and James Watmough. Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission. *Mathematical Biosciences*, 180(1):29 – 48, 2002.

[349] Heymerik A van der Grient and Jan de Haan. The use of supermarket scanner data in the dutch cpi. In *Joint ECE/ILO Workshop on Scanner Data*, volume 10, 2010.

[350] Sander P van Noort, Cláudia T Codeço, Carl E Koppeschaar, Marc Van Ranst, Daniela Paolotti, and M Gabriela M Gomes. Ten-year performance of influenzanet: Ili time series, risks, vaccine effects, and care-seeking behaviour. *Epidemics*, 13:28–36, 2015.

[351] Alessandro Vespignani. Predicting the behavior of techno-social systems. *Science*, 325(5939):425–428, 2009.

[352] Simeon Vosen and Torsten Schmidt. Forecasting private consumption: survey-based indicators vs. google trends. *Journal of Forecasting*, 30(6):565–578, 2011.

[353] Dashun Wang, Dino Pedreschi, Chaoming Song, Fosca Giannotti, and Albert-László Barabási. Human mobility, social ties, and link prediction. *KDD*, 2011.

[354] Pu Wang, Timothy Hunter, Alexandre M. Bayen, Katja Schechtner, and Marta C. González. Understanding road usage patterns in urban areas. *Nature*, 2:1001, 12 2012.

[355] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T Campbell. Studentlife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 3–14. ACM, 2014.

[356] Rui Wang, Gabriella Harari, Peilin Hao, Xia Zhou, and Andrew T Campbell. Smartgpa: how smartphones can assess and predict academic performance of college students. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*, pages 295–306. ACM, 2015.

[357] Ying Zhu Yong Sun Yu Wang. Nokia mobile data challenge: Predicting semantic place and next place via mobile data. *Work*, 80(100):120, 2012.

[358] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1998.

[359] Amy Wesolowski, Caroline O Buckee, Linus Bengtsson, Erik Wetter, Xin Lu, and Andrew J Tatem. Commentary: containing the ebola outbreak-the potential and challenge of mobile network data. *PLoS currents*, 6, 2014.

[360] Alan Wilson. Boltzmann, lotka and volterra and spatial structural evolution: an integrated methodology for some dynamical systems. *Journal of The Royal Society Interface*, 5(25):865–871, 2008.

[361] Alan G Wilson. Ecological and urban systems models: Some explorations of similarities in the context of complexity theory. *Environment and Planning A*, 38(4):633–646, 2006.

[362] Kumanan Wilson and John S Brownstein. Early detection of disease outbreaks using the internet. *Canadian Medical Association Journal*, 180(8):829–831, 2009.

[363] Kumanan Wilson, Barbara von Tigerstrom, and Christopher McDougall. Protecting global health security through the international health regulations: requirements and challenges. *Canadian Medical Association Journal*, 179(1):44–48, 2008.

[364] N Wilson, K Mason, M Tobias, M Peacey, QS Huang, and M Baker. Interpreting "google flu trends" data for pandemic h1n1 influenza: the new zealand experience. *Eurosurveillance*, 14(44):19386, 2009.

[365] J Woodall. Official versus unofficial outbreak reporting through the internet. *International journal of medical informatics*, 47(1-2):31–34, 1997.

[366] B. Wotal, H. Green, D. Williams, and N. Contractor. Wow!: The dynamics of knowledge networks in massively multiplayer online role playing games (mmorpg). *Sunbelt XXVI International Sunbelt Social Network Conference*, 2006.

[367] Rongjing Xiang, Jennifer Neville, and Monica Rogati. Modeling relationship strength in online social networks. In *Proceedings of the 19th international conference on World wide web*, pages 981–990. ACM, 2010.

[368] Ye Xu and Dan Rockmore. Feature selection for link prediction. In *Proceedings of the 5th Ph. D. workshop on Information and knowledge*, pages 25–32. ACM, 2012.

[369] Guangtao Xue, Yuan Luo, Jiadi Yu, and Minglu Li. A novel vehicular location prediction based on mobility patterns for routing in urban vanet. *EURASIP Journal on Wireless Communications and Networking*, 2012(1):222, 2012.

[370] Albert C Yang, Shi-Jen Tsai, Norden E Huang, and Chung-Kang Peng. Association of internet search trends with suicide death in taipei city, taiwan, 2004–2009. *Journal of affective disorders*, 132(1):179–184, 2011.

[371] Ning Yang, Xiangnan Kong, Fengjiao Wang, and Philip S Yu. When and where: predicting human movements based on social spatial-temporal events. In *Proceedings of the 2014 SIAM International Conference on Data Mining*, pages 515–523. SIAM, 2014.

[372] Shihao Yang, Mauricio Santillana, and S. C. Kou. Accurate estimation of influenza epidemics using Google search data via ARGO. *Proceedings of the National Academy of Sciences*, 112(47):14473–14478, nov 2015.

[373] Wan Yang, Marc Lipsitch, and Jeffrey Shaman. Inference of seasonal and pandemic influenza transmission dynamics. *Proceedings of the National Academy of Sciences*, 112(9):2723–2728, 2015.

[374] Gökhan Yavaş, Dimitrios Katsaros, Özgür Ulusoy, and Yannis Manolopoulos. A data mining approach for location prediction in mobile environments. *Data & Knowledge Engineering*, 54(2):121–146, 2005.

[375] Josh Jia-Ching Ying, Wang-Chien Lee, Tz-Chiao Weng, and Vincent S Tseng. Semantic trajectory mining for location prediction. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 34–43. ACM, 2011.

[376] Qingyu Yuan, Elaine O Nsoesie, Benfu Lv, Geng Peng, Rumi Chunara, and John S Brownstein. Monitoring influenza epidemics in china with search query from baidu. *PloS one*, 8(5):e64323, 2013.

[377] Qian Zhang, Corrado Gioannini, Daniela Paolotti, Nicola Perra, Daniela Perrotta, Marco Quaggiotto, Michele Tizzoni, and Alessandro Vespignani. Social data mining and seasonal influenza forecasts: the fluoutlook platform. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 237–240. Springer, 2015.

[378] Qian Zhang, Nicola Perra, Daniela Perrotta, Michele Tizzoni, Daniela Paolotti, and Alessandro Vespignani. Forecasting Seasonal Influenza Fusing Digital Indicators and a Mechanistic Disease Model. In *Proceedings of the 26th International Conference on World Wide Web - WWW '17*. ACM Press, 2017.

[379] Nan Zhao, Wenhao Huang, Guojie Song, and Kunqing Xie. Discrete trajectory prediction on mobile data. In *Asia-Pacific Web Conference*, pages 77–88. Springer, 2011.

[380] Changsong Zhou, Lucia Zemanová, Gorka Zamora, Claus C Hilgetag, and Jürgen Kurths. Hierarchical organization unveiled by functional connectivity in complex brain networks. *Physical review letters*, 97(23):238103, 2006.

[381] Jun Zhu, Jiaming Song, and Bei Chen. Max-margin nonparametric latent feature models for link prediction. *arXiv preprint arXiv:1602.07428*, 2016.