Universidad
Carlos III de Madrid
www.uc3m.es

TESIS DOCTORAL

Mobility and interaction patterns in social networks

Autor:

Alejandro Llorente Pinto

Director:

Esteban Moro Egido

Tutor:

Esteban Moro Egido

Departamento de Matemáticas

Leganés, junio de 2016

Universidad
Carlos III de Madrid
www.uc3m.es

(a entregar en la Oficina de Posgrado, una vez nombrado el Tribunal evaluador , para preparar el documento para la defensa de la tesis)

TESIS DOCTORAL

Mobility and interaction patterns in social networks

Autor: *Alejandro Llorente Pinto*

Director: Esteban Moro Egido

Firma del Tribunal Calificador:

Firma

Presidente: (Nombre y apellidos)

Vocal: (Nombre y apellidos)

Secretario: (Nombre y apellidos)

Calificación:

Leganés, de de

Universidad Carlos III de Madrid

Doctoral Thesis

Mobility and interaction patterns in social networks

Author:
Alejandro Llorente Pinto

Supervisor:
Esteban Moro Egido

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Mathematics

in the

Department of Mathematics

June 2016

“Son muchos los hombres que han sufrido moral y espiritualmente del mismo modo que tú ahora. Felizmente algunos de ellos han dejado constancia de su sufrimiento. Y de ellos aprenderás si lo deseas. Del mismo modo que alguien aprenderá algún día de ti si tienes algo que ofrecer. Se trata de un hermoso acuerdo de reciprocidad. No se trata de educación. Es historia. Es poesía.”

El Guardián entre el centeno, J.D. Salinger.

UNIVERSIDAD CARLOS III DE MADRID

Abstract

Department of Mathematics

Doctor of Mathematics

Mobility and interaction patterns in social networks

by Alejandro Llorente Pinto

The question of analyzing the predictability of human behavior has been widely studied in literature, to unveil how individuals move, how they can be mobilized and, more philosophically, to understand to what extent our decisions are random or whether we are free to choose. As a consequence of humans relate to each other, we also tend to live in groups at different hierarchies in a social way so it is interesting to analyze how individual features and choices affect the global structure of a society.

In this work, we explore the limits of human predictability in terms of shopping behavior, observing that, even when we are constrained to a limited set of possible places where we can make a purchase, predicting where the next purchase will happen is not accurately possible to do by only observing the past. The next question is to study how individual decisions affect emergent phenomena such as the economy or information diffusion across a country. We analyze the contents, temporal and mobility patterns extracted from users' social media publications to build a profile of the geographical regions that allow to predict the unemployment rate. Finally, we also use a mobile phone call dataset to test whether the dynamics at the urban level, how people create and destroy links within a city, affect the inter-urban diffusion of diseases, virus or rumors. Our results suggest that inter-regional structure is robust and does not vary significantly on time so diffusion processes can be well modeled in terms of static properties of the inter-urban network.

Agradecimientos

Si pienso en qué tenéis en común las personas que aparecéis en los renglones escritos a continuación, es que en algún momento, quizás en muchos, me habéis hecho sentir que estabais orgullosos de que yo estuviera haciendo esta tesis, igual no importaba mucho sobre qué, pero sí lo estabais por el esfuerzo que lleva, las horas, los cabezazos, las borderías y también, por supuesto, las alegrías que de vez en cuando daba (artículos, congresos, charlas, gráficos bonitos que os pasaba por Whatsapp, o cualquier otra circunstancia positiva). Aunque mi cara no lo expresara, cada vez que me preguntabais "¿y la tesis para cuándo?", era orgullo lo que sentía por dentro, casi siempre... ;)

Gracias, en primer lugar a mis padres, porque en cada decisión que he ido tomando a lo largo de estos años, siempre he sentido que confiabais vosotros más en mí que yo mismo. Gracias papá por cuidarme, por aguantarme los cansancios de cada noche y las pocas ganas de hablar. Gracias mamá por empujarme a hacer cualquier cosa que me apeteciera y en la que creyera, desde tesis hasta empresas. Gracias a los dos, por todo.

A continuación, no puede ir nadie más que mi director de tesis: muchas gracias, Esteban, por todos estos años (bastantes ya si hacemos la cuenta) en los que hemos trabajado codo con codo, hemos estudiado, hemos pintarrajeado pizarras, hemos publicado algo de vez en cuando, hemos conspirado y, en definitiva, nos lo hemos pasado muy bien. Gracias porque a tu lado he aprendido mucho y he crecido contigo. Gracias a mi hermana académica (y de cervezas y datos), Giovanna, porque alguna conversación que hemos tenido también me han dado un empujón a seguir peleando. Gracias también a Manuel Cebrián, con el que hemos colaborado en buena parte de la tesis, y del que no se deja de aprender cosas ni un segundo que se está con él.

Gracias a mi abuela, la cual, si todo va bien en la defensa, verá a su nieto hecho un doctor con 91 años a sus espaldas. Gracias a mi tía Marisol, muy especialmente, por ser un bastón en el que apoyarme mucho tiempo y porque creo que parte de mi vena científica viene por culpa de ella... igual en la empresarial también tiene algo que ver. Gracias a mi tío Ángel y a mis primos, Miguel, Elena y Belén, mis mayores debilidades, aunque ellos igual no lo sepan. Gracias a mi abuelo, del que recuerdo, retengo e imito gestos y humor; de haberme podido ver, no creo que hubiera habido nadie más orgulloso de mí que él.

Gracias al resto de mi familia, los que veo más y los que veo menos, aunque tengamos menos contacto, el orgullo del que hablaba al principio lo he sentido cada vez que ha sido posible.

Gracias a mis amigos. Primero el club de los "Más-De-10-Años-Aguantándome": gracias, Dide, por llevar tanto tiempo cerca y tirarte conmigo a la piscinas que se nos ponen por delante; gracias, José porque aunque ahora nos vemos menos, hemos tenido miles de conversaciones de muchos tipos, hasta de política; muchas gracias, María, por no irte nunca de mi lado, por tantos años, por tantííísimas cosas, porque nadie se ríe tanto de mis chorradas, porque es muy difícil que en la vida de uno esté alguien como tú.

Gracias a Larra, por los viajes, por hacer de representante agresivo, por las cañas, por los vídeos de la Tuerka y los tweets de xaviconde, por hacerme pensar las cosas algunas veces. Gracias, Javitxu, por tantos momentos espectaculares, lamentables en ocasiones, pero que nunca decepcionan, es más, "que no entusiasman, pero satisfacen". Gracias, Borja, por muchas risas, por los partidos y por los viajes. Gracias, Pano y Alberto, por tantas noches de sobriedad y sosiego en el Ocean. Gracias, resto de los coyotes, por el reencuentro después de tantos años y por tanta intensidad deportiva.

Gracias a PiperLab, a todos, a los socios, a los que estamos, a los que se han ido y a los que están por venir. Gracias también a todos mis ex-compañeros del IIC, porque también fueron muchos años y un montón de cosas buenas.

Si de orgullo se trata, qué voy a decir de estos... Gracias a La Gota Que Colma, otros del club de "Más-De-10-Años-Aguantándome", por tantísimos veranos, por tantísimos momentos, por tantísimas fiestas (y las que quedan). Que no se mosqueen el resto, pero gracias especialmente a Pitu, Pájaro, Toni, Manu, Pablo y Gañán. Sois muy grandes.

Y por último, gracias a ti, la última persona en llegar, aunque lo hicieras hace muchos años. Gracias, Isa, por incendiarlo todo.

Contents

List of Figures

List of Tables

A mis padres

Chapter 1

Introduction

1.1 Understanding individuals and group behavior with data

1.1.1 Anticipating individual patterns

Are we humans predictable? A possible definition of individuals' predictability is the capacity of anticipating, by any method, every actions based on the historical observations of individuals or considering other external factors. For instance, when a customer always asks for the same dish on the same day of the week in his favorite fast food restaurant, this person becomes predictable because it is likely to happen the same in future; when a buyer always acquire trousers in the same color every time he visits a merchant, this person has a predictable shopping pattern. However, besides our daily habits, human predictability must be also due to external factors: for instance, if we take the car to go to work but a non-usual traffic jam is happening in our path to the destination, our historical patterns tell nothing about whether we are going to get to work one hour later but there is an external factor which is previously known makes again our behavior predictable.

Independently we are more or less predictable, in last years our world have been invaded by the appearance of many digital sensors in our daily living that make all our activities be registered: when we use a bus card, the place where we get on the bus, the hour when we do it and our personal identification are stored in a database [1]; when we cross through a toll by car, our license number and when we do it are also registered; when we make an expense by using a credit card, the merchant where we are, when we did it and how much we spend are warehoused in a huge database [2]. If our individual patterns are predictable, we can use all these data sources to anticipate what every one will do in future and using this information to optimize our decision making.

From the business point of view, anticipating individual behavior has important and obvious advantages. If I know where my customer will be in an hour, I might send a personalized email to him offering a discount in the nearest shop of my company (this is what it is usually called as geo-marketing) [3]; if I am able to discover that an online user always click a certain kind of content, I will try to adapt my advertising on that content to attract the attention of the user [4]; if I realize that a client is visiting more frequently my merchant and he is spending more money on many products, I can infer that his wages have increased and make an offer to sell a more expensive product [5]. Summarizing, predictability on human behavior patterns allow companies to know and segment better their clients and make this information actionable to produce bigger profits.

However, are we completely predictable? Is free will an illusion? [6–8] From the most philosophical point of view, this problem has been discussed for centuries, with many different perspectives. Some of them talk about a completely determinist future but this does not imply the negation of free will since there exists a current of thought, compatibilism, claiming for the existence of free will in a determinist future. Other thoughts, like hard indeterminism, state that future is completely random so, in this context, free will cannot exist neither. Anyway, what it seems clear is that our interests, our obligations, our social network and our geography constraint the space of possible choices we can take, despite there are other physical possible solutions (if I must get to work in an hour and I have to stop by the supermarket, maybe I have to make a large deviation from my original path to go to the one I like the most, even when it is physically possible).

The challenge is to be able to analyze, understand and predict individual patterns to take advantage of them. In particular, on this manuscript, in Chapter 2 we explore the limits of human predictability in terms of shopping behavior, analyzing the balance a constrained space of possible merchants where an individual might make his next purchase but also the difficulty of predicting exactly where it will happen only observing his historical purchases. Besides, in Chapter 3, we use the digital traces of individuals in social networks to infer knowledge of the regions where they live such as the technology penetration rate, mobility between cities or their socio-educational levels.

1.1.2 The group as a complex and emergent phenomena

The way we, humans, tend to organize is in societies, groups of individuals with a certain size that agree in stating a set of rules for coexisting, leading to relationships between the members of the group that finally make them more productive and more efficient in terms of consume. [9, 10] However, through the evolution, these relationships and groups have given place to a more complex and stratified social structure, where social rules are different depending on the level of the hierarchy and where the motivations for the different levels to exist are different. In particular, the role of the individual in the different levels changes dramatically: in a family, every member is important and recognized by the rest of the group, almost irreplaceable whereas, in a city, where phenomena like economy emerge, the role every individual plays is not so clear, it is difficult to perceive so it is interesting to wonder about the individual importance in large societies [11].

Even though social relationships happen in pairs fundamentally (a family is a set of peer-to-peer relationships), when we abstract from the individual level and observe how societes relate to each others, we acquire a new level of complexity: for instance, when inter-city communication patterns in a country are analyzed, we are not observing cities making phone calls one to other but individuals living on them and calling to other person living in a different one to propagate some information. So a natural question arise: what is the impact for communication between cities if these two individuals break their relationship? Considering mobility between regions, one person traveling from one point to another might spread an infectious disease, what happens to this spreading if this person stops traveling? We might think in two possible extreme possibilities: if the scenario changes absolutely, this inter-regional network depends strongly on arbitrary individual decisions, exhibiting a weak structure; on the other side, if we find no variation after the breaking event, we might infer that there is something more than individual relationships in the inter-society (regional in the example) structure [12, 13].

Definitely, since groups at the different levels are heterogeneous (there exist great variance in the properties of neighborhoods, cities, states,...), these different properties must be due to, at least in some portion, the individual behavior of people belonging to them. For instance, a city generating a better economy might be explained because people there have a higher educational level which is, neglecting many other factors, a free individual decision; if a city is saturated by traffic every morning at the same hour, it is the result of many individual decisions of people living there. So the question is, to what extent can we infer information of societies by just analyzing information at the individual level? How the microscopic (individual) and the macroscopic (group, society, city,...) levels relate each other? Can we predict phenomena at the macroscopic level by studying individual patterns?

We answer these two last questions from two different points of view: in chapter 3, the main goal is building a predictive model of economy, in terms of unemployment rate, from features of the regions inferred from individual digital traces on social networks. On the other hand, in chapter 4, we analyze other emerging phenomena, such as diffusion processes (disease spreading, rumor diffusion, etc.) among cities in a country based on the dynamic characteristics observed within them.

1.1.3 Big Data, Computational Social Science and Social Networks

One of the major disruptions in both the business and the academic worlds is what every one has accepted to name as Big Data, a term that references the vast amount of data that is being collected since the appearance of digitalization and the Internet era, changing dramatically all the companies independently from the field they belong to. Usually, people uses the Big Data term as a mixture of concepts referring both technologies allowing to process and store this huge amount of data and the value and applications that may be extracted from this data so other terms, such as Data Science, have emerged to make a distinction between technology and knowledge from the data, the possibility to apply statistical and machine learing techniques to all this data for building predictive models, automatic segmentations, recommender systems, etc. The availability of all these data sources is not only an opportunity for companies and researchers related to technology, engineering or applied sciences but also for other fields such as human and social sciences. Why is this the case? Because the root of most of these data sources come from the human behavior, reflecting the individuals' activity, their economical capacity or their social relationships so, if researchers are able to use this data properly then hypothesis that were published many years ago and had not been strongly and universally proved, can be accepted or rejected at an unprecedented scale.

This combination between engineering, data and social sciences is what is called Computational Social Science [14], the capability of applying statistical techniques on data to verify social hypothesis at large scale. In this field, social networks play a central role. We do not mean to online social networks exclusively, like Twitter, Facebook or Instagram, but other social networks inferred from our communications, presence indicators, visited places, purchases, etc. All this data processed from the point of view of social networks reflect not only our activity but how we relate to our environment, to what extent we are constrained because of our social and geographical conditions or how we create and destroy relationships on time that allow us to find new information sources or job opportunities. Answering some of these questions is the main goal of this manuscript where we have been able to infer emerging phenomena such as economy or information diffusion from different data sources as credit card transactions, social media data or call detail record from a mobile phone company.

1.1.4 Universal patterns in mobility and social networks

From the scientific point of view, even though when human patterns are not completely determinist, researchers have found rules that appear recurrently when data is analyzed, universal patterns that are observed when some phenomena is analyzed. Typically, this universal rules are modeled by statistical distributions that accurately fit the same phenomena in different contexts. Surprisingly, this kind of universal laws have been observed at the individual level but also at aggregated level such as social networks or cities.

Analyzing individual patterns, one might think of human activity as an homogeneous phenomena, that is, there is a constant probability distribution or model fitting accurately some feature of individual behavior. However, circadian rhythms influence absolutely our lives so we do not find the same probability of observing activity depending on the time of the day. On the other hand, when people prioritize their daily tasks and activities they are setting implicitly a temporal distribution that is far from being poissonan and depends on our past activities (it is not a memory-less process). The resulting activity distribution is bursty, which means that very high intense activity periods are observed followed by long slow activity periods [15–18]. This behavior is universal because it has been widely observed in communication patterns, mobility laws or motifs of individual behaviors [16, 19, 20].

Not only individual patterns exhibit universal laws but they are also observed in the way humans connect to each other. For instance, typically in a social network, the degree distribution is very heterogeneous, that is, we find some few nodes with a high degree (named as hubs) but most of them hold low levels of connectivity. Many works have shown that degree distribution can be well fitted by a Power-Law distribution in this case [21]. This kind of networks have been named as Scale-free networks because of this main feature. Scale-free networks are ubiquitous and appear in many contexts such as online social networks [22], communications networks [19], transport and airport networks [23], biology networks [24] and many other cases. To explain why this is the case, generative and evolving network models have been purposed, as the Barabasi-Albert model [25]. In this model, a new node joining the networks connects randomly to the rest of the nodes but the probability of wiring to a particular nodes is higher for very connected nodes.

This process leads to a Power-Law degree distribution but it is unable to explain other network features like clustering. Precisely, clustering (the proportion of closed triplets in a network) and short paths between nodes (small-world networks) have been also shown to be ubiquitous in social networks so they can be also considered universal laws in this sense [26–28] .

As we mentioned before, mobility is a mixture of human activity, social relationships, geographical constraints and dynamics, which attracts interest from many researchers. Even though it is a very complex phenomena, at both individual and aggregated level, universal laws are also extensively found. At the individual level, authors have shown that many features of mobility follow a Power Law distribution such as the distance individuals travel, the time they stay in a particular place or the elapsed time for getting from one place to another. [20]. It has been also proved that commuting patterns, analysis of trips between home and work, reveal universal patterns in terms of visitation to places and commute time, despite differences in spatial features and infrastructures at the country and urban levels [29]. Also at the individual level, it has been studied whether mobility is truly predictable and the information that previous locations has in the choice of the next place is enough to explain at most the 93% of mobility [6]. But also mobility at the aggregated level is predictable, in the sense that predicting the number of people going from one place to another can be accurately modeled. In this sense, many models such as the Gravity or Radiation models have been purposed to describe universally migration patterns [30, 31].

1.1.5 Deepening on mobility and social interaction patterns

As a general purpose, this dissertation deeps on the study of big databases from different sources to build mathematical model that help us to understand and predict human behavior at both individual and group levels. Three non-solved problems have been addressed in this work:

- Are we predictable in our economic decisions? How does geography constraint our space of possible economic decisions? Does it determine where our next purchase will be made?
- Are our behaviors and social features really reflected in social media data? Can we infer group characteristics, as mobility flows, from social media? Does it contain enough information to model economy?

- How does geography affect the spreading of information in dynamic societies? What is the role of intra-urban networks in inter-urban diffusion processes? What is the role of the individual relationships in the spreading of global epidemics?

The first question is answed by analyzing whether shopping patterns are predictable, that is, to test whether it is possible to predict the next place where a customer will make his next purchase. Moreover, we explore he link between the shopping constraints given by mobility and the arbitrary random choices of customers. There exist many data sources that might be used to face this problem: in electronic commerce platform, not only purchase data is registered but also visits to products and related contents; we might have used data from social media geo-tagged in commercial areas to study whether people returns to the same places; but, in this document, we use two credit card transaction datasets (one from an European country and one from North America), where the customer ID, the merchant ID and when it was made are registered. This data is a reflection of behavioral shopping patterns embedded in a geographical context where mobility plays an essential role.

To solve the second point, going beyond the predictability analysis at the microscopic level, we study how we can infer properties of different social groups from the individual data to predict emergent phenomena at the macroscopic level. In this case, we use a dataset composed by geo-tagged tweets (social contents published by users on Twitter) and we focus our analysis in predicting the unemployment rates of geographical regions. On Twitter, users publish their opinion, interests, share information and interact to other users but there is much more information that can be extracted that has not been explicitly created by the user. For instance, we use the time tweets were created to infer the activity daily rhythms of the users and, as a consequence, of the city where they live; to analyze how regions are connected, we infer the individual mobility network which is aggregated by regions afterward to discover connectivity patterns among them; We also derive the educational level of Twitter users by examining the misspellings that they make and computing the proportion of these users over the whole population. All this information is about individuals but it can be used to profile geographical regions and eventually predict an emergent phenomena as economy based on them.

Finally, to answer the third case, we analyze the dynamical inter-regional communication patterns between regions, which is also an emergent phenomena derived from the complexity of the underlying social network composed by internal (involving nodes in the same region) and external (involving nodes in different regions) links. Even though links happen between individuals, we use them to build the macroscopic network and analyze its characteristics. Our final goal is studying how the dynamical and statical structure affect diffusion processes on this geography. To this end, we use a dataset from a major mobile phone company, containing the CDRs (Call Detail Records) of the calls in the UK where every record is formed by the IDs of the two mobile phone numbers involved in the call, when it was made and the duration of the call.

1.1.6 Summary and chapters

This manuscript is divided in the next chapters:

- Chapter 1 - Introduction: a general overview of the mail goal of this work based on different data sources as well as the description of the statistical techniques used through all the rest of chapters.
- Chapter 2 - Predictability of shopping behavior: this chapter analyzes the question about whether the patterns behind shopping behavior may be accurately predicted based on the past data of every individual. To this end, we perform an analysis of a credit card transaction dataset in two major countries. In this chapter, we study the predictability limits based on entropy measures and build Markov models to predict where an individual will make his next purchase. We get that, despite the existing regularity on the shopping patterns at the global level, there exists a huge randomness inside the space of possible merchants at the individual level.
- Chapter 3 - Geo-tagged digital traces and economical status: in this case, we test the next hypothesis: if our economical status modify our behavior and social media has been studied as a reflection of our activity, can we model the economical level of geographical regions based on social media data? We analyze a huge dataset composed by geo-tagged tweets in Spain, one of the countries that held large unemployment rates during the crisis, to unveil variables related to technology adoption, daily patterns, social connectivity and educational level and to use them as predictors of the unemployment rate. Our results exhibit that this can be done satisfactorily allowing to develop new economical indicators in future.
- Chapter 4 - Information diffusion on dynamic geographical networks: in literature, it has been widely studied how the structure of geographical networks affect diffusion processes and also how the bursy dynamics of human interactions slow down diffusion. But, what is the real effect of the bursty, vibrant dynamics of the intra-urban relationships in the inter-urban diffusion spreading? Our results indicate that, despite the network changes strongly in our temporal window (almost 20% of the edges are created and destroyed), its effect on the inter-regional diffusion is not significative, leading to conclude that individual relationships are not important at the country level in terms of diffusion processes.

Chapter 2

Predictability in shopping behavior

In literature, universal laws have been found in network structure in different fields [19], individual activity patterns [16] or mobility [20]. In general, all these works use data with high granularity since, for instance, we are continously deciding where to move in our next step (or maybe we decide to stay in the same place) showing a kind of continuity in patterns. However, is shopping behavior as predictable as the previously mentioned behaviors? Are we able to predict accurately what our next economical decision will be? In this Chapter, we use a credit card transaction dataset to analyze to what extent the shopping behavior is really predictable. Our results indicate that high regularity is exhibited in our consumer visitation patterns, with recurring visits to our favorites merchants and small values in the entropy measures as in the examples in literature. However, despite this regularity, no relevant information about customers' next decision is contained in the previous historical visits concluding that consumers vary the order of the visits to merchants in time breaking the temporal structure observed in other contexts as usual mobility.

2.1 Human mobility predictability and next-place prediction problem

The capability of analyzing and modelling human mobility is of large interest to understand how individuals behave and how geography constraints their free will, that is, since one cannot travel anywhere due to multiple factors (physicial ones like time, distance or costs to arbitrary ones like job, interests, obligations, or family), the number of possibilities individuals might go is limited. On the other side, in our daily lives, we make arbitrary choices about which our next destination will be so we also have freedom to vary our trips and position within the limited space of possibilities we mentioned before. These two features of mobility, limited number of places to go and variations in our patterns, make the study of mobility particularly interesting since it is possible to find repetitive patterns on it (for instance, most of the people travel from home to work every day) but we also have to deal with random jumps that might have not been observed before. As Song et. al studied in [6] using mobile phone users' trajectories, a striking homogeneity is observed in human mobility predictability patterns within a population, independently the number of different locations visited by the user. The authors showed that by only using historical trajectory data from the mobile phone users, the predictive power of models (even if we think of the best possible model) is limited by the Fano's inequality, giving that, in their dataset, the median of predictability is approximately 0.93, exhibiting that historical series of trajectories contain much information about where the individuals are going to be next. Moreover, what they found is the visitation order determines the next-place where the user is going to be, that is, our recent movements contain much information about where we will be soon. In fact, we only consider the proportion of time in the different places or only the number of different places a user visits but we neglect the order of the visits, the predictability of mobility decreases dramatically.

However, not everyone has access to data from mobile phone companies and there are other sources of data containing mobility information: for instance, one might use geo-tagged social contents to model mobility (as it will be explained in Chapter 3 using Twitter data) or a financial company might want to know where their clients will make their next presential purchase, that is, in which merchant a client will make the next expense. In this last context, the problem changes completely: in this new scenario, data is not as fine as in the mobile phone context is because when this latter dataset is considered, we use records extracted every few minutes with the approximate position of the mobile phone holder (inferred by triangulization using the power of the signal received by the mobile phone from every antenna around it) whereas when credit card transaction records are considered, the granularity of the data is much coarser and cardholders cannot be tracked constantly. This is the main problem we face in the present chapter: can we predict the next purchase of a cardholder only with its past transactions? Is the granularity of the data fine enough to build accurate next-place prediction models?

Forgetting the motivations behind human decisions, mobility might be simply seen as the combination of two basic quantities: on one side, we can analyze the waiting time distribution, that is, how long we stay in a particular place and, on the other side, the distance we travel to our next destination. We can also think on mobility as a kind of random walk generated by individuals, where trajectories are samples of an underlying modified Markov process in continuous time. However, fat-tailed distributions in waiting time distribution and also in displacement have lead researchers to model trajectories as Continous Time Randow Walk or Levy Flights [20]. But, as it can be expected, human mobility is driven by interests and liability and it exhibits properties that are not present in this kind of models such as i) a tendency to stop visiting new places ii) a preferential return to the top visited places. These are the facts that make human mobility predictable. As we said before, if we think of an employee person, it spends the night at home sleeping, then it wakes up more or less at the same hour every day and afterwards it goes to it job on the same way as the previous day. Of course, some kind of variability might turn out but are the daily motivations which constraints the space of places where we can go to and then mobility becomes predictable. But the question is how we can infer the motivations behind mobility, how we can combine the temporal patterns with this information and, finally, whether there are data sources allowing us to do this kind of models.

Trying to taking advantage of our daily repetitive mobility, several authors have used non-linear time series analysis to uncover patterns data. In [32] authors use the Delay Embedding Theorem to reconstruct the phase space based on location variables (longitude, latitude an attitude) and temporal variables, to recover a higher dimensional space with the same features as the original space and where traditional time series model, such as AR(p) models, might be used to predict the next values of the series. Despite the sophistication of these techniques, they only use historical trajectories and, as we have explained before, depending on data resolution, there exist limits in the accuracy of the models because of the data nature. So in order to build better models, we need to use different sources of data containing information about individuals' mobility. A very widely used information applied to this problem to increase the performance of the models is social information, that is, information about how users are related among them (friends, family, colleagues,...). If it is reasonable to think that interests and motivations constrain mobility, so it is to hypothesize that friends' mobility is somehow related. For example, using Flicker data, in [33] showed that the probability of two users being friends dramatically increases when they have been in the same place in different places. In [34] authors use non-linear time series techniques but adding variables related to Mutual Information between trajectories of different users and they show that the model is more accurate when this information is considered. Finally, in [35] Noulas et al. also build new variables considering features such as non-linear individual mobility variables but also temporal patterns and information from the rest of users who sharing places with the user we want to predict where he is going to be. As we will see in the next section, this idea is also present in the Global Markov Model to predict where a customer will buy next.

Finally, from a more practical point of view, usefulness of mobility prediction has been widely explored in different fields such as epidemic spreading modeling [36–38], to traffic analysis in cities [39–41], optimization and resource planning such as public transport [42, 43], or scaling of communication power using mobile phone antennas depending on how many people there will be in a a place in future [44].

2.2 Using entropy to understand mobility predictability

Since one might think that mobility is predictable, we need to quantify to what extent this is the case. However, it is difficult to provide one exact definition of predictability and different aspects of mobility are related to it. For instance, an individual staying always at the same place is predictable but so is an individual visiting 10 different places but staying the 90% of the time in its favorite place. To deal with this context, we proceed as in [6] and define three measures to explain different aspects of mobility:

- Random entropy: defined as the logarithm of the number of different places visited by the user N_i
$$S_i^{rand} = \log_2 N_i \tag{2.1}$$

- Temporal-uncorrelated entropy: defined as the Shannon entropy. Considering $p_{i\alpha}$ the probability of finding user i in place α (estimated by the frequency), it is defined as
$$S_i^{TU} = -\sum_{\alpha} p_{i\alpha} \log_2 p_{i\alpha} \tag{2.2}$$

- Sequence-dependent entropy: given by the Lempel-Ziv algorithm to estimate the Kolmogorov Entropy as
$$S_i^{SD} \approx \frac{\log N}{\langle L(w) \rangle} \tag{2.3}$$
where where $\langle L(w) \rangle$ is the average over the lengths of the encoded sub-chains

On the limit, considering infinite sequences, SD entropy converges TU for random for large enough random sub-chains. As it can be easily checked, when visits are completely random among the chances of our space, TU equals RA. We also use these three metrics in the next section where we analyze human mobility based on credit card transaction data, both in an European country and in the USA.

As we will see in the next section, when we analyze the mobility patterns using credit card data, all these metrics are strongly related to the data resolution and the space of places where an user might be so it is necessary to look for more variables and additional techniques to retrieve more information where the customer is going to make his or her next purchase.

2.3 Consumer visitation patterns

2.3.1 The dataset

We sample tens of thousands individual accounts from one North American and one European financial institution. In the first case we represent purchases made by over 50 million accounts over a 6-month window in 2010-2011; in the second, 4 million accounts in an 11-month window. Data from transactions included timestamps with down-to-the-second resolution.

We filter each sample to best capture actual shoppers' accounts, to have sufficient data to train the Markov models with time series that span the entire time window, and to exclude corporate or infrequently used cards. We filter for time series in which the shopper visits at least 10 but no more than 50 unique stores in every month, and makes at least 50 but no more than 120 purchases per month. We test the robustness of this filter by comparing to a set of time series with an average of only one transaction per day (a much less restrictive filter), and find similar distributions of entropy for both filters.

The 25th, 50th (median) and 75th percentiles of the number of merchants per customer in the filtered time series are 46, 64 and 87 in the North American (6 months) and 69, 101, and 131 in the European (11 months) dataset, that is, the median customer in Europe uses his credit card in almost 11 different merchants per month whereas, in the USA, this figure raises up to almost 17. In order to compare with other data sources and focusing on the North American case, we retrieve the location of a user every 2500 minutes approximately whereas in [45] authors report to observe the location of an individual every 260 minutes.

2.3.2 Detecting heterogeneity in the visitation patterns

In order to understand the underlying nature of the credit card transaction dataset, we perform a descriptive analysis of the the distribution of the proportion of times the customers visit a merchant. As in many other cases in literature such asdisplacements in travels [20], degree distribution in real-world networks [21], time until an user opens an email [46] or appearence of words in texts [47], a Power-Law distribution fits this metric accurately. Given a random variable X, the probability density function of a Power-Law variable is given by

$$f(x) \sim x^{\alpha} \tag{2.4}$$

The exponent α varies depending on the application and it is typically estimated from data by the maximum-likelihood estimator given by

$$\hat{\alpha} = \frac{1}{N} \sum \log \frac{x_i}{x_{min}} \tag{2.5}$$

where x_{min} is the mininum value that the random variable might reach and N is the number of points in our data. However, applying this method directly has been shown to produce inaccurate estimations of the exponent due to large fluctuations happening on the tail of the distribution so we use the method proposed in [48] to compute the exponents.

Typically, to check visually and qualitatively whether a variable in a dataset follows a Power-Law distribution, it is usually plotted in logarithmic scale because if one takes logarithms in both sides at Eq. 2.4, it is obtained

$$\log f(x) \sim \alpha \approx \log x \tag{2.6}$$

that is, we observe a linear density in this new scale so it is very simple to identify.

At longer time scales, shopping behavior is constrained by some of the same features that have been seen to govern human mobility patterns. We find that despite varied individual preferences, shoppers are on the whole very similar in their statistical distribution (with significant differences in the exponent of the Power-Law), and return to stores with remarkable regularity: a Zipf's law (Power-Law distribution) $P(r) \sim r^{-\alpha}$ (with exponent α equal to 0.80 and 1.13 for the North American and European datasets respectively) describes the frequency with which a customer visits a store at rank r (where $r = 3$ is his third most-frequented store, for example), independent of the total number of stores visited in a three-month period, see Figure 2.1. This holds true despite cultural differences between the North American and Europe in consumption patterns and the use of credit cards. While our main focus is not the defense of any particular functional form or generative model of visitation patterns, our results support those of other studies showing the (power law) distribution of human and animal visitation to a set of sites [7, 8, 49, 50]. As a consequence, consumers visit their single top location approximately 13% (North American) and 22% (European) of the times, which indicates a clear trend to go back to the favorite merchants but to a certain point since this proportion is not a majority of the visits.

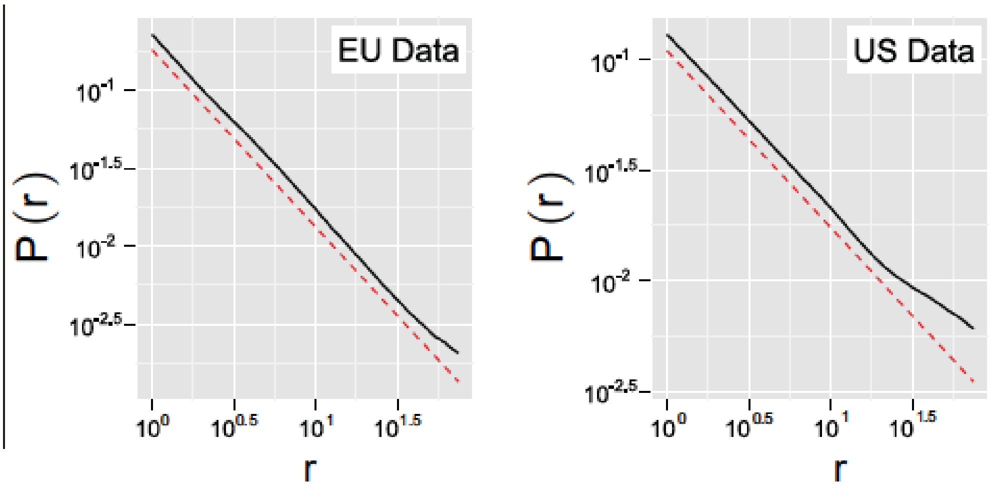


FIGURE 2.1: Probability of visiting a merchant, as a function of merchant visit rank, aggregated across all individuals. Dashed line correspond to power law fits $P(r) \sim r^{-\alpha}$ to the initial part of the probability distribution with $\alpha = 1.13$ for the European and $\alpha = 0.80$ for the North American database.

2.3.3 Predictability of consumer visitation patterns

Although just looking at the main visited merchants could give us some predictability of the mobility behavior, we now study if there is a specific hidden predictability in the temporal patterns, that is, is there any information in the sequence of visited merchants? How much information is in a shopper's time series of consecutive stores? A universal measure of individual predictability would be useful in quantifying the relative regularity of a shopper.

Informational entropy is commonly used to characterize the overall predictability of a system from which we have a time series of observations. It has also been used to show similarities and differences across individuals in a population [51].

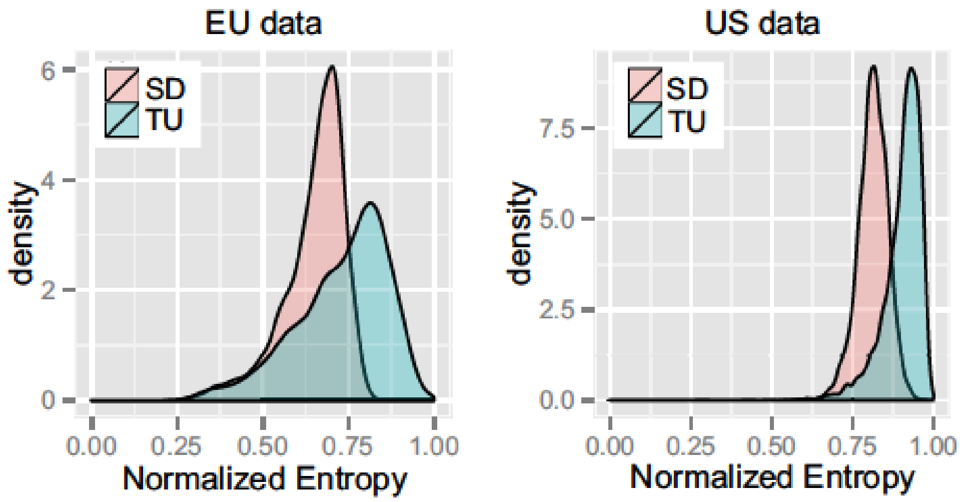


FIGURE 2.2: Normalized entropy distributions for the North American and European populations. Normalized entropy is computed by dividing the TU and SD entropies by the logarithm of the number of different merchants visited by a customer. TU entropy distributions are slightly higher for both populations

We consider two measures of entropy (see 2.2 for further explanation about how informational entropy is computed in every case in this work):

(i) The temporally-uncorrelated (TU) entropy for any individual i is equal to $S^{\alpha}_{TU} = -\sum_{i \in M_\alpha} p_{\alpha,i} \log(p_{\alpha,i})$ where $p_{\alpha,i}$ is the probability that user α visited location i. Note this measure is computed using only visitation frequencies, neglecting the specific ordering of these visits.

(ii) The sequence-dependent (SD) entropy, which incorporates compressibility of the sequence of stores visited, is calculated using the Kolmogorov complexity estimate [52, 53].

We find a narrow distribution of TU and SD entropies across each population, Figure 2.2.

Another dataset, using cell phone traces [6], also finds a narrow distribution of entropies. This is not surprising, given the similarity of the two measures of individual trajectories across space. Yet we find a striking difference between the credit card and the cell phone data. In the shopping data, adding the sequence of stores (to obtain the SD entropy) has only a minor effect of the distribution, suggesting that individual choices are dynamic at the daily or weekly level. By contrast, cell phone data shows a larger difference. Why does this discrepancy occur? A possible explanation is that shoppers spread their visitation patterns more evenly across multiple locations than do callers. Even though visitation patterns from callers and from shoppers follow a Zipf's law (figure 1), callers are more likely to be found at a few most visited locations than are shoppers. This is true, but to a point. As we mentioned before, consumers visit their top location approximately 13% in the North American case while it happens for the 22% in the European one, different from mobile phone datasets where callers exhibit more frequent visitation to top location. Yet shoppers' patterns follow the same Zipf distribution seen in the cell phone data, and the narrow distribution of temporally-uncorrelated entropy indicates that shoppers are relatively homogenous in their behaviors.

An alternative explanation for our observed closeness of temporally-uncorrelated and sequence-dependent entropy distributions is the presence of small-scale interleaving and a dependence on temporal measurement. Over the course of a week a shopper might go first to the supermarket and then the post office, but he could just as well reverse this order. The ability to compare individuals is thus limited by the choice of an appropriate level of temporal resolution (not necessarily the same for each dataset) to sample the time series. With the large-scale mobility patterns inferred from cell phones, an individual is unable to change many routines: he drives to the office after dropping off the kids at school, while vice versa would not be possible. In the more finite world of merchants and credit card swipes, there is space for routines to vary slightly over the course of a day or week.

To test the extent to which the second hypothesis explains the discrepancy between shoppers and callers, we simulate the effect of novel orderings by randomizing shopping sequence within a 24-hour period, for every day in our sample, and find little change in the measure of SD entropy. In other words, the re-ordering of shops on a daily basis does little to increase the predictability of shoppers, likely because the common instances of order swapping (e.g. coffee before rather than after lunch) are already represented in the data. We then increase the sorting window from a single day to two days, to three days, and so forth.

Yet when we sort the order of shops visited over weekly intervals, thus imposing artificial regularity on shopping sequence, the true entropy is reduced significantly. If we order over a sufficiently long time period, we approach the values seen in mobile phone data. Thus entropy is a sampling-dependent measure which changes for an individual across time, depending on the chosen window. While consumers' patterns converge to very regular distributions over the long term, at the small scale shoppers are continually innovating by creating new paths between stores.

2.3.4 Markov models

In the next section we use Markov models to predict the next place where a customer will make her new purchase so let's define them in a formal way and study how they are related to the entropy measures used before. Markov chains are used to model temporal stochastic processes, in which the present state depends only on the previous one(s). Mathematically, let X_t be a sequence of random variables such that

$$P(X_t = x_t | X_{t-1} = x_{t-1}, X_{t-2} = x_{t-2}, X_{t-3} = x_{t-3}, ...) = P(X_t = x_t | X_{t-1}) \tag{2.7}$$

then $\{X_t\}$ is said to be a Markov process of first order.

This process is summarized with transition matrix $P = (p_{ij})$ where $p_{ij} = P(X_t = x_j | X_{t-1} = X_i)$, that is, the probability of going from one state to another one in the next time (we only consider discrete time Markov chains). Markov chains can be understood an extension of a simple frequentist model in which

$$P(X_t = x_t | X_{t-1} = x_{t-1}, X_{t-2} = x_{t-2}, X_{t-3} = x_{t-3}, ...) = P(X_t = x_t) \tag{2.8}$$

applied on every observed state.

If the present transaction location depends in to some extent on the previous one, a first order Markov model would be able to predict the location with greater accuracy than a simple frequentist model.

This observation allow us to note two relationships:

- *Temporal-uncorrelated entropy and frequentist model:* both use $P(X_t = x_t)$ without additional information. Temporal-uncorrelated is a good approximation of the distribution of states, and is thus related to the performance of the frequentist model.
- *Sequence-dependent entropy and Markov model:* SD entropy is a single measure of all sub-chain frequencies, and is thus related to the accuracy of a 1st order Markov model, which represents the probability of a single set of sub-chains occuring.

Once we have created a Markov model, different measures related to its accuracy and performance must be computed to assess its quality, that is, to measure whether real world data is well represented by the model. However, depending on the predicition problem and on the type of model, different performance measures might be computed. The measure used is in this case is the percantage of hitted purchases,

$$acc = \frac{\#\text{ of hits}}{\#\text{ of predicted purchases}} \tag{2.9}$$

Even though this problem might be understood as a recommendation system problem and therefore we might have selected a softer criteria (for instance, whether the customer has bought something in the predicted merchant in the next 5 purchases), our main goal is to show that there is no significant information in the visitation sequence to predict the next place although our space of places to make a purchase is constrained because of our daily patterns. Finally, this way of evaluating next-place predictors has been extensively used in the literature [32, 54] but typically in contexts based on finer data such as mobile phone datasets.

2.3.5 Using Markov models to test predictability

In order to measure the predictability of an individual's sequence of visits, we train a set of first order Markov chain models. These models are based on the transition probabilities between different states, with the order of stores partially summarized in the first-order transition matrix. It is thus related to the SD entropy measure. We measure the probability of being at store x at time $t+1$ as $Pr(X_{t+1} = x | X_t = x_t)$ and compare the prediction values to the observed values for each individual. We build several models, varying the range of training data from 1 to 6 months of data for each individual, and compare the model output to test data range of 1 to 4 subsequent months.

We additionally compare the results of the Markov models to the simplest naive model, in which the expectation is that an individual will chose his next store based on his distribution of visitation patterns, e.g. he will always go to one of the top two stores he visited most frequently in the training window (recall that for most people this store visitation frequency is on average just 20-35%). Since this is a simply frequentist approach to the next-place prediction problem, it is strongly related to TU entropy which is computed using the probability that a consumer visits a set of stores.

Comparing the match between model and observed data, we find that using additional months of training does not produce significantly better results. Moreover, results show some seasonal dependency (summertime and December have lower prediction accuracy, for example). For fewer than three months of training data, the frequentist model does significantly better than the Markov model. This suggests the existence of a slow rate of environmental change or exploration that would slowly undermine the model's accuracy.

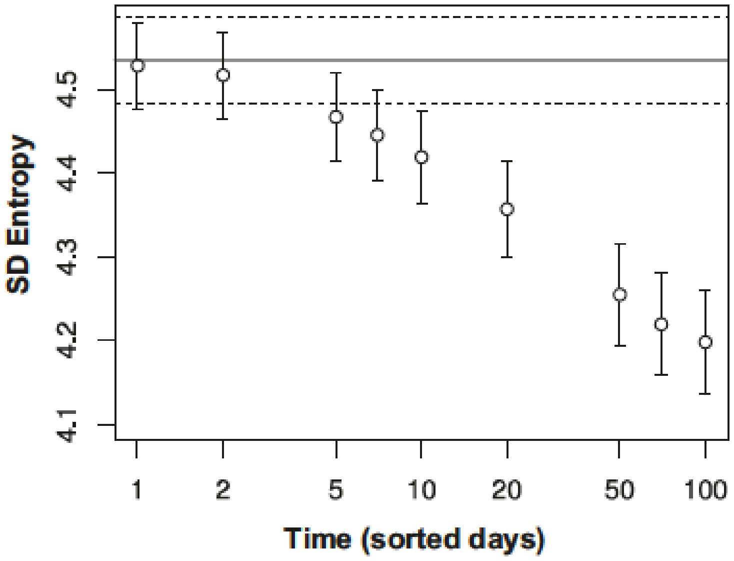


FIGURE 2.3: Sequence-dependent entropy for a number of "artificially sorted" sequences. For each window size over which the time series is sorted, we measure the sequence-dependent entropy for the population and estimate the error of the mean. The horizontal solid line at the top of the figure indicates the average SD entropy for the original data whereas the dashed lines depict the band for the error of the mean.

For each of the two populations, we next test a global Markov model, in which all consumers' transition probabilities are aggregated to train the model. We find that such a model produces slightly better accuracy that either the naive or the individual-based models (with accuracy $\approx 25-27\%$). To test the sensitivity of this result we take ten global Markov models trained with 5% of time series, selected randomly. We find the standard deviation of the accuracy on these ten models increases to 3.6% (from 0.3% using all data), with similar mean accuracy. Thus the global Markov model depends on the sample of individuals chosen (for example, a city of connected individuals versus individuals chosen from 100 random small towns all over the world), but does in some cases add predictive power.

As previous work has indicated [55], mobility patterns can be predicted with greater accuracy if we consider the traces of individuals with related behaviors. In our case, even though we have no information about the social network of the customers, we can set a relationship between two people by analyzing the shared merchants they frequent. The global Markov model adds information about the plausible space of merchants that an individual can reach, by analyzing the transitions of other customers that have visited the same places, thus assigning a non-zero probability to places that might next be visited by a customer.

Yet in almost every case, we find that people are in fact less predictable that a model based exclusively on their past behavior, or even that of their peers, would predict. In

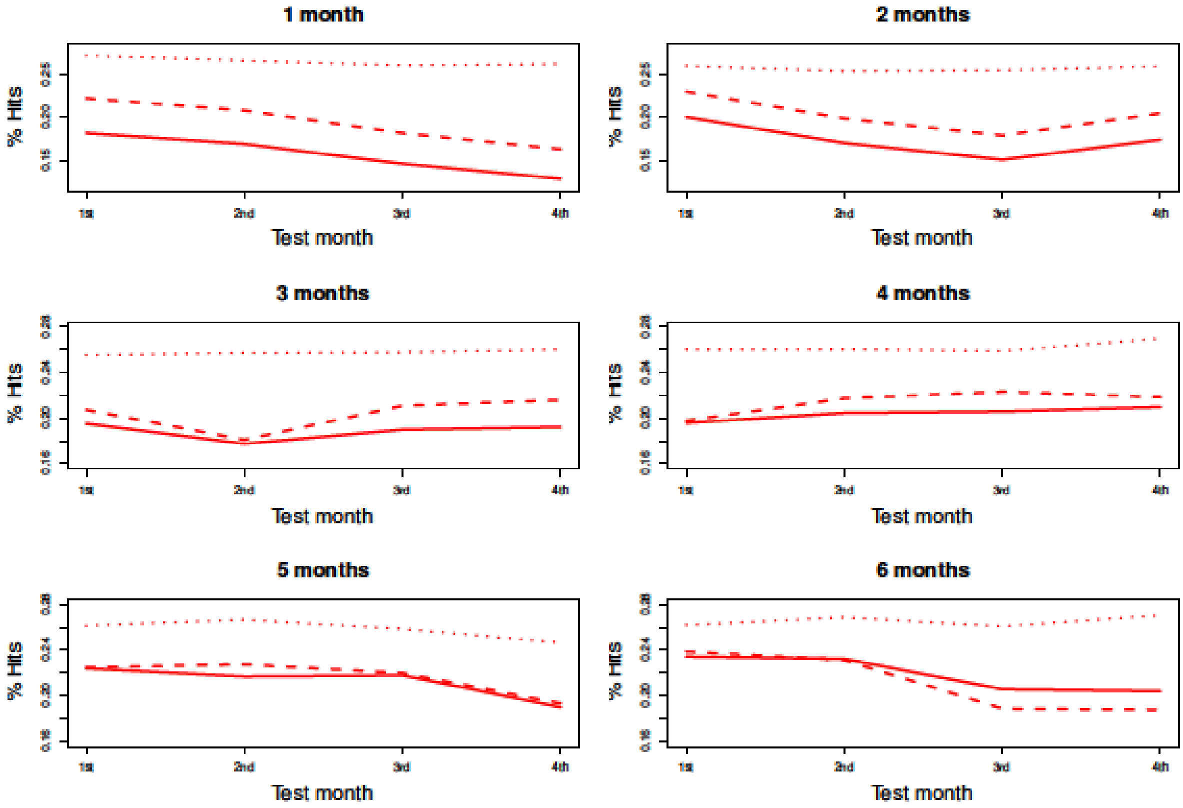


FIGURE 2.4: Markov model results for different temporal windows in training and test. The solid red line indicates hit percentage for Markov model, dashed line exhibits accuracy for the naive model and the pointed line indicates results for the Global Markov model.

other words, people continue to innovate in the trajectories they elect between stores, above and beyond what a simple rate of new store exploration would predict.

2.4 Conclusions and further discussion

Colloquially, an unpredictable person can exhibit one of several patterns: he may be hard to pin down, reliably late, or merely spontaneous. As a more formal measure for human behavior, however, information-theoretic entropy conflates several of these notions. A person who discovers new shops and impulsively swipes his card presents a different case than the one who routinely distributes his purchases between his five favorite shops, yet both time series show a high TU entropy. Similarly, an estimate of the SD entropy can conflate a person who has high regularity at one level of resolution (for example, on a weekly basis) with one who is predictable at another.

As example, take person A, who has the same schedule every week, going grocery shopping Monday evening and buying gas Friday morning. The only variation in A's routine is that he eats lunch at a different restaurant every day. On the other hand, person B sometimes buys groceries on Tuesdays, and sometimes on Sundays, and sometimes goes two weeks without a trip to the grocer. But every day, he goes to the local deli for lunch, after which he buys a coffee at the cafe next door. These individuals are predictable at different time-scales, but a global measure of entropy might confuse them as equally routine.

Entropy remains a useful metric for comparisons between individuals and datasets (such as in the present and cited studies), but further work is needed to tease out the correlates of predictability using measures aligned with observed behavior. Because of its dependence on sampling window and time intervals, we argue for moving beyond entropy as a measure of universal or even of relative predictability. As our results suggest, models using entropy to measure predictability are not appropriate for the small scale, that is, their individual patterns of consumption.

Shopping is the expression of both choice and necessity: we buy for fun and for fuel. The element of choice reduces an individual's predictability. In examining the solitary footprints that together comprise the invisible hand, we find that shopping is a highly predictable behavior at longer time scales. However, there exists substantial unpredictability in the sequence of shopping events over short and long time scales. We show that under certain conditions, even perfect observation of an individual's transition probabilities does no better than the simplistic assumption that he will go where he goes most often.

2.5 Publications, media coverage and acknowledgment

During the development of this manuscript, the analysis described in this Chapter was published in a scientific article named as "The predictability of consumer visitation patterns" in 2013. [56].

This work also gained a remarkable impact on online blogs [57–59] but also on one of the most important generalist media company in Spain [60].

The author wants to thank their effort to all the collaborators who made possible this work.

Chapter 3

Geo-tagged digital traces and economical status

Economic indicators are useful to monitor how the wealth in different regions evolve or to study how public policy making impact the real life of inhabitants. However, many of these indicators are elaborated by conducting surveys on the population, which is a costly process and does not allow continous monitoring of economy. Additionally, in some countries, where public administration and infrastructures are not able to reach all the population, these methods have been proved to be inefficient. Today, the appearance of digital technologies, sensors and social networks allow us to study a huge amount of information generated by individuals in societies that are publicly available. How can we take advantage of this information to monitor economy? In this chapter we build a predictive model of unemployment rate in geographical regions by only using variables extracted from Twitter, an online social network where users publish their opinions, interests and location. We show how this information can be used to estimate variables that are typically related to economy such as the technology penetration rate, the geo-social connectivity, the temporal patterns or the educational level. Our results indicate that social networks are a reliable source of information of human activity in such a way that economy, in terms of unemployment rate, can be accurately modeled.

3.1 Mobility, digital traces and economy

3.1.1 Digital traces and economic indicators

The pervasive usage of technology in our daily life has changed the world we live. Nowadays, we can be constantly connected to information sources and we can interact to our contacts through different channels instantly at any time. Since indivuals are permanently using technology and provided that companies store information about how users make use of their services, the digital traces generated originally for business and process monitoring must also serve to analyze how we behave, move or interact every day. Simply by observing the individuals' activity, we can analyze heterogeneity in the circadian rythms of the population and conclude different aspects such as socio-demographic differences, bursty behavior of interactions or its impact on information diffusion [18, 61, 62]. Independently this was not the original purpose for storing all this information, it is clear that digital traces can be used for studying human behavior and, going one step forward, analyzing how human behavioral patterns are related to economic indicators. In particular, in this Chapter, our main goal is to use a open available online data source, Twitter, for extracting features of regions from data at the individual level in different aspects (technology penetration, educational level, activity patterns and geo-social connectivity) and test whether all this information is related to unemployment rate in Spain.

However, this is not a completely new idea: the widely spread use of technology in developed countries made the researchers and policy makers think that information contained in digital traces might be used to compute new indicators or to compute current ones but reducing costs. A variable that has been extensively studied with similar goals is the penetration rate of different technologies such as telecommunication infrastructure [63], computer and Internet penetration [64, 65] or online social networks usage [66]. The hypothesis behind these works is quite reasonable: since population must have a certain wealth to be able to acquire technology and also the country must be wealthy enough to build infrastructures capable of supporting the needs of the mobile phone and Internet users, economical differences between regions might be encoded in how countries are able to adopt and suppor technology. But technology penetration rate has not been the only variable extracted from digital traces that has been proved to hold a strong relationship with economic development.

On other hand, researchers hypothesized how the different network structure of communications, inter-regional commercial activity or social interactions are related to economy. The theoretical idea behind this is that those areas holding a higher degree of connectivity and diversity exhibit a larger economical activity, resulting in more different sources of information for the regions, risk diversification and, eventually, holding a higher number and more diverse connections allow regions to pay attention to more opportunities to improve their development [67–71]. However, many of these works lack of empirical evidence because of the shortage of data available at the time they were published but, nowadays, many sources of information contain relational information at both individual and geographical level. For instance, Call Detail Records (CDR) from telecommunication companies store information about how people communicate each others but also geographical information of every individual (the billing address or the antenna the user was connected to at the beginning of a phone call) [51]. Also social media databases contain this kind of information since many of the users can be located in a particular place and they allow different types of interactions among users (on Twitter we can find mentions, retweets, replies and favorites; on Facebook, there exist mentions, comments, shares and likes; ...). In particular, in this chapter we analyze how Twitter data can be used to set a connection structure between regions and how the functional partition of a country can be inferred from this information.

A way of studying the complexity of the network between regions is analyzing the underlying structure in mobility data. Many works have used this kind of information to predict economical levels. In [72], authors use data from London railway to infer movements within the city and to build a model that is able to uncover the areas with low or high deprivation levels; In [51], they mainly analyze diversity in social connections and they use istto build a model of the economical regions in the UK: those areas exhibiting more diverse mobility and connections are wealthier than those with a more predictable mobility and social connectivity. However, as we will see in the next section, in our case both social and mobility diversity are not the most important factors to predict unemployment in Spain based on Twitter data; in [73] they combine information from Call Detail Records and airtime credit to understand the socio-economic state of regions in Cote d'Ivoire; from a more Machine Learning point of view, [74] use a extensive set of mobility variables to build a non-linear model to predict economic levels.

Regarding borders and geo-economical aspects, many works have focused on how artificial limits given by administrations made by humans, such as provinces, states or countries, influence mobility fluxes, making people and capital more difficult to move whereas, in other situations, where administrative limits are not very strong (limits within a country), these borders do not constraint mobility and therefore individuals' daily mobility result in more dynamical regions where economical activity lies [75–78]. We adopt these these techniques to extract the functional regions of Spain, in order to overcome the heterogeneity in municipalities and provinces in Spain, playing the role of new regions emerged from mobility graph community structure.

Finally, other data sources have been used to set relationships between insights extracted from them and the economy, for both state and financial economy. For instance, using Google query logs, [79] showed how to build a buying and selling stock strategy based on changes in the volume of searches in terms related to economy; also using Google Trends data, [80] analyzed how many searches refer to future past and how many searches refer to the future and showed that those countries where more queries were made about future hold a higher GDP; using Twitter data, as in the next section, [81] made economical indicators based on usual expressions used when people lose their job or when they get a new one.

As it can be observed, there exists large variety of data sources we can extract economical information from. The key aspect is this data might be processed in real time and therefore to be able to make decisions in smaller time scales than before. Finally, in developing countries where there are no infrastructures and even census is not completely accurate, all these methodologies might be used to build indicators they did not have until now.

3.1.2 Mobility models

As we have mentioned in the previous section, mobility data plays a key role in the analysis of inter-regional activity because it is a very good way to see how regions are connected, the intensity of these connections and what the real implications of the underlying structure are. In particular, in our work we use geo-tagged tweets to infer when a user has travelled from one city to another in the same day which is the basis to quantify the intensity of the relationship between two cities. However, a need task is to check whether this inferred mobility follows the same statistical features of other mobility data sources (which is done in following sections). In particular, many mobility models have been shown to predict accurately mobility flows between regions so, if our data is good enough to study inter-regional flows, it should be well fitted by a mobility model.

Extensively in the literature, mobility models have attracted great interest in the scientific community. The analysis of fluxes between areas is one of the most prevalent studied problems, that is, analyzing which the variables are influencing in how many trips are observed between two given places. Typically, this problem involves, on one hand, a partition of the geographical space, that is, a set S of disjoint regions of the space which union returns the whole space of analysis. On the other hand, these regions of the space are the other elements of interest because inferred variables of them are the elements composing the basic elements of the these mobility models. Traditionally, the number of trips between two regions i and j is denoted by T_{ij} and the matrix $\Theta = (T_{ij})$ is named the Origin-Destination matrix (OD matrix). The analysis of this matrix is particularly interesting in fields such as transport planning [82], migration prediction [83–85], traffic optimization in cities [39], event detection in geographical spaces [86] or epidemic spreading [36].

In 196, Zipf presented an idea based on Newton's Gravity Law to model the OD matrix [87], where the number of trips T_{ij} is approximated by

$$T_{ij} \approx T_{ij}^{gra} = \frac{P_i^{\alpha} P_j^{\beta}}{f(d_{ij})} \tag{3.1}$$

where P_i and P_j are the population of places i and j respectively and d_{ij} is the distance between both places. Many different functional forms have been proposed in the numerator to model the cost of traveling between far places as exponential or stretched functions but typically in modern works a polynomial form $f(d_{ij}) = d_{ij}^{\gamma}$ has been used (very commonly with $\gamma = 2$) as in [31, 72, 78, 88]. Despite the simplicity of the model, which is only based on population of the regions and the distance between them, it has been shown a good description of human mobility fluxes. However, other authors have detected anomalies between human migrations and the Gravity Model such as lack of formal derivation, symmetry of the model or discrepancies with real data and proposed more advanced models, such as the Radiation Model, in both discrete and continuum fashions [31, 89] using information about opportunities and also population inside the circles with radius less or equal the distance between the points. Even though Gravity Model can be outperformed in some cases, we use it in our research to model mobility fluxes based on geo-tagged social media data and to conclude that this data can be used as a proxy of human mobility. Other families of models, such as entropy based models, have been studied since the 1970s and they have also been linked to Gravity Model under certain constraints. Finally, one of the main factors to consider with Gravity models is data sources are not generally homogeneous in geographic spaces in terms of penetration so population and users rate must be corrected to avoid estimation biases.

3.1.3 Social Network Analysis

3.1.3.1 Introduction

In order to understand the underlying structure from Twitter data used to build predictive models on unemployment rate in Spain, we use a set of mathematical techniques belonging to Social Network Analysis (SNA), a field of knowledge between mathematics, statistics and computational science developed to study not only networks involving people and social interactions but also networks where the nodes can be molecules, places, text concepts, etc. In particular, in this Chapter we focus on three kind of applications of the SNA techniques:

- *Transforming user social network into a geographical network*: although Twitter data is disaggregated at the individual level and we observe interactions between users, every one can be located at a particular place, what we call its hometown, and therefore we can build a geographical network by considering relationships between users in different cities.
- *Metrics on social networks*: once we have built a network, metrics related to activity, conectivity or clustering can be measured on it, allowing us to understand it quantitatively.
- *Community Detection*: a structural way of understanding how the network is build is to study which groups are cohesively connected, that is, detecting the communities. In particular, we use community detection algorithms to extract the functional regions from mobility patterns, understood as groups of municipalities, in Spain.

3.1.3.2 Building networks at the microscopic and macroscopic level

A network is a mathematical object $G = G(V, E)$ composed by two sets: the node set (or vertices) $V = \{v_1, ..., v_N\}$ and the edge set (or links) $E = \{e_1, ..., e_K\}$ where each element e_i is a pair of nodes. Depending on the context, this pair of nodes might be ordered indicating that the relationship between nodes is directed or this order might not appear leading to a undirected network. Through this manuscript, vertices represent places where an individual make a purchase (in chapter 2), Twitter users (in chapter 3) or individuals who are mobile phone users (in chapter 4). A common factor in all these cases is the presence of geographical data since purchases occur in a particular merchant, tweets are geo-tagged and mobile phone users live in a concrete region. Once this context is present, it is natural to define the aggregated geographical network: as every individual is located in a particular region, we build the macroscopic network where nodes are these regions and two of them are connected if at least there exists a link involving users from both regions at the microscopic level. We define $E_{\alpha,\beta}$ as the set of edges of the microscopic level between regions α and β with $w_{\alpha,\beta} = |E_{\alpha,\beta}|$ the number of links between two regions. In the particular case of Chapter 3, every user is located in a municipality given by the latitude and longitude of geo-tagged tweets and two cities are connected when at least one users is observed to tweet in both cities on the same day.

3.1.3.3 Metrics on social networks

In all the analysis carried out, we focus on three main dimmensions that can be computed in our networks:

- *Activity:* in all the networks that are built on a kind of interactions (purchases, tweets, phone calls) a typical factor to analyze is the level of activity of every individual on it. This is not completely related to the number of different connections that an individual holds because we might find nodes with low values of connectivity but highly intense activity. Activity might be also computed at the relationship level; for instance, the activity of a phone call relationship is higher when a large number of calls is observed between two individuals. Hence, for a given node i, we can measure the activity w_i by

  $$w_i = |\{E_{l,k} : l = i \quad \text{or} \quad k = i\}| \tag{3.2}$$

  If the network is built by only considering structural relationships (for instance, follower network on Twitter), this measure matches the degree of the node k_i, that is, the number of total connections in the graph.

- *Connection diversity:* one might want not only to analyze the intense of the activity of nodes and relationships but to study whether the activity always happens in the same place or it is distributed among the relationships. For instance, in Chapter 2 we analyze for every credit card owner if she tends to spend all her money in some few places or she distributes her expenses equally on her space of visited merchants; in Chapter 3 we study how a region is connected to the rest of regions, that is, whether there exists some few prefferential connections or the interactions are spread over more regions. This is what it is measured by diversity: it is a combination of the number of different connections per node but also a measure of the intensity of every relationship. The most widely used diversity measure is based on Shannon entropy: given a node i, we compute its connection diversity by

  $$S_i = -\sum_{j \in \mathcal{N}_i} p_{ij} \log p_{ij} \tag{3.3}$$

  where $\mathcal{N}_i$ is the set of neighbors of node i and p_{ij} is the fraction of interactions with j over the number of interactions of j. We also use its normalized version

  $$\tilde{S_i} = \frac{S_i}{\log N_i} \tag{3.4}$$

  where N_i is the size of the neighborhood of node i.

- *Clustering*: in real-world networks is found that closed triangles are likely to appear, that is, if i and j are connected and so are j and k then i and k are likely to be connected. One can compute the proportion of times this phenomena occurs in a network which is called *clustering*. This definition can be also extended to weighted networks [90] by

$$C_i^w = \frac{1}{2s_i(k_i - 1)} \sum_{j,k \in \mathcal{N}_)} (w_{i,j} + w_{i,k}) a_{i,j} a_{i,k} a_{j,k} \tag{3.5}$$

where s_i is the sum of the link weights involving i, $k_i = |\mathcal{N}_)|$ and $a_{i,j}$ is 0 or 1 depending on the existence of one link between nodes i and j.

3.1.3.4 Community detection algorithms

As we will see in the following sections, a problem when studying data at the municipality level in Spain is the large heterogeneity in terms of population and number of Twitter users. To solve this circumstance, we extract from the mobility graph which municipalities are strongly connected among them so that we recover regions composed by municipalities with a functional relationship, there exists a significant activity among them. Algorithmically, this is performed by using community detection algorithms.

Even though there is no a global way to define communities, the idea is identifying nodes of the network that are more connected among them compared to the rest of the graph. Because of this qualitative definition, many methods to compute these communities have been purposed depending of several factors. From a machine learning perspective, hierarchical clustering based on structural similarity between nodes has been used to identify communities, that is, by setting a measure on how many neighbors share two nodes [91]. Other algorithms are based on centrality of nodes and edges such as the Girvan-Newman algorithm [92] that computes the betweeness of edges to carry out a percolation process and identify the resulting connected components as parts of the same community. However, these algorithms are computationally expensive because the measures they are based on are hard to compute. Most of modern algorithms rely on modularity optimization (a measure of how a partition describes correctly the structure of links between nodes) such as Multilevel community detection algorithm[93], or on the information enclosed in random walks on the graph such as Infomap algorithm [94].

Independently from the algorithm, the result is a vector of membership to a community, that is, a vector $B = (C_{i1}, ..., C_{iN})$ where C_k is the identifier of the community the node k belongs to. So given two different partitions of a network, how can they be compared? This can be done by means of Normalized Mutual Information (NMI) [95]: let B_1 and B_2

be the membership vectors returned by two differents community detection algorithm and C_1, C_2 the number of detected communities by each algorithm, NMI is computed as

$$NMI(B_1, B_2) = \frac{-2\sum_{i=1}^{C_1}\sum_{j=1}^{C_2} N_{ij}\log(\frac{N_{ij}N}{N_i N_j})}{\sum_{i=1}^{C_1} N_i \log(\frac{N_i}{N}) + \sum_{j=1}^{C_2} N_j \log(\frac{N_j}{N})} \tag{3.6}$$

where N_{ij} is the number of nodes in community i that also appears in community j and N is the total number of nodes.

3.2 Social media fingerprints of unemployment

3.2.1 Introduction

In the previous section, we have reviewed many situations in which different data sources are used as a reflection of human activity and, if this is the case, we can extract insights about populations' and societies' phenomena such as the economy. Nevertheless, most of them focus on some few features extracted from the data sources and exhibit the relationship to economy from a point of view. In our approach, we use one only data source, Twitter data extracted from the public Streaming API, to build variables of regions in four different dimmensions: penetration rate, educational level, temporal patterns and geo-social diverse connectivity. We select these four families of variables because, priorly, it is reasonable to think there might be a connection to economy: differences in economical development might be reflected in diffferent Twitter penetration rates because it implies the usage of a relatively costly associated technology; it is known that educational level and wealthy of regions are correlated (we are not setting or looking for any kind of causality) so if create a variable being a proxy of this, it might be related to economy; there might be differences in the temporal patterns of activivy between cities with different unemployment rates; and finally, inspired by the scientific literature [51] , diversity of connections allow individuals and regions to obtain information and opportunities from many sources which also has an impact on economic develoment. In our understanding, this is a unique holistic approach with an only data source, trying to model four different and complemmentary features of regions to model an economic indicator.

3.2.2 The dataset

Twitter is a microblogging online application where users can express their opinions, share content and receive information from other users in text messages of 140 characters long, commonly known as *tweets*. Users can interact with other users by mentioning them or retweeting (share someone's tweet with your followers) their content. Some of these tweets contain information about the geographical location where the user was located when the tweet was published; we refer to them as geo-located tweets.

To perform our analysis, we consider 19.6 million geo-located Twitter messages (tweets), collected through the public API provided by Twitter from continental Spain, ranging from 29th November 2012 to 30th June 2013. Tweets were posted by (properly anonymized) 0.57 Million unique users and geo-positioned in 7683 different municipalities. We observed a large correlation (Pearson's coefficient $\rho = 0.951[0.949, 0.953]$) between the number of geopositioned tweets per municipality and the municipality's population. On average we find around 50 tweets per month and per 1000 persons in each municipality.

In order to analyze the relationship between social media and the economical level measured by the unemployment rate, we also consider population and economical information about the municipalities from the Spanish Census (2011) [96] and unemployment figures from the Public Service of Employment (Servicio Público de Empleo Estatal, SEPE) [97]. In the former In the latter case, registered unemployment (in number of persons) is given for each Spanish municipality by gender, age, and month. To get unemployment rates we divide register unemployment by the total workforce in the municipality, estimated as the number of people with age between 16 and 65 years.

3.2.3 Twitter as mobility proxy

Considering all of the available transitions in our database, one can compute the distance between origin and destination, the elapsed time of the transition and the number of trips per user among many other statistics. All of them seems to show a Power-law distribution with a cutoff due to the finite spatial size of Spain and the constraint of considering only transitions where the origin and destination checkins are done the same day. Focusing on the log-linear part of the distributions, self-similar behaviors arise when Twitter based mobility is analyzed (see figure 3.1).

Remarkably, the statistical properties of trips coincide with those of other mobility datasets: for example, trip distance r and elapsed time δt are power-law distributed with exponents $P(r) \sim r^{-1.67}$ and $P(\delta t) \sim \delta t^{-0.67}$, very similar to those found in the literature [30, 66]. Exponents are computed by the method used in [48].

We say that there is a daily trip between municipality i and j if a user has tweeted in place i and j consecutively within the same day. In our dataset we find 1.9 million trips

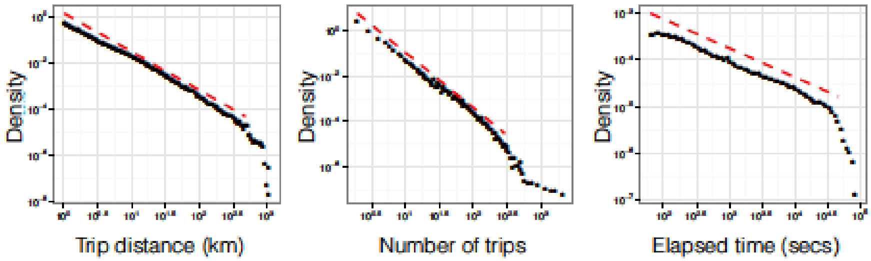


FIGURE 3.1: Probability distributions for the different properties of daily trips in the Twitter dataset. Dashed lines corresponds to a power law fit with exponents -1.67, -2.43 and -0.62 respectively

by 0.22 million users. With those trips we construct the daily mobility flux network T_{ij} between municipalities as the number of trips between place i and j (see Fig. 3.2B).

Twitter based inter-city flows can be well modelled by means of the The Gravity Law, which is one of the most extended methods to represent human mobility (see section 3.1.2 for further reading). The Gravity Model for human mobility assume that the flows between cities can be explained by the expression

$$T_{ij}^{grav} = \frac{P_i^{\alpha_1} P_j^{\alpha_2}}{d_{ij}^{\beta}} \tag{3.7}$$

where T_{ij}^{grav} is the flow, in terms of number of people, between cities i and j, d_{ij} is the geographical distance and P_i and P_j the population of every city respectively.

Given the data we can obtain the parameters of the model by Weighted Least Squares Minimization,

$$\alpha_1^*, \alpha_2^*, \beta^* = \underset{\alpha_1, \alpha_2, \beta}{\text{argmin}} \frac{1}{N} \sum_{i,j} w_{ij} \left(T_{ij} - T_{ij}^{grav} \right)^2 \tag{3.8}$$

where N is the total number of connections in the mobility graph and w_{ij} is a weight proportional to the number of observed transitions between i and j. In particular we find that taking $w_{ij} = T_{ij}^{1.3}$ gives the best performance in the model.

In our case, this model fits quite accurately the inter-city mobility based on Twitter GPS checkins (see Table 3.1). Even though we are considering T_{ij} not necessarily symmetric, the exponents of the populations are similar indicating that we are observing a similar flows in both directions between i and j. The exponents in (3.7) are very similar to those reported in other works $\alpha_i \simeq \alpha_j = 0.48$ and $\beta \simeq 1.05$ [31, 66]. These results suggest that detected mobility from geo-located tweets is a good proxy of human mobility within and between municipalities [98].

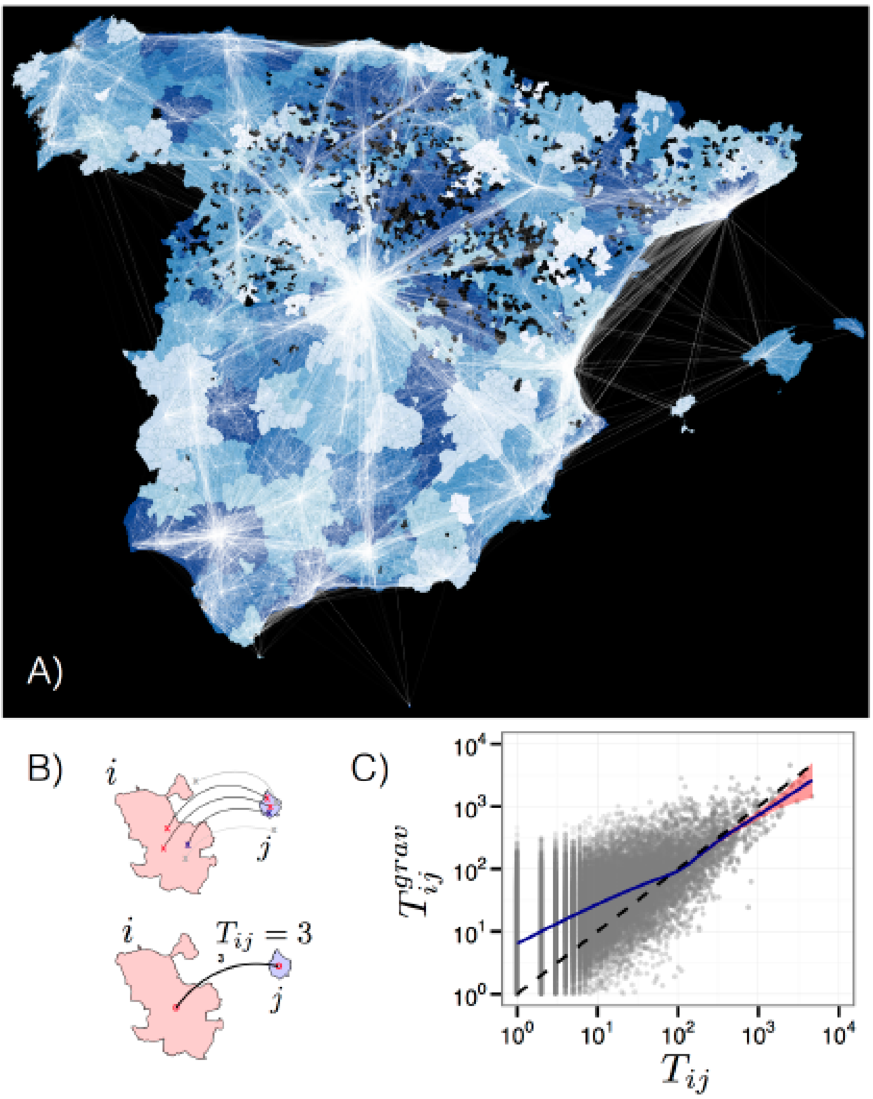


FIGURE 3.2: A) Map of the mobility fluxes T_{ij} between municipalities based on Twitter inferred trips (white). Infomap communities detected on the network T_{ij} are colored under the mobility fluxes (blue colors). B) Mobility fluxes T_{ij} between municipalities i and j are constructed by aggregating the number of trips between them. C) Correspondence between the observed fluxes T_{ij} and the fitted gravity model fluxes. Dashed line is the $T_{ij} = T_{ij}^{\mathrm{grav}}$ while the (blue) solid line is an conditional average of T_{ij}^{grav} for values of T_{ij}. Maps were created using the `maptools` and `sp` packages in the `R` environment.

Gravity Model		
Parameter	Description	Spain
α_1	Origin exponent	0.477***(0.002)
α_2	Destination exponent	0.478***(0.002)
β	Distance exponent	1.05***(0.0035)
R^2	Goodness of fit	0.797
ϕ	Correlation between T_{ij} and T_{ij}^{gra}	0.826

TABLE 3.1: Description of the parameters for the Gravity Law Model in geo-tagged social media data for Spain. $(***)$ means significance $p < 0.0001$.

3.2.4 Heterogeneity among the municipalities

Despite this high level of social media activity within municipalities, we find their official administrative areas not suitable to study socio-economical activity: administrative boundaries between municipalities reflect political and historical decisions, while economical trade and activity often happens across those boundaries. The result is that municipalities in Spain are artificially diverse, ranging from a municipality with only 7 inhabitants to other with population 3.2 million.

Although there exists natural aggregations of municipalities in provinces (regions) or statistical/metropolitan areas (for instance, NUTS areas [99]), we have used our own procedure to detect economical areas. In particular, we have used user daily trips between pairs of municipalities as a measure of the economic relatedness between said municipalities.

We use the network of daily fluxes between municipalities T_{ij} to detect the geographical communities of economical activity. To this end we employ standard partition techniques of the mobility network T_{ij} using graph community finding algorithms. This technique has been applied extensively, specially with mobile phone data, to unveil the effective maps of countries based on mobility and/or social interactions of people[76, 77, 100].

Typically, complex networks exhibit community structure, that is, there are subsets of nodes that are more densely connected among them comparing to the rest of the nodes. In mobility networks, whose nodes correspond to geographical areas, these communities are interpreted as zones with high common activity and tend to be constrained by geographical and political barriers. We check whether this is also observed in our dataset by performing 6 state-of-art community detection algorithms: FastGreedy [101], Walktrap [102], Infomap [94], MultiLevel [93], Label Propagation [103] and Leading Eigenvector [104]. These six different algorithms exhibit different community structures in terms of number of communities, average size of community or modularity (see tables 3.2 and 3.3).

Communities Size Stats			
Algorithm	$< \|N_i\| >_i$	$\max\{\|N_i\|\}$	$\|\{N_i\}\|$
Fastgreedy	309.696	1385	23
Walktrap	9.262	433	769
Infomap	21.011	143	339
Multilevel	323.772	1132	22
Label Propagation	22.052	750	323
Leading Eigenvector	1017.571	5344	7

TABLE 3.2: Size statistics of the communities $\{N_i\}$ returned by the six algorithms.

Communities Performance Stats			
Algorithm	Modularity	*NMI P*	*NMI C*
Fastgreedy	0.726	0.712	0.590
Walktrap	0.417	0.744	0.757
Infomap	0.758	0.770	0.831
Multilevel	0.800	0.717	0.599
Label Propagation	0.732	0.749	0.761
Leading Eigenvector	0.381	0.264	0.205

TABLE 3.3: Performance statistics of the communities $\{N_i\}$ returned by the six algorithms. *NMI P* refers to the comparison between communities and provinces whereas *NMI C* considers counties instead of provinces.

Members (municipalities) of the resulting communities are spatially connected except some few cases as figure 3.3 shows.

We test the statistical robustness of the obtained communities by randomly removing a proportion p of the original links and performing the algorithms on this new graph G_p. We will consider that communities are robust when the communities given for the original network G and G_p are highly similar. In order to compare two arbitrary memberships to communities, we use the Normalized Mutual Information (NMI) method described in [?] which returns 0 when two memberships are totally different and 1 when we compare two equal memberships. We compute the NMI for each chosen algorithm performed on G and G_p, for p between 1% and 10%, concluding that obtained community structures are robust because they are not broken when some randomly chosen links are removed (see table 3.4).

As other works have shown, mobility graph communities are usually interpreted in terms of geographical and political barriers and a natural question is whether the mobility based communities are related to any of these barriers. In Spain, there are different territorial divisions for administration purposes. In this work, we consider two of them: provinces, defined in 1978 Constitution, are 48 different heterogeneous aggregations of municipalities; and counties (*comarca* in Spanish terminology) which are traditional aggregations of municipalities mainly based on Spanish holography (rivers, valleys, ridges,

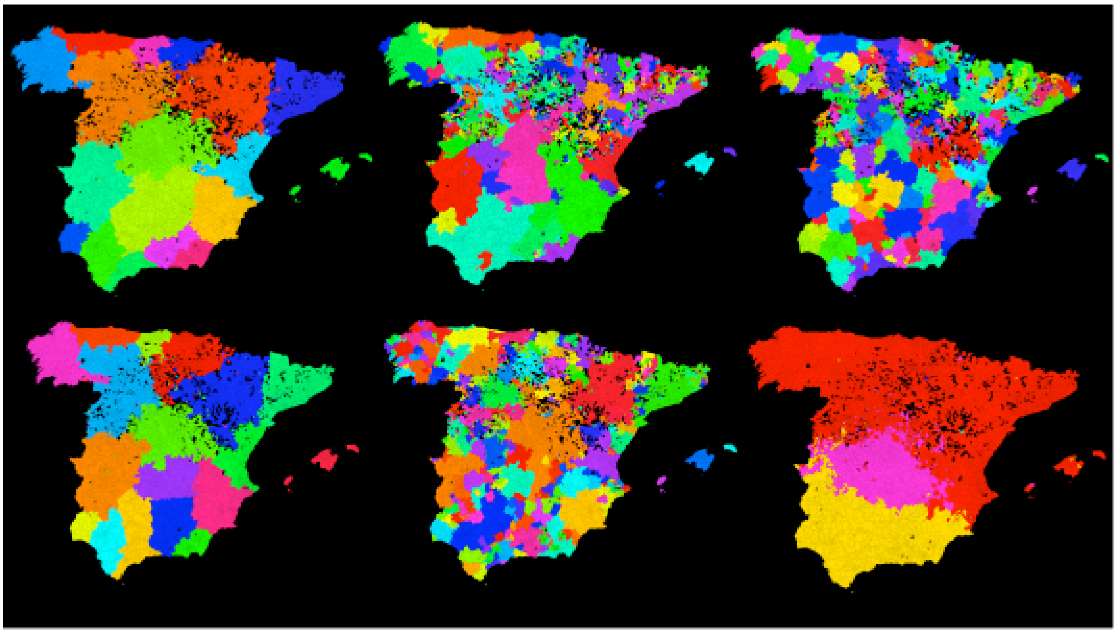


FIGURE 3.3: From left to right and from top to bottom: Fastgreedy, Walktrap, Infomap, Multilevel, Label Propagation and Leading Eigenvector communities on Twitter based mobility transitions.

NMI between G and G_p for different p			
Algorithm	$p = 0.01$	0.05	0.1
FG	0.995	0.981	0.959
WT	0.954	0.945	0.931
IM	0.988	0.978	0.966
ML	0.994	0.948	0.947
LP	0.906	0.895	0.904
LE	0.960	0.910	0.884

TABLE 3.4: NMI measure comparing G and G_p.

etc) and some of them are composed by municipalities of different provinces. We use again the NMI method to compare the communities structure given by the algorithms to the administrative limits. Except Leading Eigenvector algorithm, the rest of methods return communities that are quite related to provinces ($NMI \approx 0.7$) whereas for the county administration limits, higher variability is observed. In this last case, the algorithm providing more relationship with county limits is Infomap, $NMI \approx 0.83$. Therefore, Twitter based mobility summarizes the inter-city flows exhibiting that these flows are influenced by geographical and political barriers.

Because all of these results, we have chosen the Infomap algorithm [94] as the one giving the best partition for our problem. We have found 340 different communities within Spain. The average number of municipalities per community is 21, and the largest community contains 142 municipalities. Besides all the previously analyzed features, resulting communities conserve very interesting properties: (i) they are cohesive geographically (see Fig. 3.2), (ii) they are statistically robust against randomly removal of trips in our database (iii) modularity of the partition is very high (iv) it shows a large overlap (83% of NMI) with counties. This result shows that the mobility detected from geo-located tweets and the communities obtained are a good description of economical areas. Finally, we also mitigate the effect of analyzing very heterogeneous regions leading to a wide spectra of values in the penetration rate (see figure 3.4).

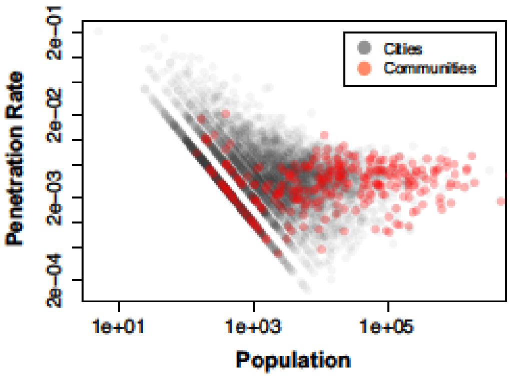


FIGURE 3.4: Penetration rates for both cities and detected communities.

In the rest of the chapter, we restrict our analysis to the geographical areas defined by the Infomap detected communities (see Fig. 3.2). For statistical reasons, we discard communities which are not formed by at least 5 municipalities. Despite this sampling, 96% of the total country population is considered in our analysis.

3.2.5 Detecting hometown and users' communities

The metrics we will explain in next sections rely on the detection of users's hometown and the community they belong to as a consequence. Instead of using information in the user profile, we analyze the places where the user has tweeted and we set as *hometown* of the user the municipality where he/she has tweeted with the highest frequency, a method usually employed in mobile phone and social media [66, 105]. To this end we select those users with more than 5 geo-located tweets in our period and which have tweeted at least 40% of their tweets in a given municipality, which we will consider their hometown. After this filtering we end up with 0.32 million users and we can then define the twitter population π_i in area i as the number of users with their hometown within area i. We obtain a very high correlation between π_i and population of the cities P_i in the national census $\rho = 0.977[0.976, 0.978]$ which provides an indirect validation of our approach with the present data.

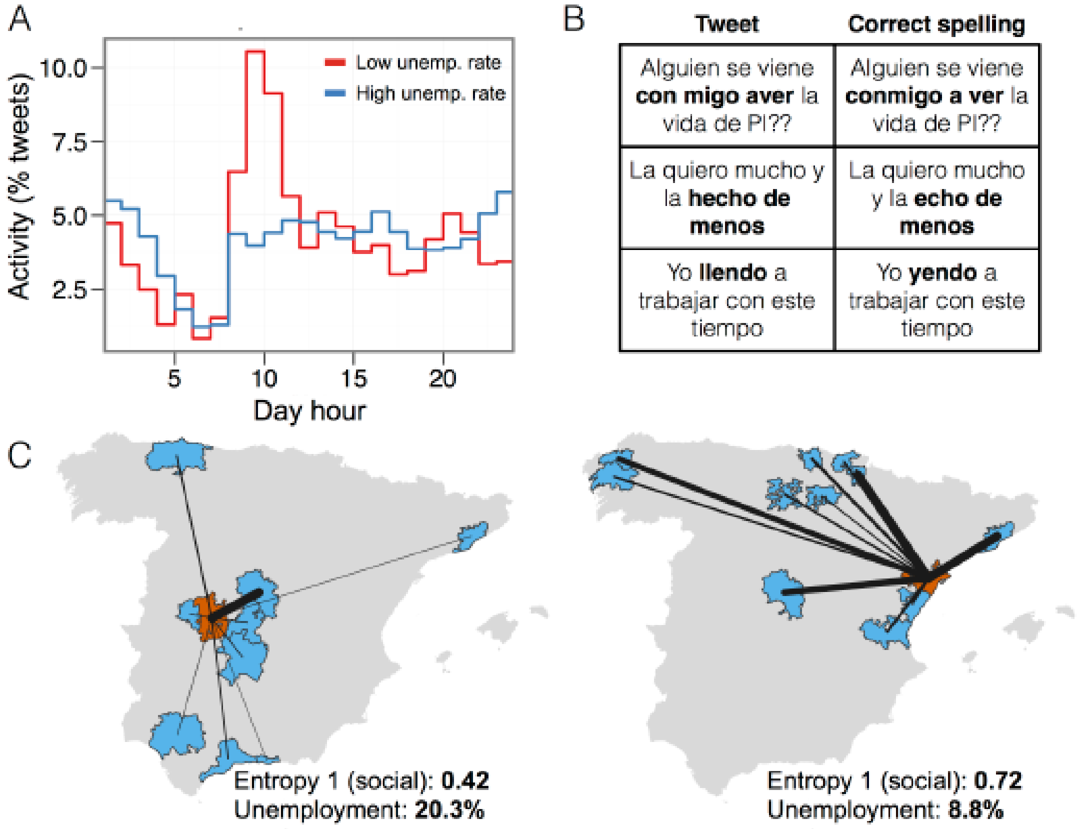
3.2.6 Social media behavioral variables

The main purpose of this chapter is to show a methodology to quantify how and what behavioral features can be extracted from social media and then related back to the to the economical level of geographical regions. To this end, we define four groups of measures that have been widely explored in other fields like economy or social sciences. Some of those metrics are already reported in the literature, but some others are introduced in this work. Specifically we consider:

- *Social media technology adoption*: we can use twitter penetration rate $\tau_i = \pi_i/P_i$ in each area i as a proxy of technology adoption. Recent works have shown that indeed there is a correlation between country GDP and twitter penetration: specifically, it was found that a positive correlation between τ_i and GDP at the country level [66].

- *Social media activity*: regions with very different economical situations should exhibit different patterns of activity during the day. Since working, leisure, family, shopping, etc. activities happen at different times of the day, we might observe different daily patterns in regions with different socio-economical status. After observing the data visually, we hypothesize that communities with low levels of unemployment will tend to have higher activity levels at the beginning of a typical weekday. This is indeed what we find: Fig. 3.5A shows the hourly fraction of tweets during workdays of two communities with very different rate of unemployment. As we can observe, both profiles are quite different and, in the case of low unemployment, we find a strong peak of activity between 8 and 11am (morning), and lower periods of activity during the afternoons and nights. We encode this finding in $\nu_{\mathrm{mrng},i}$, $\nu_{\mathrm{aftn},i}$, and $\nu_{\mathrm{ngt},i}$ the total fraction of tweets happening in geographical area i between 8am and 10am, 3pm and 5pm, and 12am and 3am respectively.

- *Social media content*: some works have observed a correlation between the frequency of words related to work conditions [81] or Google searches [80] to unemployment or economical situation of countries. In our case we also find that there is a moderate positive correlation between the fraction of tweets μ_i mentioning *job* or *unemployment* terms and the observed unemployment, while the correlation is negative for the number mentions to *employment* or the *economy*. However, we have tried a different approach by measuring the relation between the way of writing and the educational level [106]. To this end, we build a list of 618 misspelled Spanish expressions and extract the tweets of the dataset containing at least one of these words. Then, in order to decide whether a tweet has a misspelling or not,



Tweet	Correct spelling
Alguien se viene **con migo aver** la vida de PI??	Alguien se viene **conmigo a ver** la vida de PI??
La quiero mucho y la **hecho de menos**	La quiero mucho y la **echo de menos**
Yo **llendo** a trabajar con este tiempo	Yo **yendo** a trabajar con este tiempo

FIGURE 3.5: Examples of different behaviour in the observed variables and the unemployment. In A, we observe that two cities with different unemployment levels have different temporal activity patterns. Panel C show how communities (red) with distinct entropy levels of social communication with other communities (blue) may hold different unemployment intensity: left map shows a highly focused communication pattern (low entropy) while right map correspond to a community with a diverse communication pattern (high entropy). Finally, Panel B shows some examples of detected misspellings in our database using 618 incorrect expressions such as "Con migo", "Aver" or "llendo".

we need to establish some patterns to select from our set of tweets. Since we want to be sure that a detected mistake corresponds to a real misspeller, we will not consider the following cases:

- Lack of written accents. People tend to avoid writing accents when talking in a colloquial way.
- Mistakes derived from removing *unnecessary* letters. The most common cases are removing a h at the beginning of a word (in Spanish the letter h is not pronounced), or replacing the letters qu by k. We understand that these mistakes can be motivated for the limitation of length in tweets, and not for a real misspelling.
- In the same line, we neglect mistakes produced by removing letters in the middle of a word, whose pronunciation can be deduced without them.

- We do not consider either mistakes related to features of specific areas in Spain. For example, in the south the pronunciation of *ce* and *se* is the same, what produces a big amount of mistakes when writing. However, since we want to extract objective and equitable conclusion over the whole Spanish geography, we neglect those misspellings that only appear in a specific area.

Likewise, we will consider as real misspellings the following mistakes:

- Adding letters. For example, writing a *h* at the beginning of a word that starts with a vowel.
- Changing the special cases *mp*, *mb* by the wrong writings *np*, *nb*.
- Mixing up *b* with *v*, *g* with *j*, *ll* with *y*, and *ex* with *es*. These are typical mistakes in Spanish, because they have the same, or a very close, pronunciation.
- Confusing the verb *haber* with the periphrasis *a ver*.
- Separating a word into two ones, for instance, writing the word *conmigo* as *con migo*.

All of these misspellings cannot be attributed to the special features of Twitter or a specific region of Spain. Finally, since in this country several languages live at the same time, depending on the part of the country, our Twitter dataset is reduced to those tweets written in Spanish. This task is carried out using the N-gram based text categorization R library *textcat* [107]. Thus, one can expect that this selection provides an accurate and equitable method of detecting misspellers. Under these conditions, the number of users who wrote at least one misspelled word is 5.6% over the whole population (more than 27000 users).

We analyze whether misspellers have different Twitter usage behavior from that people who do not make serious mistakes when publishing a tweet. Comparing the average number of tweets, it can be observed that misspellers tend to publish a larger number of tweets than those who did not made mistakes (144.71 against 23.72). This also emerges when the mean number of misspelling given the total number of tweets is considered. For users with less than approximately 30 published tweets in the observation period, the number of misspellings is almost zero whereas for users who publish more often, the mean number of misspellings scales sub-linearly with the number of tweets ($exponent \approx 0.33$). Finally, we denote ϵ_i as the misspeller rate in population i.

- *Social media interactions and geographical flow diversity*: following the ideas in [51] which correlated the economical development of an area with the diversity of communications with other areas, we consider all tweets mentioning another user and take them as a proxy for communication between users. Then we compute the number of communications w_{ij} between areas i and j as the number of mentions between users in those areas. To measure the diversity we use the normalized informational entropy (Entropy 1, social version)

$$S_{u,i} = -\frac{1}{S_{r,i}} \sum_j p_{ij} \log p_{ij} \tag{3.9}$$

where $p_{ij} = w_{ij} / \sum_j w_{ij}$ and (Entropy 2, social version)

$$S_{r,i} = \log k_i \tag{3.10}$$

with k_i the number of different areas with which users in area i have interacted. Similarly, we also consider the number of people between areas to investigate the diversity of the geographical flows through the entropy, that is, (Entropy 1, geo version)

$$\tilde{S_{u,i}} = -\frac{1}{\tilde{S_{r,i}}} \sum_j \tilde{p_{ij}} \log \tilde{p_{ij}} \tag{3.11}$$

where $\tilde{p_{ij}} = T_{ij} / \sum_j T_{ij}$ and (Entropy 2, geo version)

$$\tilde{S_{r,i}} = \log \tilde{k_i} \tag{3.12}$$

with $\tilde{k_i}$ the number of different places visited by people living in area i.

3.2.7 Properties and correlation between Twitter behavioral variables

Heterogeneity between the values of variables constructed from Twitter is large but moderate, as histograms in figure 3.6 show. We did not find any geographical area with anomalous values in any of the variables considered. Variables are normalized in different ways: both the penetration τ_i and misspellers rate ε_i are defined as the number of users or misspellers per 100.000 persons (population); activity variables ν_i are normalized as the percentage of tweets per time interval; finally, number of tweets that mention a specific term μ_i are also given per 100.000 tweets published in the geographical area.

Variables are constructed to reflect the behavior of areas in the different dimensions of Twitter penetration, social or geographical diversity, activity through the day and content. Correlation between variables does indeed show that variables within each dimensions hold strong correlations between them. As we can see in figure 3.7 social and geographical diversities are highly correlated between them, an expected fact given the *gravity law* accurate description of flows of people between geographical areas, but also the amount of communication between them. Same behavior is found for the group of variables in the activity group, while content variables are less correlated. Finally

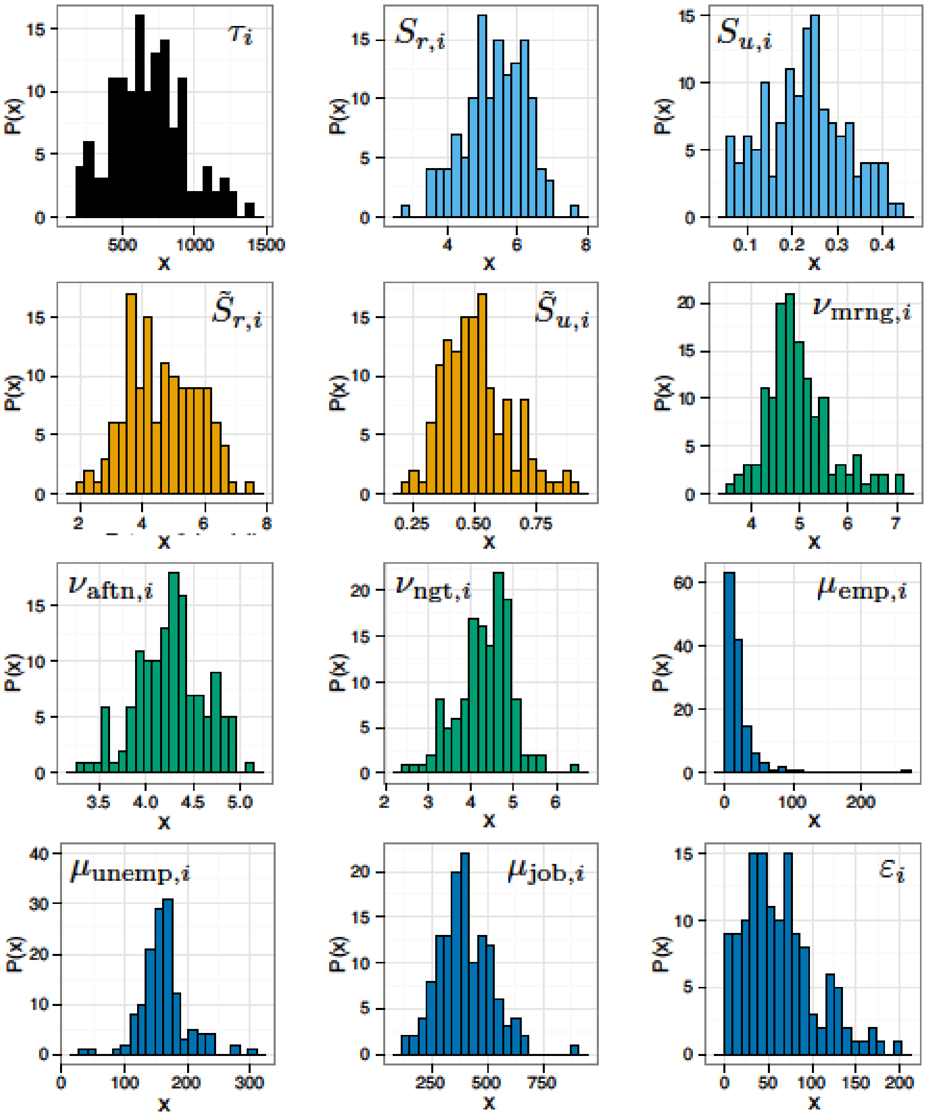


FIGURE 3.6: Frequency plots for each variable constructed from Twitter.

we find that both the penetration rate τ_i and fraction of misspellers ε_i have a strong correlation with most of the variables.

High correlation between variables might lead to collinearity effects [108] in the regression models to predict unemployment, that is, some variables with predictive variable might have non-significant weights because they explain the same part of the variance. To check this hypothesis, we perform a principal component analysis (PCA) on the independent set of variables we defined in the previous subsection. Figure 3.7 exhibits the loadings of the different variables for the considered variables. The block structure showed in 3.7 results in similar directions of the variables in the first components of the PCA. We observe some groups of variables: on the one hand, geographical and social diversity seem to explain large part of the variance; on the other hand, we find a perpendicular group of variables formed by temporal activity; finally, penetration rate and misspellers fraction seem to represent a different independent direction of data, with high collinearity between them. The structure of the correlation matrix and the PCA results show that there is indeed information in all groups of variables and thus we have take a variable in each of them for our regression models.

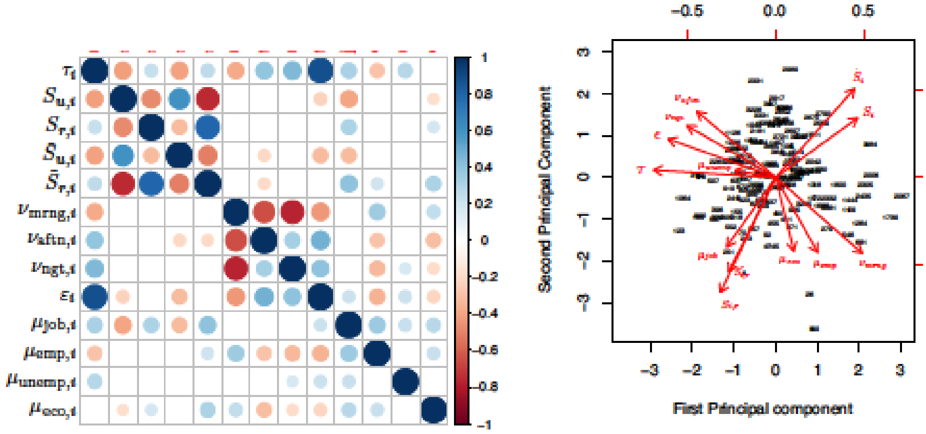


FIGURE 3.7: Left: Correlation matrix between the variables constructed from Twitter. Each entry in the matrix is depicted as a circle whose size is proportional to the correlation between variables and the sign is blue/red for positive/negative correlations. Blank entries correspond to statistically insignificant correlations with %95 confidence. Right: Variables projection on the first two principal components given by PCA. We observe different groups of variables and collinearity between some of them.

Provided these conclusions about the nature of the variables we have built, we restrict our analysis to the variables within each group with the highest correlation with the unemployment, namely the penetration rate τ_i, the social and mobility diversity variables $S_{u,i}$ and $\tilde{S}_{u,i}$, the morning activity $\nu_{\mathrm{mrng},i}$, the fraction of misspellers ε_i and fraction of *employment*-related tweets $\mu_{emp,i}$.

3.2.8 Correlation between defined variables vs unemployment and explanatory power

After the PCA study, we have selected of subset of variables containing relevant and no complemmentary information among them. However, until now, we have not dealt with the main goal of our work which is building a predictive model of unemployment. In this section, we analyze to what extent the created variables correlate with our target varaible and, eventually, are useful to predict unemployment.

Recent works have shown that there exists a correlation between country GDP and Twitter penetration rate.: specifically, it was found that a positive correlation between τ_i and GDP at the country level [66]. However, in our data we find the opposite correlation (see Fig. 3.8), namely, that the larger the penetration rate the bigger the unemployment

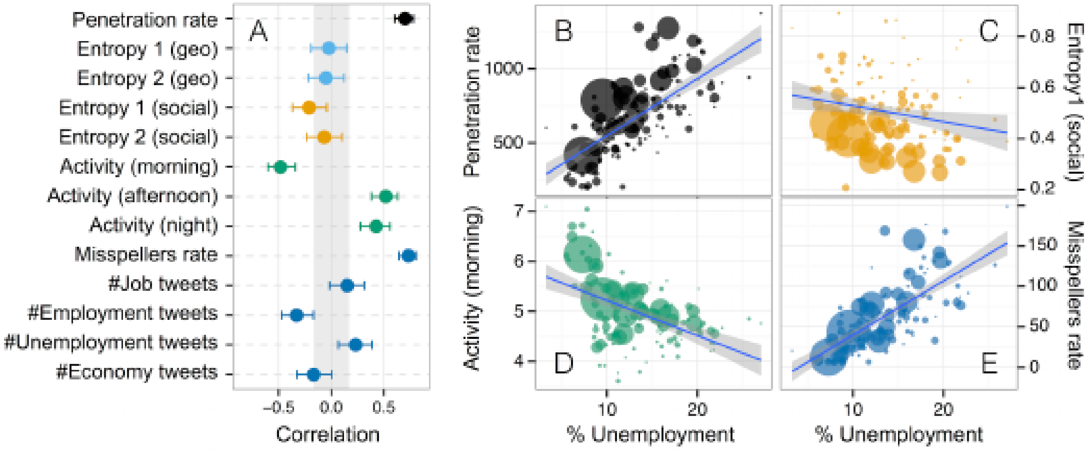


FIGURE 3.8: A) Correlation coefficient of all the extracted Twitter metrics grouped by technology adoption (black) geographical diversity (orange), social diversity (light blue), temporal activity (green) and content analysis (dark blue). Error bars correspond to 95% confidence intervals of the correlation coefficient. Gray area correspond the statistical significance thresholds. Panels B, C, D and E show the values of 4 selected variables in each geographical community against its percentage of unemployment. Size of the points is proportional to the population in each geographical community. Solid lines correspond to linear fits to the data.

is, which suggest that the impact of technology adoption at country scale is different of what happens within an (industrialized) country where technology to access social media is commoditized. Focusing on temporal patterns in regions, given by the fraction of tweets at different hour ranges of the day, Fig. 3.8 shows a strong negative correlation between $\nu_{\mathrm{mrng},i}$ and the unemployment for the communities in our database and positive correlation with $\nu_{\mathrm{aftn},i}$, and $\nu_{\mathrm{ngt},i}$, indicating that morning daily patterns are different depending on the unemployment level of the regions. Also misspellers rate ϵ_i exhibits a positive correlation with unemployment suggesting that in those regions with higher unemployment levels, a larger proportion of people misspelling on Twitter might be found. Finally, as in [51], we find that areas with large unemployment have less diverse communication patterns than areas with low unemployment. This translates in a moderate negative correlation between S_i and the unemployment, see Fig. 3.8. Similar ideas are applied to the flows of people between areas to investigate the diversity of the geographical flows through the entropy $\tilde{S}_i = -\sum_j \tilde{p}_{ij} \log \tilde{p}_{ij}/\tilde{S}_{r,i}$, where $\tilde{p}_{ij} = T_{ij}/\sum_j T_{ij}$ and $S_{r,i} = \log(\tilde{k}_i)$ with $\tilde{k}_i$ the number of different areas which has been visited by users that live in area i. Fig. 3.8 shows that as in [72], correlation of these geographical entropies is low with economical development.

The four previous groups of variables are fingerprints of human behavior reflected on the Twitter usage habits. As we observed in Fig. 3.8, all of them exhibit statistically strong correlations with unemployment. The question we address in this section is whether those variables suffice to explain the observed unemployment (their explanatory power) and also determine the most important ones among themselves (which give more explanatory power than others). Note that we are not stating a causality arrow between the measures built in the previous section and the unemployment rate but only exploring whether they can be used as alternative indicators with a real translation in the economy.

Through this thesis, we use linear models to predict economic levels by not focusing on the data complexity but in the explanatory power of the individual variables. In general, if Y is our target random variable and $X_1, .., X_N$ a family of random variables to predict Y, a dataset can be interpreted as realizations of these random variables $y_1, ..., y_k$ and $x_{i1}, ..., x_{iN}, \ i = 1, ..., k$. A linear model is a specific kind of regression model given by

$$Y = \beta + \alpha_1 X_1 + ... + \alpha_N X_N + \epsilon \tag{3.13}$$

Typically the coefficients are computed by least squares optimization or any other kind of optimization process. Despite it is a very simple model, we use it because our goal is not finding the best model to predict the phenomena we are interested in but analyzing the predictive power of the variables generated in the different contexts. In fact, linear models are also very useful to study which variables are the most important ones in the prediction leading to conclusions about the intrinsic nature of the studied processes. As result of the optimization process, some weights α_i might be close to zero, indicating that the variable X_i is not important for the prediction; on the contrary, large weights in absolute value indicate a strong presence in the prediction and the sign might indicate how the phenomena is influenced by that specific variable.

Since our main goal is to produce methods to model economic levels based on data that has not been used typically for this purpose, we have to set metrics to assess the quality of the model. A very widely used one to evaluate this kind of models is the coefficient of determination, denoted as R^2, which measures how much of the variance in the original data is explained by the model with the next formula

$$R^2 = 1 - \frac{\sum(y_i - f_i)^2}{\sum(y_i - \overline{y})^2} \tag{3.14}$$

where y_i is the value to predict, f_i is the prediction provided by the model and $\overline{y}$ is the average of the target variable. In some type of predictive models, adding more variables (degrees of freedom) typically increases the performance but also the complexity and also it should be considered in our perfomance measure; to this end, we use the Adjusted R^2 described in [109].

Fig. 3.9 shows the result of a simple linear regression model for the observed unemployment for ages below 25 years as a function of the variables which have more correlation

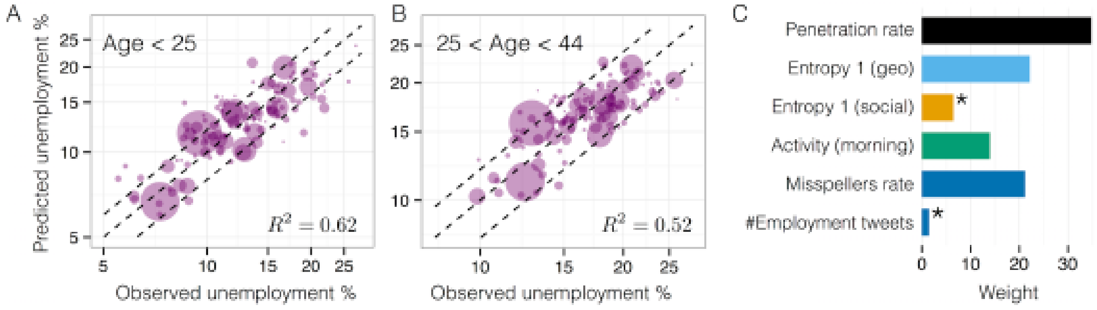


FIGURE 3.9: A) and B) Performance of the model, showing the predicted unemployment rate for ages below 25 versus the observed one, $R^2 = 0.62$ and with ages between 25 and 44. Dashed lines correspond to the equality line and ±20% error. C) Percentage of weight for each of the variables in the regression model using the relative weight of the absolute values of coefficients in the regression model (see Section J in ??). Variables marked with $*$ are not statistical significant in the model.

with the unemployment (see next section to analyze other age segments). The model has a significant $R^2 = 0.62$ showing that there is a large explanatory power of the unemployment encoded in the behavioral variables extracted from Twitter. Since we train our model with normalized data, the absolute value of the weights in the regression can be understood as their importance (predictive power) in the model and, in fact, this method has also been used as a feature selection model in the literature [110]. As one can expect, not all the variables weight equally in the model: specifically, the penetration rate, geographical diversity, morning activity and fraction of misspellers account for up to 92% of the explained variance, while social diversity and number of *employment* related tweets are not statistically significant. It is interesting to note that while social diversity obtained by mobile phone communications was a key variable in the explanation of deprivation indexes in [51, 72], the communication diversity of twitter users seem to have a minor role in the explanation of heterogeneity of unemployment in Spain.

3.2.9 Temporal, demographic and geographical variance on Twitter models

In the definition of the variables we have aggregated the Twitter activity within a 7 months time window (from December 2012 to June 2013). Since unemployment has a significant variation along time, we investigate here what is the correlation and explanatory power of the Twitter variables for the values of unemployment determined at different months through the same time window in which Twitter data was collected. Or if the variables collected in that time window are more correlated with past or future values of unemployment. Figure 3.10 shows the explanatory value of the model when the linear regression is done for values of unemployment of different months before, during and after the Twitter data time window. Although there is a small seasonal effect along

	All ages	< 24	$25 - 44$	> 44
(Intercept)	$0.11 ****$	0.10***	0.20***	0.20***
	(0.02)	(0.03)	(0.03)	(0.035)
Penetration rate	3.23*	8.57***	6.28**	2.40
	(1.41)	(2.22)	(2.17)	(2.77)
Geographical diversity	0.03	0.15***	0.08*	0.06
	(0.02)	(0.04)	(0.04)	(0.05)
Social diversity	−0.03*	−0.03	−0.05*	−0.06*
	(0.01)	(0.02)	(0.02)	(0.03)
Morning activity	−0.69*	−1.30**	−1.53***	−1.19*
	(0.26)	(0.42)	(0.41)	(0.52)
Misspellers rate	11.56	31.51*	15.46	23.60
	(8.13)	(12.78)	(12.48)	(15.94)
Employment mentions	−1.80	3.17	−9.94	2.71
	(6.27)	(9.86)	(9.64)	(12.3)
R^2	0.47	0.64	0.55	0.29
Adj. R^2	0.44	0.62	0.52	0.26

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

TABLE 3.5: Regression table for the different models in which unemployment for different age groups is fitted. The *All ages* model is the fit to the general rate of unemployment in each geographical area, while the other models are for the rates of unemployment in groups of less than 24 years, between 25 and 44 years and above 44 years.

the year, we see that the explanatory power remains around $R^2 = 0.6$, which suggest that our Twitter linear model retains its explanatory power even though unemployment changes considerably throughout the year. It is interesting to note that R^2 decays a little bit during the summer which means that our variables are less correlated with summer unemployment. There are some possible explanations for the little variation of the performance: firstly, the variation over the whole period of the global unemployment rate varies only 0.8% which does not imply little variation of the rates of every region but seems reasonable to think so. Apart from this fact, one does not expect the created variables changes dramatically in a short period of time (education, technology adoption, geographical mobility,...). Finally, unemployment used in the rest of the chapter is from June 2013, i.e. the last month in the time window used to collect the data.

On the other hand, not all demographic groups are equally represented in the our database. Twitter user demographics in Spain obtained from surveys [111] show that age groups above 44 years old are under-represented. Thus our results would mainly describe the socio-economical status of people below 44 years old. Employment analysis is then performed in different age groups: unemployment for people below 25 years old, between 25 and 44 years old and older than 44 years old. Indeed, similar but lower explanatory power is found for other age groups: $R^2 = 0.44$ for all ages and $R^2 = 0.52$

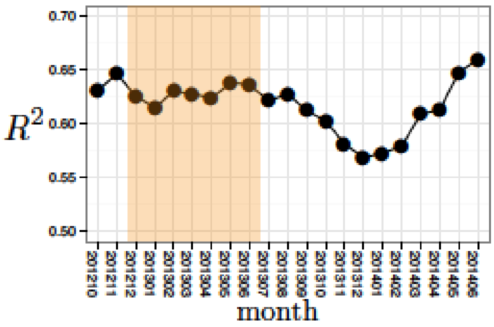


FIGURE 3.10: Explanatory power of the linear regression model when fitted against the unemployment data for different months. Gray (orange) area correspond to the time window in which Twitter data is collected and variables are constructed.

for ages between 25 and 44 years. Beyond there, the model degrades for ages above 44 years ($R^2 = 0.26$) proving that our variables mainly described the behavior of the most represented age groups in Twitter, namely those below 44 years old.

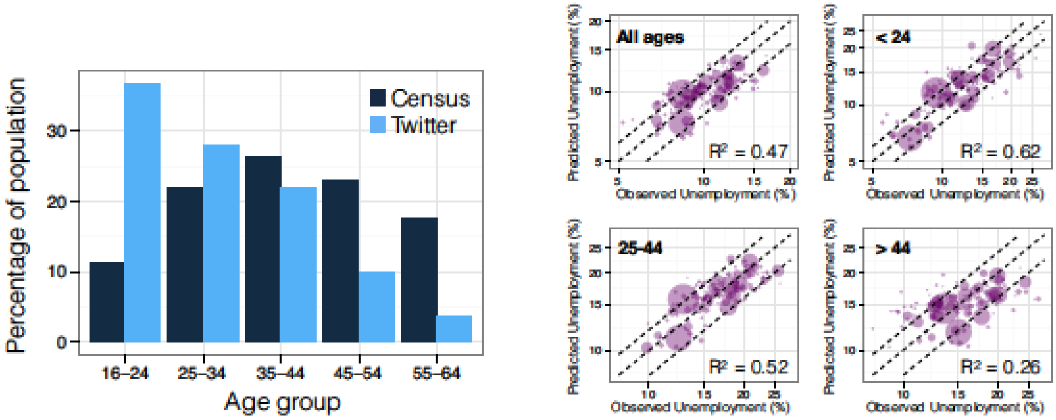


FIGURE 3.11: Left: Percentage of population in each age group from the Spanish Census (dark bars) and surveys about users in Twitter (light bars). Right: performance of the linear models for each of the age groups.

Finally, since our Twitter variables seem to describe better behavior of young people, we have investigated whether Twitter constructed variables have similar explanatory value (in terms of R^2) than simple census demographic variables for young people. If we include the young population rate in our model we get a minor improvement $R^2 = 0.65$; on the other hand a model based only on young population rate gives $R^2 = 0.24$. This semi-partial analysis shows that Twitter variables do indeed posses a genuine explanatory power away from their simple demographic representation.

Attending to geographical partitions, while municipalities are very heterogeneous demographically, other administrative areas exist in Spain at large scales that could be used for our model of unemployment. The smallest administrative division of Spain we have considered is that of the 8200 *municipalities*. At larger scales we have the 326 *counties* (*comarcas* in spanish) which are aggregations of municipalities. Finally, the

	All variables	Youth model	Twitter model (I)	Twitter model (II)
(Intercept)	0.06	−0.02	0.10***	0.09***
	(0.03)	(0.03)	(0.03)	(0.027)
Young pop. rate	0.66*	2.20***		
	(0.30)	(0.35)		
Penetration rate	8.20***		8.57***	8.62***
	(2.25)		(2.22)	(2.21)
Geographical diversity	0.14***		0.15***	0.12***
	(0.04)		(0.04)	(0.03)
Social diversity	−0.02		−0.03	
	(0.02)		(0.02)	
Morning activity	−1.42***		−1.30**	−1.28**
	(0.41)		(0.42)	(0.41)
Misspellers rate	23.95		31.51*	32.28*
	(13.09)		(12.78)	(12.71)
Employment mentions	0.34		3.17	
	(9.81)		(9.86)	
R^2	0.65	0.24	0.64	0.63
Adj. R^2	0.63	0.24	0.62	0.62

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

TABLE 3.6: Regression table for the different statistical models. The *All variables* model includes both Twitter and rate of young population variables. *Twitter model (I)* includes only the variables described in the main article, while *Twitter model (II)* only includes those variables which are significant $p < 0.05$ in *Twitter model (I)*.

largest geographical scale we considered is defined by 50 provinces (*provincias* in Spanish). In this case, we compare the performance of our Twitter model for unemployment for the variables defined in those administrative areas and relate it to the geographical communities detected and used in the main paper (see section ??). Not all the areas at different administrative divisions are considered in the model. To minimize the effect of areas in which the number of geo-tagged tweets is very small, we only consider the 1738 municipalities which have a Twitter population $\pi > 10$. Similarly, we only consider the 198 counties with $\pi > 100$. As we can see in Table 3.7 the model has a large explanatory power for areas equal or bigger than counties. As expected R^2 increases as the number of areas in the model is smaller, but the description level of the model is very low for provinces, for example. The best performance (high R^2 and high geographical description level) is attained at the level of the detected communities.

3.3 Conclusions and further discussion

This work serves as a proof of concept for how a wide range of behavioral features linked to socioeconomic behavior can be inferred from the digital traces that are left

	Communities	Municipalities	Counties	Provinces
(Intercept)	0.10^{***}	0.16^{***}	0.11^{***}	0.11^{*}
	(0.03)	(0.01)	(0.03)	(0.05)
Penetration rate	8.57^{***}	4.01^{***}	9.12^{***}	10.47^{***}
	(2.22)	(0.59)	(1.81)	(1.97)
Geographical diversity	0.15^{***}	0.02	0.12^{***}	0.08
	(0.04)	(0.01)	(0.03)	(0.07)
Social diversity	-0.03	-0.01	-0.01	-0.03
	(0.02)	(0.01)	(0.02)	0.07
Morning activity	-1.30^{**}	-1.16^{***}	-1.49^{***}	-1.03
	(0.42)	(0.14)	(0.39)	(0.88)
Misspellers rate	31.51^{*}	14.40^{***}	14.09	
	(12.78)	(2.51)	(10.02)	
Employment mentions	3.17	-0.71	2.41	-3.17
	(9.86)	(0.89)	(8.86)	(12.29)
Number of points	128	1738	198	50
R^2	0.64	0.22	0.55	0.65
Adj. R^2	0.62	0.21	0.54	0.61

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

TABLE 3.7: Regression table for the unemployment linear regression model in different levels of geographical areas. In the *Provinces* model, the misspellers rate has been removed from the model due to the large collinearity with the penetration rate.

by publicly-available social media. In particular, we demonstrate that behavioral features related to unemployment can be recovered from the digital exhaust left by the microblogging network Twitter. First of all, Twitter geolocalized traces, together with off-the-shelve community detection algorithms, render an optimal partition of a country for economical activity, showing the remarkable power of social media to understand and unveil economical behavior at a country-scale. This insight is likely to apply to other administrative definitions in other countries, specially when considering large cities with an inherent dynamical nature and evolution of mobility fluxes, and cities composed of small satellite cities with arbitrary agglomerations or division among them (e.g. London, NYC, Singapore). This result is unsurprising: it should be natural to recompute city clusters/communities of activity based on their real time mobility, which may vary considerably faster than the update rates of mobility and travel surveys [76, 77, 100].

Our main result demonstrates that several key indicators, different penetration rates among regions, fingerprints of the temporal patterns of activity, content lexical correctness and geo-social connectivities among regions, can be extracted from social media, and then used to infer unemployment levels. These findings shed light in two directions: first, on how individuals' extensive use of their social channels allow us to characterize cities based on their activity in a meaningful fashion and, secondly, on how this information can be used to build economic indicators that are directly related to the economy.

Regarding the latter, our work is important for understanding how country-scale analysis of Social Media should consider the demographic but also the economical difference between users. As we have shown, users in areas with large unemployment have different mobility, different social interactions, and different daily activity than those in low unemployment areas. This intertwined relationship between user behavior and employment should be considered not only in economical analysis derived from social media, but also in other applications like marketing, communication, social mobilization, etc.

It is particularly remarkable that Twitter data can provide these accurate results. Twitter is, among the many currently popular social networking platforms, perhaps the noisiest, sparsest, more 'sabotaged' medium: very few users send out messages at a regular rate, most of the users do not have geolocated information, the social relationships (followers/followers) contains a lot of unused/unimportant links, it is plagued by spambots, and last but not least, we have no way to identify the motive/goal/functionality of mobility fluxes we are able to extract. These limitations are not particular to our sample, but general to the sample Twitter data being employed in the computational social science community. Despite all these caveats, we are able to show that even some simple filtering techniques together with basic statistical regressions yields predictive power about a variable as important as unemployment. Other social media platforms such as Facebook, Google+, Sina Weibo, Instagram, Orkut, or Flicker with more granular and consistent individual data are likely to provide similar or better results by themselves, or in combination. Further improvements can be obtained by the use of more sophisticated statistical machine learning techniques, some of them even tailored to the peculiarities of social media data. Our work serves to illustrate the tremendous potential of these new digital datasets to improve the understanding of society's functioning at the finer scales of granularity.

The usefulness of our approach must be considered against the cost and update rate of performing detailed surveys of mobility, social structure, and economic performance. Our database is publicly articulated, which means that our analysis could be replicated easily in other countries, other time periods and with different scopes. Naturally, survey results provide more accurate results, but they also consume considerably higher financial and human resources, employing hundreds of people and taking months, even years to complete and be released — they are so costly that countries going through economic recession have considered discontinuing them, or altering their update rate in recent times. A particularly problematic aspect of these surveys is that they are "out-of-sync" i.e. census may be up to date, whereas those same individuals' travel surveys may not be, and therefore drawing inferences between both may be particularly difficult. This is a particularly challenging problem that the immediateness of social media can help ameliorate.

A few questions remain open for further investigation. How can traditional surveys and social media digital traces be best combined to maximize their predictive ability? Can social media provide a reliable leading indicator to unemployment, and in general, economic surveys? How much reliable lead is it possible, if at all? As we have found, Twitter penetration and educational levels are found to be correlated with unemployment, but this levels are unlikely to change rapidly to describe or anticipate changes in the economy or unemployment. However, other indicators like daily activity, social interactions and geographical mobility are more connected with our daily activity and perhaps they have more predicting power to show and/or anticipate sudden changes in employment. The relationship between unemployment and individual and group behavior may help contextualize the multiple factors affecting the socioeconomic well-being of a region: while penetration, content, daily activity and mobility diversity seem to be highly correlated to unemployment in Spain, different weights for each group of traces might be expected in other countries [51]. Finally, digital traces could serve as an alternative (some times the only one available) to the lack of surveys in poor or remote areas [74, 112]. Another interesting avenue of research involves the use of social media to detect mismatches between the real (hidden, underground) economy and the officially reported [113].

Most importantly, the immediacy of social media may also allow governments to better measure and understand the effect of policies, social changes, natural or man-made disasters in the economical status of cities in almost real-time [14, 114]. Our results can also be framed within the emerging phenomenon of the shift of the digital divide [115], which shows that the current gap in online presence in developed countries is not due to digital access, but to the socio-economical situation of individuals. These new avenues for research provide great opportunities at the intersection of the economic, social, and computational sciences that originate from these new widespread inexpensive datasets.

3.4 Publications, media coverage and acknowledgment

During the development of this manuscript, the analysis described in this Chapter was published in a scientific article named as ”Social Media Fingerprint of Unemployment” in 2015. [116].

This work also gained a remarkable impact on Spanish generalist media such as El Pais, Expansion, El Confidencial, Europa Press or La Voz de Galicia. [117–121] but also on international media like Clarin or Forbes [122, 123] and online blogs and specialized forums [124–128].

The author wants to thank their effort to all the collaborators who made possible this work.

Chapter 4

Information diffusion on dynamic geographical networks

The study of how the properties of cities scale with population has attracted much interest in scientific literature. Nowadays, we know that larger cities tend to produce in a more efficient ways that smaller and consuming smaller quantities of resources living in cities provide opportunities for the individual development in terms of economy and opportunities [10, 129, 130]. However, this problem has been mainly addressed from the static point of view, focusing on aggregated networks at the geographical level and neglecting the effect of dynamical aspects. For instance, are networks in larger cities more dynamical than in smaller ones? If this is the case, how does this vibrant dynamics affecet the network at the country level? In this chapter we show that the pace of creation and destruction of links in cities scales superlinearly with the population, concluding that larger cities tend to be more dynamical than smaller ones. Moreover, we analyze the stability of links as a function of the geographical distance between individuals showing that, the further people are, the more unstable the relationships are so it is reasonable to think of the global network as a very varying one. Despite this intuition, we show that intra-urban network is robust on time, exhibiting similar adjacency matrices on time, community structure or clustering. In this context where many dynamical and varying networks at the city level are connected by a almost permanent intra-urban structure, how does it influence information diffusion? To this end we simulate diffusion processes and predict the time to infection in regions. Our results suggest that this process is not influenced by the fast intra-urban mechanics but only by the inter-urban network, concluding that the main factor for information expansion is the static structure of the global network.

4.1 Diffusion processes on geographic social networks

4.1.1 Linking network dynamics and geographical diffusion

Many researchers have dealt with the problem of modelling epidemic and rumor diffusion through social or mobility networks, but traditionally, they have made it by considering static networks, that is, every time a relationship is set between two nodes (individuals in the social network case and locations in the mobility network one), this relationship does not evolve on time so it cannot be broken and it is considered that information can be transmitted at any time [131].

However, temporal dynamics within networks is a complex phenomena: interactions (for instance, calls between two people) happen in bursts and links are continuosly being created and destroyed on time [16, 132–134] which must affect dramatically how information spreads in a network or how long viruses need to take to infect a specific population. Actually, it has been proved that temporal dynamics slow down the diffusion processes on networks and, more specifically, how an individual is able to change its ego-network also affects the time it is infected [15, 62, 135, 136].

Moreover, most of the analyzed networks happens in a particular geography where nodes composing the network can be placed: for instance, when we analyze a Call Detail Record, as in this chapter, we place the individual in its billing address or the antennas its mobile phone was connected to when it was involved in a call [6, 51, 69]; in other cases, such as airline networks, one can locate the user at the origin and the destination of the flight, depending on the time that is being observed [36, 137]. Therefore, it makes sense to study how diffusion spreads at the geographical level even when the underlying network is a social network, when different aggregation levels can be analyzed to study intra and inter-regional diffusion properties [37, 138, 139]. However, how does the internal dynamics of urban areas affect the inter-regional spreading of a virus or a rumor?

This is problem we investigate in this chapter, we study whether the vibrant dynamics inside urban areas are able to predict the geographic expansion of a diffusion process. We observe that populations have heterogeneous patterns, that is, larger populations tend to create and destroy internal links at a faster pace than smaller areas in a super-scaling fashion. Since different dynamics are exhibited in urban areas, one might guess that these differences are able to predict the infection time of these regions. Surprisingly, we observe that, despite this varying underlying network (almost 20% of individiual links change in our dataset on time), the inter-regional structure preserve the same static properties. As a consequence, diffusion at the geographical level does not really depend on the internal dynamics of cities. To our understanding, this is the first time when diffusion at the macroscopic scale has been studied as a function of the internal dynamics of the urban areas.

4.1.2 Spreadings on real-world networks

Given a social network, a diffusion (or spreading) process is a temporal process where, at the beginning, a subset of nodes are initially activated (for instance, infected by a virus, informed about a rumor or reached by a marketing campaign) and, as time goes by, this activation is spread through the rest of nodes in the network. Typically, the unactivated nodes might become activated by external factors o because of viral dissemination, that is, the probability of becoming activated at time t depends on the number of activated contacts at time $t-1$. As an example, let's say an user of an online social network is reached by a piece of news and he finds it so interesting so he decides to share it to their followers. In this case, the activation phenomena in the diffusion process is sharing the content. Ideally, all his followers will read the shared content and, depending on how interesting they find it, they might also decide to share it, becoming new infected users. Since real social networks are clustered, an user might be exposed to the same content many times, increasing the possibilities to share it. This is an example of information diffusion on social networks, which is one of the most studied problems, not only on online social networks such as Twitter but also on Wikipedia, blogs or post networks [140–142].

Some diffusion processes on social networks are not completely naturally driven, that is, there exist some incentives for the users to spread the information. This is the case of viral marketing campaigns, where users are told to share a particular content in exchange of some prize. It has been shown that information on this kind of campaigns travels at a very slow pace, affected by the bursty behavior of human online patterns [46, 143] with different rates of spreading when users in sharing cascades appear in deeper generations [144]. This kind of viral actions has also been used in social mobilization processes such as in the DARPA challenge, where the purpose was to locate 10 balloons spread over the United States and the winners built a social strategy to diffuse the objective and incentivizing social sharing [114, 145].

Most of the theory and mathematical models developed to study diffusion on networks were originated by the research on the dynamics of spreading of infectious diseases and its control [146–148]. Recently, these models have been applied to study diffusion of diseases such as Ebola in Liberia [149, 150], influenza pandemic [151] or analyzing human mobility and contagious processes to model Malaria diffusion [152].

In this chapter we analyze the dynamic and static properties of a social network inferred from mobile phone call records and test to what extent the different behaviors in geographical regions are related to the information diffusion and the appearance of outbreaks in regions. As we will see in the following sections, the resulting diffusion is a combination of the inter-regional structure of the network and the vibrant dynamics at the individual level so it is a complex process involving social networks and the bursty interaction patterns of the relationships.

4.1.3 Dynamic networks and social strategies

By nature, social networks are dynamic, that is, given a relationship between two nodes, communications between them are not constant and interactions happen on time at specific and mostly instantaneous moments. For instance, two mobile phone users who are friends are not constantly communicating each other but they do it in concrete moments; if we analyze when two users in a social network are mentioning each other or talking by chat, it happens the same; when a contact network is analyzed, that is, a network where every interaction between users is a real meeting, this only happens spuriously and in a active fashion for a relatively short period, independently from a long stability of the relationship between the involved actors. In terms of information diffusion, viral contagion can only be produced in this specific interactions which means that interaction dynamics strongly affect dissemination on networks. On this chapter, we investigate how the dynamism of social relationships among individuals in the population at the country level affects the outbreak propagation through the different regions.

These punctual interactions lead to the definition of dynamic concepts such as link stability or link survival in a very natural way. We consider that a relationship between two nodes i and j of the network are opened at time t if, at least, there exist one interaction between them before t and after t. Moreover, we can define when a link is new (just opened for the first time) or when a link is destroyed: given a dataset where the timestamps of every interaction is stored, we consider that a link is created on time t if the first interaction between two nodes happens at this time and destroyed if the last one occurs at time t. However, this naive definition leads to methodological problems since, under this definition, all the relationships considered in a dataset would be created and destroyed. To solve these problems of definition at the border, typically we proceed as in [15], by defining three temporal windows that are equally large (Ω_1, Ω and Ω_2). Given this context, we formulate a better definition of new relationship by considering links that do not occur on Ω_1 but they do on Ω for the first time; analogously, a link is destroyed in Ω when the last interaction happens in Ω. Beyond these definitions, we can naturally talk about link stability: a link is considered stable when interactions appear in the three different periods or we name it instantaneous when both the first and last interaction happens in Ω. This formulation has been used in different investigations related to temporal social network analysis: for instance, this framework leads to the setting of social strategies: in social networks there exist users creating and destroying links at a faster pace (social explorers) than others (social keepers). Beyond the degree of dynamism in the creation and destruction of links, the number of opened connections on time has been shown to be almost constant.

Independently from the dynamic structure of the links, individual interactions and their temporal distribution strongly affect the diffusion of information. It has been widely proved that inter-event time patterns are very heterogeneous, that is, most of the events happens in a very short time period consecutively but, many other times, a long time happens between two events [16, 17, 132, 135, 153]. Typically, this phenomena is modeled by means of a Power-Law or Log-Normal distribution. This bursty dynamics make that defining the properties of the network from the temporal point of view make sense: the intensity of the relationship between two nodes is dynamic because the number of interactions varies on time [62]; in the same way, features like shortest paths between two nodes, the diameter of the network, node and edge centrality or spanning trees are translated into a dynamic version since only opened links are considered in the computation [154].

As we will see in the next sections, this bursty behavior is also observed in our dataset and our final goal is to understand whether the dynamics at the microscopic level of the links affect the inter-regional diffusion.

4.1.4 Modelling diffusion processes

As we mentioned in the previous sections, we simulate diffusion processes to test whether dynamic or static properties of geographical networks predict the time to infection of regions. To this end we use the so called SI model on the social network inferred from mobile phone calls, which is a version of one of the most widely used models, the SIR model.

In the SIR model [146, 147], the population is divided in three different segments (we explain it in terms of a disease diffusion): susceptible individuals, people who are not infected but might become infected in future; infected individuals, people who are able to transmit the disease; and recovered individuals, people who were infected in past but they cannot propate the virus any more. One can analyze the temporal evolution of this model by simulation numerically the next system of differential equationss

$$\begin{aligned}\frac{dS}{dt} &= -\frac{\beta IS}{N}\\ \frac{dI}{dt} &= \frac{\beta IS}{N} - \gamma I\\ \frac{dR}{dt} &= \gamma I\end{aligned}$$

where $S(t)$, $I(t)$ and $R(t)$ represent the number of susceptible, infected and recovered individuals respectively, β represents the transmision rate of the disease and γ the recovery rate. However, in this version the transmision and the recovery rate are constant, neglecting the underlying social structure or the seasonality of diseases.

It is also interesting to study the geographical diffusion of diseases, virus, rumors or information. Focusing on the structure of a country, we observe that relationships in different social networks are overlapped and interacting with each other: a country is composed by provinces which are also composed by municipalities and, at the finer level, formed by neighborhoods. Since social relationships have strong geographical contraints, people tend to hold friendships with people nearby them [155–158], one can find different dynamics happening at the same time at the different macroscopic levels aggregating geographically. In particular, contact-based models integrated in heterogeneous networks have been widely explored in the literature [37, 138, 159, 160] giving mathematically analytical and simluation approaches. This models on heterogeneous networks, that is, networks with similar features to geographical ones, mixing subsets of nodes with their own topology embedded in a larger network (as cities within a country), has been used to model propagation in intra-urban and inter-urban contexts.

A way to study diffusion on social networks is simulating stochastic processes based on data (data-driven simulations) which allows to study spreading processes not only as a function the connectivity structure of the network but also, at the same time, as a function of other features such as the temporal dynamics. This approach has been used in literature to analyze how the different features of the network affect the velocity of diffusion which is done by comparing the spread on the original to null models (for instance, if we want to test how the temporal dynamics affect the diffusion, we compare it in the original data and shuffling the timestamps of the interactions) [62, 135, 137, 161]. Moreover, it is found that the for the inter-event time distribution is heavy-tailed and the creation/destruction of ties slows down information diffusion when simulations on real data are compared with those done on shuffled data. Moreover, also individuals with a more persistent network (keepers) are infected earlier than those with a more volatile neighborhood (explorers) [15].

In our case, we consider the CDR from a mobile phone company, that is, for each call we use the origin and destination of the call and when it was made, the set of all the interactions for every link between two mobile phone users. In our framework, information can only be transmitted between two users when an interaction is observed so we are considering at the same time both the social network structure and the temporal dynamics of the interactions.

4.1.5 Network scaling on cities

For the first time in history, the majority of people around the world are living in cities. Although a deep analysis of migrations (between different countries, from rural areas to cities, due to wars, etc.) is complex and many factors play a role, living in cities provide opportunities for the individual development in terms of economy and opportunities [129, 130]. Using the biological methaphore that represent cities as consumers of resources and producers artifacts and information, it has been shown that living in cities optimize the production exhibiting super-scaling functional relationships in terms of innovation, patterns, research and development investment or total bank deposits [133, 162, 163]. Moreover, cities are not only able to produce better but optimizing the costs of infrastructures such as gasoline stations, length of electrical cables or road surface [10].

More surprisingly, it has been also proved that size of the cities is also related to the underlying social network involving the individuals inhabiting there. In particular, the number of interactions among the individuals and the number of connections grow superlinearly with the population size [164] which might be the reason behind the increase of the economic development with the city size.

Since our goal is to understand spreading processes at the geographic scale, we need to characterize the network connectivity and activity at the urban level not only in terms of the static structure of the network but also the dynamical features. In our work we replicate the results in the literature about the superlinear growth of interactions and connections between people living in larger cities and extend this result analyzing the pace of creation and destruction of relationships as a function of the population.

4.1.6 Comparing evolving adjacency matrices

As we will see in next sections, we need to analyze whether, despite intra-urban networks are changing at a very fast pace, this fact has an impact at the macroscopic (geographic) level. We use two different methods to prove the stability of the underlying structure in our network.

- *Correlation of adjacency matrices*: given two times t_1 and t_2, we consider the two sets of open links E_{t_1} and E_{t_2} and the corresponding adjacency matrices $\mathcal{A}_{t_k}$ with entries $A_{t_k,\alpha,\beta}$, that is, the number of open links between two regions at time t_k. Interpreting all the entries of the matrices as vectors, we compute the correlation between these two matrices,

  $$\rho_{t_i,t_j} = cor(\mathcal{A}_{t_i}, \mathcal{A}_{t_j}) \tag{4.1}$$

  which can be interpreted as a measure of the stability of the network.

- *Community stability*: other approach is using community detection to infer the underlying structure of the network. Our hypothesis is that network stability must be also reflected on the communities structure so that we use the Normalized Mutual Information method (NMI) [95] to compare membership vectors on time: the higher the value of NMI is between two partitions of the network, the more similar they are.

4.2 Pace of change in urban social networks

4.2.1 The dataset

To study the static and dynamical properties of the social network, we have considered the CDRs from a single mobile phone operator over a period of 19 months from February 2009 to August 2010. The data consists of the anonymized voice call records of about 20 million users that form 700 million communication ties in the United Kingdom. After filtering out all the incoming or outgoing calls that involve other operators, we only consider users that are active across the whole time period and keep only reciprocal ties. As in [15] we consider an observation window $\Omega = [0, T]$ of 7 months in the middle of the 19 months, where the 6-month time windows before and after Ω are used to asses whether a tie has been created/destroyed in Ω. Out of the 20 million users, we have considered only a group of 3.2 million users for which we know their billing address which is considered as their location up to the postcode area level in England, Scotland and Wales. We only consider calls between users in those postcode areas, which leave us with 10.5 million communication ties.

4.2.2 General and geographical properties of the microscopic network

At each time step t in Ω we consider the set of communication ties that are opened at that instant, that is, the set of ties which have communication events before and after t. Those ties constitute our social microscopic network at that given instant. However, since ties are continuously created and destroyed, that social microscopic network changes very fast. In fact we found that the microscopic network has 7.7 million links at $t = 0$ but more than 2.8 (2.9) million links are created (destroyed) in the observation period Ω. Since humans tend to balance the number of created and destroyed links [15], we find that the number of open ties at any given instant t fluctuates almost constantly around 7.7 million links. However, only 81% of the links present at $t = 0$ remain at $t = T$, showing the high pace at which the microscopic network evolves.

The geographical properties of these microscopic links are very similar to those found in previous works [100, 165, 166]. For example, if d is the distance between two individuals in a tie, we find that the probability to find a tie with distance d is $P(d) \sim 1/d^{\alpha}$ with $\alpha = 1.52 \pm 0.02$, see figure 4.1. We also find that the number of calls per link decreases with distance and reaches a minimum around $50km$. This result is similar to the one found in other countries in mobile phone communications or online social networks and seems to indicate the different role of communication at short and large distances [35, 161]. Interestingly, this distance is comparable with the average distance of ties between different postcodes ($62km$), which suggests that these geographical units enclose short distance communication behavior. Finally, in figure 4.1 we see that the persistence of ties only depends slightly on the distance in the tie, showing also a minimum around $50km$. It is interesting to see, that short-ranged and long-ranged ties are almost equally persistent, that is, distance does not penalize the stability of ties.

4.2.3 General and geographical properties of the urban network

From that microscopic network we construct the time-dependent weighted urban (postcode) network as follows: at each time t we consider all postcodes α and we quantify their social connections by measuring $w_{t,\alpha,\beta}$, the number of open social ties between postcodes α and β (see Panel A) in 4.2).

Since microscopic links are constantly created and destroyed, at any given instant the set of ties between α and β are different. In fact, we measure the persistence of the internal postcode links $P_{t,int}$, which measures the fraction of microscopic links present at $t = 0$ which are still present at time time t and both nodes belong to the same

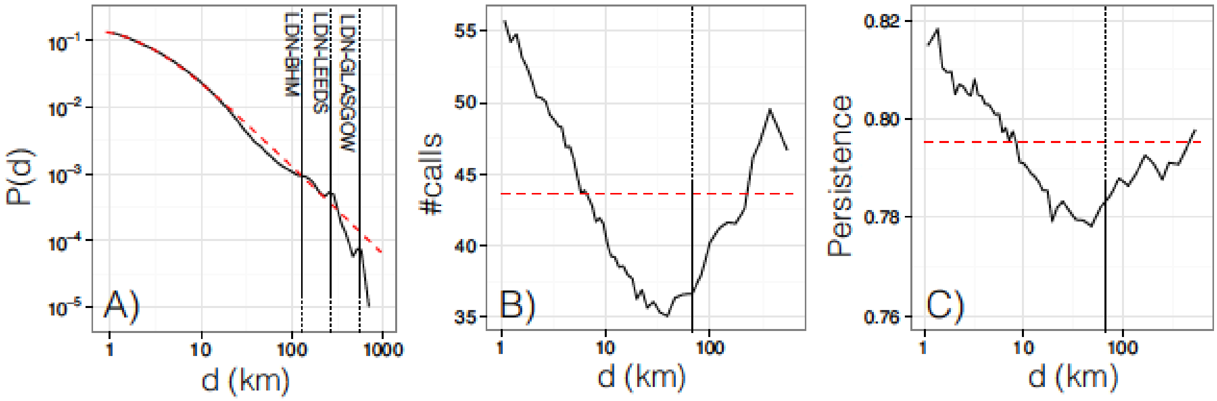


FIGURE 4.1: **Geographical properties of the microscopic network**. Panel A) shows the probability to find a tie with distance d (black line) with a non-linear fit to $P(r)$ $a/(b+r^{\alpha})$ with $\alpha = 1.77 \pm 0.05$. Dashed vertical lines indicate distances from London (LDN) to main cities in the UK: Leeds, Birmingham (BHM) and Glasgow. B) Average of the number of calls per ties with a given distance d. Dashed vertical line shows the average distance of ties between different postcodes ($62km$). C) Persistence of ties at distance d.

postcode. Formally, let $E_{t,int}$ be the set of internal links opened at time t,

$$P_{t,int} = \frac{|E_{0,int} \cap E_{t,int}|}{|E_{0,int}|} \tag{4.2}$$

We can analogously define the persistence for the external links, $P_{t,ext}$. We find that, in both cases, the persistence steadily decreases throughout Ω and reaches values of $P_{t,ext} = 0.78$ for external links and $P_{t,int} = 0.89$ for internal ones.

Given this amount of turnover, we investigate how the urban network changes in time. To quantify the network dynamics we measure the evolution of the urban network characteristics at different time instants. Firstly, we measure the time evolution of the average number of links in the network

$$\overline{w}_t = \sum_{\alpha,\beta} w_{t,\alpha,\beta} \tag{4.3}$$

In correspondence with [51] we have also measured the average connection diversity for a postcode $d_{t,\alpha}$,

$$d_{t,\alpha} =< -\frac{1}{\log|\mathcal{N}_\alpha|} \sum_{j \in \mathcal{N}_\alpha} p_{ij} \log p_{ij} >_{i \in \alpha} \tag{4.4}$$

where p_{ij} is the fraction of all the interactions involving i which also involves j.

Finally, we also compute the weighted clustering coefficient $C^{w}_{t,\alpha}$ introduced in [90] (further information in section 3.5).

As we can see in panel C) of figure 4.2 these measures averaged over the network practically do not change over the 212 days of our observation window, indicating that the urban network does not change significantly over that time. This result is in agreement with other results [100] and show that, despite the constant evolution of the network, its main global characteristics remain almost stable.

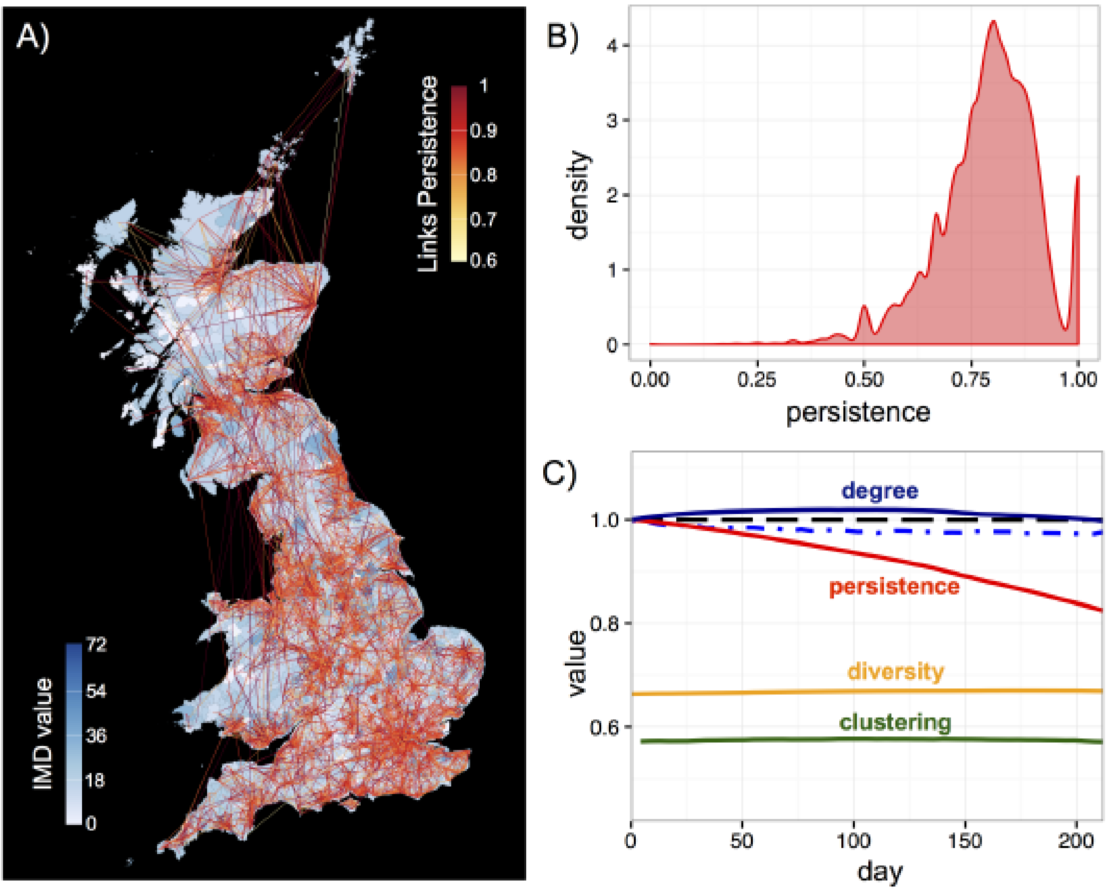


FIGURE 4.2: **The dynamics of urban social networks**. Panel A) depicts the graph of ties between different postcodes in England, Wales and Scotland. The color of each postcode is proportional to its IMD value, while the color of the links between postcodes is proportional to their persistence level. Only links between postcodes with more than 100 social ties are considered. Panel B) shows the persistence density for all the links in our postcode graph, while the inset shows the persistence of links from and to a given postcode. Panel C) shows the values of different measures over the postcode graph at different days of our observation time. Although degree, diversity and clustering almost stay constant, the persistence of links decays dramatically, showing that the apparent steady nature of the postcode graph at the macroscopic level is accompanied by a vibrant microscopic change at the user level.

However, we also found that the stability of the network goes beyond averaged quantities. In particular, if we concentrate in a single link between postcode α and β in the urban network we can see that $w_{t,\alpha,\beta}$ almost remains constant along Ω. That is, although the actual communication ties between those postcodes might be constantly changing, the number of communication between postcodes remains constant. These results is the spatial generalization of the findings in [15] where it was shown that despite many links are opened and destroyed by individuals, the number of opened social ties remains constant. Our results here demonstrate that this result also holds at the relationships between urban areas: links between two given areas are created and destroyed at the same rate so that the amount of communication ties between those areas remain constant. This probably signals that the communication between the areas reflects a functional, social or economical relationship between areas which does not depend on the actual actors behind that relationships. Or that the geographical component in link creation or destruction conditions strongly were those events might happen.

This dynamical equilibrium in the communication between areas implies that the network almost remains constant in Ω. In fact, if we consider the Pearson correlation ρ between the adjacency matrix $w_{t,\alpha,\beta}$ at different times, we see that it slightly fluctuates at $\rho = 0.999$ (see section 4.1.6 for further details).

Finally, we also carry out a community analysis on the network at three different moments: at the beginning, at the middle and at the end of Ω. We proceed by executing one of the most used community detection algorithms, Infomap [94], at these three moments and comparing the postcode membership structure to communities. Visually, we can observe in figure 4.3 that not big differences are appreciated, that is, postcodes belonging to a certain community tend to belong to the same community on time. In fact, the Normalized Mutual Information (NMI) [95] of the community structures of the postcode network found at time $t = 0$ and T fluctuates around 0.972, exhibiting a high robustness of the inter-urban connectivity structure.

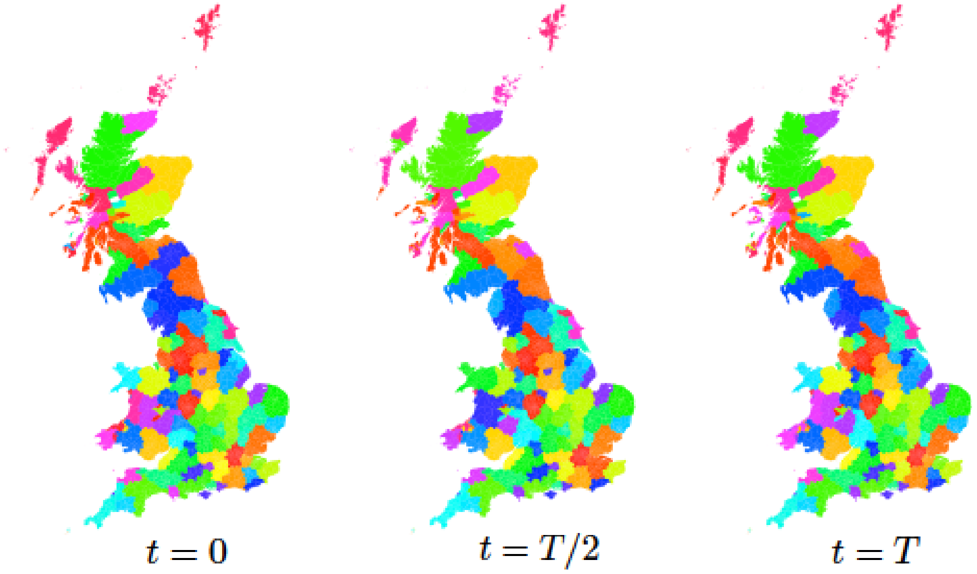


FIGURE 4.3: Comparison of the communities found in the urban social network by the Infomap algorithm at different times in the observation period Ω.

Since the urban network does not change considerably we can investigate its geographical properties at any given instant. Firstly we investigate how the number of links in a urban area and its persistence changes with the population living in that area. Recent work has found that there is a super-linear effect of the population of an area in the number of links present within the area [10, 167]. In our case we also found this super-linear effect, where the normalized number of links inside a postcode scales as $\tilde{w}_{\alpha,\alpha} \sim P_{\alpha}^{1.12\pm0.01}$ with the population in the postcode P_{α}, where $\tilde{w}_{\alpha,\alpha} = w_{\alpha,\alpha}/\nu_{\alpha}$ and ν_{α} is the penetration rate in postcode α, i.e. the ratio of the number of users to the population in that postcode. The same holds for the number of created links inside a postcode $\tilde{w}^{+}_{\alpha,\alpha} \sim P_{\alpha}^{1.24\pm0.02}$. This means that more populated areas have proportionally more links and create more links than small populated areas. On the contrary, when we proceed in the same way and consider the dependence of the number of external links with population (links connecting postcodes α and β with $\alpha \neq \beta$), we observe a linear behavior. That means that the super-linear behavior of density of links only happens for links at small distances, in this case, for those within postcodes.

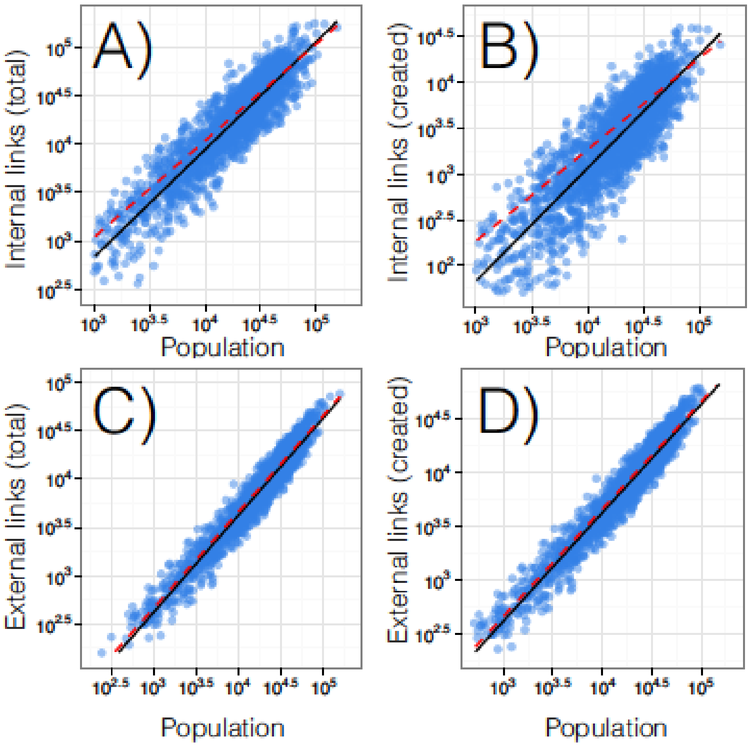


FIGURE 4.4: **Super-linear and linear number of internal and external links.** Normalized number internal ties within a given postcode (Panel A) or external with the rest of postcodes (Panel C). Solid (dashed) line is the non-linear (linear) fit to the data. Panels B) and D) show the corresponding analysis for the internal and external number of created links in a postcode.

4.2.4 Gravity Law for link creation patterns

Finally, the geographical distribution of links present at any given time t can be described by means of the gravity law (see figure 4.5), i.e. the number of links between postcodes α and β tends to satisfy the expression given by

$$w_{t,\alpha,\beta} = \frac{P_\alpha^\nu P_\beta^\nu}{d_{\alpha,\beta}^\gamma} \tag{4.5}$$

with $\nu = 0.43$ and $\gamma = 1.37$. Note that we have chosen to fit the same exponent for both postcodes because of the bidirectionality of communications. Moreover, the number of created links between postcodes α and β is also well described by a gravity law like previous equation with $\nu = 0.46$ and $\gamma = 1.125$. This leads to two main conclusions: i) not only existing ties at t can be explained by the gravity law, but also future ties between postcodes happen proportionally to the populations and inversely proportional to the distance between postcodes and ii) no large differences are observed in the population exponents (0.43 vs 0.46) whereas, the distance is a more important factor when the total number of links is considered, that is, the total number of links decays faster than the number of created links with the distance. To our understanding, this is the first time that a phenomenological law for the geographical nature of link creation/destruction is given.

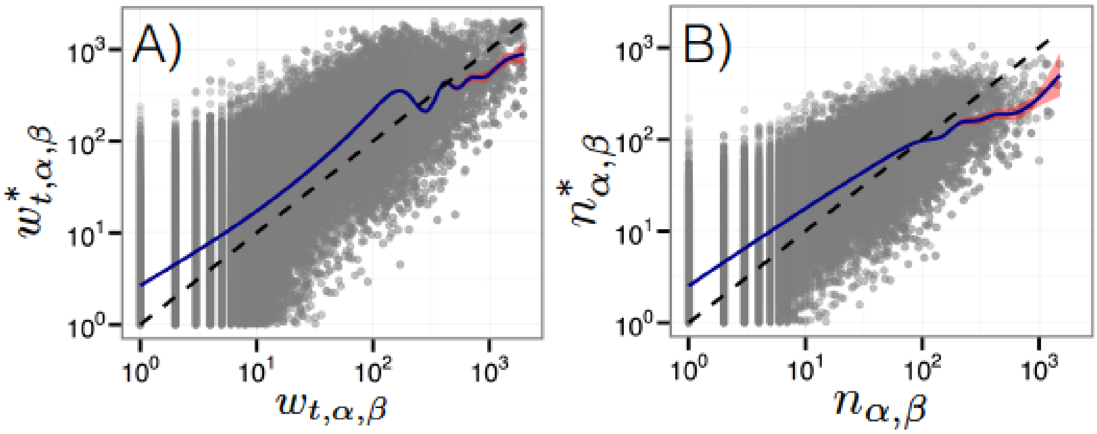


FIGURE 4.5: **The gravity law of existing and created ties.** Panel A) depicts the number of ties between different postcodes ($w_{t,\alpha,\beta}$) against the prediction given by equation (4.5). Dashed line is the $y = x$ line, while the solid blue line is a running average. B) shows the same analysis for the number of created ties during Ω between different postcodes.

4.2.5 Information diffusion

Finally, we test the effect of the urban network dynamics on the diffusion of information. To do that, we simulate an SI (Susceptible-Infected) process on the actual calls. Specifically, starting from a seed randomly chosen from the list of users, we consider that, in each call between two users, information is transmitted with probability one from the infected to the susceptible user. For statistical robustness reasons, we run 200 realizations of the SI process on our data . For each realization of the SI model and after a given time (typically, seven months in our data), the whole population becomes infected (see figure 4.6) which is perfectly reasonable regarding the network connectivity and the features of this specific spreading process. In order to compare the effect of the temporal dynamics of the urban network, we have also simulated the SI model on time-shuffled data in which the timestamp in which calls are made is shuffled across the database [15, 135]. Note that static properties of the links like number of calls or the geographical situation are conserved, but the temporal properties of the links are completely destroyed.

The question we address here is to understand whether the different pace of social networks in urban areas do confer those areas a relative advantage with respect to information diffusion in terms of being informed soon in the propagation. To do that we investigate the time to infection of the different postcodes in our database; obviously, the more populated the postcode is, the earlier information will get to it, so we concentrate on the relative time to infection τ_α in which 10% of the population in postcode α gets infected in the SI process to remove this bias. Figure 4.6 shows the average cumulative infection curves for individuals and postcodes. In every case we recover the results in [62], in which speed of information diffusion in real data is always smaller than in the time-shuffled data (informartion spreading is slower), in the case of postcodes the average time to reach 10% infection is $\langle\tau_\alpha\rangle \simeq 80$ days (when all postcodes are considered together), while $\langle\tau_\alpha\rangle \simeq 44$ days in the time-shuffled case.

More interestingly, the spreading is not homogeneous since time to infection is very different for each postcode. For example, $\langle\tau_\alpha\rangle = 56$ days for the BB1 Blackburn postcode in England, while it can be as large as $\langle\tau_\alpha\rangle = 160$ days for the CF82 Hengoed postcode in Wales (see figure 4.8). In principle, the time to infection for a given postcode might depend on the global/local static and dynamical properties of the network.

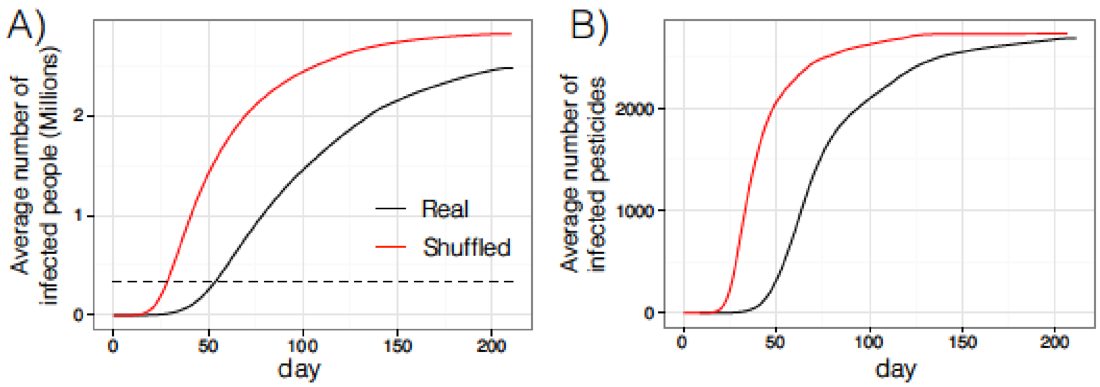


FIGURE 4.6: **Cummulative infection curves**. Panel A) shows the average cumulative infection curves for individuals in our database for the SI run on the real (black) and time-shuffled (red). Panel B) shows the average cumulative 10% population infection curves for postcodes. Colors are the same as in panel A).

4.2.6 Measuring importance of static and dynamic variables on infection times

To unveil which are the most important properties of the network in that process, we consider several static and dynamical variables for each postcode. Specifically, we consider the number of ties within and outside each postcode, $\omega_\alpha^{in} = w_{\alpha,\alpha}/n_\alpha$ and $\omega_\alpha^{out} = (\sum_{\beta\neq\alpha} w_{\alpha,\beta})/n_\alpha$ respectively. We also consider the diversity of communication for each postcode by computing the normalized Shannon entropy d_α, ranging between 1 (if α has equal number of ties with each β) and 0 (if α concentrates most of the communications with a particular postcode among the k_α postcodes). Finally we also consider dynamical variables like the normalized number of calls made by users in postcode α, c_α and the persistence of the ties in which users of postcode α are involved, P_α. It is interesting to see that due to the geographical properties of the urban network, these variables are not independent. Specifically, we find that the fraction of internal links are highly anticorrelated with external links $\rho = -0.48$ $[-0.51, -0.44]$ showing that having more external links means less internal communication. This can be interpreted as a limitation of the connection capacity of postcodes as well as it has been observed in humans since individuals cannot pay attention to all their connections equally because of finite time. On the other hand, more external links per user implies higher number of calls (obviously, $\rho = 0.59$ $[0.56, 0.62]$) and lower persistence $\rho = -0.53$ $[-0.55, -0.50]$. This last result can be partially explained by the fact that persistence of long-ranged ties is smaller (see figure 4.1) and thus, the more ties outside of the postcode, the less persistence.

Regarding the time to infection we find that, as expected, the fraction of links within the postcode, outside the postcode and the average number of calls per user are negatively correlated with τ_α ($\rho_1 = -0.27$, $\rho_2 = -0.57$ and $\rho_3 = -0.62$ respecively). Thus, higher number of calls and larger number of social ties decrease the time to infection, a result which has also been observed at the individual level in many other networks and information diffusion processes [137, 139]. On the other hand, since persistence and number of links are inversely correlated, we find that the larger the persistence of ties in a postcode is, the smaller the number of links is and thus the larger the time to infection. It is interesting to see that this result is contrary to what is found at the individual level: in [15] it was found that keepers (those with higher persistence of links) have lower time to infection than explorers (those with more volatile neighborhoods). In our simulations we found the opposite, that urban areas with more persistent networks have larger time to infection because more persistent networks means less external links.

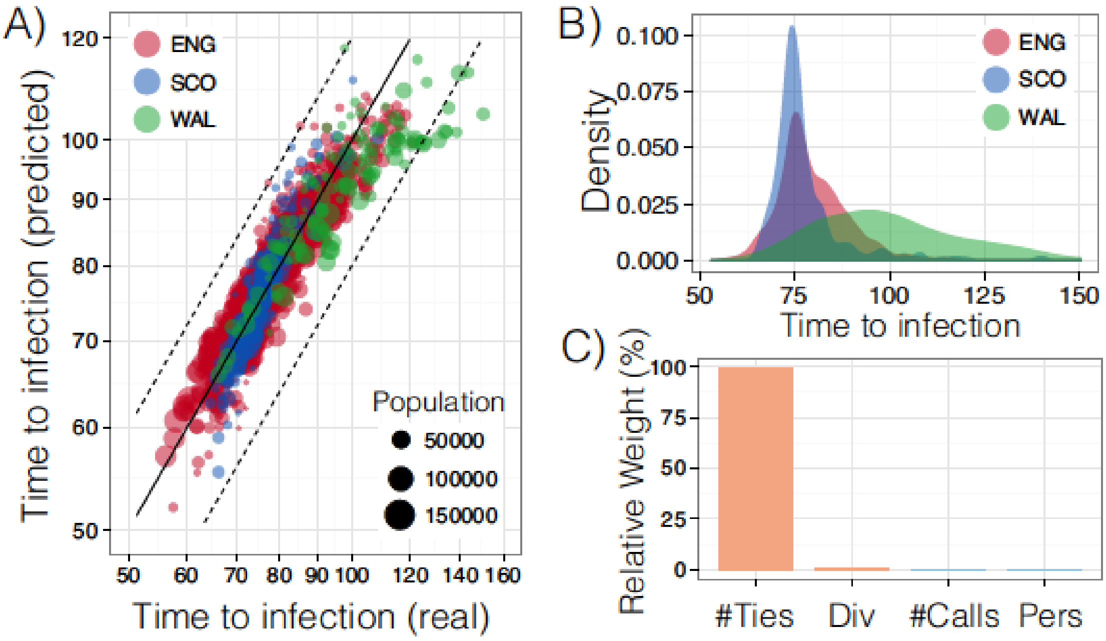


FIGURE 4.7: **Time to infection model**. Panel A) shows the goodness of the fit to the 10% infection rate in each postcode by comparing the real and the predicted time. Fit has an explanatory power of $R^2 = 0.84$. Panel B) shows the density plots for the time to infection values in each country, showing that in general WAL gets much larger values than the other countries SCO and ENG. Panel C) shows the relative importance of the variables in the fit. The number of links amounts to 99% of the relative importance in the fit.

The large correlation between time to infection and network variables anticipates that most of the heterogeneity in the latter can be explained by the variables. To show that, we have built a very simple non-linear model between the τ_α and the network static and dynamical variables

$$\log \tau_\alpha \sim \log \omega_\alpha + \log d_\alpha + \log c_\alpha + \log P_\alpha \tag{4.6}$$

were we have put together the total number of ties of a postcode $\omega_\alpha = \omega_\alpha^{in} + \omega_\alpha^{out}$. As it is shown in figure 4.7 the explanatory power of this model is large $R^2 = 0.84$. This is an interesting result, since it shows that properties of a spreading process that happen on the full network depend only on local properties for each postcode. But note that the number of links only account for 99% of the variance explained by the model, i.e. ω_α is the only relevant variable in the model. In particular we found that

$$\tau_\alpha \sim \frac{1}{\omega_\alpha^{0.64}} \tag{4.7}$$

This is reminiscent of a well known property of random networks in which time to infection goes like $1/k_i$ where k_i is the connectivity of node i (cita).

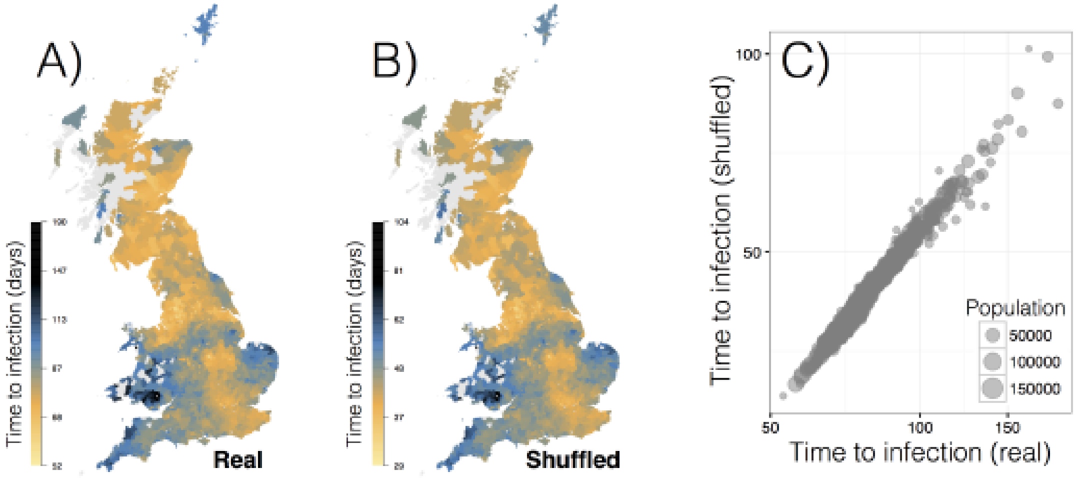


FIGURE 4.8: **Time to infection in the urban network**. Panel A) and B) depicts the time to 10% infection on the real and shuffled data. Time to infection is only calculated for postcodes with more than 200 users. Light gray areas show postcodes with less than 200 users. Panel C) shows the high correlation between the time to infection in real and shuffled data for each postcode.

Despite the dynamical properties of the network around a postcode seems to have no direct impact in the time to infection of that postcode, it could explain the different behavior between the spreading process on real and shuffled data. It is well known that tie dynamics slows down information propagation when it is compared with shuffled data and thus we could expect that postcodes with higher persistence of their ties could have an advantage (in time-to-infection) with respect to what happens when data is shuffled. However, as we can see in figure 4.8 time to infection in real and shuffled data are highly correlated ($\rho = 0.99$) at postcode level. That means that

$$\tau_\alpha^{\text{real}} \sim \tau_\alpha^{\text{shuffled}} + 36 \text{ days} \tag{4.8}$$

Thus, also models (4.6) and (4.7) are good models for the shuffled data process. In summary, our results indicate that urban network dynamics does not play any role to explain the heterogeneity in the time to infection and it only induces a global slowing-down time factor of 36 days in information spreading.

4.3 Conclusions and further work

In this work we have analyzed to what extent the dynamics of creation and destruction of relationships among the individuals in an industrialized country affect the inter-regional structure of the network at the macroscopic scale and, going one step forward, we have answered the question about which family of features about the social connectivity of the regions, static or dynamic ones, influence the most information spreading. The main conclusion in this work is the velocity of the infection in a particular region is strongly related to how it is connected to the rest of the country, in terms of number of ties, and this feature is, precisely, a very stable one on time, despite approximately 20% of the social relationships are created and destroyed in the temporal period of our study. This fact has important implications about the role of individuals and their particular relationships in contagion and geographical diffusion processes : the key insight in our analysis is it does not matter whether a particular connection is open or not at a given time but only the number of connections between regions, which is a surprisingly constant quantity on time. This might be interpreted as a emergent phenomena given by the complex relationships in a country, where the individual does not play a key role in the information diffusion and, actually, it is the structure of the inter-regional network what makes information spread faster or not leading to the conclusion that those regions with a high number of connections are more likely to become infected soon in the spreading.

4.4 Publications

At the time of the deposit of this thesis, the paper has been submitted for publication [168].

Chapter 5

Conclusions and future work

5.1 Conclusions

On this dissertation, we have discussed three problems in which the main goal is to anticipate and model mathematically different aspects and phenomena related to both individual and groups. As a general conclusion, all the available data originated by the digital traces registered in our daily lives are a reflection of our activity as individuals but also it is also encoded a huge amount of information about our behavior as societies. In particular, our results lead to conclusions about the nature of economic decisions in terms of shopping behavior, exhibiting that humans are predictable but to a point; regions can be analyzed by using information from social media in such a manner that allow us to build predictive models of emergent phenomena as economy; and, finally, in a country, many geographical networks are overlapped and active at the same time (one per neighborhood, postcode, city,...) with different properties and we have studied how the internal dynamics affect the information diffusion on the global network.

5.1.1 Consumers are predictable but not deterministic

We began this dissertation wondering whether our individual behavior are completely determined by internal and external constraints such as the geography or our social network. The answer to this question is not simple: as a first big conclusion, yes, we, consumers, are predictable, we show recurrent visitation patterns to our favorite merchants and our possible space of them where we will make our next purchase is relatively limited. Typically in scientific literature the problem of studying human predictability has been addressed by means of computing entropy measures, which is also our approach, but we have detected problems due to the data resolution. Since the temporal scale in credit card transaction data is much coarser than in other data sources like CDRs, we have observed that this might lead to methodological problems since true entropy estimation methods are strongly dependent on the temporal window. Finally, we show that, despite the predictability of consumer visitation patterns, the ordered sequence of purchases does not contain enough information to predict next one with more than 30% of accuracy.

5.1.2 Using Twitter to model unemployment

The problem of creating economic indicators from non traditional data sources has powerful potential uses for anticipating economy evolution, detecting crisis before observing their worst effects or monitoring economy at new smaller scales. This is only possible if all this data is a true reflection of our activity, interests, work patterns and daily activity. In our work, we have shown this is the case using Twitter data. In particular, we create a functional partition of an European country by considering the mobility inferred from Twitter to eliminate municipality heterogeneity and then infer potentially predictive variables to model economy. These regional variables are created as geographical characteristics, as Twitter penetration rate, but other are aggregations of individual features, such as the proportion of misspellers or the geo-social connectivity. Combining all these variables we show that unemployment rate can be well fitted by a simple linear model, concluding that the activity reflected on social networks is really a real mirror of society we can use to nowcast economy.

5.1.3 Information diffusion is tolerant to intra-urban changes

How the properties of cities scale with population has attracted much interest in scientific literature but, as far as we know, how network dynamics varies among geographical regions had not been deeply studied until now. In this dissertation we show that not only network density or economy scale superlinearly with the population when geographical regions are observed but also the number of new created internal relationships (with both individuals in the relationship belonging to the same geographical region) in time, that is, the network is more dynamical in larger cities. In general, we see that almost 20% of links are created and destroyed in our temporal window so it is reasonable to question whether this fast pace of change in urban networks affect the information diffusion at the country level. Our results indicate that, despite this varying network, the inter-urban network remains stable with the same features in terms of connectivity and community structure which is also reflected in being able to predict the time to infection of regions when diffusion processes are simluated by only using static properties of the global network.

5.2 Open problems

5.2.1 Improving next purchase prediction models by using geographical constraints

Through Chapter 2 we have analyzed the balance between the high predictability of economic decisions, because we only can reach a constrained space of merchants where making our next purchase, and the seemly random choices within this space of possibilities. Obviously, a major problem in the analyzed context, where we use credit card transaction data, is the granularity of data compared to other location sources such as mobile phone data. But, in fact, we do not even use the geographic position of merchants to predict the next place. If we used it, how much would the model accuracy be imporved? Since time between consecutive transactions is large, compared to inter-event times in other data sources, one might think that geography would not help too much to get better perfomance rates but there is also information about periodic behaviors [32] that might lead to predictive information depending on the position of the merchant: if a model detects an user is always near a concrete place with periodic patterns, maybe it is more likely to observe a transaction close to there. Other variables such as day of the week, hour, etc. might also be used to improve the model so it could be fully understood how much the impact of every family of variable is in the next-place prediction problem.

5.2.2 Using Twitter data to analyze economy in countries

In Chapter 3, we have analyzed how data can be used not only as a reflection of individuals but also as a way of studying intrinsic features of societies (in this case, municipalities and regions induced by activity). In particular, we observe that Twitter penetration rate is highly positively correlated to the unemployment rate, that is, in regions with worse economy, people tend to use Twitter more. However, this is completely contrary to results in literature where it is shown that countries with wealthier status exhibit higher technology penetration rates, even measured by the proportion of Twitter users [66]. This result show a strong difference that might depend on the geographical scale or the country that it is being observed. Can we extend our analysis to different geographical aggregations? Would we obtain analogous results in other countries? At the present time, we are developing a similar study in Indonesia. Being able to explain economy from non-traditional data sources like Twitter would have a strong impact because the administrative and public structures are not able to reach all the regions and creating economic indicators in traditional ways is costly and, in some occasions, it is not even possible.

5.2.3 Using behavioral models to anticipate economic indicators

Even though in Chapter 3, we analyze how the model get worse on time and explain the particular evolution of the error, an interesting question is whether we might really anticipate economy indicators using non-traditional data sources like Twitter or other online social networks, search query logs, navigation patterns, etc. Since some of these indicators are based on surveys that are carried out periodically, for instance, the most liable unemployment indicator in Spain, building this kind of models might be used as real-time monitoring tools of the economy. Moreover, we might analyze whether this kind of methodologies are able to predict future values of indicators what would lead to create mehtods to prevent crisis (unemployment incresings, stock markets shocks, unstabilities in the import-export balance, ...) and therefore acting consequently before all the effects of a financial shock emerged.

5.2.4 Might Twitter data help to predict black economy?

After building the predictive variables related to Twitter penetration rate, educational level, geo-social connectivity and temporal activity of regions, we build a predictive model of the unemployment that exhibits a good performance ($R^2 = 0.64$). But, obviously, it is not a perfect model and there is a high percentage of the variance that is not explained by our methodology. There are some few possible explanations for this fact: i) our variables are not good enough to predict better the economy ii) a linear model might be outperformed by some other non-linear model iii) there exist some deviation from the reality underying in the data because of some methodological mistake iv) public data does not reflect reality because not all the economy is registered in the official numbers. Focusing on the fourth reason, might the error in our model explain black economy? Since Twitter is a reflection of our activity and it has been shown to predict accurately unemployment in many regions, why does the model fail in other places? We might analyze where our model predicts lower unemployment rates than what it is officially registerd and study why this is the case and whether it is related to black economy.

5.2.5 Analyzing economy and information diffusion at lower scales

We have studied how to build economic indicators at the inter-regional level and, in Chapter 4, we have discussed the relationship between the intra-urban dynamics and the information spreading at the country level. However, in general, it is difficult to find statistics, economic indicators, socio-demographic data at lower scales such as neighborhoods. Since we have built a model for predicting unemployment at the inter-regional level, might we downscale this model to small areas of a city? This would lead to a completely new way of studying societies where we can disaggregate and monitor reduced groups of people (from cities to neighborhoods, from neighborhoods to streets,...) and analyze the reasons behind the evolution of its interests, habits and eventually economy.

Bibliography

[1] Todd Litman. Measuring transportation: traffic, mobility and accessibility. *Institute of Transportation Engineers. ITE Journal*, 73(10):28, 2003.

[2] John Philip Pettitt. Method and system for detecting fraud in a credit card transaction over the internet, February 22 2000. US Patent 6,029,154.

[3] Els Gijsbrechts, Katia Campo, and Tom Goossens. The impact of store flyers on store traffic and store sales: a geo-marketing approach. *Journal of Retailing*, 79 (1):1–16, 2003.

[4] Paul Resnick and Hal R Varian. Recommender systems. *Communications of the ACM*, 40(3):56–58, 1997.

[5] Byung-Do Kim and Sun-Ok Kim. Measuring upselling potential of life insurance customers: Application of a stochastic frontier model. *Journal of Interactive marketing*, 13(4):2–9, 1999.

[6] C. Song, Z. Qu, N. Blumm, and A.L. Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

[7] R.O. Doyle. Free will: it's a normal biological property, not a gift or a mystery. *Nature*, 459(7250):1052–1052, 2009.

[8] M. Heisenberg. Is free will an illusion? *Nature*, 459(7244):164–165, 2009.

[9] Rom Harré, Rom Harré, Rom Harré, and Rom Harré. *Social being: A theory for social psychology.* Blackwell Oxford, 1979.

[10] Luís MA 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.

[11] Edward E Sampson. The debate on individualism: Indigenous psychologies of the individual and their role in personal and societal functioning. *American psychologist*, 43(1):15, 1988.

[12] Marina Alberti, John M Marzluff, Eric Shulenberger, Gordon Bradley, Clare Ryan, and Craig Zumbrunnen. Integrating humans into ecology: opportunities and challenges for studying urban ecosystems. *BioScience*, 53(12):1169–1179, 2003.

[13] Ron Martin and Peter Sunley. Complexity thinking and evolutionary economic geography. *Journal of Economic Geography*, 2007.

[14] 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.

[15] Giovanna Miritello, Rubén Lara, Manuel Cebrian, and Esteban Moro. Limited communication capacity unveils strategies for human interaction. *Scientific reports*, 3, 2013.

[16] Albert-Laszlo Barabasi. The origin of bursts and heavy tails in human dynamics. *Nature*, 435(7039):207–211, 2005.

[17] Márton Karsai, Kimmo Kaski, Albert-László Barabási, and János Kertész. Universal features of correlated bursty behaviour. *Scientific reports*, 2, 2012.

[18] Hang-Hyun Jo, Márton Karsai, János Kertész, and Kimmo Kaski. Circadian pattern and burstiness in mobile phone communication. *New Journal of Physics*, 14(1):013055, 2012.

[19] 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.

[20] Chaoming Song, Tal Koren, Pu Wang, and Albert-László Barabási. Modelling the scaling properties of human mobility. *Nature Physics*, 6(10):818–823, 2010.

[21] Albert-László Barabási et al. Scale-free networks: a decade and beyond. *science*, 325(5939):412, 2009.

[22] Yong-Yeol Ahn, Seungyeop Han, Haewoon Kwak, Sue Moon, and Hawoong Jeong. Analysis of topological characteristics of huge online social networking services. In *Proceedings of the 16th international conference on World Wide Web*, pages 835–844. ACM, 2007.

[23] Roger Guimera and Luıs A Nunes Amaral. Modeling the world-wide airport network. *The European Physical Journal B-Condensed Matter and Complex Systems*, 38(2):381–385, 2004.

[24] Albert-Laszlo Barabasi and Zoltan N Oltvai. Network biology: understanding the cell's functional organization. *Nature reviews genetics*, 5(2):101–113, 2004.

[25] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002.

[26] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393(6684):440–442, 1998.

[27] Stanley Milgram. The small world problem. *Psychology today*, 2(1):60–67, 1967.

[28] Luıs A Nunes Amaral, Antonio Scala, Marc Barthelemy, and H Eugene Stanley. Classes of small-world networks. *Proceedings of the national academy of sciences*, 97(21):11149–11152, 2000.

[29] Kevin S Kung, Kael Greco, Stanislav Sobolevsky, and Carlo Ratti. Exploring universal patterns in human home-work commuting from mobile phone data. *PloS one*, 9(6):e96180, 2014.

[30] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[31] Filippo Simini, Marta C González, Amos Maritan, and Albert-László Barabási. A universal model for mobility and migration patterns. *Nature*, 484(7392):96–100, 2012.

[32] Salvatore Scellato, Mirco Musolesi, Cecilia Mascolo, Vito Latora, and Andrew T Campbell. Nextplace: a spatio-temporal prediction framework for pervasive systems. In *Pervasive computing*, pages 152–169. Springer, 2011.

[33] David J Crandall, Lars Backstrom, Dan Cosley, Siddharth Suri, Daniel Huttenlocher, and Jon Kleinberg. Inferring social ties from geographic coincidences. *Proceedings of the National Academy of Sciences*, 107(52):22436–22441, 2010.

[34] 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.

[35] Salvatore Scellato, Anastasios Noulas, Renaud Lambiotte, and Cecilia Mascolo. Socio-spatial properties of online location-based social networks. *ICWSM*, 11: 329–336, 2011.

[36] Vittoria Colizza, Alain Barrat, Marc Barthélemy, and Alessandro Vespignani. The role of the airline transportation network in the prediction and predictability of global epidemics. *Proceedings of the National Academy of Sciences of the United States of America*, 103(7):2015–2020, 2006.

[37] Vittoria Colizza, Romualdo Pastor-Satorras, and Alessandro Vespignani. Reaction–diffusion processes and metapopulation models in heterogeneous networks. *Nature Physics*, 3(4):276–282, 2007.

[38] Duygu Balcan, Vittoria Colizza, Bruno Gonçalves, Hao Hu, José J Ramasco, and Alessandro Vespignani. Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences*, 106(51): 21484–21489, 2009.

[39] Pu Wang, Timothy Hunter, Alexandre M Bayen, Katja Schechtner, and Marta C González. Understanding road usage patterns in urban areas. *Scientific reports*, 2, 2012.

[40] Dirk Helbing. A section-based queueing-theoretical traffic model for congestion and travel time analysis in networks. *Journal of Physics A: Mathematical and General*, 36(46):L593, 2003.

[41] Juan Carlos Herrera, Saurabh Amin, Alexandre Bayen, Samer Madanat, Michael Zhang, Yu Nie, Zhen Qian, Yingyan Lou, Yafeng Yin, and Meng Li. *Dynamic estimation of OD matrices for freeways and arterials*. Institute of Transportation Studies, UC Berkeley, 2007.

[42] Anastasios Noulas, Salvatore Scellato, Neal Lathia, and Cecilia Mascolo. Mining user mobility features for next place prediction in location-based services. In *Data mining (ICDM), 2012 IEEE 12th international conference on*, pages 1038–1043. IEEE, 2012.

[43] Andreas Kaltenbrunner, Rodrigo Meza, Jens Grivolla, Joan Codina, and Rafael Banchs. Urban cycles and mobility patterns: Exploring and predicting trends in a bicycle-based public transport system. *Pervasive and Mobile Computing*, 6(4): 455–466, 2010.

[44] Tong Liu, Paramvir Bahl, and Imrich Chlamtac. Mobility modeling, location tracking, and trajectory prediction in wireless atm networks. *Selected Areas in Communications, IEEE Journal on*, 16(6):922–936, 1998.

[45] Francesco Calabrese, Mi Diao, Giusy Di Lorenzo, Joseph Ferreira, and Carlo Ratti. Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. *Transportation research part C: emerging technologies*, 26: 301–313, 2013.

[46] José Luis Iribarren and Esteban Moro. Impact of human activity patterns on the dynamics of information diffusion. *Physical review letters*, 103(3):038702, 2009.

[47] Wentian Li. Random texts exhibit zipf's-law-like word frequency distribution. *Information Theory, IEEE Transactions on*, 38(6):1842–1845, 1992.

[48] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

[49] GM Viswanathan, S.V. Buldyrev, S. Havlin, MGE Da Luz, EP Raposo, and H.E. Stanley. Optimizing the success of random searches. *Nature*, 401(6756):911–914, 1999.

[50] F. Bartumeus, MGE Da Luz, GM Viswanathan, and J. Catalan. Animal search strategies: a quantitative random-walk analysis. *Ecology*, 86(11):3078–3087, 2005.

[51] Nathan Eagle, Michael Macy, and Rob Claxton. Network diversity and economic development. *Science*, 328(5981):1029–1031, 2010.

[52] A. Lempel and J. Ziv. On the complexity of finite sequences. *Information Theory, IEEE Transactions on*, 22(1):75–81, 1976.

[53] M. Li and P.M.B. Vitányi. *An introduction to Kolmogorov complexity and its applications*. Springer-Verlag New York Inc, 2008.

[54] Juha K Laurila, Daniel Gatica-Perez, Imad Aad, Olivier Bornet, Trinh-Minh-Tri Do, Olivier Dousse, Julien Eberle, Markus Miettinen, et al. The mobile data challenge: Big data for mobile computing research. In *Pervasive Computing*, number EPFL-CONF-192489, 2012.

[55] M. De Domenico, A. Lima, and M. Musolesi. Interdependence and predictability of human mobility and social interactions. In *Proceedings of the Nokia Mobile Data Challenge Workshop.*, 2012.

[56] Coco Krumme, Alejandro Llorente, Manuel Cebrian, Esteban Moro, et al. The predictability of consumer visitation patterns. *Scientific reports*, 3, 2013.

[57] All real-world shoppers are zombies. http://www.lifehacker.com.au/2013/04/all-shoppers-are-zombies/, 2013.

[58] Shopping hapinness is bursts of spontaneity within the routine. http://analyzingmedia.com/2013/shopping-happiness-is-bursts-of-spontaneity-within-the-routine/, 2013.

[59] Shopping habits predictable to a point. http://www.abc.net.au/science/articles/2013/04/19/3740905.htm, 2013.

[60] Animales de costumbres a la hora de la compra. http://www.abc.es/sociedad/20130507/abci-consumo-estudio-carlosiii-201305062009.html, 2013.

[61] Taha Yasseri, Robert Sumi, and János Kertész. Circadian patterns of wikipedia editorial activity: A demographic analysis. *PloS one*, 7(1):e30091, 2012.

[62] Giovanna Miritello, Esteban Moro, and Rubén Lara. Dynamical strength of social ties in information spreading. *Physical Review E*, 83(4):045102, 2011.

[63] Lars-Hendrik Röller and Leonard Waverman. Telecommunications infrastructure and economic development: A simultaneous approach. *American economic review*, pages 909–923, 2001.

[64] Menzie D Chinn and Robert W Fairlie. Ict use in the developing world: an analysis of differences in computer and internet penetration. *Review of International Economics*, 18(1):153–167, 2010.

[65] Menzie D Chinn and Robert W Fairlie. The determinants of the global digital divide: a cross-country analysis of computer and internet penetration. *Oxford Economic Papers*, 59(1):16–44, 2007.

[66] Bartosz Hawelka, Izabela Sitko, Euro Beinat, Stanislav Sobolevsky, Pavlos Kazakopoulos, and Carlo Ratti. Geo-located twitter as the proxy for global mobility patterns. *arXiv preprint arXiv:1311.0680*, 2013.

[67] Jean Dreze, Amartya Sen, et al. India: Economic development and social opportunity. *OUP Catalogue*, 1999.

[68] Christopher Jencks et al. Who gets ahead? the determinants of economic success in america. 1979.

[69] Mark EJ Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.

[70] Scott E Page. *The difference: How the power of diversity creates better groups, firms, schools, and societies.* Princeton University Press, 2008.

[71] MS Granovetter. Sociological theory, volume 1, chapter 7. *The strength of weak ties: A network theory revisited*, pages 201–233, 1983.

[72] Chris Smith, Daniele Quercia, and Licia Capra. Finger on the pulse: identifying deprivation using transit flow analysis. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 683–692. ACM, 2013.

[73] Thoralf Gutierrez, Gautier Krings, and Vincent D Blondel. Evaluating socio-economic state of a country analyzing airtime credit and mobile phone datasets. *arXiv preprint arXiv:1309.4496*, 2013.

[74] Victor Soto, Vanessa Frias-Martinez, Jesus Virseda, and Enrique Frias-Martinez. Prediction of socioeconomic levels using cell phone records. In *User Modeling, Adaption and Personalization*, pages 377–388. Springer, 2011.

[75] Carlo Ratti, Stanislav Sobolevsky, Francesco Calabrese, Clio Andris, Jonathan Reades, Mauro Martino, Rob Claxton, and Steven H Strogatz. Redrawing the map of great britain from a network of human interactions. *PloS one*, 5(12): e14248, 2010.

[76] Paul Expert, Tim S Evans, Vincent D Blondel, and Renaud Lambiotte. Uncovering space-independent communities in spatial networks. *Proceedings of the National Academy of Sciences*, 108(19):7663–7668, 2011.

[77] Stanislav Sobolevsky, Michael Szell, Riccardo Campari, Thomas Couronné, Zbigniew Smoreda, and Carlo Ratti. Delineating geographical regions with networks of human interactions in an extensive set of countries. *PloS one*, 8(12):e81707, 2013.

[78] Christian Thiemann, Fabian Theis, Daniel Grady, Rafael Brune, and Dirk Brockmann. The structure of borders in a small world. *PloS one*, 5(11):e15422, 2010.

[79] Tobias Preis, Helen Susannah Moat, and H Eugene Stanley. Quantifying trading behavior in financial markets using google trends. *Scientific reports*, 3, 2013.

[80] Tobias Preis, Helen Susannah Moat, H Eugene Stanley, and Steven R Bishop. Quantifying the advantage of looking forward. *Scientific reports*, 2, 2012.

[81] Dolan Antenucci, Michael Cafarella, Margaret C Levenstein, Christopher Ré, and Matthew D Shapiro. Using social media to measure labor market flows. Technical report, National Bureau of Economic Research, 2014.

[82] Avishai Ceder. *Public Transit Planning and Operation: Modeling, Practice and Behavior*. CRC Press, 2015.

[83] Peter A Rogerson and David A Plane. Modeling temporal change in flow matrices. In *Papers of the Regional Science Association*, volume 54, pages 147–164. Springer, 1984.

[84] David K Foot and William J Milne. Net migration estimation in an extended, multiregional gravity model*. *Journal of Regional Science*, 24(1):119–133, 1984.

[85] James P LeSage and R Kelley Pace. Spatial econometric modeling of origin-destination flows*. *Journal of Regional Science*, 48(5):941–967, 2008.

[86] Anukool Lakhina, Mark Crovella, and Christophe Diot. Mining anomalies using traffic feature distributions. In *ACM SIGCOMM Computer Communication Review*, volume 35, pages 217–228. ACM, 2005.

[87] George Kingsley Zipf. The p 1 p 2/d hypothesis: on the intercity movement of persons. *American sociological review*, 11(6):677–686, 1946.

[88] James Truscott and Neil M Ferguson. Evaluating the adequacy of gravity models as a description of human mobility for epidemic modelling. *PLoS Comput Biol*, 8 (10):e1002699, 2012.

[89] Filippo Simini, Amos Maritan, and Zoltán Néda. Human mobility in a continuum approach. *PloS one*, 8(3):e60069, 2013.

[90] Alain Barrat, Marc Barthelemy, Romualdo Pastor-Satorras, and Alessandro Vespignani. The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(11):3747–3752, 2004.

[91] Robin Sibson. Slink: an optimally efficient algorithm for the single-link cluster method. *The Computer Journal*, 16(1):30–34, 1973.

[92] 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.

[93] 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.

[94] Martin Rosvall and Carl T Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105 (4):1118–1123, 2008.

[95] Leon Danon, Albert Diaz-Guilera, Jordi Duch, and Alex Arenas. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(09):P09008, 2005.

[96] Spanish 2011 census.. http://www.ine.es/censos2011_datos/cen11_datos_inicio.html.

[97] Spanish registered unemployment. http://www.sepe.es/contenidos/que_es_el_sepe/estadisticas/index.htm.

[98] Maxime Lenormand, Miguel Picornell, Oliva G Cantu-Ros, Antonia Tugores, Thomas Louail, Ricardo Herranz, Marc Barthelemy, Enrique Frias-Martinez, and Jose J Ramasco. Cross-checking different sources of mobility information. *arXiv preprint arXiv:1404.0333*, 2014.

[99] Nomenclature of territorial units for statistics (nuts). http://ec.europa.eu/eurostat/web/nuts/national-structures-eu, 2016.

[100] Marc Barthélemy. Spatial networks. *Physics Reports*, 499(1 - 3):1 – 101, 2011. ISSN 0370-1573. doi: 10.1016/j.physrep.2010.11.002. URL http://www.sciencedirect.com/science/article/pii/S037015731000308X.

[101] Aaron Clauset, Mark EJ Newman, and Cristopher Moore. Finding community structure in very large networks. *Physical review E*, 70(6):066111, 2004.

[102] Pascal Pons and Matthieu Latapy. Computing communities in large networks using random walks. *J. Graph Algorithms Appl.*, 10(2):191–218, 2006.

[103] Usha Nandini Raghavan, Réka Albert, and Soundar Kumara. Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, 76(3):036106, 2007.

[104] Mark EJ Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical review E*, 74(3):036104, 2006.

[105] Zhiyuan Cheng, James Caverlee, Kyumin Lee, and Daniel Z. Sui. Exploring Millions of Footprints in Location Sharing Services. In *Proceedings of the Fifth International Conference on Weblogs and Social Media*, Menlo Park, CA, USA, July 2011. AAAI. URL http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2783.

[106] James RA Davenport and Robert DeLine. The readability of tweets and their geographic correlation with education. *arXiv preprint arXiv:1401.6058*, 2014.

[107] Ingo Feinerer, Christian Buchta, Wilhelm Geiger, Johannes Rauch, Patrick Mair, and Kurt Hornik. The textcat package for n-gram based text categorization in r. *Journal of statistical software*, 52(6):1–17, 2013.

[108] Svante Wold, Arnold Ruhe, Herman Wold, and WJ Dunn, III. The collinearity problem in linear regression. the partial least squares (pls) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing*, 5(3):735–743, 1984.

[109] Henry Theil. Economic forecasts and policy. 1958.

[110] Yvan Saeys, Iñaki Inza, and Pedro Larrañaga. A review of feature selection techniques in bioinformatics. *bioinformatics*, 23(19):2507–2517, 2007.

[111] ADigital. Uso de twitter en españa 2012, 2013. URL http://www.adigital.org/servicios/uso-de-twitter-en-espana-2012. [Online; accessed 1-November-2014].

[112] Christopher Smith, Afra Mashhadi, and Licia Capra. Ubiquitous sensing for mapping poverty in developing countries. *Paper submitted to the Orange D4D Challenge*, 2013.

[113] Friedrich Schneider, Andreas Buehn, and Claudio E Montenegro. Shadow economies all over the world: New estimates for 162 countries from 1999 to 2007. *Handbook on the shadow economy*, pages 9–77, 2011.

[114] Alex Rutherford, Manuel Cebrian, Sohan Dsouza, Esteban Moro, Alex Pentland, and Iyad Rahwan. Limits of social mobilization. *Proceedings of the National Academy of Sciences*, 110(16):6281–6286, 2013.

[115] Alexander JAM van Deursen and Jan AGM Van Dijk. The digital divide shifts to differences in usage. *New media & society*, 16(3):507–526, 2014.

[116] Alejandro Llorente, Manuel Garcia-Herranz, Manuel Cebrian, and Esteban Moro. Social media fingerprints of unemployment. *PloS one*, 10(5):e0128692, 2015.

[117] What makes spain tweet? http://elpais.com/elpais/2014/11/17/inenglish/1416217012_371379.html, 2014.

[118] Big data hasta para predecir el paro. http://www.expansion.com/empresas/tecnologia/2015/10/14/561e6a0822601de8768b4571.html, 2015.

[119] Los datos no mienten: donde se tuitea con más faltas de ortografía, hay más paro. http://www.elconfidencial.com/tecnologia/2014-11-14/los-datos-no-mienten-donde-se-tuitea-con-mas-faltas-de-ortografia-hay-mas-paro_455021/, 2014.

[120] Los viajes de un día en españa, resumido en menos de 2 minutos. http://www.europapress.es/sociedad/noticia-viajes-dia-espana-resumidos-video-menos-minutos-20141113183732.html, 2015.

[121] El nuevo mapa de españa según twitter. http://www.lavozdegalicia.es/video/vidadigital/2014/11/13/nuevo-mapa-espana-segun-twitter/00311415898851478818412.htm, 2014.

[122] Cómo la gente usa twitter revela datos claves de la economía. http://www.clarin.com/sociedad/Twitter-revela-datos-claves-economia_0_1248475442.html, 2014.

[123] Interesting research using twitter; the unemployed get up late and can't spell. http://www.forbes.com/sites/timworstall/2014/11/22/interesting-research-using-twitter-the-unemployed-get-up-late-and-cant-spell/#60cef5f7264e, 2014.

[124] Twitter "exhaust" reveals patterns of unemployment. https://www.technologyreview.com/s/532746/twitter-exhaust-reveals-patterns-of-unemployment/, 2014.

[125] Tweets tell whether you have a job. http://www.informationweek.com/government/big-data-analytics/tweets-tell-whether-you-have-a-job/d/d-id/1317636, 2014.

[126] What twitter can tell us about unemployment. http://www.citylab.com/work/2014/11/what-twitter-tells-us-about-unemployment/382840/, 2014.

[127] You can use twitter activity to track unemployment. http://www.engadget.com/2014/11/21/twitter-activity-unemployment-tracking/, 2014.

[128] Là où le chômage est fort, les utilisateurs de twitter sont plus nombreux. http://lexpansion.lexpress.fr/high-tech/la-ou-le-chomage-est-fort-les-utilisateurs-de-twitter-sont-plus-nombreux_1623535.html, 2014.

[129] Stanley Milgram. The experience of living in cities. *Crowding and behavior*, 167:41, 1974.

[130] Peter Crane and Ann Kinzig. Nature in the metropolis. *Science (New York, NY)*, 308(5726):1225–1225, 2005.

[131] Alain Barrat, Marc Barthelemy, and Alessandro Vespignani. *Dynamical processes on complex networks*. Cambridge University Press, 2008.

[132] Alexei Vázquez, Joao Gama Oliveira, Zoltán Dezsö, Kwang-Il Goh, Imre Kondor, and Albert-László Barabási. Modeling bursts and heavy tails in human dynamics. *Physical Review E*, 73(3):036127, 2006.

[133] Diego Rybski, Sergey V Buldyrev, Shlomo Havlin, Fredrik Liljeros, and Hernán A Makse. Scaling laws of human interaction activity. *Proceedings of the National Academy of Sciences*, 106(31):12640–12645, 2009.

[134] Gueorgi Kossinets and Duncan J Watts. Empirical analysis of an evolving social network. *science*, 311(5757):88–90, 2006.

[135] Márton Karsai, Mikko Kivelä, Raj Kumar Pan, Kimmo Kaski, János Kertész, A-L Barabási, and Jari Saramäki. Small but slow world: How network topology and burstiness slow down spreading. *Physical Review E*, 83(2):025102, 2011.

[136] Alexei Vazquez, Balazs Racz, Andras Lukacs, and Albert-Laszlo Barabasi. Impact of non-poissonian activity patterns on spreading processes. *Physical review letters*, 98(15):158702, 2007.

[137] Aurélien Gautreau, Alain Barrat, and Marc Barthelemy. Global disease spread: statistics and estimation of arrival times. *Journal of theoretical biology*, 251(3): 509–522, 2008.

[138] Vittoria Colizza and Alessandro Vespignani. Invasion threshold in heterogeneous metapopulation networks. *Physical Review Letters*, 99(14):148701, 2007.

[139] Marc Barthélemy, Alain Barrat, Romualdo Pastor-Satorras, and Alessandro Vespignani. Dynamical patterns of epidemic outbreaks in complex heterogeneous networks. *Journal of theoretical biology*, 235(2):275–288, 2005.

[140] Jure Leskovec, Mary McGlohon, Christos Faloutsos, Natalie S Glance, and Matthew Hurst. Patterns of cascading behavior in large blog graphs. In *SDM*, volume 7, pages 551–556. SIAM, 2007.

[141] Stefan Stieglitz and Linh Dang-Xuan. Emotions and information diffusion in social media—sentiment of microblogs and sharing behavior. *Journal of Management Information Systems*, 29(4):217–248, 2013.

[142] Seth A Myers, Chenguang Zhu, and Jure Leskovec. Information diffusion and external influence in networks. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 33–41. ACM, 2012.

[143] José Luis Iribarren and Esteban Moro. Branching dynamics of viral information spreading. *Physical Review E*, 84(4):046116, 2011.

[144] José Luis Iribarren and Esteban Moro. Affinity paths and information diffusion in social networks. *Social networks*, 33(2):134–142, 2011.

[145] Manuel Cebrian, Lorenzo Coviello, Andrea Vattani, and Panagiotis Voulgaris. Finding red balloons with split contracts: robustness to individuals' selfishness. In *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*, pages 775–788. ACM, 2012.

[146] William O Kermack and Anderson G McKendrick. A contribution to the mathematical theory of epidemics. In *Proceedings of the Royal Society of London A: mathematical, physical and engineering sciences*, volume 115, pages 700–721. The Royal Society, 1927.

[147] Norman TJ Bailey et al. *The mathematical theory of infectious diseases and its applications.* Charles Griffin & Company Ltd, 5a Crendon Street, High Wycombe, Bucks HP13 6LE., 1975.

[148] Roy M Anderson, Robert M May, and B Anderson. *Infectious diseases of humans: dynamics and control*, volume 28. Wiley Online Library, 1992.

[149] WHO Ebola Response Team et al. Ebola virus disease in west africa—the first 9 months of the epidemic and forward projections. *N Engl J Med*, 371(16):1481–95, 2014.

[150] Stefano Merler, Marco Ajelli, Laura Fumanelli, Marcelo FC Gomes, Ana Pastore y Piontti, Luca Rossi, Dennis L Chao, Ira M Longini, M Elizabeth Halloran, and Alessandro Vespignani. Spatiotemporal spread of the 2014 outbreak of ebola virus disease in liberia and the effectiveness of non-pharmaceutical interventions: a computational modelling analysis. *The Lancet Infectious Diseases*, 15(2):204–211, 2015.

[151] Andrea Apolloni, Chiara Poletto, and Vittoria Colizza. Age-specific contacts and travel patterns in the spatial spread of 2009 h1n1 influenza pandemic. *BMC infectious diseases*, 13(1):176, 2013.

[152] Caroline O Buckee, Amy Wesolowski, Nathan N Eagle, Elsa Hansen, and Robert W Snow. Mobile phones and malaria: modeling human and parasite travel. *Travel medicine and infectious disease*, 11(1):15–22, 2013.

[153] Albert-László Barabási. *Bursts: the hidden patterns behind everything we do, from your e-mail to bloody crusades.* Penguin, 2010.

[154] Petter Holme and Jari Saramäki. Temporal networks. *Physics reports*, 519(3): 97–125, 2012.

[155] George A Akerlof. Social distance and social decisions. *Econometrica: Journal of the Econometric Society*, pages 1005–1027, 1997.

[156] Mark S Granovetter. The strength of weak ties. *American journal of sociology*, pages 1360–1380, 1973.

[157] Gautier Krings, Francesco Calabrese, Carlo Ratti, and Vincent D Blondel. Urban gravity: a model for inter-city telecommunication flows. *Journal of Statistical Mechanics: Theory and Experiment*, 2009(07):L07003, 2009.

[158] John R Hipp and Andrew J Perrin. The simultaneous effect of social distance and physical distance on the formation of neighborhood ties. *City & Community*, 8 (1):5–25, 2009.

[159] A Arenas, J Borge-Holthoefer, S Meloni, Y Moreno, et al. Discrete-time markov chain approach to contact-based disease spreading in complex networks. *EPL (Europhysics Letters)*, 89(3):38009, 2010.

[160] Yamir Moreno and Alexei Vazquez. Disease spreading in structured scale-free networks. *The European Physical Journal B-Condensed Matter and Complex Systems*, 31(2):265–271, 2003.

[161] Renaud Lambiotte, Vincent D Blondel, Cristobald De Kerchove, Etienne Huens, Christophe Prieur, Zbigniew Smoreda, and Paul Van Dooren. Geographical dispersal of mobile communication networks. *Physica A: Statistical Mechanics and its Applications*, 387(21):5317–5325, 2008.

[162] Hernán D Rozenfeld, Diego Rybski, José S Andrade, Michael Batty, H Eugene Stanley, and Hernán A Makse. Laws of population growth. *Proceedings of the National Academy of Sciences*, 105(48):18702–18707, 2008.

[163] Rosario N Mantegna and H Eugene Stanley. *Introduction to econophysics: correlations and complexity in finance*. Cambridge university press, 1999.

[164] Markus Schläpfer, Luís MA Bettencourt, Sébastian Grauwin, Mathias Raschke, Rob Claxton, Zbigniew Smoreda, Geoffrey B West, and Carlo Ratti. The scaling of human interactions with city size. *Journal of the Royal Society Interface*, 11 (98):20130789, 2014.

[165] Vincent D Blondel, Adeline Decuyper, and Gautier Krings. A survey of results on mobile phone datasets analysis. *EPJ Data Science*, 4(1):1–55, 2015.

[166] Pierre Deville, Chaoming Song, Nathan Eagle, Vincent D. Blondel, Albert-László Barabási, and Dashun Wang. Scaling identity connects human mobility and social interactions. *Proceedings of the National Academy of Sciences*, 2016. doi: 10.1073/pnas.1525443113. URL http://www.pnas.org/content/early/2016/06/01/1525443113.abstract.

[167] Wei Pan, Gourab Ghoshal, Coco Krumme, Manuel Cebrian, and Alex Pentland. Urban characteristics attributable to density-driven tie formation. *Nature communications*, 4, 2013.

[168] Alejandro Llorente, Giovanna Miritello, Manuel Cebrian, and Esteban Moro. Pace of change in urban social networks. *Submitted for publication to Scientific Reports*, 2016.

[169] Sven Erlander and Neil F Stewart. *The gravity model in transportation analysis: theory and extensions*, volume 3. Vsp, 1990.

[170] Anastasios Noulas, Salvatore Scellato, Renaud Lambiotte, Massimiliano Pontil, and Cecilia Mascolo. A tale of many cities: universal patterns in human urban mobility. *PloS one*, 7(5):e37027, 2012.

[171] Salvatore Scellato, Anastasios Noulas, and Cecilia Mascolo. Exploiting place features in link prediction on location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1046–1054. ACM, 2011.

[172] Giovanna Miritello. *Temporal patterns of communication in social networks*. Springer Science & Business Media, 2013.