UNIVERSIDAD POLITÉCNICA DE MADRID
ESCUELA TÉCNICA SUPERIOR DE INGENIEROS AGRÓNOMOS

APLICACIÓN DE LA TEORÍA DE REDES COMPLEJAS A PROCESOS DINÁMICOS EN LA SOCIEDAD

Francisco Javier Borondo Benito

Licenciado en CC Físicas
Master en Física de los Sistemas Complejos

TESIS DOCTORAL

Marzo de 2015

ESCUELA TÉCNICA SUPERIOR DE INGENIEROS AGRÓNOMOS

APLICACIÓN DE LA TEORÍA DE REDES COMPLEJAS A PROCESOS DINÁMICOS EN LA SOCIEDAD

Francisco Javier Borondo Benito

Licenciado en CC Físicas
Master en Física de los Sistemas Complejos

Director:
Juan Carlos Losada González
Doctor en CC Físicas

Marzo de 2015

Abstract

Society can be defined as a complex system that emerges from the cooperation and coordination of billions of individuals and hundreds of countries. Thus, we do not live in social vacuum and the social networks in which we are embedded inevitably shapes our behavior.

Here, we present an analytical model and several empirical studies in which we analyze dynamical social systems through a network science perspective. First, we introduce a model to explore how the structure of the social networks underlying society can limit the meritocracy of the economies. Conversely to meritocracy, in this work we introduce the term *topocracy*. We say that a system is topocratic if the compensation and power available to an individual is determined primarily by her position in a network. Our model is perfectly meritocratic for fully connected networks but becomes topocratic for sparse networks-like the ones in society. In the model, individuals produce and sell content, but also distribute the content produced by others when they belong to the shortest path connecting a buyer and a seller. The production and distribution of content defines two channels of compensation: a meritocratic channel, where individuals are compensated for the content they produce, and a topocratic channel, where individual compensation is based on the number of shortest paths that go through them in the network. We solve the model analytically and show that the distribution of payoffs is meritocratic only if the average degree of the nodes is larger than a root of the total number of nodes. Hence, in the light of our model, the sparsity and structure of networks represents a fundamental constraint to the meritocracy of societies.

Next, we present several empirical studies that use data gathered from Twitter to analyze online human behavioral patterns. In particular, we focus on political conversations such as electoral campaigns. We found that the collective attention is highly heterogeneously distributed, as there is a minority of extremely in uential accounts. In fact, the ability of individuals to propagate messages or ideas through the platform is constrained by the structure of the follower network underlying the social media and the

position they occupy on it. Hence, although people have argued that social media can allow more voices to be heard, our results suggest that Twitter is highly topocratic, as only the minority of well positioned users are widely heard. This minority of in uential accounts belong mostly to politicians and *traditional media*. Politicians tend to be the most mentioned, while media are the sources of information from which people propagate messages.

We also propose a methodology to study and measure the emergence of political polarization from social interactions. To this end, we first propose a model to estimate opinions in which a minority of in uential individuals propagate their opinions through a social network. The result of the model is an opinion probability density function. Next, we propose an index to quantify the extent to which the resulting distribution is polarized. Finally, we illustrate our methodology by applying it to Twitter data.

In a world where personal data is increasingly available, the results of the analytical model introduced in this work can be used to enhance meritocracy and promote policies that help to build more meritocratic societies. Moreover, the results obtained in the latter part, where we have analyzed Twitter, are key to understand the new data-driven society that is emerging. In particular, we have presented relevant information that can be used to benchmark future models for online communication systems or can be used as empirical rules characterizing our online behavior.

Resumen

Podemos definir la sociedad como un sistema complejo que emerge de la cooperación y coordinación de billones de individuos y centenares de países. En este sentido no vivimos en una isla sino que estamos integrados en redes sociales que in uyen en nuestro comportamiento.

En esta tesis doctoral, presentamos un modelo analítico y una serie de estudios empíricos en los que analizamos distintos procesos sociales dinámicos desde una perspectiva de la teoría de redes complejas. En primer lugar, introducimos un modelo para explorar el impacto que las redes sociales en las que vivimos inmersos tienen en la actividad económica que transcurre sobre ellas, y mas concretamente en hasta qué punto la estructura de estas redes puede limitar la meritocracia de una sociedad. Como concepto contrario a meritocracia, en esta tesis, introducimos el término topocracia. Definimos un sistema como topocrático cuando la in uencia o el poder y los ingresos de los individuos vienen principalmente determinados por la posición que ocupan en la red. Nuestro modelo es perfectamente meritocrático para redes completamente conectadas (todos los nodos están enlazados con el resto de nodos). Sin embargo nuestro modelo predice una transición hacia la topocracia a medida que disminuye la densidad de la red, siendo las redes poco densas-como las de la sociedad- topocráticas. En este modelo, los individuos por un lado producen y venden contenidos, pero por otro lado también distribuyen los contenidos producidos por otros individuos mediando entre comprador y vendedor. La producción y distribución de contenidos definen dos medios por los que los individuos reciben ingresos. El primero de ellos es meritocrático, ya que los individuos ingresan de acuerdo a lo que producen. Por el contrario el segundo es topocrático, ya que los individuos son compensados de acuerdo al número de cadenas mas cortas de la red que pasan a través de ellos. En esta tesis resolvemos el modelo computacional y analíticamente. Los resultados indican que un sistema es meritocrático solamente si la conectividad media de los individuos es mayor que una raíz del número de individuos que hay en el sistema. Por tanto, a la luz de nuestros resultados la estructura de la red social puede representar una limitación para la meritocracia de una

sociedad.

En la segunda parte de esta tesis se presentan una serie de estudios empíricos en los que se analizan datos extraídos de la red social Twitter para caracterizar y modelar el comportamiento humano. En particular, nos centramos en analizar conversaciones políticas, como las que tienen lugar durante campañas electorales. Nuestros resultados indican que la atención colectiva está distribuida de una forma muy heterogénea, con una minoría de cuentas extremadamente in uyente. Además, la capacidad de los individuos para diseminar información en Twitter está limitada por la estructura y la posición que ocupan en la red de seguidores. Por tanto, de acuerdo a nuestras observaciones las redes sociales de Internet no posibilitan que la mayoría sea escuchada por la mayoría. De hecho, nuestros resultados implican que Twitter es topocrático, ya que únicamente una minoría de cuentas ubicadas en posiciones privilegiadas en la red de seguidores consiguen que sus mensajes se expandan por toda la red social. En conversaciones políticas, esta minoría de cuentas in uyentes se compone principalmente de políticos y medios de comunicación. Los políticos son los mas mencionados ya que la gente les dirige y se refiere a ellos en sus tweets. Mientras que los medios de comunicación son las fuentes desde las que la gente propaga información.

En un mundo en el que los datos personales quedan registrados y son cada día mas abundantes y precisos, los resultados del modelo presentado en esta tesis pueden ser usados para fomentar medidas que promuevan la meritocracia. Además, los resultados de los estudios empíricos sobre Twitter que se presentan en la segunda parte de esta tesis son de vital importancia para entender la nueva "sociedad digital" que emerge. En concreto hemos presentado resultados relevantes que caracterizan el comportamiento humano en Internet y que pueden ser usados para crear futuros modelos.

Contents

1. Introduction **13**

2. Network science **19**
2.1. Introduction 19
2.1.1. Network science in the century of complexity 21
2.1.2. Characteristics of network science 23
2.2. Networks and graphs 23
2.2.1. Choosing the proper representation 24
2.3. Degree & degree distribution 24
2.3.1. Degree and average degree 24
2.3.2. Degree distribution 25
2.4. Adjacency matrix 26
2.5. Paths and distances in networks 26
2.6. Clustering 28
2.7. Centrality Measures 29
2.7.1. Betweeness Centrality 31
2.7.2. Closeness centrality 31
2.7.3. Eigenvector centrality 31
2.7.4. Pagerank centrality 32
2.8. Degree correlations and the friendship paradox 33
2.8.1. The friendship paradox 35
2.9. Community Structure 36
2.9.1. Mapequation 38
2.10. Network models: From random to scale-free networks 39
2.10.1. Random Networks 39
2.10.2. We live in a small-world 41
2.10.3. Real networks are scale-free 42
2.10.4. The Barabási-Albert model 44

3. The Meritocracy and topocracy of embedded markets 47
3.1. Introduction . . . 47
3.2. Modelling a networked market . . . 50
3.2.1. The Model . . . 50
3.2.2. Meritocratic and topocratic Regimes - Limit Cases . . . 51
3.3. Transition threshold . . . 55
3.3.1. General case . . . 55
3.3.2. Alternative sharing rule: comissions . . . 57
3.4. Payoff Distributions . . . 59
3.5. The statistical Meritocracy and Topocracy of Networks . . . 62
3.6. The scale-free case . . . 63
3.6.1. The meritocracy of SF networks . . . 65
3.6.2. Payoff distribution . . . 66
3.7. Generalization To Arbitrary Prices . . . 67
3.8. Discussion . . . 67

4. Twitter: An online social network 71
4.1. Introduction . . . 71
4.2. Twitter as a multilayer social network . . . 72
4.2.1. Follower Layer . . . 74
4.2.2. Mention Layer . . . 75
4.2.3. Retweet Layer . . . 75
4.3. Datasets . . . 76
4.3.1. Data gathering . . . 76
4.3.2. Datasets . . . 77
4.4. Twitter data limitations . . . 79

5. Twitter and its predictive power 81
5.1. Introduction . . . 81
5.2. Literature review . . . 82
5.3. 20N Time Series . . . 83
5.4. Further evidence . . . 89

6. Mapping the online Spanish political landscape 95
6.1. Introduction . . . 95
6.2. Network Properties . . . 97
6.2.1. The $20N$ follower network . . . 97
6.2.2. $20N$ User Interactions . . . 100
6.3. Information spreading on Twitter: Where a minority rule . . . 104
6.3.1. Universality . . . 108
6.3.2. The Model . . . 111

6.3.3. Results . . . 112
6.4. Community Structure . . . 118
6.5. Statistical properties of the c-networks . . . 120
6.6. Politicians, the main characters; Traditional media, still the main source of information . . . 122
6.7. Twitter as a multilayer social network . . . 124
6.7.1. Is Twitter a rich-club like social media? . . . 124
6.7.2. Multiple leaders emerge at the different layers . . . 125
6.8. Discussion . . . 130

7. The impact of diverse languages of the political landscape 133
7.1. Language detection . . . 133
7.2. Language Polarization . . . 133
7.3. Language Preferences . . . 135
7.4. Discussion . . . 138

8. Characterizing politicians activity 139
8.1. Introduction . . . 139
8.2. Who follows who? . . . 143
8.3. Communication among political parties . . . 143
8.4. Community structure . . . 147
8.5. A model that reproduces political Interactions . . . 147
8.6. Discussion . . . 149

9. Measuring political polarization 151
9.1. Introduction . . . 151
9.2. Estimating Opinions . . . 153
9.3. Introducing a new measure of polarization in opinion distributions: the polarization index . . . 155
9.4. Twitter data: The Venezuelan case . . . 157
9.5. Twitter shows the two sides of Venezuela . . . 161
9.6. Conclusions . . . 163
9.7. Additional Methods: Networks . . . 165

10. Conclusions 171

A. The Spanish political system 179

Bibliography 179
*

Preface

In this dissertation we present the main findings of the research that I conducted as a Complex Systems graduate student at Grupo de Sistemas Complejos in the Universidad Politecnica de Madrid. This Dissertation has been formatted to satisfy the requirements of the Universidad Politecnica de Madrid and is eligible for International PhD.

Best Wishes

Acknowledgements

I would like to thank Juan Carlos Losada, for accepting me as a PhD student, having confidence in me from the very beginning, and for guiding me, while giving me freedom to conduct the research projects. I also want to thank Juancar for the numerous discussions, advices and the time spent discussing several different scientific and life issues with me. All in all, without Juancar's support I would not have been able to have such a great experience during this PhD

I also want to acknowledge all the members of Grupo de Sistemas Complejos (GSC) for their support. In particular I would like to acknowledge the head of the group Rosa Maria Benito and the professors: Ana Tarquis, Javier Galeano, Florentino Borondo, Javier Arranz, Ramon Alonso, Miguel A. Porras, Fabio Revuelta, Luis Seidel, Mary Luz Mouronte, Carlos Mejía and Juan Manuel Pastor. I also want to thank my colleagues Alfredo Morales, Izaskun Oregui, Pedro Benitez, Henar Hernandez, Samuel Martinez, Juanjo Martin Sotoca, Ivan Gonzalez, Johan Martinez and Maxi Fernandez.

I also would like to thank Cesar Hidalgo for being my advisor at M.I.T., and for the numerous discussions and advices. He has greatly inspired my research and many of the contributions that I present on this thesis. I also want to thank all the members of his group, Macro Connections, for the great time at the Media Lab and for the numerous inspiring discussions. Finally, I also acknowledge Consejo Social-UPM for the grant I received to fund my research stay at M.I.T.

Chapter 1

Introduction

In this document we present our contribution to a few problems in the field of network science and its applicability to enhance our knowledge about society. Network science, an interdisciplinary field by nature, allows us to study an infinite variety of complex systems by abstracting them into networks composed by nodes and links. One among these many systems is society. Society, in a simplistic way can be defined as a group of people who share a defined territory and a culture. However, sociology takes this definition further by arguing that a society is also the relationships between the people and the institutions within that group. In fact, one of the classic questions in sociology is how our behavior is affected by our social relations.

From a network science perspective we can study society as a complex system that emerges from the cooperation and coordination of billions of individuals and hundreds of countries. Hence, in this thesis we abstract society as a network to improve our understanding about how the social networks in which we are embedded shape our behavior. By doing so, we are able to take advantage of network science to analyze dynamical social systems through a mathematical and computational perspective.

Stating that we live in society implies that we are embedded in the social networks that form our society. This issue is vital to understand the dynamical processes that occur within society, such as economic activity. However, the consequences of embeddedness have frequently been neglected by economic theories. Neoclassical economic theories take an atomized approach, in which individual's choices and actions are generated independently of the actions and expected behavior of other actors. However, as Granovetter [Gra85] pointed out more than two decades ago, our economy is embedded in social networks. These networks beget commercial interactions, and are begot by them. For Granovetter, the cultivation of personal relationships between traders and customers assumes an equal or higher importance than the

economic transactions involved. In fact, economic exchanges are not carried out exclusively among strangers, but often incorporate individuals involved in long-term continuing relationships.

In this document we take Granovetter's approach to society, assuming that most of our behavior is closely embedded in the networks of interpersonal relations that are part of society. Such approach largely differs from an atomized approach. To illustrate the conceptual difference between them let's consider the following situations:

1) Juan is the most talented guitar player in the whole US, and CA/DA is the most promising rock band of the country. However, Juan and CA/DA do not know each other and do not have friends in common.

2) Juan is a reasonably good guitar player, while CA/DA is a promising rock band in the country. Juan does not know any of the members of CA/DA, however, Juan's music teacher is the agent CA/DA.

Under which of the two situations is it more probable that Juan will join CA/DA? If we take an atomized approach, option 1 is surely the most likely to occur. According to neoclassical approaches we decide as atoms outside a social context. Hence, the fact that the band is connected to Juan through their agent is not relevant. In contrast, we assume that both parts have perfect information and since option one is the operation that creates the highest surplus it is the most probable to happen.

On the other hand, if we take Granovetter's approach, it is much more likely that Juan will join the band under the second situation. The embeddedness argument stresses the role of personal relations and the structure of the social network in generating trust. Thus, in the second situation the fact that Juan's guitar teacher is the agent of the band would be the decisive factor. This common neighbor will increase Juan's confidence on the band and vice-versa, enhancing the chances that Juan joins the band. The explanation is straightforward, we prefer to deal and trust with individuals with a high reputation. Under Granovetter's view we build the reputation of other individuals as a function of wether ourselves or our own contacts have had satisfactory dealings with them in the past, preferentially not relying on general reputations.

Inspired by the idea of embeddedness one of the goals of this thesis is to explore how the structure of the social network can limit the meritocracy of societies. To achieve this objective, we analyze the redistributive consequences of the networks underlying economic activity by introducing a model with tunable embeddedness. By taking advantage of network science, we are able to explore the relationship between the structure of social networks and the meritocracy of the economies that are embedded in them. We focus on exploring the mechanics of a model that help us elucidate the conditions un-

der which a society would be meritocratic or not. Our model is described as follows. Consider a world where each one of us has two sources of income. One, that is associated with our ability to produce content, and another one that is associated with our ability to intermediate transactions. Now assume that people differ in their ability to produce content. Some people, like Steve Jobs and John Lennon, have the magic ability to produce content that everyone likes. Others are not that lucky. Now, assume that people live in a network, and that they can transact directly only with whom they have a link. To transact with those that they are not connected, they need a path, and this path is provided by connections between other members of the network who collect a percentage of the transaction.

We show that the meritocracy of the model defined as the compensation that individuals receive based on their contributions decays as the network becomes sparse, giving rise to a topocratic regime, in which the compensation received by individuals is explained primarily by the position they occupy in the network. We show that the structure and connectivity of the networks where markets are embedded represent the main factor determining whether the system is in a topocratic or meritocratic regime.The results imply that theories that assume away the existence of networks are implicitly assuming away the possibility that markets can be non-meritocratic.

Understanding the redistributive consequences of networks is important in a world where markets are composed of a mix of socially embedded links and commercial arm-length relationships [Uzz96, Uzz97]. Yet, even in a world where arm-length relationships are dominant, the assumption of fully connected networks is too hopeful. Possible transactions might not take place because individuals are uncertain about the quality of the goods being offered [Big93, LM99, AA70] or due to search frictions [RW87, Won11]. These market failures are partially compensated by the emergence of middlemen who are experts at reducing information asymmetries and search frictions, but who also act as hubs controlling information ows in the network. As Ronald Burt points out, the position that middleman occupy in the network is a source of advantage, as intermediating positions constitute part of what he has termed the social capital of structural holes [Bur09, Bur04].

The fact that in our model the transition between meritocracy and topocracy depends predominantly on the density of the network has an important implication. The model predicts that meritocracy increases in societies that become better connected. This is an important implication given current changes in technology. Recent changes in communication technologies have increased the connectivity of our society, by reducing the cost of both social and commercial interactions. Most studies have emphasized the role of communication technologies on social participation and collaboration. Our

results suggests that this technological change might also have important long term effects on the meritocracy of economies. Content producers, whether these are musicians or artists, can now market their content directly to a large number of individuals, even though this causes an information overload [Sim71, Gle11] that puts us far from the idealized limit of fully connected networks. Nevertheless, in the light of this model, changes in communication technology should increase the meritocracy of markets when holding population size constant. So the good news is that recent changes in technology should help make our society more meritocratic.

The consequences of embeddedness do not limit to economic activity. In contrast it significantly contributes to how we can understand a variety of social behaviors in a wide range of contexts. The key premise of this theory is that individuals behave in uenced by the likely choices of others, meaning that their social connections have an impact on their behavior. The social network in which we are embedded also affects our opinions, political views or voting behavior. In fact, opinions are not formed in a social vacuum, but on a social network where those around us affect what we think. A classic topic in political science is how social interactions shape individuals political views, and whether the social network in which they are embedded has an in uence on their voting behavior [HS95] [Kno90] [BLM54] [Huc09].

A limitation of classic sociological studies has been the unavailability of large datasets to analyze the structure of the social networks in which societies are embedded. However, nowadays the availability of human driven digital traces, such the records of telephone calls or the data gathered from online social networks, represents an opportunity to reproduce and characterize the social networks in which we are embedded.

Another goal of this thesis is to use data gathered from Twitter to proxy social networks and analyze online political conversations to gain understanding on how our opinions are shaped. Twitter features several interaction mechanisms to facilitate the communication among users. These mechanisms establish different layers through which users can communicate and exchange information. Hence, Twitter can be seen as a multiplex or multilayer social network [BBC+14] composed by the follower, mention and retweet layers. The first interaction mechanism, is the ability of people to follow and be followed by the rest of users. This mechanism is a passive mechanism that allows users to receive the messages written by their followees at real time. By the same token they automatically deliver their posted messages to their followers. Thus, this mechanism establishes the followers layer, where users are connected among each other, according to who follows who. The links at this layer establish the substratum through which messages are delivered. We analyzed how information travels among users through Twitter, and fol-

lowing the concept of embeddednes users tend to retransmit messages from those who they follow. Hence, if Paco to whom I am not connected posts a tweet announcing that Fernando Alonso leaves Ferrari I will probably not retweet it. However, if Sara who is connected to both of us retweets the message it is much more probable that I will do the same. Thus, we found that mostly we retweet those users to whom we are linked in the followers network, or those to whom we are not directly connected when their message reaches us via a common neighbor. Moreover, on this document we show how the structure of the network in which we are embedded on Twitter will determine our ability to efficiently propagate messages.

We also explore the structure of the online social networks in which individuals are embedded when discussing politics. This is an increasingly relevant topic since online social networks and social media platforms, such as Twitter, are the latest new medium being exploited by politicians for decisive competitive advantage. Thus, today's culture is changing, Internet and social media represent a new channel through which information and ideas can quickly ow [LPA+09], bringing people a wider (and cheaper) variety of information. Today's new culture sees value in sharing information, and relies on collective wisdom [AZBA08]; just take Wikipedia [KR11] as an example. Social media fit perfectly this new context as they are about listening and being heard, about sharing information with those you trust, and about having a variety of sources of information at hand from where to choose. So, nowadays, when trying to understand the opinion formation process of individuals, we have to take into account not only their face to face relations or the propaganda coming from traditional mass media, but also the online communications that are increasingly taking place through social media platforms such as Twitter. In fact, recent research has brought evidence to show that political mobilizations in an online social network can in uence real world voting behavior [BFJ+12].

In this document we combine the sociological concept of embeddedness developed by Granovetter with network science to gain understanding of our society. By doing so we have been able to propose and solve a model that relates the structure of the social network in which a society is embedded and the meritocracy of its economy. Due to the mathematical formalism of network science we have solved the model analytically and calculated the network density threshold in which a society transits from meritocracy to topocracy. Simple calculations using the equations derived from the model suggest that our society is likely to be topocratic. Next, by gathering data from the online social network Twitter, we map social relations. We introduce a model that we use to study the impact that the structure of the twitter followers network has on the ability of users to spread messages. The model

shows that we will more likely adopt the ideas of those closer to us in the network, and that the position that we occupy in the network conditions our efficiency to propagate information. Finally, we show that politicians and traditional mass media compound the elite that rules political conversations on Twitter.

Chapter 2

Network science

2.1. Introduction

In this chapter we introduce the basic concepts of network science. To introduce the chapter we will start explaining how the U.S. Army captured Saddam Hussein and explaining how this example illustrates many insights of network science. The example is illustrated on Figure 2.1.

In March 19, 2003 the american forces started the invasion of Iraq. They found little military resistance and quickly took control of the country. However, Saddam Hussein and many of his officials initially escaped and avoided being captured. In order to capture Saddam and the regime high ranking officials the coalition forces designed a deck of cards, where each card corresponded to one of the most wanted officials. The strategy initially worked, as by the end of May over 25 officials were captured or under custody. Yet the whereabouts of Hussein the ace of spades remained unknown.

The intelligence officials thought that Hussein's high ranking officials would know where he was hiding. However, with the capture of his trusted personal secretary it became obvious that it was not the case. Hence, relying on the regime lines of power was of little help to find Hussein's whereabouts. Instead, they arrived by common sense and intuition to use network theory to solve the problem. Col. James Hickey, in charge of a series of raids known as Operation Desert Scorpion, decided they had to find out the relationships between everyone who was killed or captured. One of the officers in charge of this task, Brian Reed, had some knowledge on social networks. Thus, he reconstructed the social network surrounding Saddam Hussein by linking people according to gossip and family trees rather than government documents. Using these network diagrams to guide the raids gave instant results as they found millions of dollars, weaponry and Saddam's family photo album. Yet

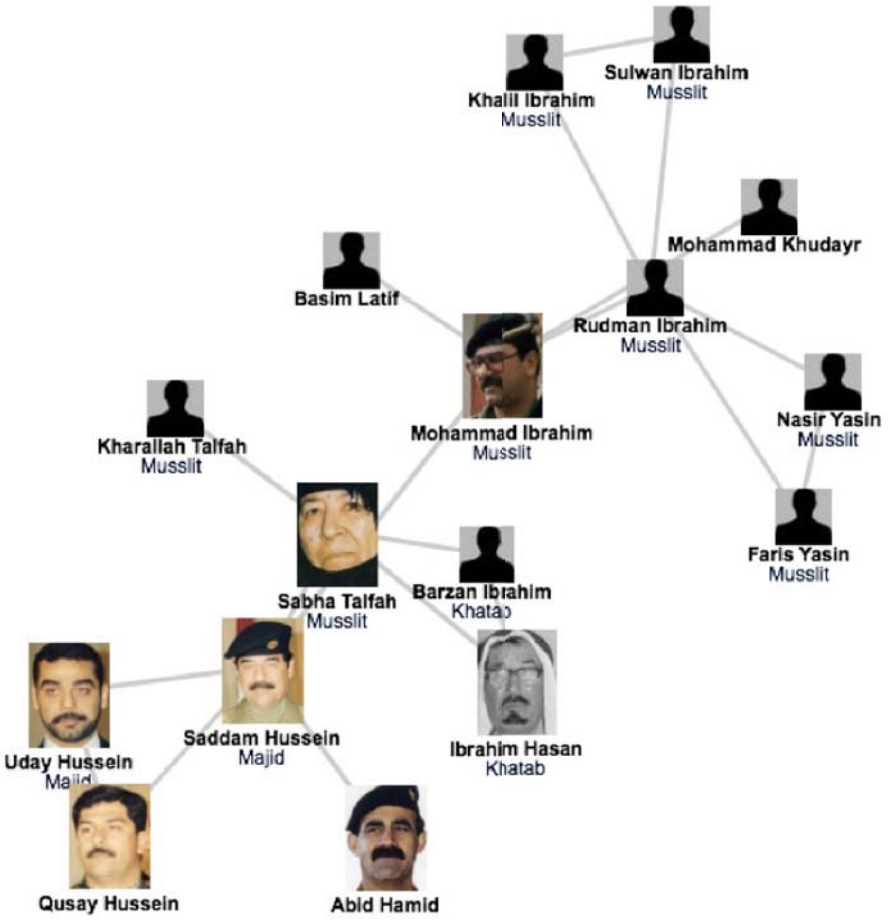

Figure 2.1: Visualization of Saddam Hussein social network. Obtained from[Bar12].

the most valuable among the three was the photo album that provided faces of those that the dictator and his family really trusted. This information helped to significantly improve the network diagrams of Saddam's social network, pointing to two key men, Rudman Ibrahim and Mohammed Ibrahim. These two individuals were Saddam's second-level bodyguards, serving as his driver, cook, or mechanic. Rodman was of little help as he died from a heart attack a few hours after being captured. However, Mohammad revealed the hideout of Hussein.

The way in which the military forces captured Saddam Hussein was mainly based on intuition and guesswork, rather than on real network science. However, the story illustrates the power and simplicity of network science and many concepts related to the discipline that we will discuss in the latter part of this chapter. The first issue that this story re ects is the predictive power of networks. Even non experts in the field such as soldiers were capable of interpreting the maps of Saddam's social network to spot the key individuals that revealed the dictators whereabouts. It also shows the stability of networks, it was the pre-invasion social ties that lead to finding him. Moreover, it also illustrates the importance of choosing the right network abstraction of the system under study. The military forces initially built a map according to military or governmental relationship that was of little help to find the dictator. In contrast, using social relationships to build the diagram was much more helpful. Finally, this example also re ects the difficulty of building precise maps that provide high resolution information.

2.1.1. Network science in the century of complexity

In words of Stephen Hawking: "I think the next century will be the century of complexity". Thus, let's start by defining the word complex. According to the dictionary[1], there are three possible definitions: 1) composed of many interconnected parts; compound; composite: a complex highway system. 2) characterized by a very complicated or involved arrangement of parts, units, etc.: complex machinery. 3) so complicated or intricate as to be hard to understand or deal with: a complex problem.

Thus, we live in a world where we are surrounded by systems that respond to this definition. In fact, our society is a complex system that results from the cooperation and coordination of billions of individuals and hundreds of countries. Each of us can be also seen as a complex system. Humans are able to reason or coordinate movements thanks to the activity of billions of neurons in our brain that interact between them. The current technological

[1]Dictionary.com

systems are also complex, as the correct functioning of the air transportation system requires the connection and coordination among thousands of airports that operate over 100.000 ights per day.

Network science can be considered as the science of the 21^{st} century, as behind each of these complex systems there is a complex network that represents each system by abstracting it into nodes and links that represent interactions between them. For example, the brain can be abstracted into a network where each neuron is represented as a node and links represent the interaction among neurons. In the same way, we can abstract the power grid as a network where the generators are the nodes and the transmission lines are abstracted as links. To fully understand these complex systems that surround us, we need to gain a deep understanding of the networks behind them. Moreover, networks are the architecture behind many revolutionary technologies or products such as Facebook, Google, Twitter or Cisco. The idea behind Facebook is to give online support to our social network. Hence, Facebook uses the web in order to connect us to our friends, so that we can communicate, and share news or pictures with them . In addition to this, it employs network based algorithms to suggest us new friends or applications we might like.

However, graph theory, is a prolific subfield of mathematics, that focuses on the analysis of networks since 1735. Hence, why does network science emerge in the 21^{st} century rather than 200 years ago? What happens in this century to justify this enhancement of network science 200 years after the appearance of graph theory?

One may think it was the appearance of new systems to study, such as the cell-phone calls network, that caused this explosion of network science. Yet it was not the case as most of the systems under study are by no means new. For example, metabolic networks date back to the origins of life, sociologists have been studying social networks for decades and the Internet is now over four decades old. In contrast the reason for this explosion of network science is availability of large amounts of data. In the past our ability to build accurate maps was limited by the absence of high resolution data. Hence, our lately acquired ability to store and share data prompted the evolution of network science. Nowadays Internet and the emergence of cheap digital storage have enhanced the storage and share of data, permitting its analysis in the shape of networks.

Another key issue that has promoted network science is that, despite most complex systems largely differ in their nature goals and scope, after abstracting them as a network they share many properties. The differences among these systems are obvious, as for example in a social context nodes are humans and links professional, friendship or intimate relationships, while in

the metabolic network nodes represent tiny molecules and links are chemical reactions. Given the diversity of the systems studied by network science, one would expect that their properties will largely differ. However, one of the most important discoveries of network science was that this is not the case. Though, the network emerging from diverse systems are rather similar to each other, allowing us to use a common formalism and mathematical tools to explore them. Hence, the universality of networks is a fundamental principle that has guided the explosion of network science on the 21^{st} century.

2.1.2. Characteristics of network science

Network science is an interdisciplinary science that offers a common formalism and methodology for very diverse disciplines such as biology, sociology or medicine. A sociologist and a biologist need to characterize their systems, cleaning data and representing them into interpretable maps from which to extract information. Yet, in many cases disciplines face similar challenges, and advances from one discipline can be extrapolated to others. For example, the scale-free property of networks that raised when analyzing the WWW network, plays a crucial role in increasing the robustness of the power grid.

The discipline has a quantitative and mathematical nature that puts together concepts from many other fields. Initially, it inherited the conceptual framework to deal with randomness and universal organizing principles from statistical physics and the formalism to deal with graphs from graph theory. More recently it has been complemented with concepts from control and information sciences, data mining and statistics. However, network science also presents a strong computational character, as one of its main challenges is to deal with large amounts of data.

In contrast to classic graph theory, network science is a data driven discipline that focuses on data and utility. Hence, network science does not limit to develop mathematical or statistical tools, but it also tests them on real data. Thus, the value of the mathematical formulation of network science resides in the value it offers about the systems structure and evolution.

2.2. Networks and graphs

In order to fully understand a complex system, we need a map of its wiring diagram. A network is an abstraction of a system in which we represent its elements as nodes or vertices and the interactions between them as links or edges. Hence, network theory offers a common language to analyze a wide

variety of systems that largely differ in their nature but that share many global properties. These systems include cell-phone calls, the power grid, or protein interactions among many others.

Number of nodes, N, represents the number of elements in the system. We usually refer to N as the size of the network.

Number of links, L, quantifies the number of interactions between the nodes.

The links of a network can be directed or undirected, depending on the system and the nature of the interaction they represent. For example, links accounting for phone calls, where one person calls the other, are directed pointing to the person who receives the call. On the other hand, links that represent romantic relationships are undirected. For a network to be labeled as directed, all of its links must be directed. By the same token, for a network to be labeled as undirected all of its links must be undirected. However, there are also networks, such as the metabolic network, that simultaneously have directed and undirected links.

2.2.1. Choosing the proper representation

When representing a complex system as a network, we will always have many choices at hand. Choosing the right representation for the problem we want to solve will determine our probability of success. Let's consider that the system under study is our country and its population. By connecting individuals that regularly interact with each other in the context of work, we will obtain the profesional network of the country. Alternatively, if we link individuals that have an intime relationship we obtain the sexual network. Hence, in an organizational research context the first representation would be the right choice. In contrast, if we were concerned with the spread of sexually transmitted diseases-like AIDS-the second choice would be the best representation.

2.3. Degree & degree distribution

2.3.1. Degree and average degree

The main property of each node is its degree (k). The degree of a node represents the number of links that the nodes has to other nodes in the network. Thus, in the context of cell-phone calls it quantifies the number of different persons she has talked to.

An important property that characterizes a network is its *average degree*, that for undirected graphs is:

$$\langle k \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i = \frac{2L}{N} \tag{2.1}$$

For directed networks, we can extend the definition of degree by distinguishing between in-degree (k_i^{in}) and out-degree (k_i^{out}). The first one measures the number of links that point to node i, while the second one measures the number of links that point from node i. The sum of a node in and out degree is equal to its degree:

$$k_i = k_i^{in} + k_i^{out} \tag{2.2}$$

Resuming the example of the cell-phone call network, the in-degree of a person represents the number of other persons that have called her, and the out-degree the number of person to whom she has made a call.

The average degree for a directed network can be re-expressed as:

$$\langle k^{in} \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i^{in} = \langle k^{out} \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i^{out} = \frac{L}{N} \tag{2.3}$$

2.3.2. Degree distribution

The degree distribution, $p(k)$, represents the probability of randomly selecting a node with degree k. Hence, in a network of size N, $p(k)$ is given by the following expression:

$$p(k) = \frac{N_k}{N} \tag{2.4}$$

where N_k represents the number of nodes in the network with degree k.

The degree distribution plays a key role in network science, as it is a requirement to calculate many network properties, such as the average degree. Moreover, It became of increasing importance since the discovery of scale-free networks [BA99] and the implications they have in a wide range of phenomena, such as robustness or the spread of deseases.

2.4. Adjacency matrix

A network with N nodes can be represented by a matrix $A = \ A_{ij} \quad 1$ $i \ j \quad N$. Given two nodes, i and j, the element A_{ij} of the matrix represents the number of edges going from the first one to the second one. Note that for undirected graphs the adjacency matrix is symmetric.

The degree of a node can be easily obtained from the adjacency matrix. In the simplest case of an undirected network the degree (k_i) of node i is either the sum over the rows or columns of the matrix:

$$k_i = \sum_{j=1}^{N} A_{ij} = \sum_{i=1}^{N} A_{ij} \tag{2.5}$$

For directed networks we can define the in-degree and out-degree as the sum over rows and columns of the matrix, respectively. Thus, k^{in} and k^{out} can be expressed as:

$$k_i^{in} = \sum_{j=1}^{N} A_{ij} \qquad k_i^{out} = \sum_{i=1}^{N} A_{ij} \tag{2.6}$$

However, most real networks are sparse and only a minority of the elements of the matrix are nonzero. Hence, when storing a large network on a computer, we rarely store its full adjacency matrix. Instead, we store the network as a list of links ($i \ e \ $ elements for which $A_{ij} = 0$), as a huge fraction of the A_{ij} elements are zero.

2.5. Paths and distances in networks

The path in a network runs along its links, thus the path between nodes i and j is any set of adjacent links that connect i to j. The distance of a path is defined as the number of links that the path contains.

The **shortest path** (or geodesic path) between two nodes refers to the path with fewer links that connects the two nodes. The shortest path between nodes i and j is denoted as d_{ij}. Thus, we can define the shortest path length, frequently labeled as distance, as the number of links contained in the shortest path. Note, that there can be more than one shortest paths between two nodes. Figure 2.2 highlights the difference between a path and a shortest path.

For undirected networks the shortest path from node i to j, d_{ij}, is the same as the shortest path from j to i, d_{ji}. However for directed networks d_{ij} is not necessarily equal to d_{ji}. In fact, the existence of d_{ij} does not guarantee the existence of d_{ji}.

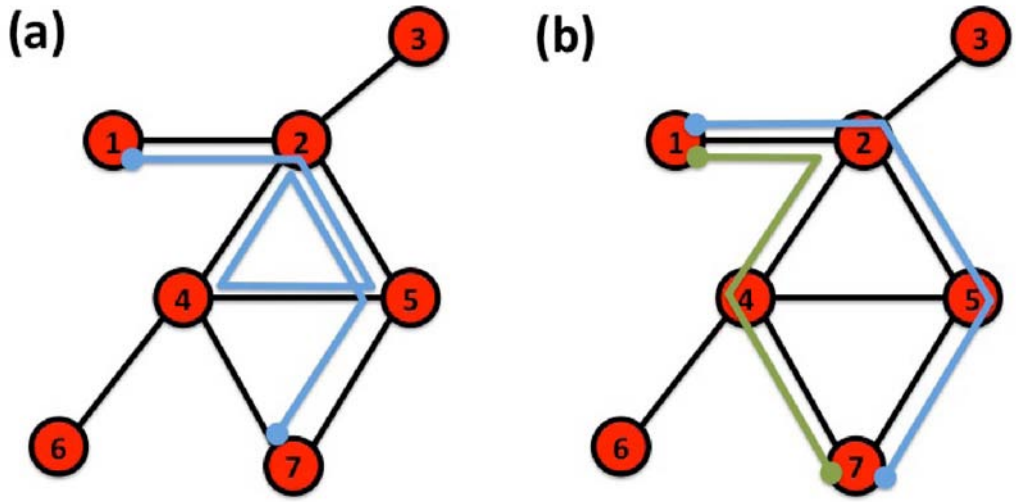

Figure 2.2: (a) Path between nodes 1 and 7. The highlighted path is:1-2-5-4-2-5-7. (b) The shortest path between nodes 1 and 7. There are two possible shortest path of distance three, one marked in blue and another in green. Obtained from[Bar12].

The **network diameter**, d_{max}, is the maximal shortest path length between any pair of nodes in the network.

The **average path length**, $\langle l \rangle$, is the average shortest path length between all pairs of nodes in the network. Hence, to calculate it we need to first compute all the shortest paths in the network and secondly average their distances. For directed networks it is given by the following expression:

$$\langle l \rangle = \frac{1}{N(N-1)} \sum_{i,j=1,N} d_{ij} \tag{2.7}$$

For undirected networks we need to correct the previous equation by introducing a factor of two in the numerator of eq.2.7

2.6. Clustering

The clustering or transitivity of a network measures the total number of closed triangles in a network. In the context of social networks, a high clustering indicates that my friends are also friends. On the other hand, a low clustering means that my friends do not know each other. Random networks, such as the Erdös-Rényi network, do not present a clustering structure, while

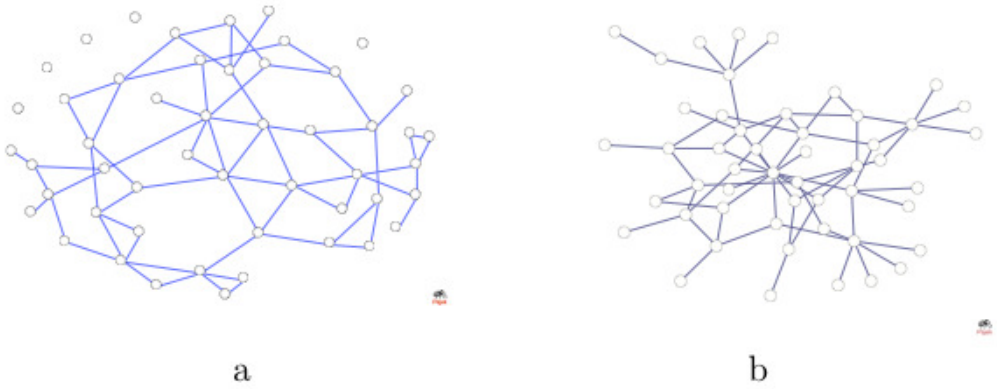

Figure 2.3: Clustering. Example of a network (a) with high clustering and another one (b) with low clustering.

most of the complex networks present in nature do. Figure 2.3 shows an example of a network with high clustering (a) and another of low clustering (b). The clustering of a network can be defined as:

$$C = \frac{3 \quad \text{number of triangles in the network}}{\text{number of connected triples of nodes}} \tag{2.8}$$

,where a "connected triplet" consists of three nodes that are connected by two (open triplet) or three (closed triplet) undirected links.. In other words, the clustering coefficient measures the fraction between the actual triangles in the network and the potential triangles that could be closed. Hence, a clustering of 1 implies that all connected triple are actually closed triplets, while 0 means that none of them actually form a complete triangle.

Watts y Strogatz CITA415 proposed an alternative measure of local clustering. The local clustering measures the degree to which the neighbors of a node are laso linked. Thus, for a node of degree k_i the local clustering can be expressed as:

$$C_i = \frac{2L_i}{k_i(k_i - 1)} \tag{2.9}$$

where L_i represent the number of links among the neighbors of i. $C_i = 0$ implies that none of the neighbors of i link to each other, while $C_i = 1$ implies that the neighbors of i form a complete graph they all link to each other.

Hence, the average clustering coefficient of the network can be calculated

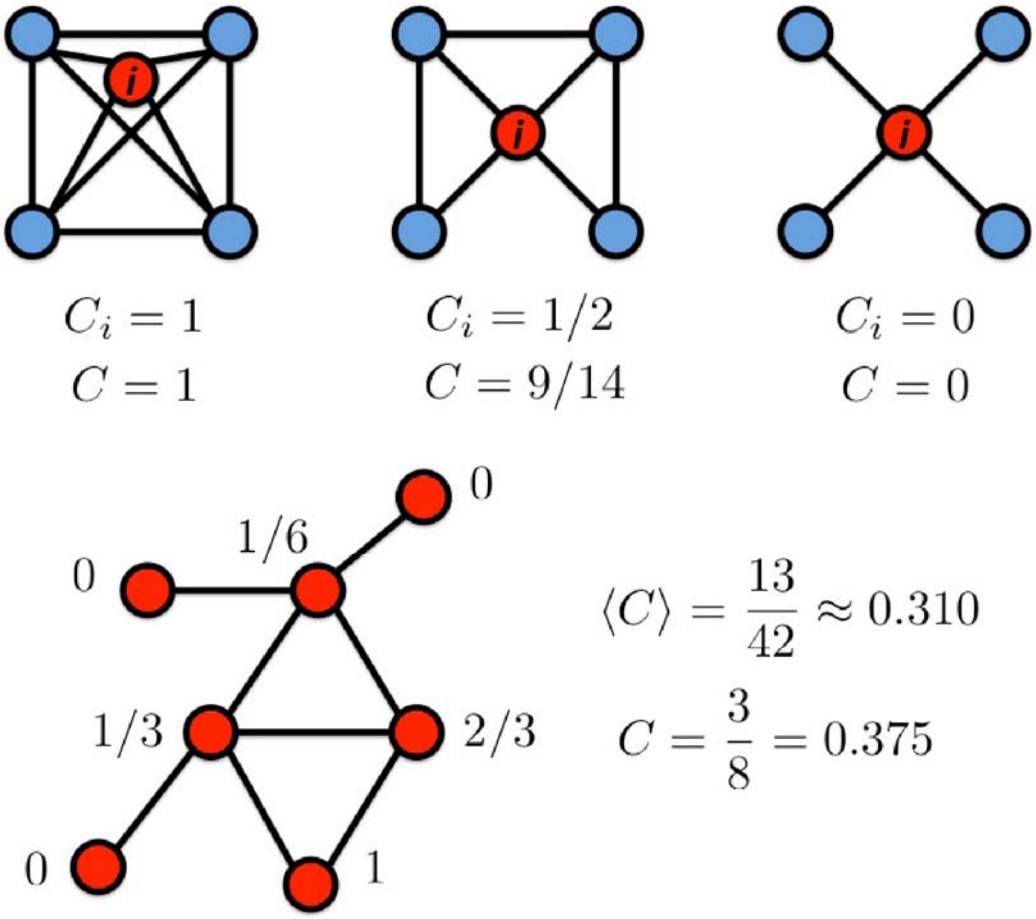

Figure 2.4: Several examples illustrating how to calculate the local clustering of nodes, and the average clustering of a network. Obtained from[Bar12].

as the average of C_i for all nodes:

$$\langle C \rangle = \frac{1}{N} \sum_{i=1}^{N} C_i \tag{2.10}$$

Thus, $\langle C \rangle$ is the probability that two neighbors of a randomly selected nodes are linked to each other. Figure 2.4 illustrates how the local clustering of nodes and the average clustering of a network is calculated.

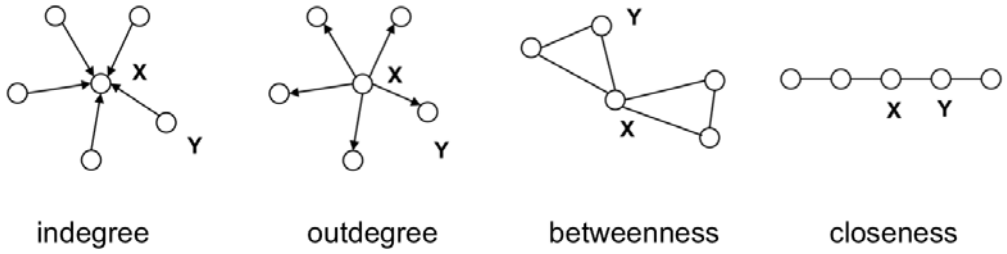

Figure 2.5: Figure illustrating the differences between centrality measures. X is more central than Y.

2.7. Centrality Measures

In the context of network science there are several measures that quantify how central each node is in a given network. The goal of centrality is to determine the importance or in uence of a given node in the network. This concept was first introduced in sociology to quantify the in uence of an individual in the whole social network. The most basic measure of centrality is the degree. Since we have already introduced the concept of degree we refer the reader to section 2.3 for a full explanation of it. The main advantage of the degree is that it is a property each node that can be computed with just local information from the node. Hence, it is a measure of direct in uence with a very low computation cost.

Other measures of centrality include closeness centrality, betweeness centrality, eigenvector centrality and page rank centrality. We have illustrated and compared them on Figure 2.5 where X is always more central than Y.

2.7.1. Betweeness Centrality

Betweenness is a centrality measure of in uence of a node within a networks (it can also be defined for links). This measure quantifies the number of times a node acts as an intermediary along the shortest path between two other nodes. It was first introduced by Freeman [Fre96] to quantify the control that an individual can achieve on the communication between other humans in a social network.

2.7.2. Closeness centrality

This measure quantifies how close (or far) is a node from the rest of the nodes in the network. Intuitively, it represents how long it will take a node to spread information to all other nodes. In connected networks we can define a natural distance between all pairs of nodes. This distance is given by the length of the shortest path connecting each pair of nodes. Thus, the closeness of a node is defined as the inverse of the sum of its distances to all other nodes. Therefore, the more central a node is the lower its total distance to all other nodes.

2.7.3. Eigenvector centrality

Eigenvector centrality measures the in uence of each node in a network. This measure assigns scores to nodes in the network based on the concept that connections to central nodes contribute more to the score of the node in question than the same number of connections to peripheral nodes. Google's PageRank is based on this same concept.

It can be computed from the adjacency matrix in the following way. For a given graph $G := (N\ L)$ with L number of vertices let $A = (a_{i,j})$ be the adjacency matrix, the centrality score of node i is defined as:

$$C_i = \frac{1}{} \sum_{j \in M(v)} x_j = \frac{1}{} \sum_{j \in G} a_{i,j} x_j \tag{2.11}$$

where $M(i)$ is the set of neighbors of i and is a constant. With a small rearrangement this can be rewritten in vector notation as the eigenvector equation

$$\mathbf{Ax} = \mathbf{x} \tag{2.12}$$

In general, there will be many different eigenvalues for which an eigenvector solution exists. However, all the entries in the eigenvector need be positive since negative centralities make no sense. Hence, this implies (by the Perron Frobenius theorem) that only the eigenvector associated to the greatest eigenvalue results in the desired centrality measure[20]. Thus, the centrality score of node i is given by the i^{th} term of the related eigenvector.

2.7.4. Pagerank centrality

In the scope of network analysis, the centrality of a node determines the relative importance of a node within the network. Therefore, in a social context it represents popularity or how in uential a person is in the social network. Despite there are numerous widely used measures of centrality, in the latter part of this thesis we will frequently use pagerank centrality.

The pagerank centrality, a network-based diffusion algorithm initially proposed to rank web pages [BP98] [LM06], has arisen as one of the preferred methods to rank a vast amount of data in different types of networks. For instance, it has been used to effectively determine the relevance of scientists in the citation network [CXMR07] [WXYM07], as well as to rank streets [WXYM07], ecological species [AP09] or leadership groups on social networks [PMGV13]. We could think about it as if a 'random surfer' surfs the net by following links between nodes, eventually the surfer decides to jump to a randomly chosen node and continue the process. The probability of the surfer visiting each node is determined by its pagerank. Hence, a node has a high pagerank when highly connected or while attached to leading ones. In the general case, the PageRank value for any node u can be expressed as:

$$PR(u) = \sum_{v \ B_u} \frac{PR(v)}{L(v)} \tag{2.13}$$

i e the PageRank value for a node u is dependent on the PageRank values for each node v contained in the set B_u (the set containing all nodes linking to node u), divided by the number $L(v)$ of links from page v.

2.8. Degree correlations and the friendship paradox

What are the chances that myself, a typical PhD student, will date a celebrity like Miley Cyrus? If we rely on random chances the answer is straight forward. In a world of about one billion individuals (10^9), where around 1,000 of them are celebrities the probability would be of (10^{-6}). If we redo the same reasoning for a celebrity like Fernando Alonso the probability will remain the same. Hence, according to random probabilities the chances that Fernando Alonso and myself will date a celebrity are the same. However, this is obviously not the case as it is widely assumed that celebrities marry each other. Harrison Ford and Calista Flockhart, Michael Douglas

and Catherine Zeta-Jones, Tom Cruise and Katie Holmes, Richard Gere and Cindy Crawford are some examples of celebrities couples.

Hence, this example is telling us that social networks do not present a random structure. In contrast, social networks are structured networks with the existence of celebrities that know many people and that are known by an even larger number of individuals. In network science these ultra connected individuals are known as hubs. The previous example is telling us that in society hubs tend to connect among themselves. However, this structure where hubs tend to be over connected among themselves does not appear in all real networks. For example, the protein interaction network presents the opposite structure, where hubs are not connected among themselves. Conversely, hubs are mainly connected to one and two degree proteins.

In the previous examples we did not take into account the fact that hubs will be more connected among themselves by the sole reason that they posses more links. Hence, we say that a network displays degree correlation if the number of links between nodes of degree k and k deviate from what would be expected by chance,

$$p_{k,k'} = \frac{kk}{2L} \tag{2.14}$$

Degree correlations represent a methodology to capture the relationship between the degree of nodes that are linked. Thus, a first way of measuring it is to compute for each value of k the average connectivity of all the neighbors of nodes with degree k:

$$k_{nn}(k) \quad \sum_{k'} k\, p(k \;\; k) \tag{2.15}$$

where $p(k \;\; k)$ is the conditional probability that following a link of a k'-degree node we reach a degree-k node. For neutral networks, without degree correlation, we have

$$p(k \;\; k) = \frac{e_{kk'}}{\sum_{k'} e_{kk'}} = \frac{e_{kk'}}{q_k} = \frac{q_{k'} q_k}{q_k} = q_{k'} \tag{2.16}$$

where $e_{kk'}$ represents the probability of finding a link between nodes of degree k and k and q_k the probability that there is a degree-k node at the end of the randomly selected link. Thus $k_{nn}(k)$ can be expressed as

$$k_{nn}(k) = \sum_{k'} k\, q_{k'} = \sum_{k'} k\, \frac{k\, p(k\,)}{\langle k \rangle} = \frac{\langle k^2 \rangle}{\langle k \rangle} \tag{2.17}$$

Hence, in neutral networks the average connectivity of the neighbors of a given node does not depend on the node degree. In fact, the value is constant and equal to $\langle k^2 \rangle \ \langle k \rangle$ and therefore, when plotting $k_{nn}(k)$ against k the result will be a horizontal line. This is exactly the pattern observed for the power grid network.

In networks that present degree correlations, we can distinguish two types: 1) **assortative networks**, hubs tend to connect to other hubs; and 2) **disassortative networks**, hubs tend to connect to low degree nodes. In assortative networks $k_{nn}(k)$ increases with k as observed in social networks such as the scientific collaboration network. On the other hand, for disassortative networks $k_{nn}(k)$ decreases with k, such as in the protein interaction network.

A way of measuring the degree of assortativity of a network is by assuming that $k_{nn}(k)$ follows a power law and approximating it by the following equation [PSVV01]

$$k_{nn}(k) = ak \tag{2.18}$$

Thus the assortatitivty or dissasortativity of the network would be determined by the sign of . > 0 implies that the network is assortative. For example, $\sim 0\ 37$ for the scientific collaboration network. A value of zero indicates and absence of degree correlations and therefore the network can be considered as neutral. < 0 implies the network is disassortative. Indeed, for the metabolic network we obtain $\sim -0\ 76$.

An alternative way to measure the assortativity of a network, by means of a single number, is the degree correlation coefficient introduced by Newman [New03]. This method assumes that $k_{nn}(k)$ is a linear function of k. The assortativity can be characterized by the following formula:

$$r = \sum_{jk} \frac{jk(e_{jk} - q_j q_k)}{\overset{2}{r}} \tag{2.19}$$

with

$$\overset{2}{r} = \sum_{k} k^2 q_k - [\sum_{k}^{jk} k q_c]^2 \tag{2.20}$$

It ranges between $-1 \quad r \quad 1$, for $r > 0$ the network is assortative, for $r = 0$ the network is neutral and for $r < 0$ the network is disassortative.

2.8.1. The friendship paradox

Do we have more friends than our friends? Our ego will convince us to answer yes, and indeed believe that we are more popular than them [ZJ01]. If we attend to logic we will probably answer that on average we have as many friends as our friends do. However, the friendship paradox, discovered by Scott L. Feld [Fel91], states that on average my friends are more popular than me.

To understand it let's return to eq 2.17. This equation tells us that the average degree of our friends is not simply $\langle k \rangle$ but depends on $\langle k^2 \rangle$. For the simple case of an ER network $\langle k^2 \rangle = \langle k \rangle^2 + \langle k \rangle$ and therefore the average connectivity of my friends is $\langle k \rangle + 1$.

The difference between $\langle k \rangle$ and the average connectivity of my neighbors becomes considerably larger for scale-free networks. For example, the average connectivity of the actors collaboration network is of 28 7, while $\langle k^2 \rangle \ \langle k \rangle =$ 565 7. Hence, in this network the average connectivity of my friends is by far lager than the average connectivity. The explanation for this paradox is simple, according to the probability laws we are more likely to be friends of hubs than of low degree individuals. Hence, our neighborhood is inevitably biased towards hubs.

2.9. Community Structure

In this section we will focus on the large-scale structure of networks. The most common and appropriate approach to study the mesoscale of complex networks is by performing a community structure analysis. It is widely assumed that almost all social networks present a community structure. A community can be defined a subgraph of densely connected nodes, within a larger and sparser network. In a social network context, a community would correspond to a group of closely related friends, while in a political context it would correspond to a political party. Figure 2.6 shows the network of political blogs. This network was analyzed in Adamic et al [AG05] where they showed that the network exhibited highly segregated community structure divided by political party.

Communities are of interest for a number of reasons. The first one, is that communities may correspond to functional units within a networked system. For example, in the scientific collaboration network a community might cor-

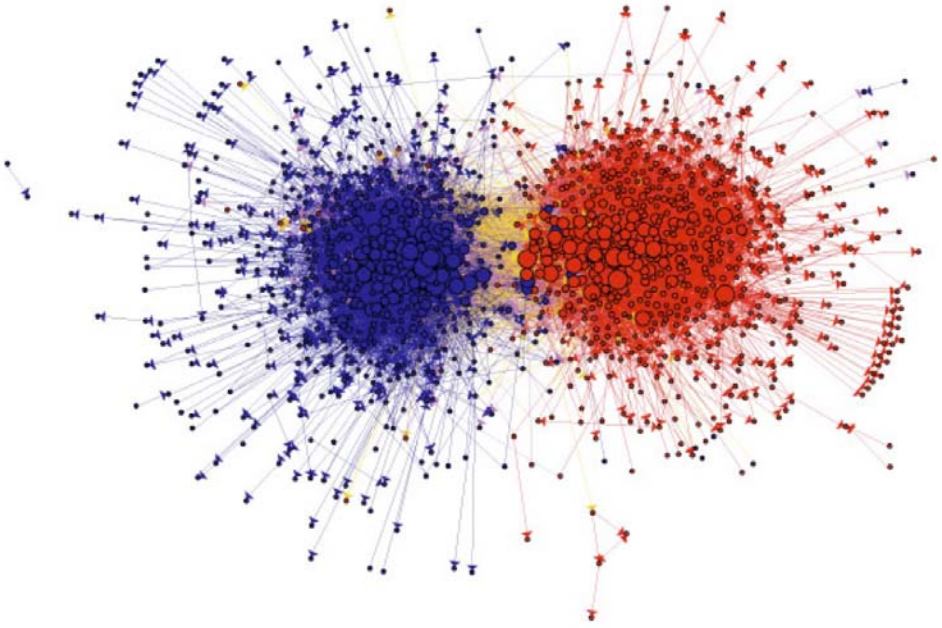

Figure 2.6: Visualization of the U.S. political blogs network as of 2004. The figure illustrates the polarization between the two political parties. Obtained from [AG05]

respond to a scientific discipline. Or in a metabolic network the network of chemical reactions within a cell a community corresponds to a circuit, or motif that carries out a specific function, such as synthesizing a vital chemical product. However, there is another reason, some times forgotten, that makes community structure analysis useful. Many networks, present significantly different properties in each community. Let's consider the scientific collaboration network of within a university. Communities should correspond to research groups. However, there might be important structural differences that affect the research output of each group. For example, some groups might exhibit a loos-knit structure with researchers closely collaborating together, while other groups may exhibit a star structure where all the collaborator are organized around a leader. Such differences can be crucial, and will only become apparent when analyzing each community separately.

For small networks, a good visualization of the network may be enough to visually distinguish the communities. However, when analyzing loads of data clear visualizations become impossible and we must rely on algorithms to detect the community structure. In fact, the development of such algorithms has been a very active scientific field during the last decade.

The community detection problem does not have an unique solution, as there is an absence of a single and precise formulation of community. In this sense, the network science community has proposed a wide variety of definitions. With this variety of definitions comes a variety of detection algorithms based on counts of edges within and between communities, counts of paths across networks, spectral properties of network matrices, or random walks among many others. In order to validate a community detection algorithm, there are two methods. The first one, corresponds to test them against well-known networks from which the community is globally known and accepted. The second one, involves creating artificial networks with a prefixed community structure.

The traditional approach to identify communities in a network has been hierarchical clustering. This method is based on a very simple principle: if we can measure how strongly nodes in a network are connected together, then by grouping the more strongly connected ones we can obtain the communities. An alternative and more modern method was proposed by Newman and Girvan [GN02]. This approach is based on the betweeness centrality of edges. First, one has to compute the centrality of each link and iterativey remove the link with higher centrality. Note that the centrality measures need to be recomputed at each step. The process finishes when the network is divided into two isolated subgraphs that correspond to the main communities. To obtain more communities one can repeat the process over the obtained communities.

Another of the most widely used methods for community detection is modularity maximization [New06]. Modularity is a benefit function that measures the quality of a particular partition of a network into communities. A particular division of a network into communities has a high modularity if the number of links across communities is minimum compared to the number of links within the communities. The modularity maximization algorithm detects communities by searching over possible divisions of a network for one or more that have particularly high modularity. Since exhaustive search over all possible divisions is intractable, practical algorithms are based on approximate optimization methods such as greedy algorithms, simulated annealing, or spectral optimization. The most popular modularity maximization algorithm is the *Louvain* method, which iteratively optimizes local communities until global modularity can no longer be improved given perturbations to the current community state[BGLL08].

2.9.1. Mapequation

The mapequation [RB08] uses the probability ow of random walks on a network to understand how information ows in a system and divide it into communities by compressing a description of the probability ow. Hence, this method groups nodes among which information ows quickly into communities, and creates links between them that represent the channels through which information ows among the communities.

The mapequation algorithm is specially appropriate for systems in which links represent mobility patterns or information exchange between nodes. For this reason we decided to use this algorithm in the analysis we will present on chapter 6, where we pretend to identify the political communities and understand how information ows among the different political alignments.

2.10. Network models: From random to scale-free networks

2.10.1. Random Networks

One of the goals of network theory is to develop models that reproduce the network structure of the complex systems that we encounter in nature. Real networks do not present a regular crystal structure, such as the lattice forming the net of a goal. In contrast, the networks we find in nature present a much more random structure.

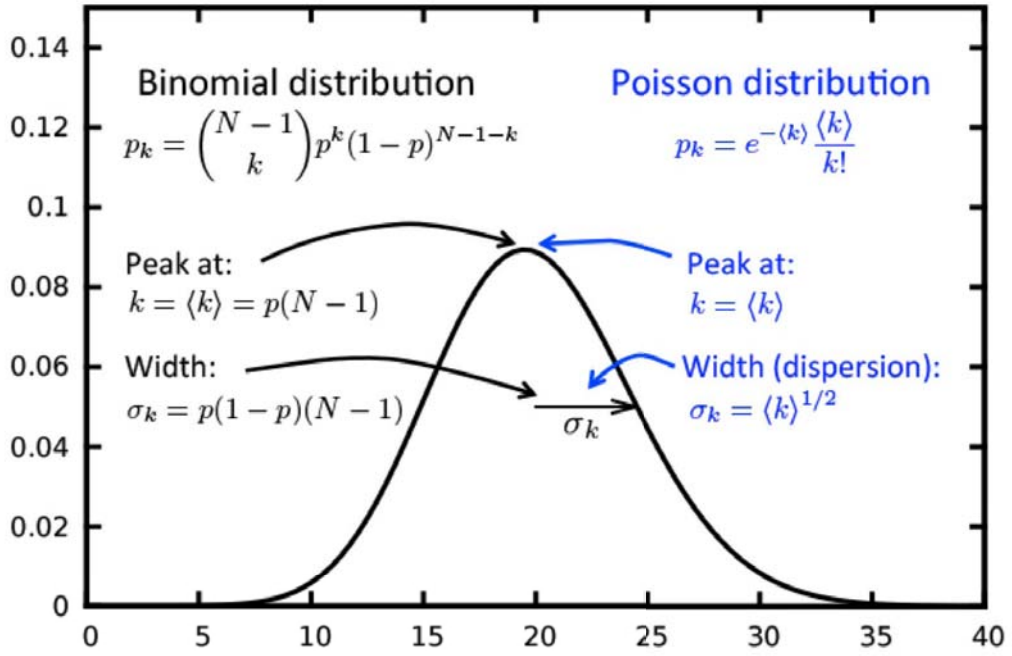

Figure 2.7: Visualization comparing the binomial and Poisson distributions. Obtained from [Bar12].

A random network can be defined as a set of nodes N where each pair of them is connected with the same probability p. In other words, a random network consists of N nodes with L randomly placed links. Thus, we can easily build a random network in the following way. We first create a set of N nodes. Next, we select all possible pair of nodes and generate a random number between zero and one for each pair. If the number is smaller than a fixed probability p we link the pair of nodes, otherwise we do not connect them.

Pál Erdös and Alfréd Rényi [Erd60], significantly contributed to improve our understanding on the properties of random networks. As a consequence we usually refer to random networks as Erdös-Rényi networks (ER).

The average degree $\langle k \rangle$ of a random network is determined by the size of the network N and the parameter p that controls the density of the network. Hence, $\langle k \rangle$ results from the product of p and the maximum number of links a node can have $(N-1)$. It can be expressed in the following way

$$\langle k \rangle = \frac{2L}{N} = p(N-1) \tag{2.21}$$

The degree distribution of a random network is given by the binomial distribution

$$p_k = \binom{N-1}{k} p^k (1-p)^{N-1-k} \tag{2.22}$$

Therefore, by using the properties of the binomial distribution we can calculate the average degree and its standard deviation. However, most real networks are sparse ($i\,e\ \ \langle k \rangle << N$), and on this limit we can approximate the degree distribution by the Poisson distribution given by the following formula

$$P_k = e^{-\,k}\,\frac{\langle k \rangle^k}{k!} \tag{2.23}$$

The Poisson distribution is only an approximation to the degree distribution of a random network. However, due to its analytical simplicity in comparison to the binomial distribution, it is usually the preferred form for p_k. Hence, during these thesis it will be the considered distribution when referring to ER networks. In Figure 2.7 we show a comparison between the binomial and poisson distributions.

2.10.2. We live in a small-world

An individual living in our same city is only a few handshakes away from us. This does not surprise any of us. However, another individual living in a different continents is no further than six hand-shakes away from us. This phenomenon is know as the small-worlds effect, or the six degrees of separation and states that the distance between two chosen randomly nodes is surprisingly short.

The phenomenon was first reported by Stanley Milgram in the 1960s. Milligram conducted an experiment in which a series of individuals on Kansas and Nebraska had to send letters to individuals who they didi not know living in Boston, Massachusetts. The letters passed from person to person and surprisingly were able to reach the designed target in a few steps-around six.

How can we explain the existence of these short distance among individuals? To further understand the six degrees of separation, let's consider a random network. In a random network the number of individuals at distance d from us is given by the following expression:

$$N(d) \quad 1 + \langle k \rangle + \langle k^2 \rangle + \quad + \langle k^d \rangle = \frac{\langle k \rangle^{d+1} - 1}{\langle k \rangle - 1} \tag{2.24}$$

Since $N(d)$ can not exceed the total number of nodes in the network $N(d_{max})$ N. Thus, by assuming $\langle k \rangle >> 1$ we can neglect the (-1) term in the denominator and nominator of eq. 2.24 and obtain:

$$\langle k \rangle^{d_{max}} \quad N \tag{2.25}$$

Hence, the diameter of a random network follows the following equation:

$$d_{max} \quad \frac{logN}{log\langle k \rangle} \tag{2.26}$$

As the diameter of the network is governed by a minority of extreme exceptions, for most networks eq 2.26 represents a better approximation of the average path length of the network. Thus, the average distance between nodes on a random network is defined by

$$\langle l \rangle \quad \frac{logN}{log\langle k \rangle} \tag{2.27}$$

This equation represents a quantitative formulation for the small-world effect. Thus, it provides an understandable interpretation of the effect. The depence of $\langle d \rangle$ with $logN$ rather than N implies that the distance among nodes in a random network is orders of magnitude smaller than the size of the network. Hence, the small-world effect implies that the distance among nodes in a network depends logarithmically on the size of the system. Finally, the dependence with $\langle k \rangle$ means that the denser the network the smaller the average distance between nodes.

2.10.3. Real networks are scale-free

If the Internet (visualized on Figure 2.8) were to be a random network following a Poisson degree distribution, its distribution should have a discretion of $_k = 2\ 52$. This quantity is by far less than what the measurement indicate $_{Internet} = 14\ 14$. In fact, just a visual inspection to its degree distribution

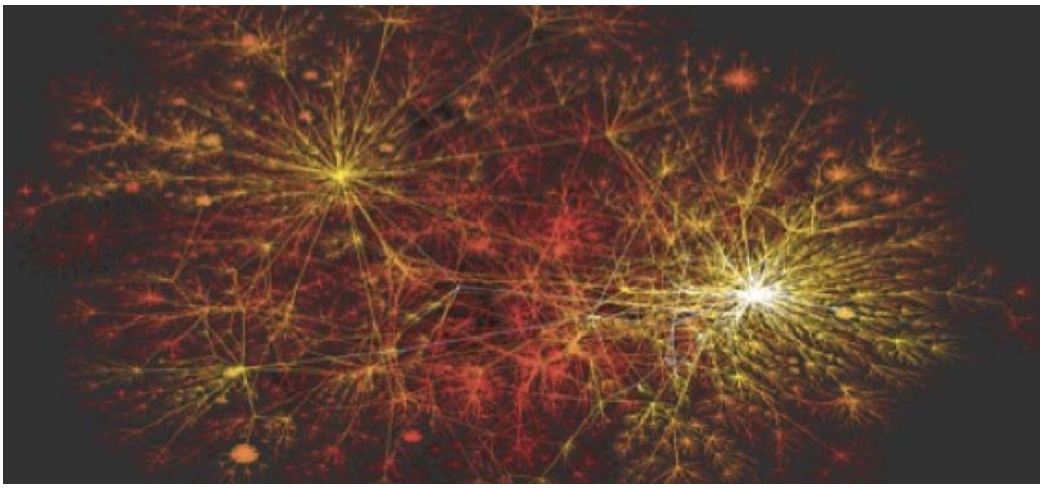

Figure 2.8: Visualization of the Internet network. Obtained from[Bar12].

reveals that the WWW network does not follow a Poisson degree distribution (ses Figure 2.9). Moreover, the WWW is no the only real system that deviates from the Poisson distribution. Almost all complex systems that we can encounter in nature present a degree distribution that is much closer to a power-law than to the Poisson distribution. As the Internet network presents linear dependence between $log(p_k)$ and $log(k)$, its degree distribution is best approximated by

$$p_k \sim k^{-\gamma} \tag{2.28}$$

where γ represents the slope between $log(p_k)$ and $log(k)$.

The main difference between these scale-free networks and random graphs appears in the tail of the degree distribution. While the probability of finding a node with degree k rapidly decreases for connectivities larger than $\langle k \rangle$ the pattern is significantly different in the scale-free case. For scale free networks the probability of finding high degree nodes, or hubs, is several orders of magnitude higher than in the random network.

2.10.4. The Barabási-Albert model

The degree distribution and the existence of hubs represents the major difference between random networks and those observed in nature. But why are power-laws and hubs absent from the random network degree distribution?

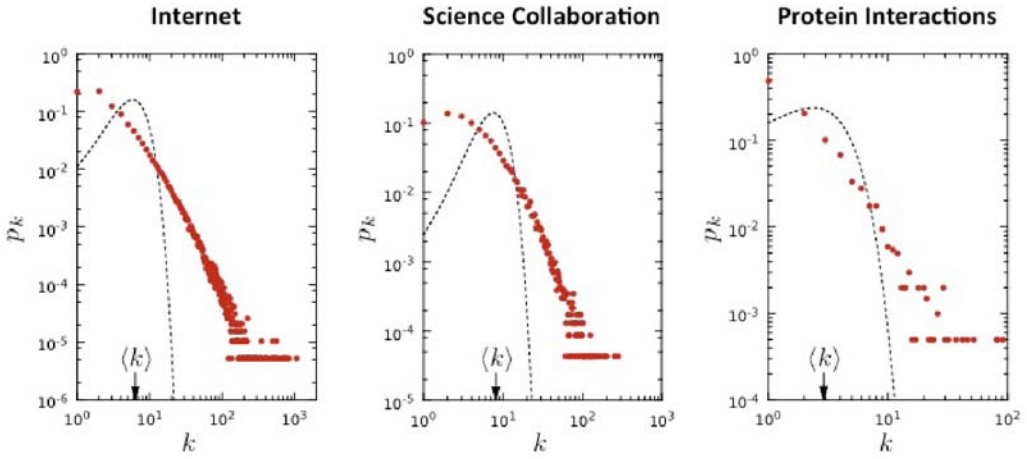

Figure 2.9: Degree distribution for the Internet, science collaboration, and protein interaction networks. The dashed lines corresponds to the Poisson prediction obtained by measuring $\langle k \rangle$. Obtained from[Bar12].

There are two basic assumptions in the ER model that do not occur in real life. The first one, is that networks evolve and are not static. The ER model fixes the number of nodes that does not vary during the forming process of the network. This is not true in nature, where systems continually grow with the addition of nodes. Let's consider the example of the WWW network. When it first appeared in 2001, it had a single node and today accounts trillion webpages.

The second one is that the ER model assumes that we randomly choose with whom we interact. However, this is not true, as new nodes prefer to connect with those that are already highly connected, hubs. This is known as the preferential attachment rule. Any new document that appears in the Internet will be rapidly attached to Google, while the probabilities that it will be linked to my blog, with just a hundred of visits per month, is minimum.

In summary the ER model fails to explain the growth of real networks and the emergence of hubs. Yet, these properties emerge from preferential attachment rules. The first network growth model based on preferential attachment was proposed by Price, although preferential attachment became widely accepted and known since the introduction of the Barabási-Albert model. The model proposed by Barabási and Albert is defined as follows :

1)At each time step a new node is created with m links that will connect the new node to the previoulsy existing ones.

2)The probability that the new node becomes connected to a node i of degree k_i is

$$(k_i) = \frac{k_i}{\sum_j k_j} \tag{2.29}$$

The model is based on a probabilistic rule that determines that a node can connect to any other node in the network, however, it will be more probable that it connects to a hub than to a low degree node. Thus, the model generates scale-free networks of exponent $\gamma \sim 3$ that present the small-world effect. After t steps the network will be formed by $m_0 + t$ nodes that are connected through $m_0 + mt$ links. m_0 represents the initial number of nodes.

To summarize, the Barabási-Albert model shows that two simple mechanisms, growth and preferential attachment, are responsible for the emergence of networks with a power-law degree distribution and small-world effect. Hence, the emergence of hubs and the power-law behavior results from the preferential attachment growth.

Chapter 3

The Meritocracy and topocracy of embedded markets

A system is said to be meritocratic if the compensation and power available to individuals is determined by their abilities and merits. A system is topocratic if the compensation and power available to an individual is determined primarily by her position in a network. Here we introduce a model that is perfectly meritocratic for fully connected networks but that becomes topocratic for sparse networks-like the ones in society. In the model, individuals produce and sell content, but also distribute the content produced by others when they belong to the shortest path connecting a buyer and a seller. The production and distribution of content defines two channels of compensation: a meritocratic channel, where individuals are compensated for the content they produce, and a topocratic channel, where individual compensation is based on the number of shortest paths that go through them in the network. We solve the model analytically and show that the distribution of payoffs is meritocratic only if the average degree of the nodes is larger than a root of the total number of nodes. We conclude that, in the light of this model, the sparsity and structure of networks represents a fundamental constraint to the meritocracy of societies.

3.1. Introduction

In the ideal world of Arrow-Debreu, every transaction that creates a surplus takes place. Unfortunately, we don't live in that world. An important difference between our world and that of Arrow-Debreu is that, in our world, every pair of individuals is not connected directly, but indirectly via networks of intermediaries, agents and middleman who expect to benefit from

their intermediating role.

As Granovetter [Gra85] pointed out more than two decades ago, our economy is *embedded* in social networks. These are networks that beget commercial interactions, and that are begot by them. For Granovetter, the cultivation of personal relationships between traders and customers assumes an equal or higher importance than the economic transactions involved. Economic exchanges are not carried out exclusively among strangers, but often incorporate individuals involved in long-term continuing relationships.

The embeddedness of markets is particularly important when links are costly. If links were costless society would behave similar to a fully connected network, and we will be back to the idealized world of Arrow-Debreu. When links are costly, however, embeddedness becomes extreme and markets are restricted by the structure of the social networks that co-exist with them.

In this chapter, we explore the redistributive consequences of the networks underlying economic activity by introducing a model with tunable embeddedness. The model separates the income of agents into two sources, the income obtained from the content agents produce, and the income that agents obtain from their intermediation role. We solve the model analytically and show that as networks become sparser, the model transitions from the meritocratic regime of Arrow-Debreu to what we call a *topocratic* regime, where the position of an individual in a network becomes the most important factor determining the compensation it receives.

Understanding the redistributive consequences of networks is important in a world where markets are composed of a mix of socially embedded links and commercial arm-length relationships [Uzz96, Uzz97]. Yet, even in a world where arm-length relationships are dominant, the assumption of fully connected networks is too hopeful. Possible transactions might not take place because individuals are uncertain about the quality of the goods being offered [Big93, LM99, AA70] or due to search frictions [RW87, Won11]. These market failures are partially compensated by the emergence of middlemen who are experts at reducing information asymmetries and search frictions, but who also act as hubs controlling information ows in the network. As Ronald Burt points out, the position that middleman occupy in the network is a source of advantage, as intermediating positions constitute part of what he has termed the social capital of structural holes [Bur09, Bur04].

In recent decades the social and economic role of networks has received an increased level of recognition [SBn03]. Economists have modeled the networks that emerge from strategic interactions [JW96, Jac10, EK10, GGJ+10], as well as the inequality in the distribution of payoffs expected in these equilibrium networks [SVR07, HS08, KISB11]. Our model contributes to this literature by separating the content producing role of an agent from its role

as an intermediary. This separation allows us to study the conditions under which the payoffs received by an individual are determined by the content she produces, or by her position in a social or professional network. To distinguish between these two payoff distribution regimes, we label the outcome of the system as meritocratic, when the distribution of payoffs is determined primarily by an agent's ability to produce quality content, and *topocratic* when the distribution of an agent's payoffs is determined primarily by her position in the network. We find that the transition to topocracy is mediated by network density, with topocracy becoming the dominant regime of sparse networks. In general, we find that the critical connectivity required to transition from topocracy to meritocracy goes as a root of the size of the network (N^a), with $a < 1$. This non-linear relationship means that the transition point is highly sensitive to both, the structure of the network and the algorithm used to distribute payoffs among individuals.

The implications of a root-rule of this kind can be explained by looking at numerical examples. Consider a network with as many nodes as people in the United States ($N = 3 \quad 10^8$). In this case an $N^{1\ 2}$ rule implies that meritocracy kicks in for connectivities above 17,320 links per node. This is certainly too large, meaning that an $N^{1\ 2}$ rule would imply that the U.S. is topocratic. An $N^{1\ 4}$ rule, on the other hand, implies a minimum average connectivity of only 131 links per node, which represents a reasonable number of social connections [Dun98]. Hence, when the transition from meritocracy to topocracy is mediated by an $N^{1\ 4}$ rule the implications for the U.S. would be that this is likely to be meritocratic.

The fact that in our model the transition between meritocracy and topocracy depends predominantly on the density of the network has two important implications. The first one is that the strong dependence of meritocracy on density makes the results of the model robust to different network formation mechanisms. Here, we can separate between two possibilities. First, there is the world in which the connectivity of individuals is determined largely by processes that are exogenous to them. This is a world where connectivity begets connectivity, such as in the case of the Barabasi Albert Model [BA99], the Yule Process [Yul25], the Price Model [Pri76], Merton's Cumulative Advantage (or Matthew effect) [Mer68], or Herbert Simon's modified version of the Yule Process [Sim55]. The second possibility is one in which the position that an individual occupies in a network is determined endogenously, for instance through strategic interactions [JW96, Jac10, EK10, GGJ+10]. Yet, when the density of the network is bounded due for instance to the high cost of links differences between link formation mechanisms, whether endogenous or not, should not introduce substantial changes to the meritocratic properties of the system. In other words, when the density of the network is

the main feature determining whether the markets embedded in them are meritocratic or not, the forces of endogenous network formation will only be able to modify this outcome slightly (we note that this is not true for an endogenous network formation process with full information. Yet, assuming full information in the network formation is equivalent to assuming a fully connected network).

The second implication that we would like to highlight is that the model predicts that meritocracy increases in societies that become better connected. This is an important implication given current changes in technology. Recent changes in communication technologies have increased the connectivity of our society, by reducing the cost of both social and commercial interactions. Most studies have emphasized the role of communication technologies on social participation and collaboration. Our results suggests that this technological change might also have important long term effects on the meritocracy of economies. Content producers, whether these are musicians or artists, can now market their content directly to a large number of individuals, even though this causes an information overload [Sim71, Gle11] that puts us far from the idealized limit of fully connected networks. Nevertheless, in the light of this model, changes in communication technology should increase the meritocracy of markets when holding population size constant. So the good news is that recent changes in technology should help make our society more meritocratic.

3.2. Modelling a networked market

3.2.1. The Model

Consider a world where individuals produce, distribute and consume content. The content produced by individuals can be abstracted as widgets of low marginal production cost, or cultural goods, such as books, films or music. To simplify the discussion we label the content producing role of an individual as her *Rockstar* role, since the payoffs collected via this role depend on the popularity of the content she produces. We call the intermediation role of an individual as her *Middleman* role, since the payoffs collected via this role are proportional to the transactions that she helps complete. The model is fully specified by three sets of assumptions:

1. **Initial Conditions:** The model begins with an exogenously determined network in which each node represents an individual. Each individual is endowed with a single parameter T representing its *Talent*, or ability to contribute to society. The talent T is fixed for the duration

of the model and determines the fraction of individuals that are willing to purchase the content produced by an individual. For example, an individual with $T = 0\ 3$ will sell its cultural good to 30% of all other individuals in the network. For simplicity, we draw T from a uniform distribution ($U \sim [0\ 1]$).

2. **Value Generation:** At each time step each node produces a new good (i.e. song, book, article, movie, etc.). These are non-rival goods, meaning that copies can be made at no cost. The goods made by an individual in her *Rockstar* role are purchased by a fraction T of all individuals. For simplicity, we choose the price to be constant and equal for all purchases. We later show that our results do not depend on prices.

3. **Value Distribution (a.k.a. Payo Structure):** If a *Rockstar* is not connected directly to an individual willing to purchase her content, the purchase is completed through the shortest path. In this case, the total revenue of the sale is distributed equally between each of the individuals in the path. For example, a purchase completed through a path of length three will give 1 3 of the payoff to the individual producing the content (the *Rockstar*), and to each intermediary (the *Middlemen*). Later, we generalize the model to other profit sharing rules.

The assumptions and the model are explained diagrammatically in figure 3.1 A.

Assumptions (1) to (3) define a model in which individuals collect payoffs by either producing content, or by being intermediaries in the distribution of the content produced and consumed by others. Hence, individuals have two sources of income: one that depends on talent, which we call meritocratic, and one that depends on their position in the network, which we call topocratic. We note that the topocratic channel defined by an individual's *Middleman* role is useful for the system, since without it transactions would not be completed. We note that the market is completed only as long as there are indirect paths between every pair of nodes.

3.2.2. Meritocratic and topocratic Regimes - Limit Cases

In a fully connected network, the model has a trivial solution where payoffs come solely from the production of content, since intermediation is not

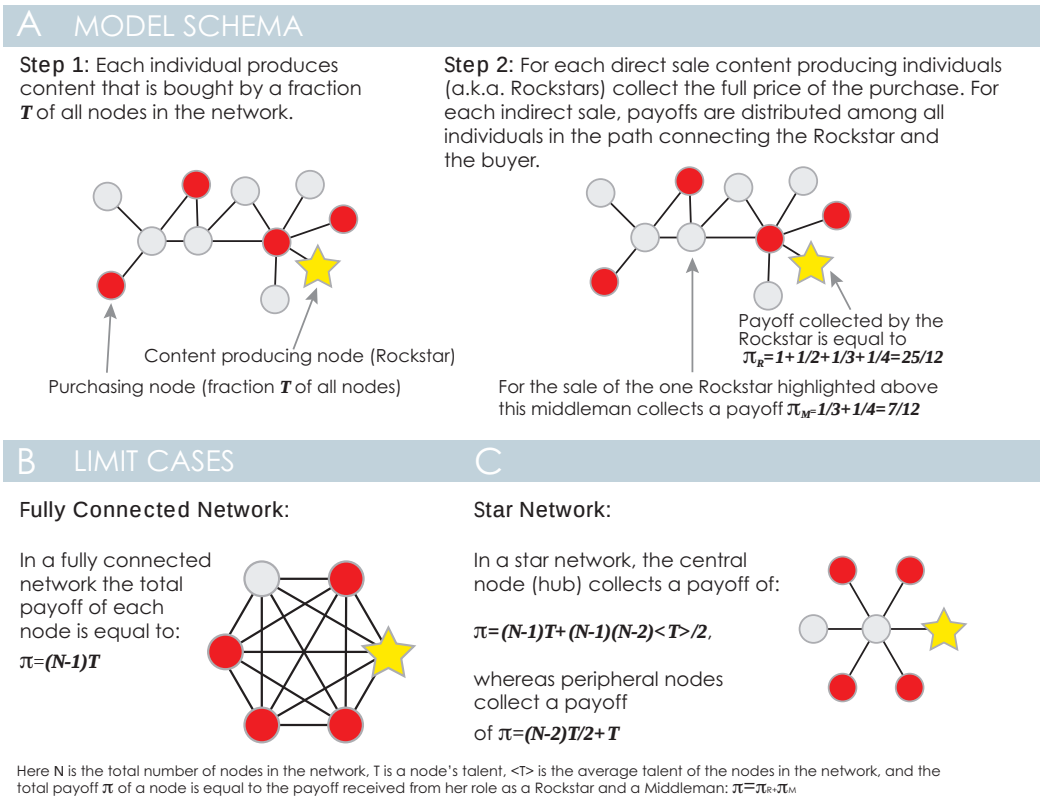

Figure 3.1: A Schematic representation of model. B Solution for a fully connected network. C Solution for a star-network

necessary, and hence, avoided. In this case, the payoffs ($_i$) follow Talent (T_i) exactly, since an individual with talent T_i receive a payoff $_i = T_i(N-1)$, where N is the number of nodes in the network (Figure 3.1 B). Hence, when the network is fully connected, the model describes a system that is perfectly meritocratic, i.e., payoffs are perfectly correlated with talent.

In a world where the network is not fully connected the income of an individual will depend not just on its talent, but also on the betweenness centrality of an individual (the betweenness centrality of a node is the number of shortest paths going through that node). Here, the extreme case is a star-network with one central node, or hub (Figure 3.1 C). In this case, a node in the periphery of the star network receives an income equal to:

$$_i = (N-2)T_i \;\; 2 + T_i \tag{3.1}$$

On the other hand, the hub receives a payoff equal to

$$_i = \langle T \rangle (N-1)(N-2) \;\; 2 + T_i(N-1) \tag{3.2}$$

where $\langle T \rangle$ is the average talent of the system, the first term represents the revenue intermediated by the hub, and the second term represents the probability that the hub is a direct buyer of the peripheral node's content.

We note that the maximum possible payoff that can be obtained from intermediation is $\sim N^2$, whereas the maximum possible payoff that can be obtained from producing content is $\sim N$. This large difference emerges because the payoff that an individual gets from its *Middleman* role grows with the number of possible links in the system, which is quadratic on the number of nodes N, whereas the maximum income for the *Rockstar* role is bounded by the number of nodes in the network (N). This difference does not depend on our choice of distribution of talents, or the way in which revenue is distributed along the chain, so it is a fundamental difference between the intermediation role of *Middlemen* and the production role of *Rockstars*. It therefore represents a fundamental constraint to any model considering this duality of behaviors.

Determinants of individual payo s

We begin by splitting the payoff collected by an individual through each of her two roles

$$_i = \ _{R_i} + \ _{M_i} \tag{3.3}$$

where $_{R_i}$ and $_{M_i}$ indicate, respectively, the payoffs from the *Rockstar* and *Middleman* behavior.

In an Erdös-Rényi network (ER)[Erd60] the degree of a node is well approximated by the average degree of the network. Hence, we can approximate the payoff that an individual gets from her *Rockstar* role by her talent times the number of individuals at distance d from her discounted by the length of the chain connecting her to each individual. Hence, in a random network where the average connectivity or degree is equal to $\langle k \rangle$, $_R$ can be approximated by:

$$_{R_i} = T_i \ \ \langle k \rangle + \frac{\langle k \rangle^2}{2} + \frac{\langle k \rangle^3}{3} + \quad + \frac{\langle k \rangle}{\quad} \ \ = T_i \sum_{j=1} \frac{\langle k \rangle^j}{j} \tag{3.4}$$

where the cutoff is equal to the average path length of the network, which in a random network is well approximated by:

$$\frac{\ln N}{\ln \langle k \rangle} \tag{3.5}$$

The income that individuals earn from their behavior as *Middlemen*, $_M$ can be obtained by noticing that in every non-direct sale conducted through a chain of length d, the $d-1$ *middlemen* participating get an equal share of the purchase. In this way, the total income in the network that is collected by *Middlemen* can be written similarly to (3.4), as

$$_M = \sum_j \ _{M_j} = N\langle T\rangle \ \frac{\langle k\rangle^2}{2} + \frac{2\langle k\rangle^3}{3} + \frac{3\langle k\rangle^4}{4} + \quad + \frac{(\ -1)\langle k\rangle}{} \quad (3.6)$$

$$_M = N\langle T\rangle \sum_{j=2} \frac{j-1}{j}\langle k\rangle^j \quad (3.7)$$

The payoff collected by a single *middleman* can be obtained by taking the share of shortest paths going through an agent with degree k_i. In a random network the number of shortest paths going through a node is given by $k_i^2 \quad _j k_j^2$ [Bar04]. Hence, the average payoff collected by a *Middleman* is:

$$_{M_i} = N\langle T\rangle \ \frac{k_i^2}{_v k_v^2} \sum_{j=2} \frac{j-1}{j}\langle k\rangle^j \quad (3.8)$$

This expression can be simplified by replacing the normalizing factor in the denominator by its expected value, $\ _v k_v^2 = N\langle k^2\rangle$. This reduces (3.8) to

$$_{M_i} = k_i^2 \frac{\langle T\rangle}{\langle k\rangle^2 + \langle k\rangle} \sum_{j=2} \frac{j-1}{j}\langle k\rangle^j \quad (3.9)$$

Finally, we can use equations (3.4) and (3.9) to write down a general formula for the payoff of individual i as a function of both, its talent T_i and its connectivity k_i. This formula takes the general form:

$$_i = CT_i + Bk_i^2 \quad (3.10)$$

where both C and B are independent of the individual and depend only on the average talent, and the average degree of the network ($\langle k\rangle$).

$$C = \sum_{j=1} \frac{\langle k\rangle^j}{j} \quad \text{and}$$

$$B = \frac{\langle T\rangle}{\langle k\rangle^2 + \langle k\rangle} \sum_{j=2} \frac{j-1}{j}\langle k\rangle^j \quad (3.11)$$

We note we have assumed all nodes to belong to the network's giant component there are no isolated nodes or clusters in the network. For networks made of several components our results are still valid by considering N to be the number of nodes in the component in question.

For sparse networks, as Figure 3.2 illustrates, high payoff individuals concentrate on the core of the network irrespective of their talent. Conversely, when the network becomes denser, the opposite becomes true: talented individuals, irrespective of their position in the network, are the ones collecting the highest payoffs.

3.3. Transition threshold

3.3.1. General case

Since the model is perfectly meritocratic for a fully connected network, and highly topocratic for a star network, the natural question to ask is when does the transition between meritocracy and topocracy takes place. We can do this by using equation (3.10) to calculate the total payoffs associated with the Rockstar role:

$$\sum_{i=1}^{N} CT_i = C\langle T\rangle N \tag{3.12}$$

Dividing the Rockstar payoffs ($_R$) by the total payoffs paid in the entire system ($= N(N-1)\langle T\rangle$) shows that the fraction of payoffs assigned through the Rockstar channel is equal to:

$$\frac{}{} = \frac{C}{N-1} \tag{3.13}$$

Finally, we ask for this ratio to be larger than 1 2 to obtain the condition

$$C > \frac{N-1}{2} \tag{3.14}$$

We can solve this condition analytically for a random ER network by assuming a large network ($N >> 1$) and an average degree larger than one ($\langle k\rangle > 1$). This implies that we can approximate C by the largest term of the sum in eqn. (3.7). Using the expression for in equation (3.5) we can re-express (3.14) as:

$$\frac{N-1}{2} < \sum_{i}^{l} \frac{\langle k\rangle^i}{i} \qquad \frac{\langle k\rangle}{l} = \frac{\langle k\rangle^{\ln N \ \ln k}}{\ln N \ \ln\langle k\rangle} \tag{3.15}$$

Finally, using the change of variable $u = \ln\langle k\rangle \ \ln N$ in (3.15), and approximating $(N-1) \ N \quad 1$ the condition becomes

$$\frac{1}{2} < u \tag{3.16}$$

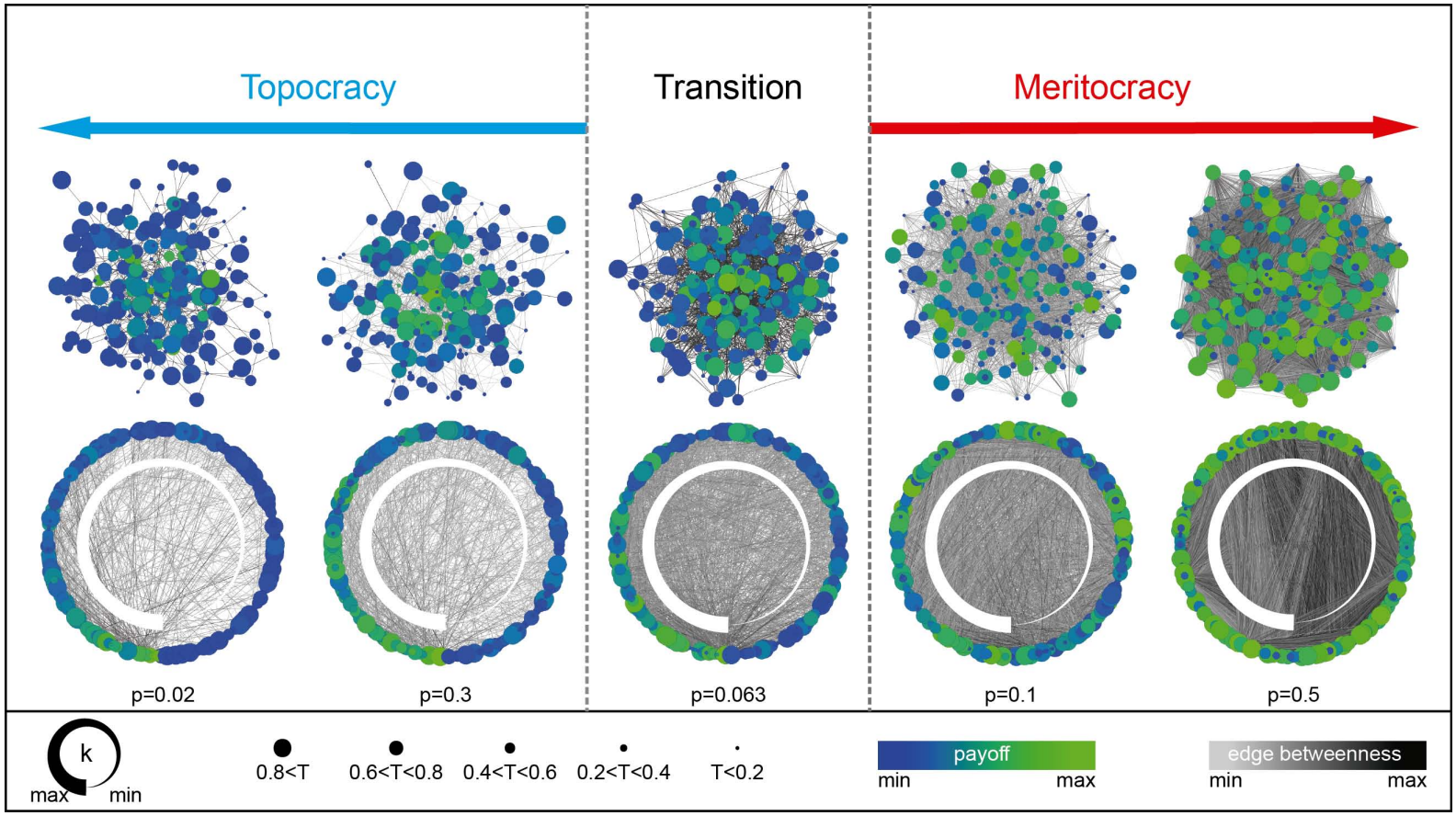

Figure 3.2: Payoff distribution for different levels of average connectivity on a network of size $N = 250$

which is equivalent to:

$$\langle k \rangle > N^{1\ 2} \tag{3.17}$$

Figure 3.3 shows how the wealth generated in the model is distributed between the two possible activities, *Rockstars* and *Middlemen*, as a function of the average network connectivity. The transition point from topocracy to meritocracy ($\langle k \rangle = N^{1\ 2}$) is indicated with a discontinuous vertical line. We note that for small values of $\langle k \rangle$ the total payoff of the system decreases, since the network becomes fragmented and there are transactions that are not completed.

Finally, we note that there is also a structural interpretation for the $\langle k \rangle = N^{1\ 2}$ threshold. When the average connectivity of the network is equal to the square root of the total number of nodes, the average distance in the network is two, meaning that individuals are no further than two hops away. Hence, the $\langle k \rangle = N^{1\ 2}$ rule obtained in this case is equivalent to saying that topocracy emerges in random networks when the average path length is larger than two, and hence, that six-degrees of separation imply a highly topocratic system.

3.3.2. Alternative sharing rule: comissions

Next, we extend the model to a payoff sharing rule in which *Middlemen* get a percentage of the total transactions in which they participate. For instance, by getting a 10 percent commission of the transactions in which they are directly involved. We note that this is not the same as a ten percent commission of the entire purchase. For example, in a purchase of an item of price one completed by a chain of length three, the first middlemen gets 0.1, the second gets (1-0.1)x0.1=0.09 and the Rockstar gets 1-0.1-0.09=0.81

With these assumptions, the payoff of a Rockstar is given by

$$_{R_i} = T \sum_{i=1} \langle k \rangle^i (1 - \quad)^{i-1} \tag{3.18}$$

and hence the total payoff collected by Rockstars (when $k(1 - \quad) > 1$) can be approximated by:

$$_{R} = N\langle T \rangle \sum_{i=1} \langle k \rangle^i (1 - \quad)^{i-1} \qquad N\langle T \rangle \langle k \rangle \ (1 - \quad)^{\ -1} \tag{3.19}$$

The total payoffs paid by the system ($= N(N-1)\langle T \rangle$) are not changed by the payoff sharing rule. Hence, the fraction of the payoff collected by the

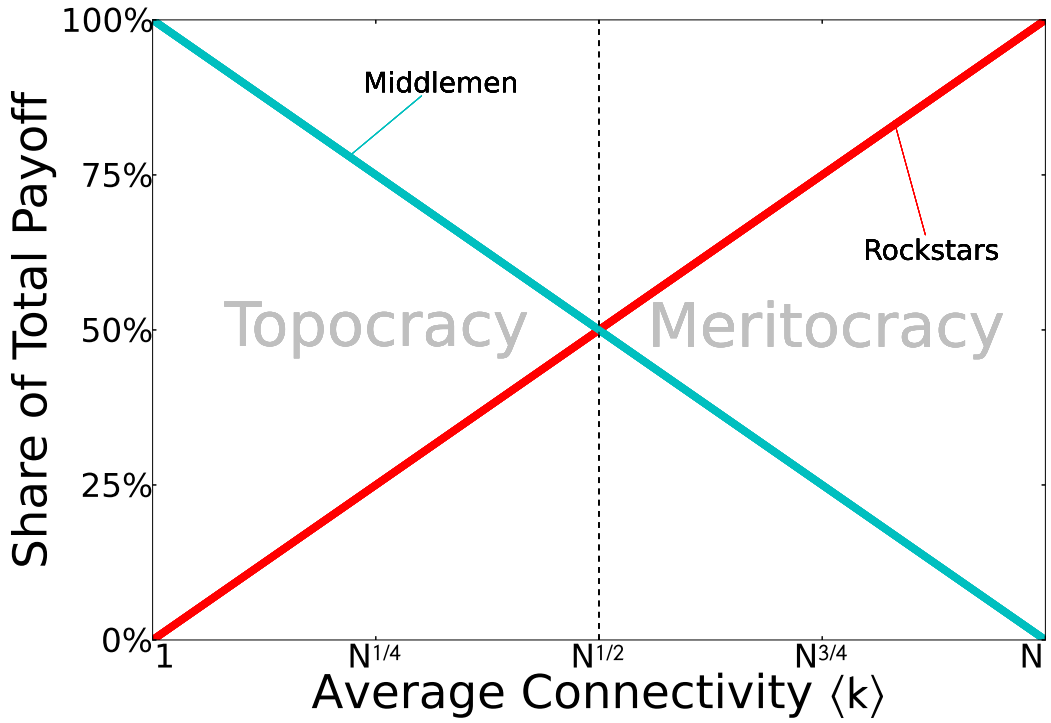

Figure 3.3: Share of total payoffs (eq. 3.13) as a function of the average network connectivity for a random Erdös-Rényi network and a proportional payoff sharing rule.

Rockstars through the meritocratic channel can be obtained similarly than before as:

$$\frac{R}{} = (1 - \quad)^{\frac{\ln N}{\ln\langle k\rangle}-1} \tag{3.20}$$

Finally, imposing $_R \quad > \quad$ gives an average connectivity of:

$$\langle k\rangle = N^{\frac{\ln(1-\alpha)}{\ln\beta(1-\alpha)}} \tag{3.21}$$

Figure 3.4 shows the share of total payoffs collected by *Rockstars* and *Middlemen* as a fraction of the average connectivity ($\langle k\rangle$) and the percentage collected by each *middlemen* ( ). When is small, the transition to meritocracy takes place at low connectivities. For instance, when $= 0\,1$ the transition to meritocracy takes place at $\langle k\rangle = N^{0\,13}$. For a network with $N = 10^7$ nodes this implies a threshold connectivity of just $\langle k\rangle = 8$, meaning that very little connectivity is required for the system to become meritocratic. For large percentages, however, the transition to meritocracy

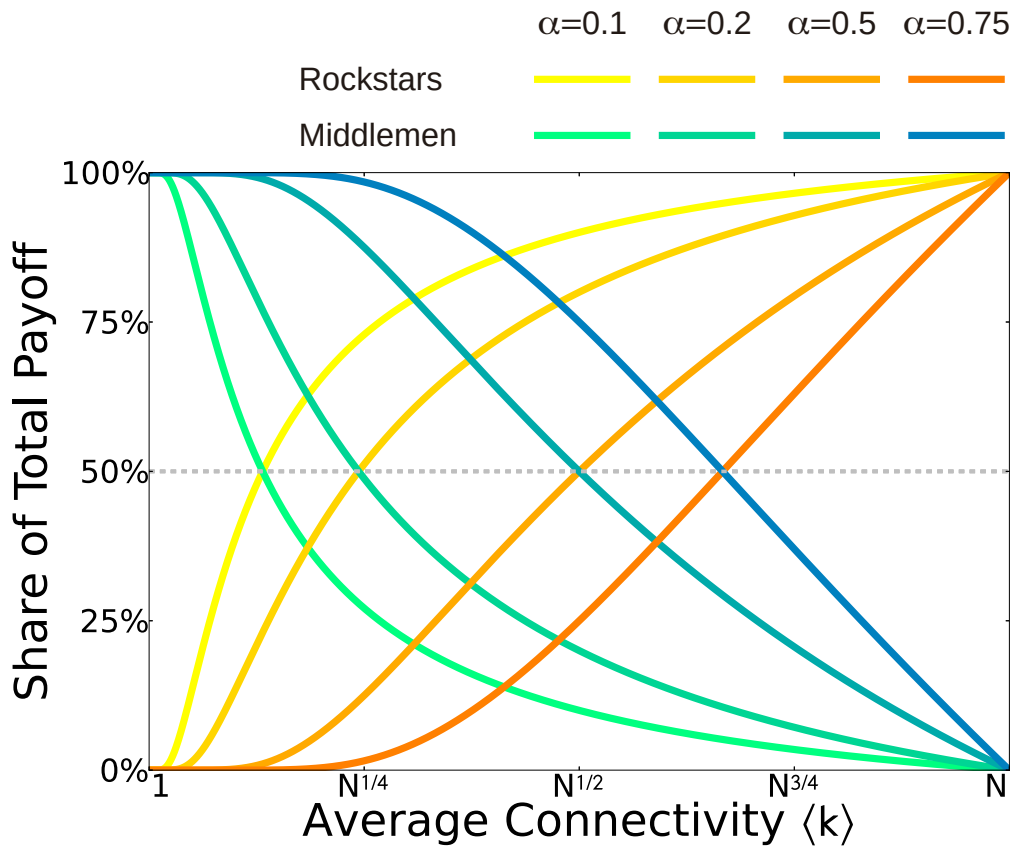

Figure 3.4: Share of total payoffs as a function of average network connectivity for a random Erdös-Rényi network and a sequential payoff sharing rule according to equation(3.20).

is not that easy. When = 1 2 the transition is once again at $\langle k \rangle = N^{1\ 2}$. The $N^{1\ 2}$ rule might seem counterintuitive when considering a percentage rule, since in a percentage rule a 50% commission would imply an extremely fast decay on the remainder received by the *Rockstar*. Yet, when $\langle k \rangle = N^{1\ 2}$ all individuals are on average, no more than two hops away from each other, and hence the payoffs are not distributed via long chains.

3.4. Payo Distributions

We can also use the results obtained to calculate the payoff distribution $P(\ )$ implied by the model. Since nodes have two sources of payoffs, we can calculate the distribution of payoffs associated with each role ($P(\ _R)$ and

$P(\ _M))$.

The distribution of payoffs can be calculated by noticing that the linear dependence of $\ _R$ on T implies that $\ _R$ is uniformly distributed between 0 and C.

$$P(\ _R) = \text{Uniform}(0\ C) \tag{3.22}$$

with C given by equation (3.11).

The distribution of income coming from an individual's connectivity is related to the degree distribution of the network $P(k)$ by the identity:

$$P(\ _M) = P[k(\ _M)]\ \ \frac{dk}{d\ _M} \tag{3.23}$$

and since $\ _M = Bk^2$, this implies that:

$$P(\ _M) = \frac{1}{2\ \overline{B\ _M}}\ P\ \ \ \overline{\frac{\ _M}{B}} \tag{3.24}$$

We can solve this for an ER network by using the fact the degree distribution is a Poisson Distribution [Poisson($\langle k\rangle$)], In this case, the model predicts that the normalized distribution of payoff will follow:

$$P(\ _M) = \frac{1}{2\ \overline{B\ _M}}\ \frac{e^{-\ k}\ \langle k\rangle\ \ \overline{\ _M\ B}}{\overline{\ _M\ B}\ !} \qquad \text{for}\ \ \ _M > B \tag{3.25}$$

where the factorial in the denominator can be made a continuous function by approximating it with the Stirling approximation: $n!\ \ \ \ \overline{2\ n}\ (n\ e)^n$. We note that for $\ _M >> \langle k\rangle$ equation (3.25) implies that $P(\ _M)$ decays exponentially. This can be easily seen by approximating the Poisson distribution by a Gaussian with mean $\langle k\rangle$ and standard deviation $\ \ \overline{\langle k\rangle}$,

$$P(\ _M) = \frac{1}{2\ \ \overline{\langle k\rangle B\ _M}}\ \mathrm{e}^{-\ \ \ \overline{\ _M\ B} - \ k\ \ ^2\ 2\ k} \qquad \text{for}\ \ \ _M > B \tag{3.26}$$

(where in this sole case $\ $ is the ratio between a circle circumference and its diameter, should not be confused with the total payoff $\ $.) The square root over $\ _M$ implies that this decay is slower than that of the Poisson or Gaussian describing the network's degree distribution.

Finally using equations (3.22) and (3.25) and taking into account that $\ _R$ and $\ _M$ are independent variables we obtain the distribution of total payoffs ($\ \ = \ _R + \ _M$) which is given by the following expression:

$$P(\ \ ) = \begin{array}{ll} \frac{1}{C}\frac{(\ \ \ \overline{\ \ B}+1,\ k\ )}{(\ \ \ \overline{\ \ B}+1)} & \text{if}\ \ \ 0 < \ \ < C \\ \frac{1}{C}\frac{(\ \ \ \overline{\ \ B}+1,\ k\ )}{(\ \ \ \overline{\ \ B}+1)} - \frac{(\ \ \overline{\frac{\pi-C}{B}}+1,\ k\ )}{(\ \ \overline{\frac{\pi-C}{B}}+1)} & \text{if}\ \ \ C < \end{array}$$

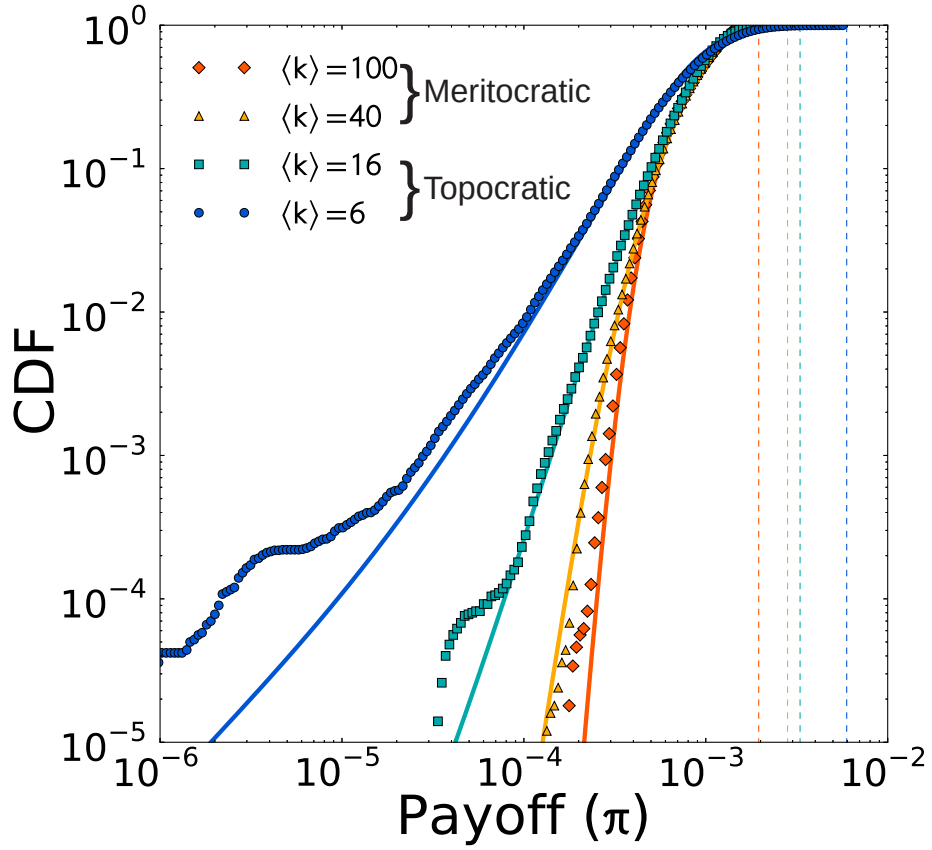

Figure 3.5: Total income cumulative distribution for random Erdös-Rényi networks with different average connectivities $\langle k \rangle$. The analytical results from equation (3.27) are plotted in full line, while dots represent the results obtained from numeric simulation using 500 realizations of the network with $N = 1000$ nodes. The maximum payoff obtained from the simulations for each network has been indicated with dashed lines. Note that payoffs have been normalized by the total payoff of the system ($\langle T \rangle N(N-1)$).

Figure 3.5 compares equation (3.27) with numerical simulations of the model showing that as the network becomes more sparse, the payoffs distribution becomes more heterogeneous. This is re ected in the increase of the maximum possible payoff and the decrease of the minimum payoff observed.

To conclude, we note that the model implies that the distribution of payoffs will always be more heterogeneous that of the underlying network's degree distribution. This illustrates that, in general, the intermediation mech-

anism described in the model makes the inequality larger than the inequality of connectivity that is already present in the network [equation (3.24)].

3.5. The statistical Meritocracy and Topocracy of Networks

Here, we study an alternative methodology to estimate the *meritocracy* $\mathcal{M}$ and topocracy $\mathcal{T}$ of the system. Instead of looking at a threshold of $_R$, we define the meritocracy $\mathcal{M}$ as the correlation between an individual's contribution to the network, represented by its talent T_i, and the total payoff $_i$ collected by the individual. Formally, we define meritocracy in terms of Pearson's correlation as:

$$\mathcal{M} = \text{corr}(T \quad) \tag{3.27}$$

By definition, when the network is fully connected, the system is perfectly meritocratic, since in that limit $_i = (N-1)T_i$ and hence $\mathcal{M} = \text{corr}[T \ (N-1)T] = 1$. In general we can express the total payoff of an individual as the sum of its contributions coming from the creation and distribution of content. Hence, we can rewrite (3.27) as:

$$\mathcal{M} = \text{corr}(T \quad _R + \quad _M) \tag{3.28}$$

Next, we decompose the correlation function in its covariance and standard deviation components to obtain:

$$\mathcal{M} = \text{corr}(T \quad _R + \quad _M) = \frac{\text{cov}(T \quad _R) + \text{cov}(T \quad _M)}{T \quad _R + \ _M} \tag{3.29}$$

Finally we use the properties of the variance and covariance, and the fact that $\text{cov}(T \quad _M) = 0$ to obtain:

$$\mathcal{M} = \frac{C \ ^2_T}{T \quad \sqrt{^2_R + \ ^2_M}} \tag{3.30}$$

We can further simplify (3.30) by noticing that T and $_R$ are uniformly distributed. For this reason, $^2_T = 1\ 12$ and $^2_R = C^2\ 12$. With this, equation (3.30) simplifies to:

$$\mathcal{M} = \frac{C}{\sqrt{C^2 + 12 \ ^2_M}} \tag{3.31}$$

By the same token, we can estimate the *topocracy*, of the system as the correlation between k^2 and the total payoff

$$\mathcal{T} = \mathrm{corr}(k^2 \quad _R + \quad _M) = \frac{\quad _M}{\sqrt{C^2 \;\; 12 + \quad ^2{}_M}} \tag{3.32}$$

Finally, we note that both $\mathcal{M}$ and $\mathcal{T}$ can be evaluated analytically for a sparse ER network by taking into account that in this case

$$_M = B^2\langle k\rangle(1 + 6\langle k\rangle + 4\langle k\rangle^2) \tag{3.33}$$

The last two expressions are illustrated in Figure 3.6, and show that the meritocracy of the system, $\mathcal{M}$, decreases (from 1 to 0) as the network becomes sparse and $^2{}_M$ increases, while the opposite is true for the *topocracy*, $\mathcal{T}$, which approaches one for $^2{}_M >> C^2 \;\; 12$. We note, however, that there is a decrease in *topocracy* at low connectivities that is not accompanied by an increase in meritocracy. This is due to the fact that as $\langle k\rangle \quad 1$ the network becomes disconnected, and purchases are only completed in the connected components.

3.6. The scale-free case

In the previous sections, a mathematical analysis of the model introduced in Section 3.2.1 was presented. In particular, equations (3.10)-(3.11), (3.25), and (3.31)-(3.33) give, respectively, analytical expressions for the individuals payoff, the corresponding distribution, and the meritocracy and topocracy parameters of the network for the sparse ER case.

In this section similar expressions are derived for a scale-free (SF) network, which is different from the ER case due to the structural properties of both networks. First, the diameter of a SF network is comparatively smaller than that of an ER with the same average connectivity. For neutral SF networks we can approximate the average neighborhood connectivity of a given node by $\quad = \langle k^2\rangle \;\; \langle k\rangle$ [CH03]. As a result, $\quad _R$ for an individual with talent, T, and degree, k, should now be written as:

$$_R = Tk \;\; 1 + \frac{1}{2} \quad + \frac{1}{3} \;\; ^2 + \frac{1}{4} \;\; ^3 + \qquad = Tk \sum_{i=0} \frac{^i}{i+1} \tag{3.34}$$

The cutoff $\quad$ in (3.34) is equal to the network average path length, $\langle l\rangle$, which in the SF case with an exponent between $2 < \gamma < 3$ can be approximated by:

$$\langle l\rangle = \frac{\ln\ln N}{\ln(\gamma - 1)} \tag{3.35}$$

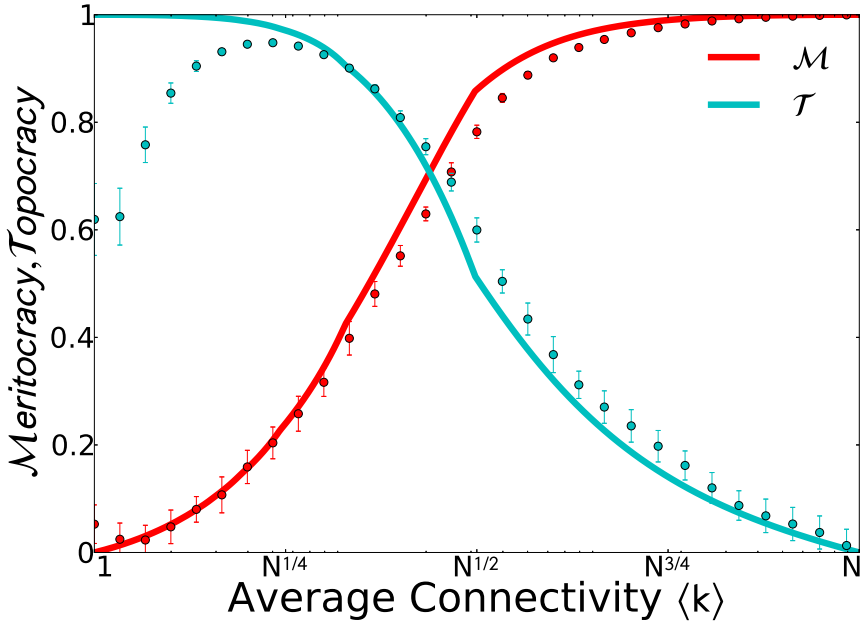

Figure 3.6: Meritocracy (full red line) and topocracy (full blue line), as defined in equations (3.31) and (3.32) as a function of the average network connectivity. The corresponding values obtained by numerical simulation using 500 realizations in a random Erdös-Rényi network with $N = 1000$ nodes are shown for comparison with red and blue dots, respectively.

An expression for the average payoff collected by a *middleman* can be obtained following a similar derivation to that made in the ER case, but taking into account that for a SF network of exponent γ, the betweenness centrality is proportional to k [GKK01] with

$$= \frac{\gamma - 1}{ - 1} \tag{3.36}$$

where is the exponent of the power law followed by the betweenness centrality distribution. The final expression for $_M$ is

$$_M = k \frac{\langle T \rangle \langle k \rangle}{\langle k \ \rangle} \sum_{i=2} \frac{i-1}{i} \frac{\langle k^2 \rangle}{\langle k \rangle}^{i-1} \tag{3.37}$$

and then the corresponding B coefficient is given by

$$B = \frac{\langle T \rangle \langle k \rangle}{\langle k \ \rangle} \sum_{i=2} \frac{i-1}{i} \frac{\langle k^2 \rangle}{\langle k \rangle}^{i-1} \tag{3.38}$$

3.6.1. The meritocracy of SF networks

In order to calculate the meritocracy and topocracy of SF networks we begin by assuming that on average most of the transactions happen through chains whose distance is approximately equal to the average path length of the network. Thus, similarly as in section 3.3.1 we can approximate the total payoff associated to rockstars as:

$$_R \quad T\frac{N(N-1)}{\langle l\rangle} \tag{3.39}$$

Since the total wealth of the system is still the same that in the ER case ($TN(N-1)$), the meritocracy for SF networks can be expressed as:

$$\_^R \quad \frac{1}{\langle l\rangle} \tag{3.40}$$

Therefore, for SF networks the decay of meritocracy depends on the size of the network and exponent of the distribution.

For the extreme case of $\gamma = 2$ the degree of the biggest hub in the network grows linearly with the size of the network $k_{max} \sim N$. Thus, the network takes a hub configuration where the nodes are not further than two hops away. Therefore the meritocracy of a SF network with $\gamma = 2$ does not decay with the size of the network. As we said before in a star network approximately all transactions take place through a single intermediary (the hub), and therefore, 50% of the wealth would go for the rockstars and the other half for the middlemen. However, such a world would not be ideally meritocratic, as the inequality among the hub centering the network and the rest of nodes would reach its maximum value.

For networks with exponent between two and three, the average distance among nodes is given by equation 3.35. By combining this equation with equation 3.40 we obtain:

$$\_^R =\sim \frac{\ln(\gamma-1)}{\ln\ln N} \tag{3.41}$$

This equation predicts that the meritocracy of societies decreases as $\ln\ln N$, a significantly slower dependence than the $\ln N$ we derived earlier for ER networks. Networks in this regime are called ultra-small, as the emergence of hubs radically reduces the path length. They do so by linking to a large number of small degree nodes, creating short distances between them. Hence, in this regime a SF network of the same size and average connectivity as an ER network redistributes a significantly higher fraction of the wealth through

the rockstar channel. However, on such networks the power available for the few middlemen hubs would be significantly higher than the one available for rockstars; and the resulting inequality would be comparatively higher than in the ER case.

SF networks with a critical exponent of $\gamma = 3$ such as the BA network have a different behavior in which the $\ln N$ dependence of the ER network returns. Yet the calculations in [CH03] propose a $\ln \ln N$ correction. Hence, in this case the meritocracy decays as:

$$\underline{\quad}_R \sim \frac{\ln \ln N}{\ln N} \tag{3.42}$$

Finally, SF network with $\gamma > 3$ follow the same small-world effect as the ER case and we recover the decay of meritocracy we obtained for such networks on section 3.3.1.

3.6.2. Payo distribution

The corresponding expression for the *middlemen* income distribution can be easily derived by taking into account that now the degree distribution follows $P(k) \sim k^{-\gamma}$, this giving a normalized [1] payoff distribution of:

$$P(\ _M) = m^{\gamma-1}(\gamma-1)\quad \frac{N}{N-1}\quad ^{-1}\ B^{\frac{\gamma-1}{\eta}}\quad _M^{\frac{1-\gamma-\eta}{\eta}} \tag{3.43}$$

that is a power-law in $\ _M$ with an exponent equal to $(1-\gamma-\ )$ [GKK01].

Moreover, expressions (3.31) and (3.32) are valid for $\mathcal{M}$ and $\mathcal{T}$ also for a SF network [assuming that $\mathrm{cov}(k\ \ _R)$ is negligible compared to $\mathrm{cov}(k\ \ _M)$]. In this case the distribution width is given by

$$^2{}_M = B^2\ (\langle k^2\ \rangle - \langle k\ \rangle^2) \tag{3.44}$$

Hence

To conclude, it should be stated that all moments of the SF degree distribution appearing in the different equations of this section are given by

$$\langle k^M \rangle = -m^M\quad \frac{\gamma-1}{M-\gamma+1}\quad \frac{N}{N-1}\quad N^{\frac{M-\gamma+1}{\gamma-1}} - 1 \tag{3.45}$$

Note that this momentum can diverge for large networks with exponent between two and three.

[1] The normalization factor of the distribution has been obtained using the limits $\pi_M^{\min} = m^\eta B$ and $\pi_M^{\max} = m^\eta N^{\eta/(\gamma-1)} B$, corresponding to $k^{\min} = m$ and $k^{\max} = K = mN^{1/(\gamma-1)}$, respectively [GKK01].

3.7. Generalization To Arbitrary Prices

Do prices have an effect on the connectivity threshold separating the meritocratic and topocratic regime of the system that was discussed in Section 3.2.2? To explore this question we assume that each individual sells her content at a price s_i. Under this assumption the total payoff collected by rockstars is equal to:

$$_R = \sum_i \ _{R_i} = C \sum_i T_i s_i \tag{3.46}$$

whereas the total payoff of the system is given by

$$= \ _R + \ _M = \sum_i \sum_{j=i} T_i s_i = (N-1) \sum_i T_i s_i \tag{3.47}$$

Hence, the threshold condition $_R \ = C \ (N-1)$ is identical than the one found when the prices are equal for all individuals. This means that the threshold separating the meritocratic and topocratic regimes ($k = N^{1\ 2}$) is valid for an arbitrary vector of prices, and is therefore true even for a system where prices are not in equilibrium. Moreover, we note that the equality $_R \ = C \ (N-1)$ depends only on $C(k)$, which is a structural parameter of the network. Hence, generalizations of these results to alternative network topologies will always be independent of prices and are reduced to determining the relationship between N and k that satisfies the condition $C \ (N-1) \quad 1\ 2$

3.8. Discussion

For inequality to exist, there must be a story justifying why those at the top are entitled to more than those at the bottom. Centuries ago, European monarchs used divinity to justify their privileged positions. It was their connection to God what made them special, and, ultimately legitimized their special status [Sti12]. In our modern era, justifying inequality based on divine right is no longer acceptable and a number of scientific dictums have emerged to fill the societal role once filled by holy explanations. Marx and Friedman pronounced themselves in this area, and although they did not share their view on economics, they shared the sense of poetry in their expression. In The Critique of the Gotha Program, Marx famously said From each according to his ability, to each according to his need [Mar08]. Through this phrase Marx attempted to convey what he thought should be the economic relationship between an individual and society: individual's contribute according to their

ability, but should receive according to their need. More than 100 years later Friedman used Marx phrasing to voice what he thought was the right interpretation of this relationship to each according to what he and the instruments he owns produce [Fri80].

Economies, however are made of more than talents and property. As the social capital and embeddedness theory has often remarked[Uzz96, Uzz97, Bur09, Bur04, Gra85, Col88], people are structured in social networks. These networks can be instrumental drivers of inequality in centrally planned economies, but also in free markets, where connections to business elites can take the role that connections to party leaders have in autocratic regimes. Networks, thus, affect the functioning of decentralized economies and limit the often desired equality of opportunity since they help determine the information and resources available to each individual. In the context of equality of opportunities, John Rawls argues that equality of fair opportunity will only be satisfied in a society where the same native talent and the same ambition have the same prospects of success [Raw99, Raw01]. Policy-makers who adhere to Rawls ideas have emphasized the field-leveling role of inheritance taxes, education and anti-discriminatory policies in the labor market. Yet, opportunities are not constrained only by talents, education and property, but also by the connections available to each individual, which cannot be taxed. Hence a thorough understanding of the meritocracy of market mechanisms cannot be achieved without understanding the effects of an individual's position in a network and its relative effect with respect to other forms of advantage where field leveling policies do exist.

In a 21st century context the results of this paper also speak about the social changes that are implied by recent changes in technology. In recent years the emergence of the internet has given rise to a world in which it is much easier for individuals to market directly to each other, or at least, through one large intermediator (such as iTunes, Amazon or eBay). Our model predicts that these changes should increase the meritocracy of society since they help reduce the long chain of intermediations that consume valuable payoffs in a poorly connected society.

But does this mean that denser networks are unambiguously preferable to sparser networks? Not quite. Making such a judgement would require weighing the effects that network density has on meritocracy with its effect on other social and economic outcomes. Social networks do not only affect the distribution of payoffs among content producers and middlemen, but also are known to affect the outcome of coordinated collective action. For instance, evolutionary game theory suggests that cooperative strategies are more likely to emerge in networks that are not highly connected. In a public good game sparse network prevents free-riders from prospering be-

cause free-riders cannot sustain enough links to exploit multiple neighbors [SSP08, PS08, GGnCFM07]. Thus, in a sparse network, the same agents that benefit from their position as middleman might be the same agents that play a crucial role enhancing cooperation. Therefore, making a judgement on whether a denser or sparser network is more beneficial for society in general, is a matter that cannot be answered easily, since it requires weighing the effects of the network structure on meritocracy and cooperation, but also, on other relevant outcomes, from the preservation of cultural diversity to the spread of disease.

Finally, the model also invites us to explore a number of different generalizations. Two generalization seem particularly interesting. First, is the development of an endogenous model in which individuals can invest in the creation of new links, or could modify their talent, for instance, by investing in education. The ability of such a dynamic process to restore the meritocracy of the system, will be limited whenever the maximum connectivity of nodes is bounded. This is likely to be true due to time constraints and the limited cognitive capacities of individuals, but it would nevertheless be interesting to explore the strategies that can help balance meritocracy in a limited setting. The other generalization involves the use of a model of this kind to explain the properties of real world networks. In particular, one could venture that the organization of society around small social groups might be a way for large groups of people to form structures that can ensure meritocracy in the local context of a group of peers. More research will be required to answer these questions and help us elucidate the role that networks play in defining the boundaries between meritocracy and topocracy.

Chapter 4

Twitter: An online social network

4.1. Introduction

Recent changes in technology are radically changing the communication patterns, and how information reaches the vast majority of the population. The number of users engaged to online social networks, social media, or blogs, is rapidly growing all around the globe. Nowadays, Twitter is one of the most popular social media platforms and its main feature consists in allowing people to post and exchange text messages limited by 140 characters. This platform is specially suitable to conduct computational social science analysis [LPA+09], as it represents a wide variety of communications, going from personal to those coming from traditional mass media [WHMW11].

On Twitter all messages may be identified using keywords called hashtags [Pös11]. This mechanism generates the trending topics, and people use them to discuss and exchange ideas without the necessity of having any explicit relation. Regarding research, the analysis of hashtag usage has helped to predict social relations [RTU11] or collective attention [LGRC11].

In addition to hashtags, Twitter features several interaction mechanisms to facilitate the communication among users. These mechanisms establish different layers through which users can communicate and exchange information. Hence, Twitter can be seen as a multiplex or multilayer social network composed by the follower, mention and retweet layers. Multiplex networks [BBC+14] [DDSRC+13] [CGGZ+13] can help our understanding of a myriad of complex systems, ranging from social networks to biological systems, as most of them do not operate in isolation but through multiple interconnected layers. Thus, the main advantage of this new formalism is that it incorpo-

rates multiple channels of connectivity, what makes it specially suitable to describe systems where the properties and neighbors of each node vary across layers.

In this chapter, we present Twitter as a multilayer social network defined by the follower, mention and retweet interaction channels. The first interaction mechanism, is the ability of people to follow and be followed by the rest of users. This mechanism is a passive mechanism that allows users to receive the messages written by their followees at real time. By the same token they automatically deliver their posted messages to their followers. Thus, this mechanism establishes the followers layer, where users are connected among each other, according to who follows who. The links at this layer establish the substratum through which messages are delivered. Previous research have shown the complex properties in this network [KLPM10]. For example, it presents a scale-free [BA99] degree distribution, the small world effect [Mil67] [WS98], and a modular structure [New06] with users clustered around leaders. Although having a large number of followers increases the visibility of the tweets posted by users, it not necessarily makes them in uential [CHBG10]. Twitter also allows users to retransmit or retweet messages posted by someone else. The retweet mechanism allows individual messages to propagate and travel throughout the social network, and also serves as a way for people to endorse their point of view over specific subjects [BGL10]. Finally, the third available interaction mechanism is the mention. By mentioning someone's username in the message text, people are able to send directed messages to other users. Whenever a user is mentioned on a tweet, she gets notified about it, as it appears in her private in-box, significantly increasing her chances of reading it. This mechanism is used either to establish conversations between users, through the exchange of messages; or to refer somebody in the messages text [HH09]. Overall, a high number of followers implies more visibility for the messages. However, it does not make a user in uential in the active layers, as this depends on the value of the tweets content (retweets), or the name value of the user (mentions) [CHBG10].

In this chapter we begin by defining the the three different Twitter layers: follower, mention and retweet. Next, we describe the methodology followed to download Twitter data, and characterize the datasets that we use on this thesis. Finally, we discuss the possible limitations that one needs to consider when conducting research with Twitter data.

4.2. Twitter as a multilayer social network

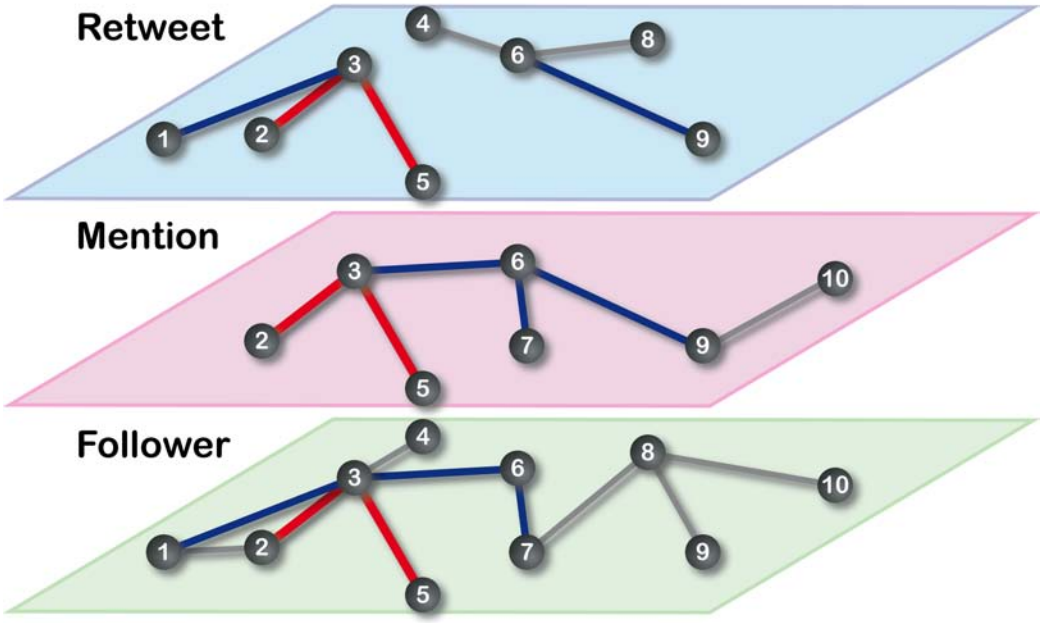

Figure 4.1: Schematic representation of Twitter as a multilayer social network. The follower, mention and retweet layers have been represented at different levels. Links occurring on a single layer are colored in gray, while those occurring in two of them are colored in blue, and finally those present in all layers are colored in red.

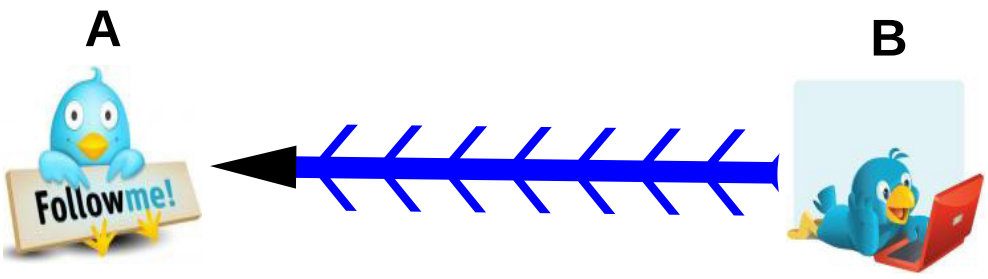

Figure 4.2: Example of a link in the followers network. B follows A and therefore receives the messages posted by A.

By distinguishing among the different interaction mechanisms available on Twitter, we can define this platform as multilayer online social network of three layers: follower, mention, and retweet. We have illustrated this definition on Figure 4.1. In it, the bottom layer (green) represents the follower layer, where links represent who follows who. The middle layer (pink), represents the mention layer. On it, links indicate who mentioned who on her tweets, and the weight [BBPSV04] [SBV09] of links represent the number times it occurred. Finally, the top layer (blue) represents the retweet layer. At this level links indicate who retweeted whom, and the weight quantifies the number of retweets going from one user to another. As the figure shows, not all users have to be present on all layers, although all of them will be present in the follower layer, by the sole fact of participating on the conversation. Similarly, links are not necessarily repeated on more than one layer, although mentions and retweets tend to occur through the followers layer. On the figure, gray links, indicate that the link took place on a sole layer, while those in blue occurred in two layers, and red ones were repeated in all the layers.

4.2.1. Follower Layer

The first and most basic Twitter interaction layer is the Follower network. We identify this layer as a passive layer that represents the social substratum through which most of the information ows. In it, a new edge is created whenever a user, B, decides to follow another user, A, (direction B A). An scheme of the follower mechanism is illustrated on Figure 4.2. Thus, the edges on this layer represent who follows who. In other words, the opposite direction of edges indicate who received whose messages and therefore, the

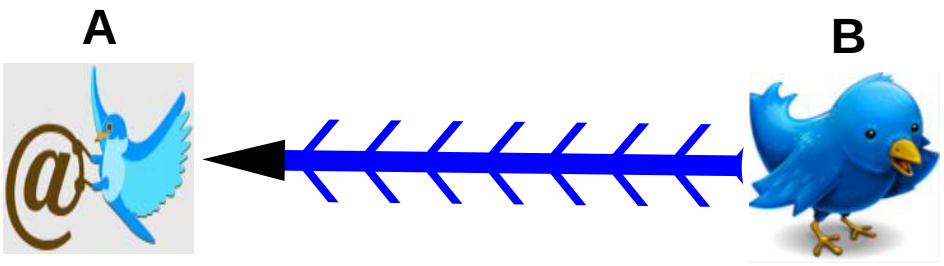

Figure 4.3: Example illustrating a link in the mention network. B mentions A in one of her tweets. Hence, the tweet appears in A's private inbox

direction in which information travels. As a consequence, the follow interaction is a nonreciprocal relation, and the network resulting from this layer is asymmetric directed and non weighted. At this layer, the in-degree of a given node accounts for how many people follow her, *i e* how many people receive her messages. On the other hand, the out-degree measures the number of users that a certain user follows, which indicates from how many users she receives messages.

4.2.2. Mention Layer

The second considered Twitter layer is the mention one. By mentioning someone's username in the message text, people are able to send directed messages to the mentioned user's inbox. This mechanism is often used to establish conversations between users, or just to refer somebody in the messages text [HH09]. Hence, in this layer a new edge appears when a new message posted by user, B, contains a mention to another user, A, (with direction $B \quad A$). A is notified about the new tweet, what significantly increases her chances of reading it. In this layer the weight of links indicate the number of mentions going on among users. An schema explaining the mention mechanism is illustrated on Figure 4.3.

4.2.3. Retweet Layer

The last considered layer is the retweet layer. The structure and heterogeneity of the follower layer has a big impact on retweets, as it raises a high level of disparity in the reception of the messages, and consequently in the information spreading process. This layer is considerably smaller and

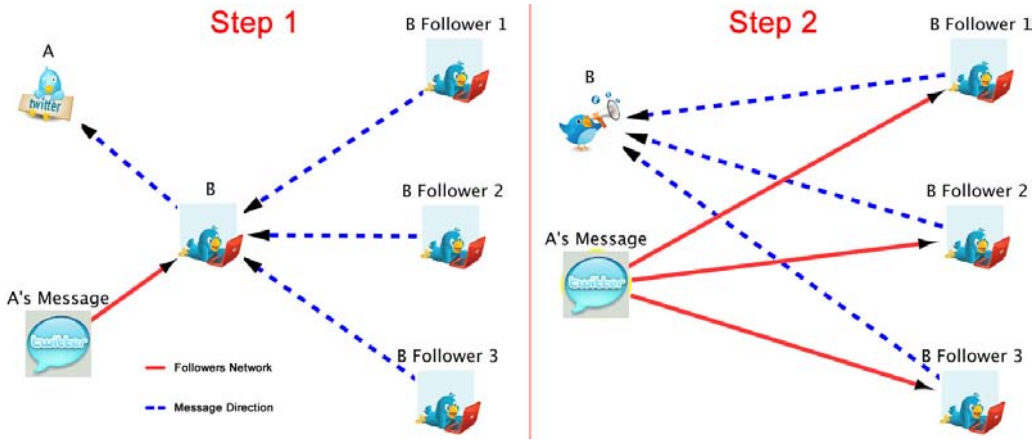

Figure 4.4: Example illustrating retweets. B retweets a message posted by A. Thus, B's followers also receive the tweet (Originally posted by A).

sparser than the followers one. This fact evidences that users are much more selective when actively spreading information, than when just receiving or reading it [HRW08]. To further understand how users retweet, we analyzed the emergent retweet network from the studied conversations on the following chapters. At this layer edges are created whenever a user retransmits a message originally posted by someone else. Hence, edges are directed and their weight indicates the number of times users retweeted each other, plus the number of subsequent propagators that retweeted the same message. Most of the ux at this layer occurs through links of the followers graph. An scheme of the retweet mechanism is illustrated on Figure 4.4.

4.3. Datasets

4.3.1. Data gathering

Twitter has several Application Program Interfaces (API) to access and download Twitter public information, such as tweets or profiles. These APIs were used in this PhD thesis to gather the data that will be analyzed in the following chapters. There are three main Twitter APIs:

1. The Search API [1] queries messages from a temporal index of recent tweets, posted within a lapse of a week old. To perform queries, users must specify a keyword to search within the message's text. Alternatively, once specify the coordinates of the twees. Its limitations depend on the complexity

[1] https://dev.twitter.com/docs/using-search

and frequency of the queries, rather than on a percentage of the total main stream.

2. The Stream API [2] delivers real time tweets, limited to a maximum of a 1% of the total main stream. Users can filter the queries by specifying keywords, or coordinates for geolocated tweets. In such way, one can download the total tweets regarding an specific topic or region as long as it does not represent more than 1% of the Twitter main stream.

3. Finally the REST API [3] provides programmatic access to read and write Twitter data, or follow people. It also allows to download user-related information like profiles or followers lists.

4.3.2. Datasets

Using the Twitter Search API and Twitter Stream API, we have built several datasets from public access messages. The empirical study presented on chapter 9 regards the late Venezuelan President Hugo Chávez. However, most of the work that we present in the remaining chapters refers to the 2011 Spanish general elections, and 2012 Catalan elections datasets. Their properties can be found in Table 4.1. We also supplemented the analysis with other datasets related to events, like political protests, electoral campaigns or historical announcements. To build the datasets we have queried the Twitter databases by looking for messages that contain keywords (or hashtags) that identify the events. In this section we will describe each of these datasets.

The first dataset that we use is constructed from public access messages posted in Twitter, related to the 2011 Spanish general elections. We downloaded all the messages that included the keyword *20N*, in a three week period including the election day. We chose this keyword because is an ideologically neutral identifier, used by all the political parties during the campaign and voting day. In summary we analyzed over 370.000 messages, written by over 100.000 users. We found that 40% of the messages were retweets, and over 25% contained at least one mention. This fact makes the event quite relevant, since it has been reported that retweets represent about 4% of the overall messages [Pea09]. From this point onwards we will refer to this dataset as $20N$.

The second one, labeled as $25N$, regards the last Catalan elections that took place on the 25th of November of 2012. We downloaded all the tweets searching for the specific keyword 25N that identified this conversation. To build this dataset corresponding to the Catalan elections, we downloaded

[2] https://dev.twitter.com/streaming/overview
[3] https://dev.twitter.com/rest/public

all the messages that included the keyword $25N$ posted in a seven week period including the official electoral campaign and voting day. We chose this tag, $25N$, for being ideologically neutral identifiers, used all around Spain and by all the political parties when referring to these elections. To better understand these datasets we have included, in Appendix A, a brief description about the 2012 Spanish political situation.

In chapter 9, we analyze messages from the online social network Twitter related to the Venezuelan political polarization. We queried for messages mentioning the name of the late Venezuelan President Hugo Chávez, during the events that surrounded his disease and death in 2013. We considered a two month period from February 4th, 2013 (29 days before the death announcement) to April 4th, 2013 (26 days after the death announcement). In summary, we downloaded 16,383,490 messages posted by 3,173,090 users from more than 159 countries (according to the 0.4% of geographically located messages). The Venezuelan Internet penetration represents about 40% of the population, where most of users belong to middle and middle-low class [Ten12]. Online social networks are very popular in this country. Around 33% of Venezuelans use Facebook [Ten12] and almost 10% use Twitter [Sem12]. In fact, Venezuela ranks thirteenth out of all countries in number of Twitter users [Sem12]. Moreover, Venezuela has the highest proportion of mobile Internet in Latin America at over 30% of total connections, due to the popular use of social media from mobile phones [Gsm13]. The political usage of Twitter in Venezuela is of great importance and has played a fundamental role in the recent Venezuelan history [MV12, JN12]. The late President Hugo Chávez was considered to be the second most in uential world leader on Twitter [Cou12], preceded only by the US President Barack Obama. The collective who opposes the late President, also finds on social media a channel to freely speak to their supporters and protest against the Government [MLB12].

In order to generalize results, we have also considered other datasets related to conversations of diverse nature such as sports, news, political protests and electoral campaigns. The first of these datasets is related to a Venezuelan political protest that took place exclusively by digital means at December 16th, 2010. The event consisted in posting messages identified with the hashtag $\#SOSInternetVE$. We downloaded all the messages that included this hashtag between December 14th-19th, 2010 (two days before and after the protest). At total we found 421.602 messages, written by 77.706 users. It is remarkable that 42% of messages where retweets and 60% were sent from smart mobile phones. Another of these datasets is related to a political scandal that took place on the Spanish parliament on 2012 due to some unappropriated comments from a congresswoman that had a big impact on social

Table 4.1: General information about the networks built from the two studied datasets concerning the Spanish and Catalan elections, including the number of nodes, number of edges and assortativity by language (r_L) of each network.

		Retweet Network			Mention Network		
Dataset ID	Region	Nodes	Edges	r_L	Nodes	Edges	r_L
$20N$	Spain	75546	153549	0.43	39631	86029	0.38
$25N$	Catalonia	102959	289486	0.70	29088	72129	0.37

media. This dataset was built by downloading 35.835 messages from 23.498 users, using the hashtag $\#Andreafabra$, from July 12th, 2012, to July 23th, 2012. Another dataset concerns a conversation about a Venezuelan baseball team. It was built by downloading 142.808 messages that contained the team's name *leones*, posted by 46.608 users during a 3 weeks period from Dec. 22th, 2010, to Jan. 12th, 2011. We have also constructed a dataset regarding the announcement of the Spanish separatist band, ETA, declaring the end of the armed struggle. We downloaded 617.545 messages posted by 241.292 users during a ten days period from Oct, 10th to 25th, 2011. We have also built another dataset related to the 2011 Arab Spring, by downloading 7.433.542 messages that contained the keyword (and hashtag) Egypt, posted by 1.80.715 users during a 5 weeks period, from Jan. 12th, 2011, to Feb. 17th, 2011. During this period the former Egyptian president Mubarak was overthrown by the social demonstrations. Another dataset regards the 2012 US presidential elections, for this dataset we download all the messages that contained the word Gingrich during a week period from Feb. 29th, 2012, to Mar. 3rd, 2012. This dataset is compound by 93.063 messages and 43.061 users. Finally, another dataset regarding the same elections was built by collecting messages mentioning Obama around the first televised debate from Oct. 3rd, 2012, to Oct. 5th, 2012. This dataset is compound by 6.818.782 messages and 2.265.799 users.

Most of these datasets are related to events that occurred o ine, such as televised debates, electoral processes or historical happenings, but also had a big repercussion in the online social media.

4.4. Twitter data limitations

Twitter is a social media platform where millions of people interact with each other on a daily basis. Hence, the availability of the data represents a good opportunity to further understand and analyze the structure of the

social network of actual societies. However, this data is not exempt of limitations that we need to consider in order to conduct research with Twitter data. For example, due to its technological nature, younger people and urban areas tend to be over represented in Twitter. We also need to take into account that its use is not equally extended all around the globe. For instance, it is banned from countries like China or Iran. Therefore, when extrapolating conclusions derived from Twitter to the whole society we need to take these limitations into consideration .

Chapter 5

Twitter and its predictive power

5.1. Introduction

The electoral campaign is a period preceding elections where political parties do an organized effort so that their candidates garner supporters. Maximizing the in uence of their messages over voters is the main objective. In this way, politicians use different techniques to transmit their messages in the most effective way to their potential voters, such as mass meetings, rallies, husting or media management. Understanding and exploiting in a more efficient way the available resources for information ow than your opponent can make the difference.

Over the last century mass media has been monopolized by old media , such as televisions or newspapers. However, nowadays we are attending to a transition where a new interactive online social media world is settling its bases. Online social networks, such as Twitter with over 200 million users, have become ideal platforms for information ows. This has been noted in [HH09] where they reported that these tools may serve as a framework for discussion. Other studies have been directed towards identifying in uential users [RGAH11] or discovering its commercial usage [JZSC09]. Moreover the percentage of population using online social networks has increased in recent years, reaching in Spain a 42% of the population, quantity that is almost duplicated (82%) for young adults between 18-29 years old [Pew11].

Following the idea one must be where people are , politicians are now present in the most popular online social networks. However, some politicians do not have a defined strategy for the usage of these tools and the rest are still far of exploiting all the available potential. The importance and

popularity of social media in politics became clear with Obama's campaign for the 2008 U.S. Presidential elections and his famous tweet: "This is history...", posted just after winning the elections. This fact attracted not only popular, but also scientific attention, making political conversations in Twitter a popular subject for research. Lately, the data gathered from Twitter has been used as a social sensor to predict election outcomes [TSSW10]. Other studies have focused in analyzing the interactions between different political communities [CRF+11b], and finally a proof-of-concept-model has been developed [LSAA11] to predict candidate's victory.

In this chapter we introduce a new parameter that measures the ratio of the support in Twitter between two candidates, which we call the Relative Support (RS), and apply it to the 2011 Spanish Presidential elections, to show how it can be used to indicate and quantify which candidate and in which proportion is getting more benefits from events occurring o ine. Next, we extend our results by applying the RS parameter to the French presidential elections, where Hollande obtained a very close victory. Finally, we review a research paper [CCP+14] where the authors apply the RS parameter that we propose on [BMLB12] to the last Italy national elections.

5.2. Literature review

In today's society, journalists and politicians are increasingly using Twitter to propagate information. Hence, there is a political discourse and debate taking place on Twitter that can be tracked on real time. For this reason Twitter data is emerging as an alternative to traditional polls to estimate political sentiment and voting intention [CCP+14]. An important advantage of Twitter is that the voting intention can be measured in real time. Moreover, the number of individuals discussing politics on Twitter is orders of magnitude larger than the sample sizes of traditional polls. Finally, another advantage of this new way of tracking political sentiment, is that users posting on Internet are not conscious that they are being tracked, hence, they express themselves spontaneously.

Despite some researchers doubt about the validity of this approach[GA12], during the last few years many researchers have began to analyze political conversations taking place on Twitter and used it to estimate voting intention. On [TSSW10] Tumasjan et al. used this platform to analyze the 2009 German federal election. They found that Twitter is indeed used extensively for political deliberation and that the mere number of messages mentioning a party re ects the election outcome. They also showed that their data re ected the later formed political ties and coalitions. Regarding Spain, the

2010 Catalan elections were analyzed finding a good correlation between the number of times a political party was mentioned on Twitter and the number of votes it received [CFME11]. On another study [OBRS10] the authors analyze the 2008 US presidential elections. To this end, they perform sentiment analysis by means of a semantic scrutiny of the tweets, and obtain a high correlation coefficient between polls and tweets. Similarly, another paper [OBRS10] regarding the election of the U.S. House of Representatives, found that the time series of mentions of Republican candidates correlates to the difference between the number of votes they received with respect to the Democratic candidates. Finally, another paper analyzing the 2010 US elections used network centrality measures of the candidates to predict their victory or loss [LSAA11]. The researcher achieved over an 80% rate of success.

However, election outcomes is only one among the many issues that online data can help to anticipate. For example, using Twitter to estimate the center of earthquakes works better than the previous comparable methods [SOM10]. When an earthquake occurs people post many tweets related to the earthquake, which enables detection of earthquake occurrence promptly. Online data is also being used to anticipate the spreading of u [GMP+09] [Cul10] [AMM11]. In a similar way, query logs can be used to anticipate stock-trading volumes [BBC+12]. Finally, Facebook data can reveal the opinions, choices and tastes of people [LKG+08].

Finally, when using online data, such as data gathered from Twitter, to predict, anticipate or estimate o ine events we need to take into consideration a series of limitations. First, a fraction of tweets is due to the automatized activity of robots. However, this activity can be detected and a significant part of the Twitter content is still produced by genuine users [RCM+11a] [RCM+11b]. In particular, when estimating election outcomes we need to consider that Twitter users is a biased representation of the voting population of a country. First, because young people under the voting age can be posting politicized tweets, and secondly, because young and urban individuals are over represented on this platform. However, the information regarding opinion orientation that can be extracted from the platform is still valuable [MLA+11].

5.3. 20N Time Series

We can begin to understand how the Spanish political landscape is re- ected in Twitter by comparing the number of times that each political party has been mentioned during the *20N* discussion and the number of votes it

Table 5.1: Results by political party for the votes obtained, the mentions on tweets and the messages sent from official accounts (Activity).

Political Party	**Acronym**	% **Votes**	% **Tweets**	**Activity**
Partido Popular	PP	44,62	39,92	1228
Partido Socialista Obrero Español	PSOE	28,73	26,33	1819
Izquierda Unida	IU	6,92	5,03	451
Unión Progreso y Democracia	UPyD	4,69	11,8	1852
Convergencia i Unio	CIU	4,17	4,51	208
AMAIUR	AMAIUR	1,37	2,76	11
Partido Nacionalista Vasco	PNV	1,33	2,20	11
Ezquerra Republicana de Catalunya	ERC	1,05	1,47	113

obtained in the elections. Previous studies show that there is a correlation between the number of times a political party is mentioned during an electoral campaign on Twitter and the number of votes the political party obtains [TSSW10]. These results are backed up by our study, where we find tweets to be a quite accurate survey. We prove this statement by ordering the political parties with at least a 1% of votes by the number of votes they obtained and comparing it to the number of times they were mentioned during the *20N* conversation. The results are presented in Table 5.1, where the name of these parties and their acronyms can be found. We observe that the only deviation from the predicted order is the swap of positions between UPyD and IU. Despite IU obtaining more votes, UPyD was mentioned more times. This can be explained by the much more active Twitter campaign done by UPyD in comparison to IU, that barely used this media to campaign, as it can be seen in Table 5.1.

Since in Spain there are two main political parties that out stand on top of the others, we focused our study on them. We analyze the time series of the accumulated tweets mentioning at least one of these parties, PP and PSOE, or their candidates, Rajoy and Rubalcaba. Looking at Figure 5.1A we can state two things: Firstly tweets contain more mentions to the political parties rather than to their candidates; secondly the more conservative party, PP and its candidate Rajoy, were much more mentioned than PSOE and Rubalcaba.

One of the most important results we have obtained studying the *20N* Twitter conversation, is that the time series of the accumulated tweets mentioning political parties or candidates present piecewise linear growth, as it is showed in Figure 5.1B. On top of that the points where the slope increases coincides with important events occurring outside Twitter that fuel the activity of the conversation, which occur at the same time for both parties. This fact makes us think about the ratio between the rate at which the cumulative mentions to two political parties grow as a better indicator of the outside political support to each party, than just the raw number of mentions. Following this idea we define an instant indicator of the support in Twitter between two political parties, which we call the Relative Support parameter RS_B^A, given by the following expression:

$$RS_B^A = \frac{m_A}{m_B} \tag{5.1}$$

where m_A and m_B are the slopes for the accumulated mentions to the A and B political parties.

From our point of view there are two days of special relevance in our study: the debate between the two main candidates that took place on November 7th, and the voting day on November 20th. We did a further analysis of

these two days.

During the debate, people's attention was completely focused on the two candidates, Rajoy and Rubalcaba. This provoked that, contrary to what happened during the whole campaign, the candidates were more mentioned than their corresponding political parties, as it can be seen in Figure 5.1A. Therefore, for this period, we studied the time series of the accumulated mentions to the candidates rather than the parties. The majority of tweets about *20N* posted on this day are concentrated on the two hours that lasted the debate, with a total of 2.733 messages mentioning Rubalcaba and 4.150 mentioning Rajoy. In Figure 5.1C we present a detail of the time series of the accumulated tweets for the candidates during the debate. We can observe that, in accordance with what we said before, both series present linear growth, being the slopes for both candidates constant during the whole debate, and changing its value at the end of it. The Relative Support during the debate was $RS_{Rub}^{Raj} = 1\ 53$, Rajoy over Rubalcaba. This value is pretty close to the relation between the votes (1,55) obtained by the two candidates thirteen days after (see Table 5.2).

The election day survey is one of the most relevant and reliable surveys to predict election outcomes. This makes us believe that a further study on this day must be done when analyzing election results. As it can be seen in Figure 5.1A, the major increase of political mentions occurred during the voting day, what reinforces our idea about the importance of this day. In Figure 5.1B we show a detail of the time series of the accumulated messages from 8:00 to 21:20 for the Spanish political parties. In correspondence with our theory of piecewise linear growth, in Figure 5.1B we can distinguish three important regions (D, E, F) of the space-time for this day: Voting time , Waiting for results , Results release , that we further discuss, and present in panels D, E and F.

VOTING TIME (8:00-19:00). This panel covers the entire voting period, from the opening of the electoral colleges to the closure. Over 7.500 tweets containing either PP or PSOE were posted. From the four panels studied in detail (Figure 5.1C, D, E, F), this one presents by far the lowest activity per hour, what makes it the less representative sample. The Relative Support took a value of $RS_{PSOE}^{PP} = 2\ 31$ in favor of PP, indicating that PP users were much more enthusiastic than PSOE.

WAITING FOR RESULTS (19:00-20:00). This period lasts only one hour, starting with the closure of polls and ending when the first news were released. At this point the news informed about the participation statistics,

Table 5.2: Comparison between the ratio of votes and the Relative Support parameter for the two main political parties.

Votes Ratio	**Debate**	**Voting Time**	**Waiting for Results**	**Results Release**
$\frac{PP}{PSOE} = 1\ 55$	$RS^{PP}_{PSOE} = 1\ 53$	$RS^{PP}_{PSOE} = 2\ 31$	$RS^{PP}_{PSOE} = 1\ 64$	$RS^{PP}_{PSOE} = 1\ 41$

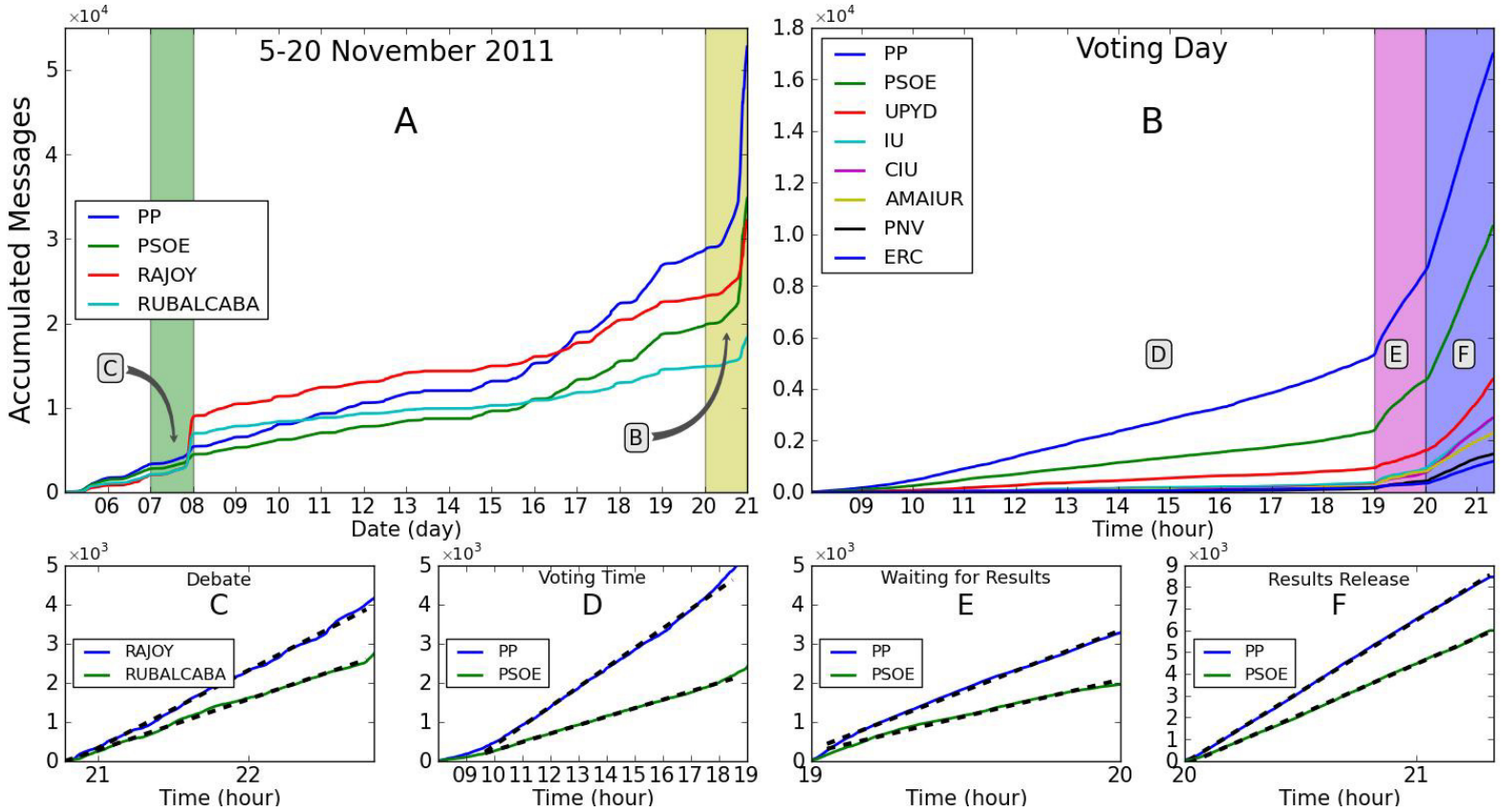

Figure 5.1: Time series of the accumulated tweets mentioning political parties and candidates for the entire campaign (A), voting day (B), debate (C), voting time (D), waiting for results (E) and results release (F). The dashed lines (C, D, E, F) represent a linear fit. At all panels, the order of the labels in the legend corresponds to the same order of the curves at their final value in descendant way. In the horizontal axis, hours are given in UTC time.

and gave provisional results for a 5% scrutiny. Over 5.000 tweets mentioning either of the two main parties were posted during this hour. During this period the Relative Support parameter estimated quite accurately the upcoming results, taking a value of, $RS^{PP}_{PSOE} = 1\ 64$ in favor of PP.

RESULTS RELEASE (20:00-21:30). This region covers the entire period in which the results were given, starting with a 5% of scrutiny and ending with an 85%, point at which the politicians made their first speeches. It was the period with more activity per hour in Twitter of the whole study, with more than 13.000 tweets posted mentioning PP or PSOE. The measure of the Relative Support while results were given was of $RS^{PP}_{PSOE} = 1\ 41$, pretty close to the relation between the votes of the two parties, as it can be seen in Table 5.2.

Summarizing, we have centered our study on the two dominant parties of the Spanish political landscape and observed that in this system the relation in votes and tweets between them coincides quite precisely (Table 5.1). We introduce a new measure to study political support in Twitter, the Relative Support between two parties RS^A_B, which we see as a useful tool to study future elections or to determine how Twitter users react to external events, and who gets more popular with them. In our study we identify the debate between the two candidates (7th of November) as the key point in Twitter. This was the point where users of the social network began to actively participate in the *20N* conversation, and during the two hours of debate people re ected their preferences in Twitter, $RS^{PP}_{PSOE} = 1\ 54$. The lack of external critical political events during the campaign and the firmness of people's vote intention, maintained the ratio of tweets constant around this value along the whole campaign. Although future work should be done in applying the RS parameter to other elections, we believe that this parameter is capable of revealing election outcomes even when o ine events occurring at the last minute change voter support. In this way it would have detected Zapatero's victory against forecast in the 2004 Spanish Presidential election, that took place four days after the 11M terrorist attack.

5.4. Further evidence

In this section we show the estimations of the relative support parameter *RS* for elections in France and Italy. In the 2012 French presidential elections, the support of voters was equally divided between the two candidates Hollande and Sarkozy. In Figure 5.2 we have visualized the accumulated

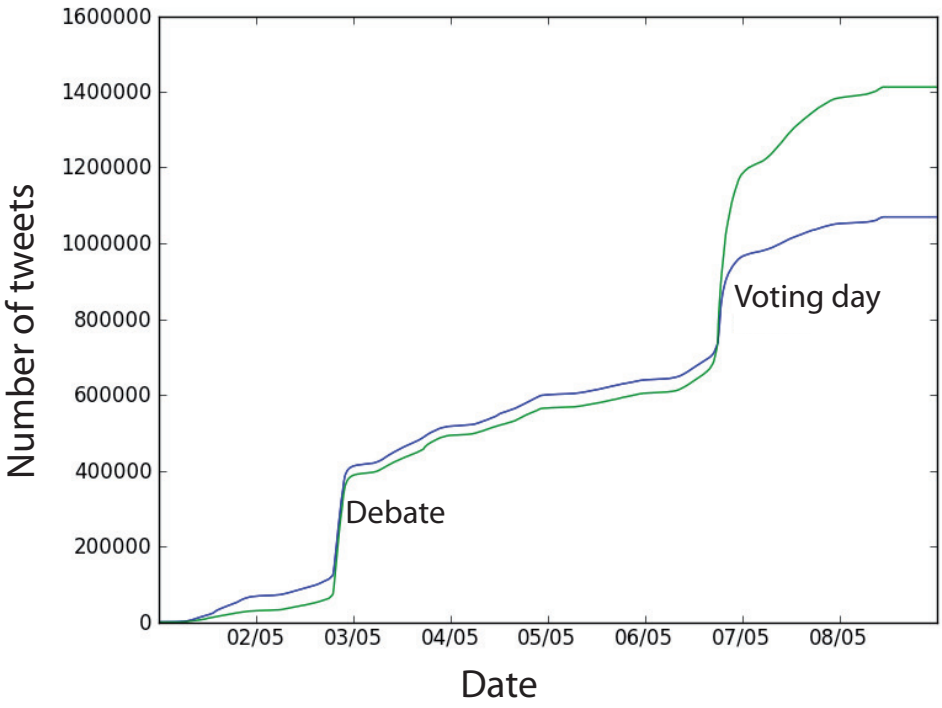

Figure 5.2: Time series of the accumulated tweets mentioning political the two candidates standing for 2012 French presidential elections. Hollande in green and Sarkozy in blue.

mentions to both candidates. If we were to estimate the results by the total volume of tweets mentioning each candidate we would have forecasted a victory for Sarkozy, as just before the voting day he was the candidate that accumulated most mentions. However, if we apply the relative support parameter at key events, such as the debate, we would have forecasted a tight victory of Hollande, who ended up winning the elections. In fact, similarly to the Spanish case the RS during the televised debate was in very good agreement to the final results. During the debate $RS = 1\ 08$ in favor of Hollande, while the ratio among their final votes was of 1 06. Moreover, during the voting day the RS parameter re ected again that Hollande was getting more votes. During this day, as can be seen in Figure 5.2, users were mentioning more frequently Hollande than Sarkozy.

Next, we summarize a research article [CCP+14] where the authors adopted our RS parameter and applied it to the 2013 Italy national elections. In general, it exhibited a high performance providing a very good proxy of the final results. Figure 5.3 visualizes their main results. The top panel shows the RS parameter for each pair of possible parties, while the lower panel displays the accumulated mentions for each party. A summary of the results is shown in Figure 5.4. As can be seen the authors recovered most of the correct relative strengths of one party against the others by calculating the respective RS. Hence, the RS parameter, seems to be a fair predictor of the relative strengths of political parties.

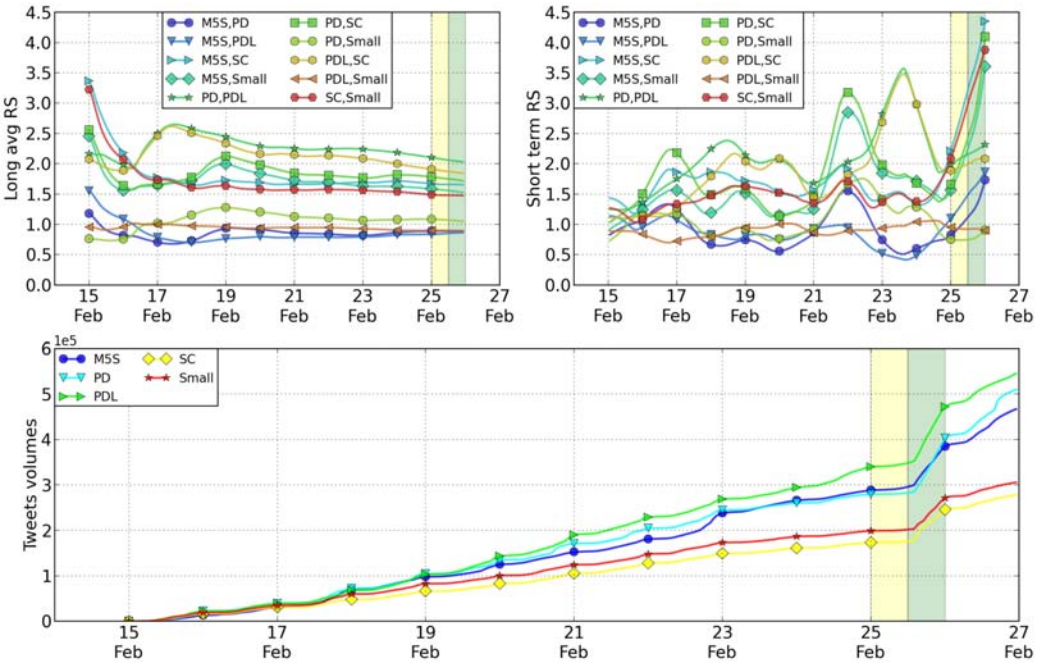

Figure 5.3: Results of RS applied to Italy elections 2013. Top left: the plot of the RS parameter computed on an average of 10 days centered on the specific day on the x-axis; top right: the values of the RS parameter with no averaging day after day. The parties have been ordered according to their final rank at the elections, with the exception of the ratio SC-SMALL. Note how most of the values are above 1, correspondingly SC-SMALL is below one. Below we have a detailed plot of the tweet volume in the final days before the elections, where changes of linear slopes are especially evident during voting days (yellow in all the figures). The yellow and green vertical lines represent the election day and the day after respectively (when exit-polls are released). Obtained from[CCP+14]

Parties A,B	$RS_A^B(48)$	Ratio of Votes
M5S-PD	**1.1±0.3**	1.0052
M5S-PDL	**0.9±0.3**	1.1850
M5S-SC	1.8±0.5	3.0769
M5S,Small	1.6±0.5	2.2208
PD-PDL	**0.8±0.2**	1.1789
PD-SC	1.7±0.1	3.0610
PD-Small	1.5±0.1	2.2093
PDL-SC	2.0±0.2	2.5966
PDL-Small	1.8±0.3	1.8741
SC-Small	**0.9±0.1**	0.7217

A correct prediction means that the ratio of votes and the RS should be either less or more than 1. We note that this parameter recovers all the the ranks with the exception of M5S-PDL, PD-PDL where no prediction is given at any rate because within the error the ratio can be both larger and smaller than one. We indicate this exception with a bold font number.
doi:10.1371/journal.pone.0095809.t002

Figure 5.4: Results of RS applied to Italy elections 2013. A correct prediction means that the ratio of votes and the RS should be either less or more than 1. We note that this parameter recovers all the the ranks with the exception of M5S-PDL, PD-PDL where no prediction is given at any rate because within the error the ratio can be both larger and smaller than one. We indicate this exception with a bold font number.Obtained from[CCP+14]

Chapter 6

Mapping the online Spanish political landscape

6.1. Introduction

Opinions are not formed in a social vacuum, but on a social network where those around us affect what we think. A classic topic in political science is how social interactions shape individuals political views, and whether the social network in which they are enmeshed has an in uence on their voting behavior [HS95] [Kno90] [BLM54] [Huc09]. In fact, back more than 60 years ago, Lazarsfeld et al. [LBG44] pointed out in The People's Choice that the impact social contacts have on voting decision is bigger than that from mass media or politicians. This resulted on the Two-step ow of communication [Rob76] [Kat57] [Wea82]. According to this theory, ideas ow from mass media (or politicians in an electoral context) to opinion leaders, and from them to the population. More recently, studies have shown that political participation is affected by the social environment; as friends, family members, or co-workers exhibit similar behavior [Ken92] [HS95] [NF00]. The explanation for this in uence is a prevalent theme of research. Burt [Bur87], distinguished between *contagion by cohesion* and *contagion by equivalence*. According to the first one, people's political preferences are directly in uenced by their social network. This theory sees social in uence as a result of intimacy within primary social grouping, and is referred in literature as direct political in uence, political assimilation, or socialization [HS95] [JSB09] [JN68]. Alternatively contagion by equivalence proposes a structural equivalence model to explain this in uence, where people base their behavior on what they observe from others that occupy a similar position to them [Dow57b] [Pop91] [HS91].

In this chapter, we intend to explore the structure of the online social

networks in which individuals are embedded when discussing politics. This is an increasingly relevant topic since online social networks and social media platforms, such as Twitter, are the latest new medium being exploited by politicians for decisive competitive advantage. Thus, today's culture is changing, Internet and social media represent a new channel through which information and ideas can quickly ow [LPA+09], bringing people a wider (and cheaper) variety of information. Today's new culture sees value in sharing information, and relies on collective wisdom [AZBA08]; just take Wikipedia [KR11] as an example. Social media fit perfectly this new context as they are about listening and being heard, about sharing information with those you trust, and about having a variety of sources of information at hand from where to choose. So, nowadays, when trying to understand the opinion formation process of individuals, we have to take into account not only their face to face relations or the propaganda coming from traditional mass media, but also the online communications that are increasingly taking place through social media platforms such as Twitter. In fact, recent research has brought evidence to show that political mobilizations in an online social network can in uence real world voting behavior [BFJ+12]. Moreover, the availability of the data represents a big opportunity to study social phenomena, such as politics [BMLB12] [CGFM12], viral marketing [LAH07], information diffusion [GJE+12], or social in uence [CF09].

The target of this chapter is to map the communication patterns behind the political conversations taking place on social media to uncover possible constrains in the online political discussion, and to answer questions such as: Do really social media platforms represent a channel through which more voices can be heard and encourage political discussion?

To this end, we begin by characterizing the main properties of the three different Twitter layers: follower, mention, and retweet. For this purpose, we review our main findings related to two political conversations: the 2010 Venezuelan protest [MLB12] [MBLB14] and the Spanish general elections of 2011 [BMLB12] [BMBL14]. We found that the structure of the follower layer conditions the retweet layer, as having a low number of followers represents a constrain to effectively propagate information on the retweet level. Next, we analyzed the rich-club ordering [CFSV06] of the global multiplex network. Moreover, we explored the in uence gained by two differentiated type of accounts, *traditional media* and *politicians*, in the three available layers. Both types of accounts are supposed to have a high value name and should produce tweets with high value content. We found both of them to have a high visibility, as they were the top followed accounts. However, their in uence on the active layers significantly differed. While politicians captured most of the collective attention, by having the highest in-degree in the mention

layer; media accounts were the top in uential on the retweet layer, as their accounts were the top retweeted. Next, we analyzed how users clustered around these in uential accounts in the three layers and show how the large and dense follower communities brake down into smaller and more segregated communities in the mention and retweet layers. Finally, we discuss the major implications of our results.

6.2. Network Properties

6.2.1. The $20N$ follower network

The $20N$ followers graph represents all the users who participated on the conversation during the considered period. On it, each node represents a user, while links represent who follows whom. Thus, edges indicate who receives whose messages and therefore, establishes the direction in which information travels. At the end of the conversation the $20N$ followers graph was compound by 110,717 nodes and 6,031,076 edges.

The in-degree of a given node accounts for how many people follow her, $i\,e$ how many people receive her messages. On the other hand, the out-degree measures the number of users that a certain user follows, which indicates from how many users she receives messages. Both the in and out degree distributions are presented on Figure 6.1. The in-degree presents a very heterogeneous distribution, reaching extremely high values as the top accounts reached over 30,000 followers. The out-degree distribution presents a less heterogeneous behavior, reaching lower maximum values. In terms of the in degree, the distributions indicate that over 50% of the users are followed by less than 15 users, while just around 1% of the users have over 1000 followers. This fact, shows the existence of a minority of ultra connected users followed by a vast majority. A large number of followers, enhances the visibility of the messages posted by the user. However, it does not necessarily makes the user more in uential, in terms of retweets or mentions gained. The presence of hubs with an extremely high in-degree or out-degree, together with the density of the network drives this layer to an average path length of 3 2. This value indicates a small world [Mil67] behavior or even an ultra small world [CH03]. This phenomenon was first reported by Stanley Milgram when he detected that on average two randomly chosen people could be linked through 6 hops (intermediary persons). Previous studies performed on the Twitter global follower graph state that the mean distance

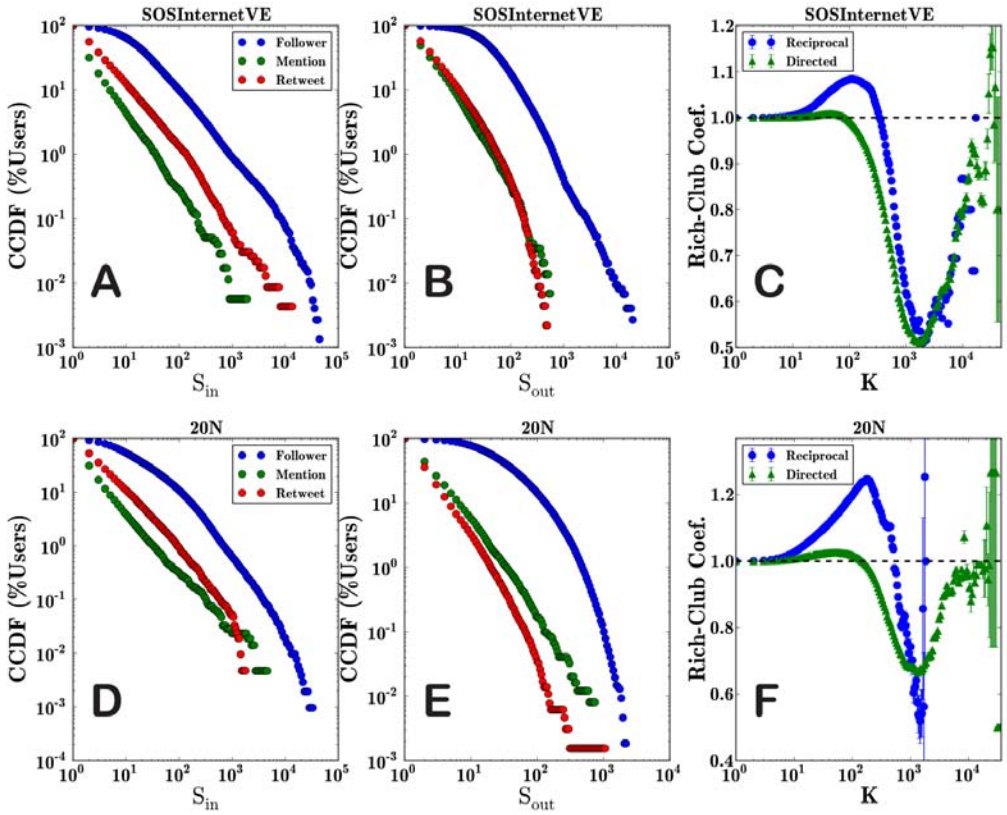

Figure 6.1: (A) Complementary cumulative distribution of the in-strength at the follower (blue), mention (green) and retweet (red) layers for the $SOSInternetVE$ dataset. (B) Complementary cumulative distribution of the out-strength at the follower (blue), mention (green) and retweet (red) layers for the $SOSInternetVE$ dataset. (C) Rich-club coefficient for the directed (green) and reciprocal (blue) multiplex networks of the $SOSInternetVE$ dataset.(D) Complementary cumulative distribution of the in-strength at the follower (blue), mention (green) and retweet (red) layers for the $20N$ dataset. (E) Complementary cumulative distribution of the out-strength at the follower (blue), mention (green) and retweet (red) layers for the $20N$ dataset. (F) Rich-club coefficient for the directed (green) and reciprocal (blue) multiplex networks of the $20N$ dataset.

between users is around 4 [JSFT07]. Comparatively, our dataset presents a shorter average path length. The explanation for this behavior, is that we are studying a specific conversation, instead of the global Twitter Network. Accordingly, our sample corresponds to an specific community of the global

Twitter network, and therefore, the users engaged on it are closer among themselves than to the remaining users.

At this level, communities are large and contain several in uential accounts, related to a same collective-like Mass Media, Political Parties, Social Activism, or geographical region. The community structure of this network is visualized on figure 6.2. For example, the largest community holds several important Spanish media (panel C). As it can be seen, various hubs stand out above the crowd. These hubs correspond to accounts of the main Spanish media, such as Europa Press, El Pais, or ABC. Another important community was formed around popular politicians and media from Catalonia (panel B). The main characteristic of this community is the use of the Catalan language. This community holds a majority of users ($\sim 66\%$) that preferentially tweeted in Catalan. Other large communities clustered together users holding the similar political ideology. These communities were formed around a single political party accounts, and therefore exhibit a highly segregated partisan structure. Despite a small fraction of links across opposed ideology communities, users tend to interact with those holding a similar ideology. This phenomenon is illustrated in Panel D that visualizes the communities of the two dominant parties: Partido Popular (PP) and Partido Socialista Obrero Español (PSOE). In this panel PP has been colored in blue, while PSOE has been colored in red. As can be appreciated there is a high political polarization. This behavior is similar to that observed for the United States, on Twitter [CRF+11a] and blogs[AG05].

6.2.2. $20N$ User Interactions

In the previous chapter we showed that the user activity is correlated with the $20N$ election results. In this section we will further analyze such activity and characterize its emergent structural and dynamical patterns based on the Twitter interaction mechanisms. First of all, we have analyzed the cumulative probability distribution for the user activity, that we define as the number of messages posted by user. This distribution follows a power law in the form $P(x > x\ ) = x^{-}\ $, where $\quad = 1\ 275 \quad 0\ 002$, as shown in Figure 6.3A. Such exponent is within the expected values for scale-free human activity phenomena [New05]. This implies a very high heterogeneity level in user behavior. In fact, we found that half of the messages were posted by only 7% of the participants, who were the most active users and posted from 8 to over 4.000 messages each, while the other half of the messages were posted by the

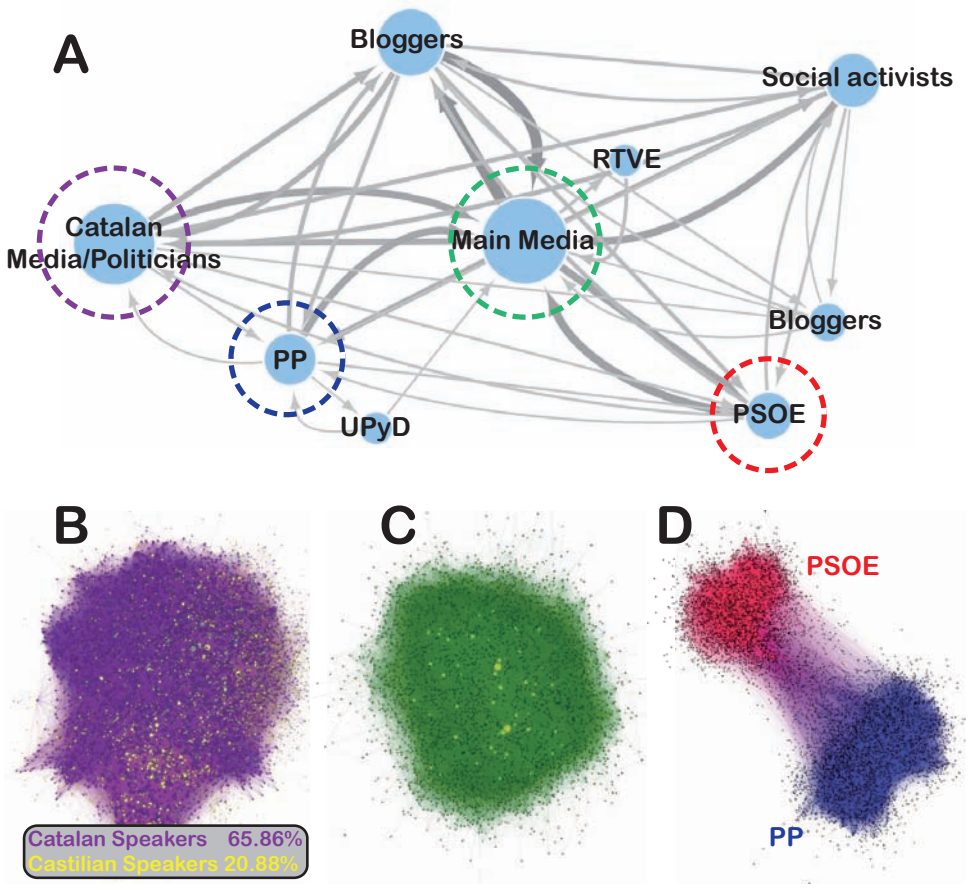

Figure 6.2: For the 20N conversation: (A) Visualization of the network of communities in the follower layer. Nodes represent communities, and links account for relations among them. (B, C, D) Visualization of the inside structure of several communities at the follower level.

remaining 93% of users, who posted less than 8 messages each. Similar results were obtained in the study of the 2009 German elections [TSSW10] and in the study of the 2005 Canadian elections [KJ09], who concluded that the political discussions during the campaign in social media were controlled by a very small fraction of the participants. However, it is unclear whether this activity really represented an actual discussion or debate. To answer this question we will next analyze the user activity taking into account the way participants interacted with each other, either by the mention or retweet mechanisms. Therefore we have built two networks according to who mentioned who and who retweeted (or retransmitted) who . Both networks have directed and weighted edges, whose weight is directly proportional to the number of times that a user has mentioned or retweeted another user. In total, the mention network has over 39.631 nodes and 86.029 links, while the retweet network has over 75.546 nodes and 153.549 links. In Table 6.1 we present the networks' main properties, some of which we will discuss next

In Figure 6.3B we present the in strength cumulative distribution for both networks. The in strength indicates the number of mentions received by user, and the number of retweets gained by user, respectively. Both measures are related to the level of collective attention that users may gather along the conversation. The in strength distributions follow power laws in the form $P(x > x\ ) = x^{-}$ where $_{M} = 1\ 14 \quad 0\ 01$ and $_{R} = 1\ 051 \quad 0\ 008$. Once more, such distributions display a high heterogeneity level found in the users profiles. As a matter of fact, we found that just 1 04% of the users were target for half of the total mentions and 2 24% of the users wrote the messages that caused half of the total retransmissions. These results show that both mechanisms are highly elitist, since a remarkably small fraction of users, mainly compound by media and politicians, concentrate half of the collective attention, while the large majority individually attracted only a few. Such collective attention is built out of adding individual efforts, which are characterized in the out strength distributions, shown in Figure 6.3C. These distributions indicate the amount of mentions or retransmissions made by user, respectively. For the mention network we found a power law distribution in the form $P(x > x\ ) = x^{-}$ where $_{M} = 1\ 438 \quad 0\ 001$, and for the retweet network we found that the data fit better to an exponentially truncated power law in the form $P(x > x\ ) = x^{-}\ e^{-x\ c}$ where $_{R} = 1\ 479 \quad 0\ 005$ and $c = 130 \quad 30$. As we found on the overall user activity distribution, the out strength distributions show that a small fraction of users (7 71% for mentions and 12 41% for retweets) concentrated over half of the activity, while the majority of users who concentrated the other half (92 29% for mentions

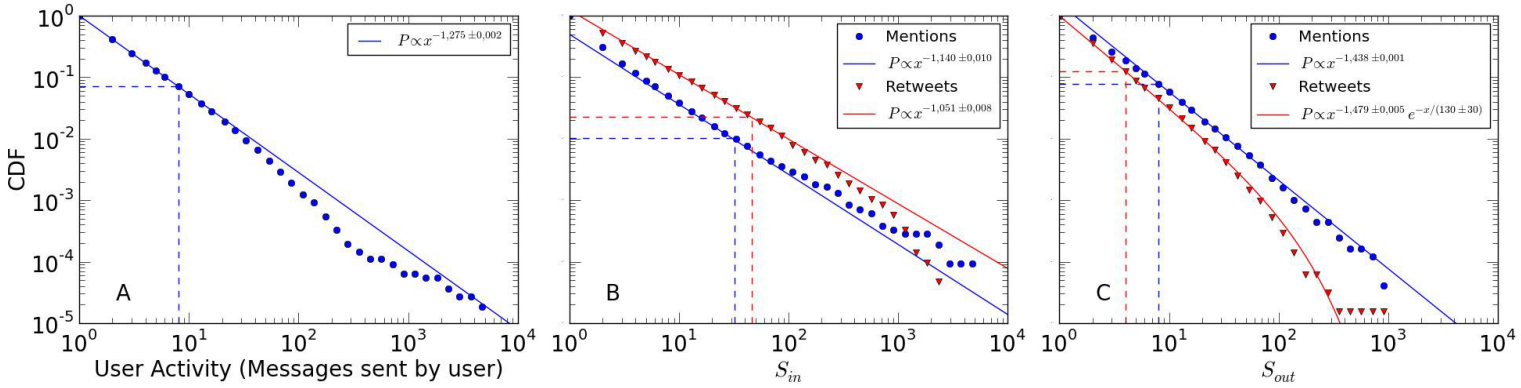

Figure 6.3: User cumulative distribution for user activity (A), mention and retweet networks in strength (B) and mention and retweet networks out strength (C). The solid lines represent the best fitted curve for each distribution. The dashed lines indicate the percentage of users that posted 50% of the messages (A), received 50% of the mentions or retweets (B) and made 50% of the mentions or retweets (C).

Table 6.1: Topological properties of the mention and retweet networks. r represents the assortativity by degree coefficient combined by in and out degrees.

Property	**Mentions**	**Retweets**
Nodes	39 631	75 546
Edges	86 029	153 549
Bidirectional Edges	2 17%	0 99%
$r_{out,out}$	−0 039	0 087
$r_{out,in}$	−0 141	−0 107
$r_{in,in}$	−0 021	−0 043
$r_{in,out}$	−0 005	0 017

and 87 59% for retweets), mentioned less than 8 users and retweeted less than 4 messages.

In order to unveil how such heterogeneous users interacted with each other, we have also calculated the assortativity by degree coefficient for both networks [New03]. As our networks have directed edges, we have calculated this measure by splitting it into combinations of in and out degree pairs [FFGP10]. As shown in Table 6.1, we found the out-in and the in-in pair to be slightly disassortative for both networks ($r^M_{out,in} = -0\ 141$, $r^R_{out,in} = -0\ 107$, $r^M_{in,in} = -0\ 021$, $r^R_{in,in} = -0\ 043$). These results reinforce the asymmetric shape detected of these networks, where the hubs that concentrate much of the incoming links, are often targeted by regular users, who neither mention nor retweet too much, and receive few of the collective attention. The out-out pair seems to be slightly assortative for the retransmission network ($r^R_{out,out} = 0\ 087$), which implies that users who retransmit a lot, also target users who also retransmit a lot. This result is related to the fact that retransmissions occur in cascades [WG10], contrary to mentions that do not imply an explicit propagation process, and actually presents a minor degree of disassortativity ($r^M_{out,out} = -0\ 039$). Finally, we found the in-out pair to be a little assortative for the retransmission network ($r^R_{in,out} = 0\ 017$), and practically not correlated for the mention network ($r^M_{in,out} = -0\ 005$). This implies that the highly targeted people, do not tend to target an specific kind of user.

Previous works on network assortativity [New03], state that social networks tend to be assortative, as popular people want to be friend with popular people, and regular people are usually friends among the regular people. However our measures indicate the opposite. This was already reported by Hu and Wang [HW09], who detected that most online social networks are

disassortative and in the same order as the networks of our study. The reason for this result is that online relations are different from real life ones. For example in Twitter, regular people are now able to relate and communicate with popular accounts, either by following, mentioning or retweeting their messages. This new kind of interactions are responsible for the changes in the structural and dynamical patterns previously reported on social networks.

We have also carried out a community structure analysis for both networks based on a random walk algorithm [RB08] and found that this conversation also presents a modular structure. In fact the largest modules in the mention network are formed around politicians, while mass media accounts centered the largest modules in the retweet network. In an effort to further discover the role that politicians and mass media accounts have played during the election campaign, we have analyzed how users mentioned and retweeted these accounts. We found that most mentions were targeted to politicians (77 83%) while the most retweeted were mass media accounts (63 24%), as it can be seen in Figure 6.4A. The importance of mass media accounts in the retweet network becomes even more clear during important o ine events that produce activity bursts (e.g. Election Day shown in the panel B of Figure 6.4). Therefore we confirm that retweets are used by users to propagate news, originally posted by the traditional media, and to endorse individual opinions.

In summary, we found that most of people participated by posting only a couple of messages that were mostly targeted to a minority part of the participants. Such elitist group is mainly compound by popular accounts, like media and politicians, and provokes the emergence of communities around them due to their high in uence. Furthermore, most of the interactions occurred in one way only, since just a small fraction of the edges, 2 17% for mentions and 0 99% for retweets, were bidirectional (see Table 6.1). This lead us to suggest that users do not actually discuss with each other. In fact, the interaction mechanisms were mainly used to campaign rather than debate, as we will see in the following section.

6.3. Information spreading on Twitter: Where a minority rule

The structure and heterogeneity of the follower layer has a big impact on retweets, as it raises a high level of disparity in the reception of the messages, and consequently in the information spreading process. The retweet layer is

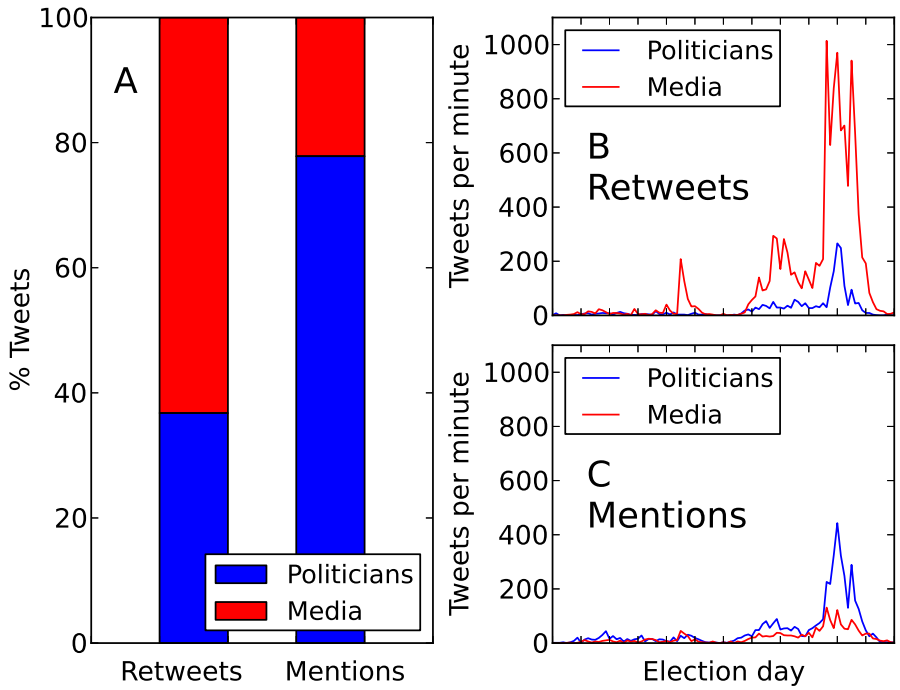

Figure 6.4: Comparison of the percentage of retweets and mentions targeted to politicians and mass media official accounts (A). Comparison of the rate of retweets (B) and mentions (C) targeted to politicians and mass media official accounts during the voting day.

considerably smaller and sparser than the followers one, as not all users retweet or are retweeted. This, evidences that users are much more selective when actively spreading information, than when just receiving or reading it [HRW08]. As we described on 4.2.3, retweets are retransmissions of a message originally posted by someone else. Hence, edges on the retweet network are directed and their weight indicates the number of times users retweeted each other, plus the number of subsequent propagators that retweeted the same message. Most of the ux at this layer occurs through links of the follower graph. This is illustrated on Figure 6.5A, where we have visualized a subset of the retweet layer (green edges), superimposed on the followers network (gray edges). In it, nodes who posted an original message are colored in red, while those who propagate it are colored in yellow. A fraction ($\sim$ 38%) of the total retweets were done by users not directly connected to the author of the original tweet. This happens because retweets tend to occur in cascades [MBLB14] that emerge when a single message is retransmitted by any user to its followers, allowing them and their own followers to do the same. We have sketched this phenomenon on Figure 6.5B, where we have visualized an example of a retweet cascade. On it, node 0 (colored in red) posts a new tweet. This tweet, travels through the follower layer to the followers of node 0 (at 1 step distance from the source) enabling them to retweet the message. Nodes in white do not retweet it, while yellow ones do. Similarly, in the second time step, the tweet reaches the followers of the nodes who retransmitted it on the first instance (at 2 steps distance from the source), giving them the opportunity to retweet it or not. The process would continue until no more nodes retweet the considered message. Retweet cascades tend to be small, as more than half were formed by only two users besides the author and just a minority of them involve a large amount of users. The reason for this behavior is that information looses its attraction when farther from the author's social surroundings [WHAT04].

As we have seen in the previous sections only a minority of users achieve a high number of retweets, as it is very hard to get retweeted. This fact suggests that the majority of users would need to post an enormous amount of tweets to gain a significant number of retweets. Figure 6.6 visualizes the relationship among the activity, retweets gained, and number of followers/followees of users in the $20N$ conversations. Thus, we address the next question: what is the relation between the in uence users gain and the effort they must employ to do so? To answer this question, in [MBLB14], we proposed a measure to rank users according to their efficiency to propagate information. Accordingly, we defined *user efficiency*, , as the ratio between retweets gained and the activity employed for it. Thus, can be expressed in the following way:

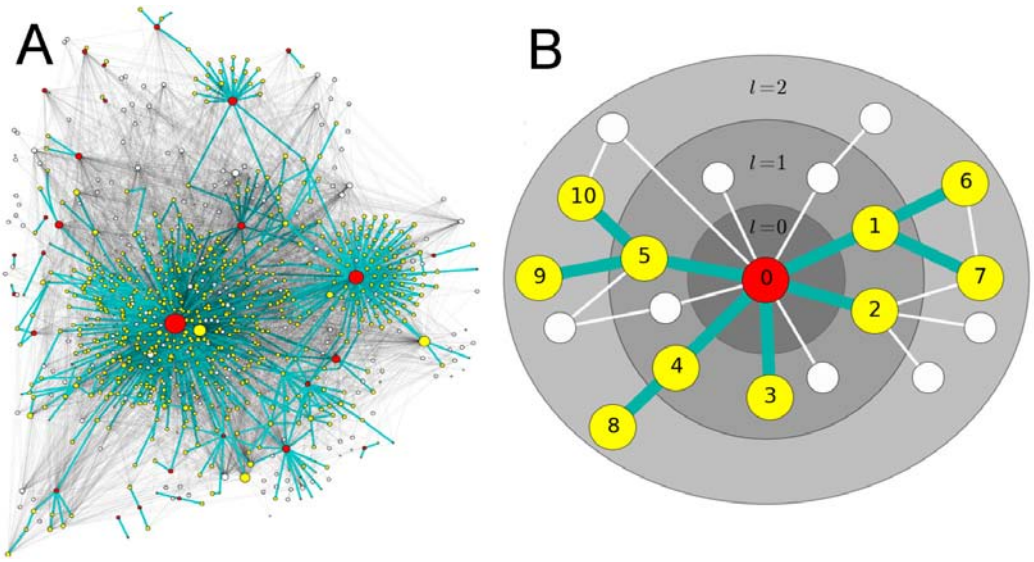

Figure 6.5: (A) Visualization of a subset of the retweet layer superimposed on the follower layer for the $SOSInternetVE$ dataset. Nodes in red posted an original message, and those in yellow retweeted a message. Links colored in green correspond to retweets, while those in gray to following relations. (B) Schema explaining how retweet occur in cascades. On it, we visually explain how the message posted by the red node travels to her followers allowing them to propagate it. In the next step, the followers of nodes who propagate the tweet receive it and are able to propagate it.

$$= \frac{R_{in}}{A} \tag{6.1}$$

where A represents the user activity or the total number of messages she posted, and R_{in} the number of retweets gained. Hence, $= 1$ establishes the threshold from inefficient to efficient (more retweets gained that activity employed). In average most of the users who get retweeted, gain as many retransmissions as messages posted. However, a minority of users accomplish a very high level of retransmission with little effort.

To further understand the distribution, we have superimposed in Fig. 6.7D the correspondent lognormal curve, with the mean and variance taken from the empirical observations of the $20N$ dataset. Lognormal distributions arise from multiplicative growing processes, like branching processes, as they are explained by the central limit theorem in the logarithmic scale [Mit04]. An example of these processes are found in viral marketing campaigns [IM11b, IM11a], where the number of leaves grow multiplicative as the branches split. As can be seen in the figure, the initial part of the distribution fits well to a lognormal curve, but after reaching its maximum the distribution changes its scaling behavior, shifting towards a power law (also superimposed in Fig. 6.7D). This means that the maximum observed efficiency is significantly higher than what a lognormal distribution predicts. These minority of extremely efficient users correspond to those with highest in-degree in the followers layer, as can be observed in Figure 6.6. This fact, shows that a lack of followers is a constrain to efficiently become in uent at the retweet layer, while occupying a privileged position on the followers layer helps being highly retweeted. Hence, topocracy (the compensation for individuals is primarily determined by the position they occupy in a network) [BBRSH14] seems to play a relevant role on Twitter.

6.3.1. Universality

In order to identify whether this distribution is constrained to the present case study or rather represents a consequence of an universal feature of the interaction mechanism, we have calculated the user efficiency distribution $p(\ )$ for other Twitter conversations. Specifically, we performed the analysis over six different datasets described in section 4.3. The datasets regard diverse topics and largely differ in their sizes measured as the volume of tweets. In figure 6.7 we present the emergent distributions for all the

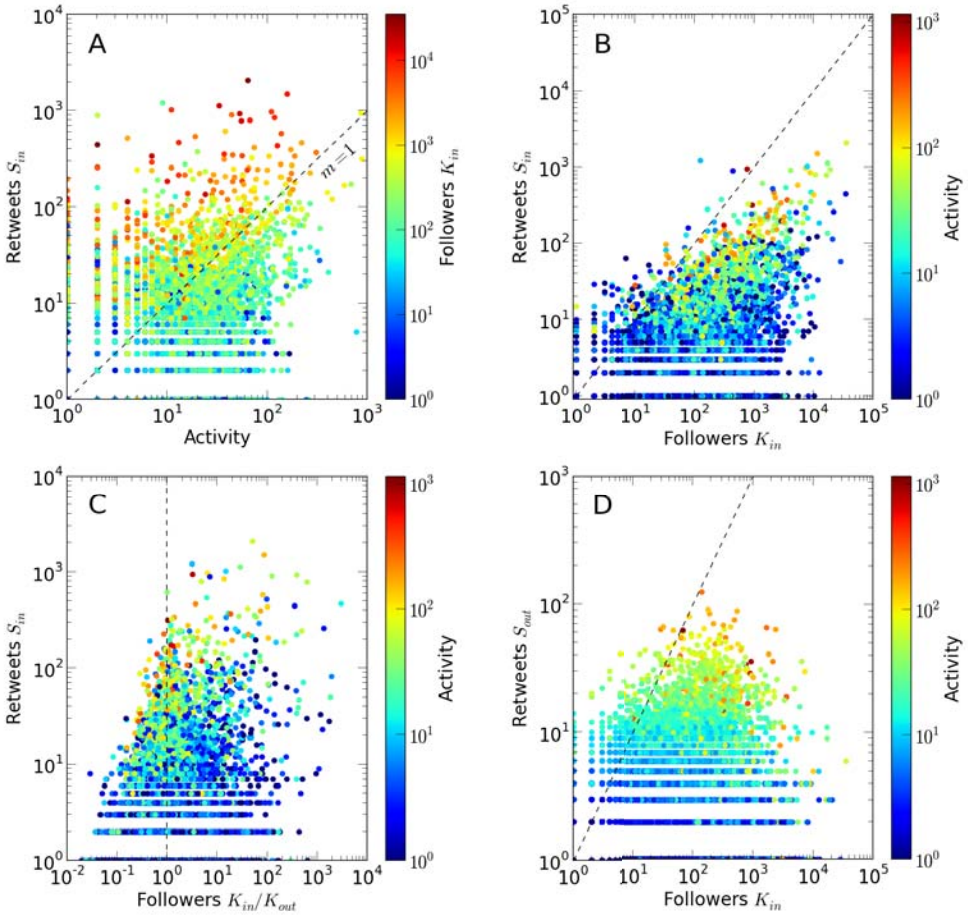

Figure 6.6: Characterization of the properties of users who participated on the $20N$ conversation.

datasets, sorted by size in ascending order(from A to F). We must note that the lognormal distribution with a minority of extremely high efficient nodes that deviate from it, emerges in all conversations, including the smallest ones (Fig. 6.7A-B). However, as the size of the conversation increases, the emergence of highly efficient users that deviate from the efficiency predicted by the lognormal becomes more significant.

The analyzed datasets are very diverse as they differ in various aspects. First, the size of the datasets ranges from four to six orders of magnitude. Moreover, the datasets consider topics of different nature, and that are discussed in several languages. Hence, it is remarkable that still the resulting distributions present a very similar pattern. This ubiquity of the resulting distributions, strongly suggests the existence of a universal behavior in the relation between individual activity, and the collective reaction to such activity. Such relation can depend on the structure of the underlying followers network. So we open the following question: what factors cause the emergence of ultra-efficient users? Is the structure of the follower network what determines the heterogeneity of the distribution? In the next section we propose and simulate a model to explain the emergence of the observed distribution.

6.3.2. The Model

In order to model the propagation of retweets, we propose a simple spreading model based on independent cascades [GLM01] that take place through the followers network. In this model, nodes can post messages. Whenever a node posts a message their followers receive it, and they can choose between retweeting it or not. By the same token if they retweet it, their followers will receive the message and also have the opportunity to retweet it. The process finishes when no more nodes retweet the message. Each message may trigger an independent cascade despite the author's previous activity. Besides, nodes may participate in several cascades at the same time. Hence, the model can be described as follows:

1)The model begins with an exogenously determined followers network. In addition to this, each node is endowed with a parameter (A_0) that determines the number of original messages the node will post. This parameter is determined by a probability density function $P(A_0)$.

2)At each time step the followers of the nodes who posted an original message or retweeted someone else message will receive the corresponding

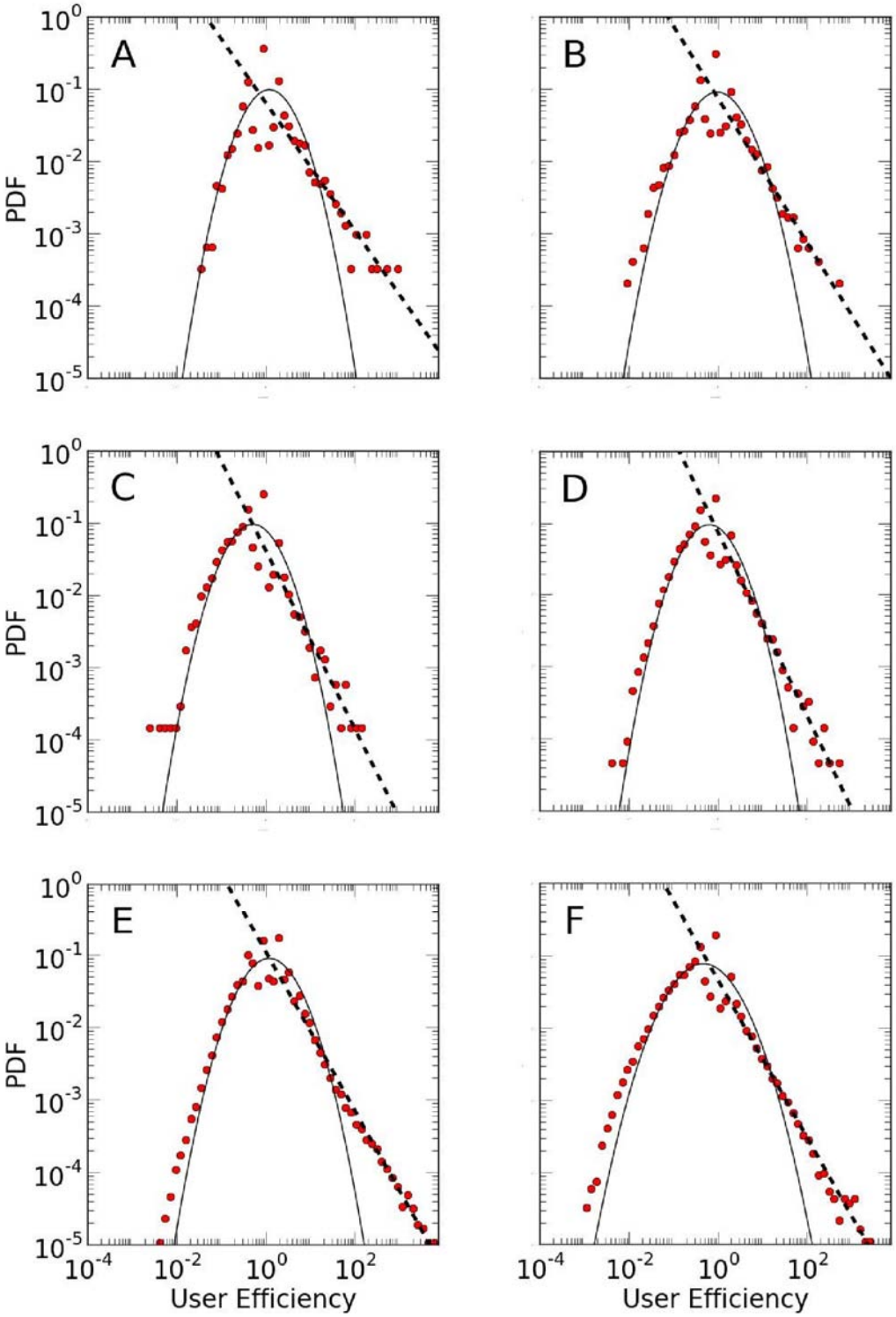

Figure 6.7: Probability density function of the user efficiency for several Twitter conversations. (A-F): (A) Andreafabra, (B) Gringich, (C) Leones, (D) 20N, (E) Obama, and (F) Egypt. The black solid lines repesent the lognormal fit, the black dashed lines represent the power-law fit and the red dots correspond to empirical distributions.

message/ messages. Thus, they will retweet it or not with a probability that depends on their distance (l) measured as number of layers to the original author of the message.

3)At the end of the process we calculate the total activity and retweets gained for each user. The total activity of a given node will be the sum of the original messages she posted A_0 and the total number of retweets she made. Thus, it can be expressed as:

$$A_i = A_{i,0} + \sum_{l=1}^{d_{max}} A_{i,l} \tag{6.2}$$

where d_{max} is the length of the deepest cascade, while $A_{i,l}$ represents the number of times node i retweeted messages that were originally posted by users at distance l from her. By the same token we can calculate the number of retweets gained by each user as:

$$R_i = \sum_{l=0}^{d_{max}-1} R_{i,l} \tag{6.3}$$

where $R_{i,l}$ represents the number of times that the tweets i retweeted at distance l from the original author, were subsequently retweeted. In the limit of $l = 0$ it represents the number of times the messages she originally posted were retweeted. This implies that a node can gain retweets either from the messages she originally posted ($R_{i,0}$), or from the messages he retweeted. Hence, the in uence of nodes does not only depend on their connectivity in the follower graph, but also on the degree of their neighbors. Finally, after all the cascades extinct, we calculate the efficiency for each user according to eq. 6.1, as well as $p(\ )$.

In order to simulate the model, we first define the underlying followers network through which the propagation process will take place. Next, we also define the initial user activity distribution $P(A_0)$ and the retransmission rate at each layer $_l$.

6.3.3. Results

We applied the model to two followers networks from the considered datasets. One of these networks corresponds to the present case study *0N* and the other one is constructed from the *#SOSInternetVE* dataset. The results of the user efficiency and retweet distributions are displayed at the top and bottom panels of Fig. 6.8, respectively. These results have been averaged over 50 model realizations. For both conversations, the system has

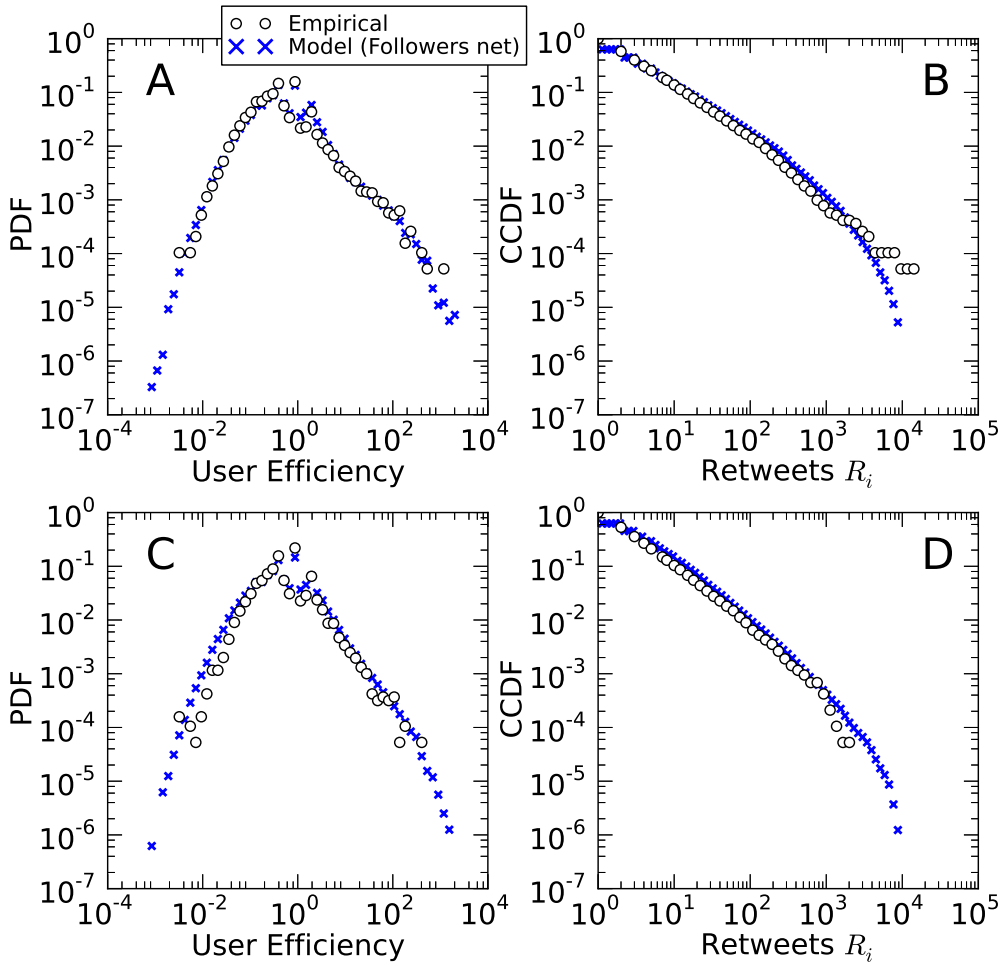

Figure 6.8: Model results to the user efficiency distribution (left column) and retweets gained by user distribution (right column), with the empirical results. The model has been applied to the followers network from the #SOSInternetVE dataset (top panel) and the 20N dataset (bottom panel).

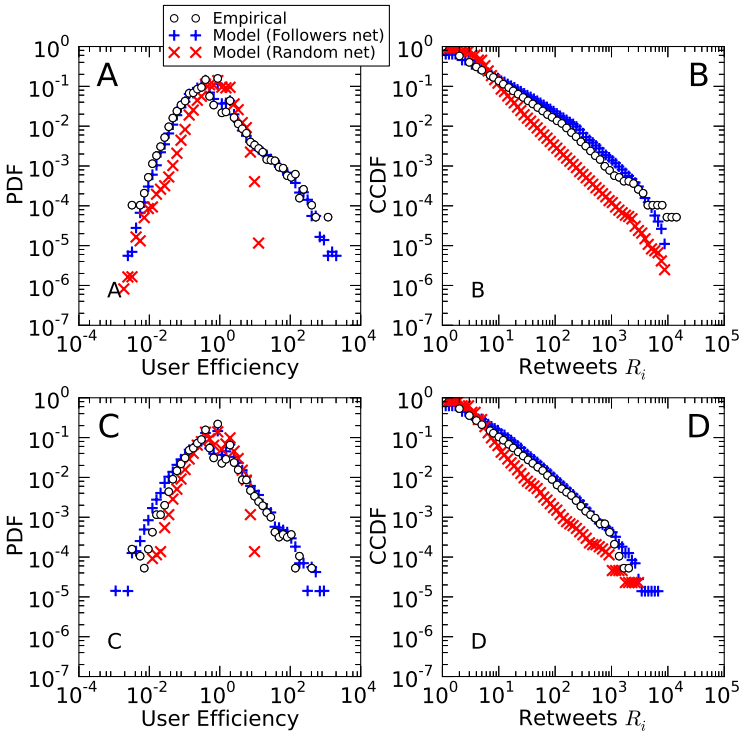

Figure 6.9: Effects of the underlying network topology on the model results in terms of the user efficiency distribution (left column) and retweets gained by user distribution (right column). The model has been applied to the followers network (blue crosses) and their randomized versions (red x symbols). Two datasets have been considered: #SOSInternetVE (top panel) and 20N (bottom panel). In all cases, an heterogeneous initial activity distribution $P(A_0)$ $A_0^{-1\,4}$ has been considered.

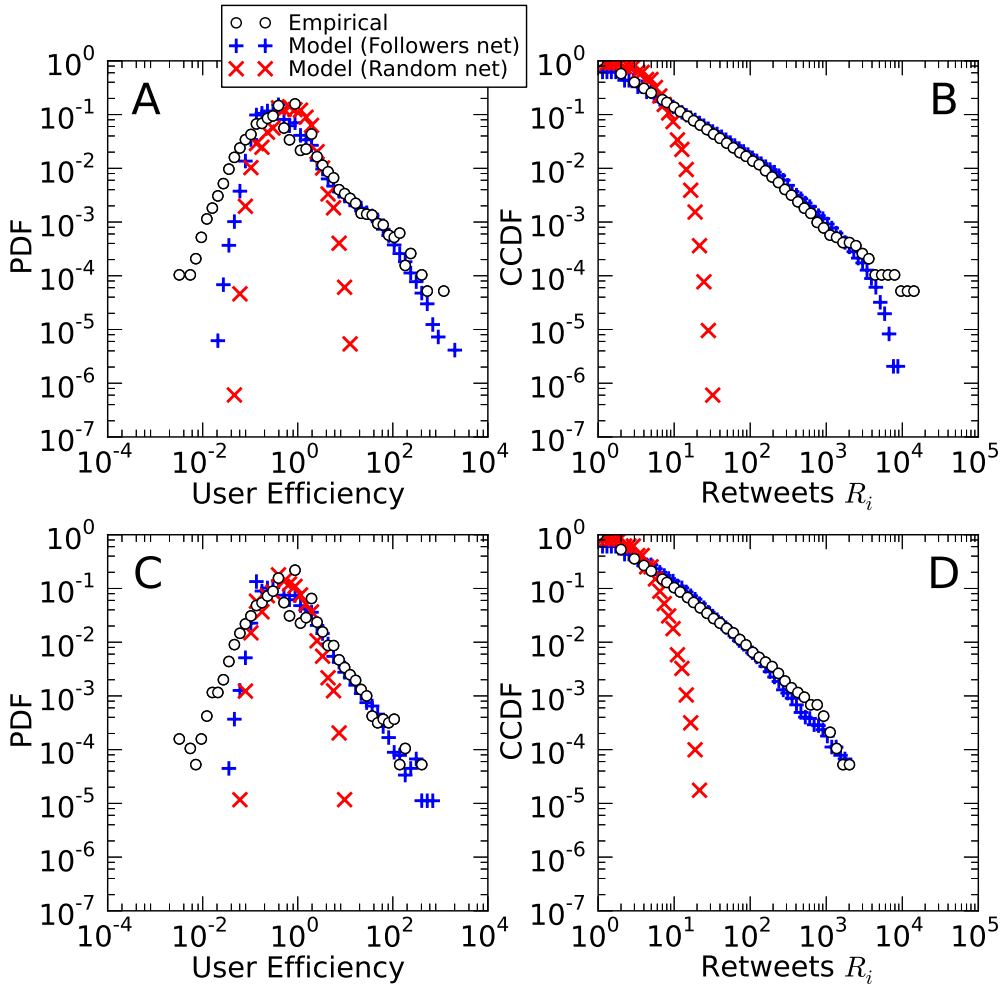

Figure 6.10: Effects of the individual user behavior on the model results in terms of the user efficiency distribution (left column) and retweets gained by user distribution (right column). The model has been applied to the followers network (blue crosses) and their randomized versions (red x symbols). Two datasets have been considered: $\#SOSInternetVE$ (top panel) and $20N$ (bottom panel). In all cases, an homogeneous activity distribution $P(A_0) = 1\ \ 6$ where $A_0 \quad [1\ \ 6]$ has been considered.

been initially excited using an heterogeneous user activity distribution in the form: $P(A_0)$ $A_0^{-1\,4}$, and the spreading probabilities at each layer were calculated from the empirical cascades. We must note that the resulting efficiency distributions plotted in Fig. 6.8A and C (blue crosses) are in very good agreement with the empirical data (open circles) for both cases. In fact, the distributions also presents the different scaling behavior at the right side of the curve. Besides, the resulting retweet distributions in Fig. 6.8B and D (blue crosses), are also in very good agreement with the empirical data (open circles). These results show that the distributions emerge from the dynamical process behind the spread of messages on Twitter.

After having validated the spreading mechanism, we will use the model to control for the effect that individual activity and network topology have on user efficiency. First, we analyze the effect that the heterogeneity of underlying network topology has on the Twitter spreading process. For this matter we applied the model over two different networks: the original followers networks, from the datasets *#SOSInternetVE* and *20N*, and their respective ER networks. When building their respective ER networks we maintained the average connectivity of the network. These random networks were built to eliminate the presence of hubs and create more homogeneous follower networks, that follow a Normal curve instead of a power law. The resulting distributions after having excited the system with the same heterogeneous $P(A_0)$ are plotted by red x symbols in Fig. 6.9A and C . We observe that the distributions resulting from these homogeneous networks present a different behavior than the empirical data. There is a slightly lower density of inefficient users, but more importantly, the maximum obtained values are almost two orders of magnitude lower than in the empirical cases. In this case the distribution does not deviate from the lognormal curve for high values. In contrast, the retweets distribution shown in Fig. 6.9B and D (red x symbols) still presents power law behavior, due to the heterogeneity of $P(A_0)$. However, the probabilities of finding highly retweeted nodes are lower than for the empirical observations. Hence, a society structured in a random network would present users that gain a large number of retweets, but only by means of employing an enormous amount of activity to do so.

Secondly, to study the impact of individual activity in the distribution we applied the model to both followers networks (the *#SOSInternetVE* and the *20N* dataset) and their corresponding ER versions, but considering an homogeneous distribution of the initial activity $P(A_0)$, in the form: $P(A_0) = 1\ 6$ where A_0 [1 6], instead of the heterogeneous one previously considered. The results of applying this homogeneous user behavior to the heterogeneous followers networks are presented by blue crosses in Fig. 6.10. It can be noticed that the resulting user efficiency distributions in Fig. 6.10A

and C, present the same behavior on the right side of the curve as the empirical observations (open circles), even though the considered user behavior is radically different than the empirical one. Besides, the retweets distribution (Fig. 6.10B and D) also coincide quite well with the empirical observations and hardly changes in comparison to the distributions obtained when users posted messages in a heterogeneous way. However, if we change the substrata to their randomized versions, the model results no longer reproduce the empirical behavior and all the distributions loose their heterogeneity (red x symbols in Fig. 6.10). This confirms that the emerging patterns are not dependent on the way users post original messages, but instead a consequence of their heterogeneous connections on the underlying network.

On Twitter, the follower network represents the channel through which information and ideas may ow. On this basis, the proposed model has shown that if the social network in which we are embedded presents an heterogenous degree distribution there will always appear highly efficient persons capable of retransmitting their ideas to the majority. Moreover, this will occur with independence of the activity strategy (number of messages) that each individual employs ($i\ e$ it does not matter wether $P(A_0)$ is drawn from a uniform or a power-law distribution). On the other hand, if Twitter users were embedded in a homogenous social network, as the ER network, the retweets gained by user would re ect the frequency and amount of posted messages, and their efficiency to propagate messages or ideas would be strongly limited. However, despite the fact that in an homogeneous society it would be more difficult to find extreme cases of high efficient users, the density of extremely low efficient users also decreases when the attention is shared homogeneously among the Twitter collective. Therefore, this shows that in order for some users to gain more collective attention, others must loose it at the same time. This is because individual attention is limited.

In summary, we have been able to model the efficiency of users to spread their opinions though Twitter, and found that the emergent patterns are remarkable in uenced by the underlying network topology. We have shown an evidence of the *robust but vulnerable* property of complex networks. In the sense that complex networks appear to be robust for most of the external excitations, as most of people post messages that do not travel at all, but vulnerable for selected excitations, as the activity performed by the highly efficient users have a remarkable impact in the resulting patterns [Wat02]. This effect is also measured through the macroscopical property of the percentage of retweets on the overall posted messages. In the protest 47% of the messages were retweets, while our simulations gave 45 3% when simulating over the followers network and 40 3 0 1% over the ER version. This additional 5% of retransmissions were only possible due to the complex and

scale-free organization of the network.

6.4. Community Structure

Online social networks provide us information on how people interact with each other through the Internet. The best approach to analyze the upper hierarchical level of such networks is community structure [New11] [NG03]. Within complex networks, a community can be defined as a subgroup of densely connected nodes inside a larger graph. Thus, in social networks a community would correspond to a group of strongly related friends. Similarly, in a politicized context, such as the present study, one expects people to group themselves by ideology. Following this idea communities correspond to the different political parties, as already noted for blogs [AG05]. Analyzing the $20N$ dataset by means of community structure we intend to map the political communication strategies, identifying large groups of users among which information ows quickly and understand the user behavioral patterns engaged among them. In order to perform the community structure analysis we used the mapequation algorithm [RB08], based on random walks, and the intuitive idea that if a community exists the random walker tends to get trapped inside it. We decided to use this method for considering it specially appropriate to study systems where links represent information exchange among nodes, such as the mention and retweet networks built from Twitter data. In this way, we build networks of communities (c-networks) [GHKV07], where each community is represented by a c-node and the ow of information going on across communities constitute the links.

In the top panels of Figure 6.11 we present, for the $20N$ conversation, both c-networks: the mention c-network (A) and retweet c-network (B). In order to obtain a clear and simplified map from where to extract useful information, we only display the 30 most representative (higher aggregated pagerank) c-nodes, and the links representing at least the 0 005% of the total inter-community ow for each c-network. Despite the small number of communities displayed, the map captures the most relevant information, as it represents over 30% of the total activity. Moreover, these most popular communities attracted the attention of users belonging to many other communities without necessarily directing tweets at them. In the figure the name of each c-node corresponds to the username with highest popularity (pagerank) inside it, and the size to the pagerank of the community (aggregated pagerank of all nodes within the community). Communities are

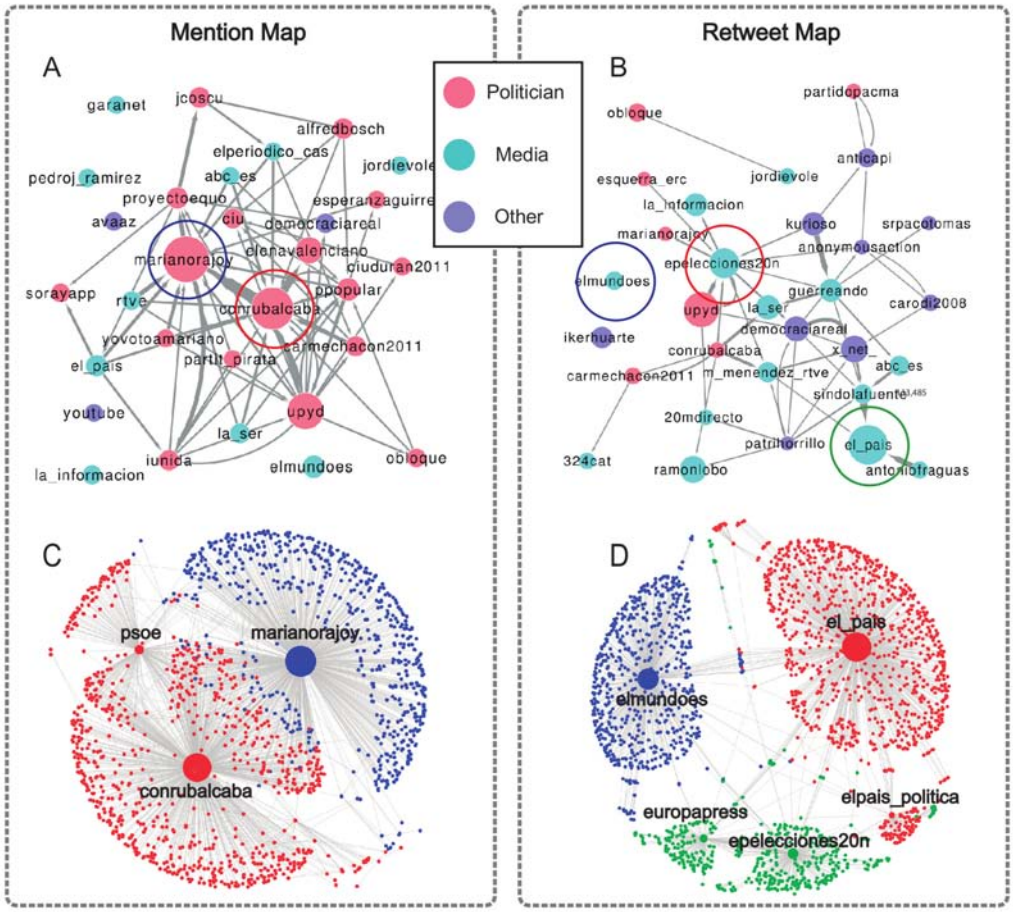

Figure 6.11: Map of the 20N mention (A) and retweet(B) c-networks, showing the 30 most important communities. The color of communities indicate their class (politician, media, others), and the size their pagerank centrality. For a clearer representation we only show the links representing at least the 0 005% of the total inter-module ow. A detail of the inside structure of some important communities (marked with circles) is displayed in (C) for the mention c-network and (D) for the retweet c-network. Nodes size represent their centrality and their color the community to which they belong.

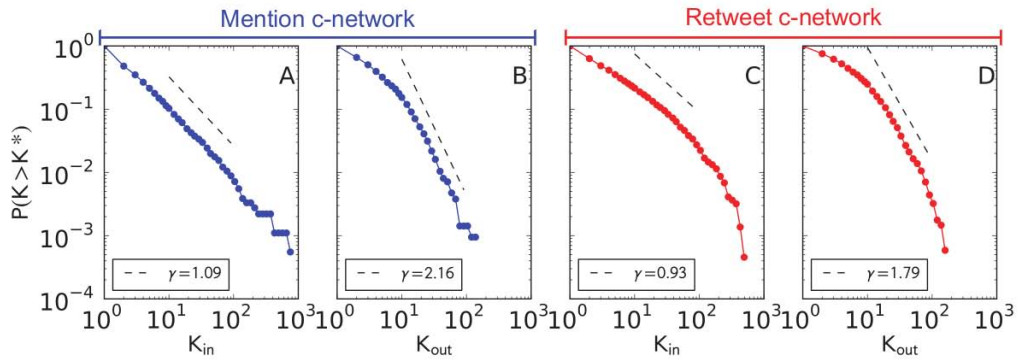

Figure 6.12: Indegree (A, C) and outdegree (B, D) complementary cumulative distribution functions for the mention (A, B) and retweet (C, D) c-networks of the $20N$ conversation. Dashed lines and their corresponding γ values are included for comparison. They do not represent fitted curves.

colored according to their classification: politicians (red), media (blue) and others (purple). The thickness of the links represent the communication ow going from one community to another. For more details about the statistical properties of the c-networks see section 6.5.

6.5. Statistical properties of the c-networks

In this section we characterize the statistical properties of the two constructed c-networks: the mention c-network and the retweet c-network. In Figure 6.12 we compare the complementary cumulative distributions of the in and out degree for both c-networks. The in and out degree measures for a community define the number of inter-community edges pointing to (indegree) and from (outdegree) the community. In other words the indegree quantifies the attention a community received from other communities, and the out degree the attention that the community paid to other communities. The indegree distribution for the mention c-network can be well fitted to a power law, in the form $P(k > k\ ) = k^{-\gamma}$ where $\gamma = 1\ 09$, meaning that the growth process of this c-network follows the preferential attachment rules: popular communities become more popular [BA99]. However the retweet c-network presents a slightly less heterogeneous distribution not reaching so extreme values. Outdegree distributions follow similar decrements for both of the c-networks. These distributions are less heterogeneous than those observed for the indegree, as they present an initial smooth decadent to promptly decrease for values larger than 10.

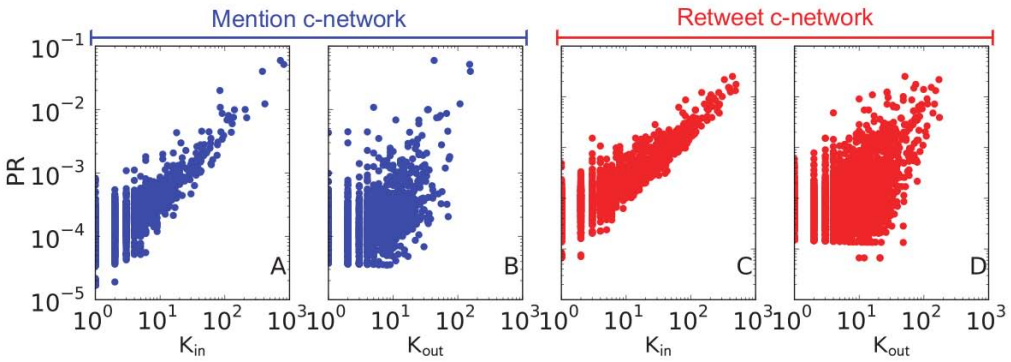

Figure 6.13: Comparison of the correlation between the aggregated pagerank (PR) and degree (measured from the c-netowrk) of communities resulting from mentions (A, B) and retweets (C, D) for the $20N$ conversation. Aggregated pagerank plotted versus indegree (A, C), and versus outdegree (B, D).

To further understand the structure of these c-networks and how the collective attention has been distributed, we have calculated the popularity, measured as the aggregated pagerank, of each community and compare it to its indegree and outdegree. Note that the aggregated pagerank of a community is calculated as the sum of the pagerank (in the mention or retweet networks) of all the nodes inside it. Thus, it is not a measure resulting from the c-network structure, but from the network at the lowest hierarchical level. Contrary the indegree and outdegree of a community are calculated from the c-network. In Figure 6.13 we present each community's indegree versus its popularity in a logarithmic scale for the mention (A) and the retweet (C) c-networks. The results re ect that the popularity of each community is well correlated to its indegree in the corresponding c-network. Meaning that the small fraction of very popular communities have been frequently targeted from users belonging to many others. This is usually explained by the presence of a famous account inside it, as discussed in section 6.6. This result differs to what we observe when plotting the outdegree versus the popularity of each community, where as we can see in panels B and D of Figure 6.13, correlation is lost. Overall, we can affirm that the most popular communities have been preferentially mentioned or retweeted by members belonging to many other communities. However, such important communities have not necessarily directed tweets to other communities in order to attract this attention.

6.6. Politicians, the main characters; Traditional media, still the main source of information

Politicians have drawn most of the attention, as their accounts were the most mentioned. This means that users addressed their messages to them by referring to their accounts when tweeting. This is illustrated in Figure 6.11A (mention c-network map), that show how the majority and more significant communities centered their attention at them. For example, the map is centered around the two most voted candidates c-nodes (*marianorajoy* and *conrubalcaba*, belonging to the two dominant parties: PP and PSOE respectively), as their accounts were the most popular.

To further understand the mention c-network and how the collective attention has been directed to politicians, we have explored the inside structure of the politicized communities. We found that communities grow around a single political alignment official accounts, as all the politicians inside a same community belong to the same party (as can be seen in Table 6.2). Politicians accounts, although representing a very small fraction of the total members; have attracted most of the inside collective attention, concentrating a very high fraction of the total pagerank (Table 6.2). To illustrate this phenomena, Figure 6.11C visualizes the inside structure of the two most central communities corresponding to the dominant parties PP and PSOE. Both of them are large communities (over 600 nodes), with a high pagerank (5 9% and 5 2%), and with a small number of politicians (7 and 11). As it can be seen, the vast majority of users only mentioned a single political party accounts. However, a minority of intermediaries linked the two sides of the conversation mentioning both candidates.

Despite being the main characters regarding mentions, politicians' loose importance when considering retweets in favor of media and anonymous people such as bloggers. This resulted in traditional media accounts prevailing in the retweet map, where they were the top retweeted in the most important communities (can be checked in Figure 6.11B and Table 6.2). This means that media accounts were the preferred source of information from where users propagated information regarding the elections. For example, the media account *epelecciones*20*n* has not only attracted many retweets within its community, but also from other communities. We must note that this account was created by the national news agency (Europapress) to report about the 20*N* elections.

When exploring the inside structure of the retweet communities, (Figure 6.11D) we found that users spread information from a minority of official

Table 6.2: Characterization of the inside structure of the top 5 communities (c-nodes) for each network of the 20/N dataset. We present them ordered, firstly by their network type (NW): mention (M) or retweet (RT) and secondly by their total pagerank (PR). In the table each community is identified by its c-node name, that corresponds to the leading account of the community, and classified according to its class (politician or media). Following we show the size of each c-node in terms of number of nodes (N) and PR; the number of official accounts inside it together with their affiliation; the dominant affiliation; the relative PR of the official accounts belonging to it (PR affiliation).

NW	Class	C-Node	Size(N/PR)	Official Accounts	Affiliation / Media	PR Affiliation
M	Politician	marianorajoy	612/5,90%	7(PP)	PP	44,94%
M	Politician	conrubalcaba	804/5,20%	11(PSOE)	PSOE	54,72%
M	Politician	upyd	533/4,00%	22(UPyD)	UPyD	47,09%
M	Politician	elenavalenciano	2/2,00%	1(PSOE)	PSOE	55,00%
M	Politician	ppopular	303/1,20%	8(PP)	PP	53,09%
RT	Media	el_pais	986/2,50%%	5(El Pais)	El Pais	74,52%
RT	Politician	upyd	863/2,20%	30(UPyD)	UPyD	41,72%
RT	Media	epeleccciones20n	461/1,80%	3(EUR)	Europapress(EUR)	52,78%
RT	Media	laser	774/1,30%	5(Ser)	Cadena Ser (Ser)	74,15%
RT	Media	la info	140/1,22%	2(LI)	La Informacion(LI)	87,50%

accounts belonging to a same media. This fact together with the limited number of retweets going on across communities re ects the fidelity of people to their preferred media, as they don't compare different sources of information. Or if they do, they just propagate news from their favorite one. Whether each community hosts a single hub or various, depends on each media online communication strategy. For example, media using various accounts to publish news on Twitter (El Pais or Europapress), present various hubs inside their community; whereas those using just a single global account (El Mundo), present an unique hub. All this is illustrated in Figure 6.11D, where the internal structure of the mentioned communities together with their inter-links are displayed.

6.7. Twitter as a multilayer social network

6.7.1. Is Twitter a rich-club like social media?

Next, we explore whether the two studied Twitter conversations present a rich-club structure ($i\ e$ highly connected nodes tend to connect among themselves) [CFSV06] or not. To this end, we measure the rich-club coefficient of the global multiplex network-the aggregate of the three layers. A first definition of the rich-club phenomenon was introduced by Zhou and Mondragon [ZM04] and can be expressed as:

$$ (k) = \frac{2E_{>k}}{N_{>k}(N_{>k} - 1)} \tag{6.4} $$

where $N_{>k}$ represents the number of nodes with degree higher than k and $E_{>k}$ denotes the number of edges among them. Hence, (k) measures the fraction between the number of actual edges and maximum number of edges that can exist among nodes with degree larger than k. This equation can be easily generalizable for directed networks, where we define the rich nodes as those with higher in-degree. Thus, eq. 6.4 can be rewritten as:

$$ (k_{in}) = \frac{E_{>k_{in}}}{N_{>k_{in}}(N_{>k_{in}} - 1)} \tag{6.5} $$

However, hubs will be naturally more densely connected among themselves than nodes with lower degree. Thus, to properly interpret whether a network presents rich-club ordering we need to compare the rich-club coefficient with its randomized case. We randomize the networks to obtain uncorrelated networks with the same degree distribution of the original one.

Hence, we can define the normalized rich-club coefficient $_{ran}$ as:

$$_{ran}(k) = \frac{(k)}{_{ran}(k)} \tag{6.6}$$

where $_{ran}$ is the rich club coefficient of the randomized network with the same degree distribution $P(k)$ of the original one. Note that for the directed case we preserved the in and out degree distributions. Values of $_{ran}$ larger than one indicate that the network presents a rich-club ordering, as the increase in the interconnectivity among large degree nodes is larger than what could be expected in the randomized case. In contrast, values below one evidence a lack of connectivity among hubs.

In order to analyze the rich-club ordering of Twitter we have first calculated $_{ran}$ for the total directed multiplex network of each conversation. Next, we have filtered these networks by only remaining reciprocal edges, and measured $_{ran}$ for the reciprocal cases. The results are presented in panels C ($SOSInternetVE$) and F ($20N$) of Figure 6.1. The global directed network does not present a rich-club structure. In fact, it presents a similar structure to the protein network [CFSV06], where hubs are not densely connected among themselves. This result, indicates that users with a high global in-degree are presiding large communities of regular users. This absence of rich-club ordering goes in agreement with the results presented in the previous sections and in [BMLB12]. In these sections, we reported the dissasortative nature of Twitter, where hubs with a large in-degree tend to be followed, metioned and retweeted by regular users. Despite, regular users can direct their attention to famous accounts on Twitter, these rich accounts do not interact with them. Hence, the reciprocal network does present a rich-club organization. The rich-club coefficient reaches its maximum around $k \sim 200$ and disappears for connectivities over 500. Hence, when considering reciprocal interactions the rich-club ordering emerges on Twitter and hubs preferentially interact among themselves in a similar way as elite scientists do in the scientific collaboration network [CFSV06].

6.7.2. Multiple leaders emerge at the di erent layers

In this section we explore whether the in uence of two different elite collectives, such as politicians and mass media, is stable through layers, or if it varies with the considered interaction. To this end, we first study the role played by politicians and mass media in each layer and how their in uence

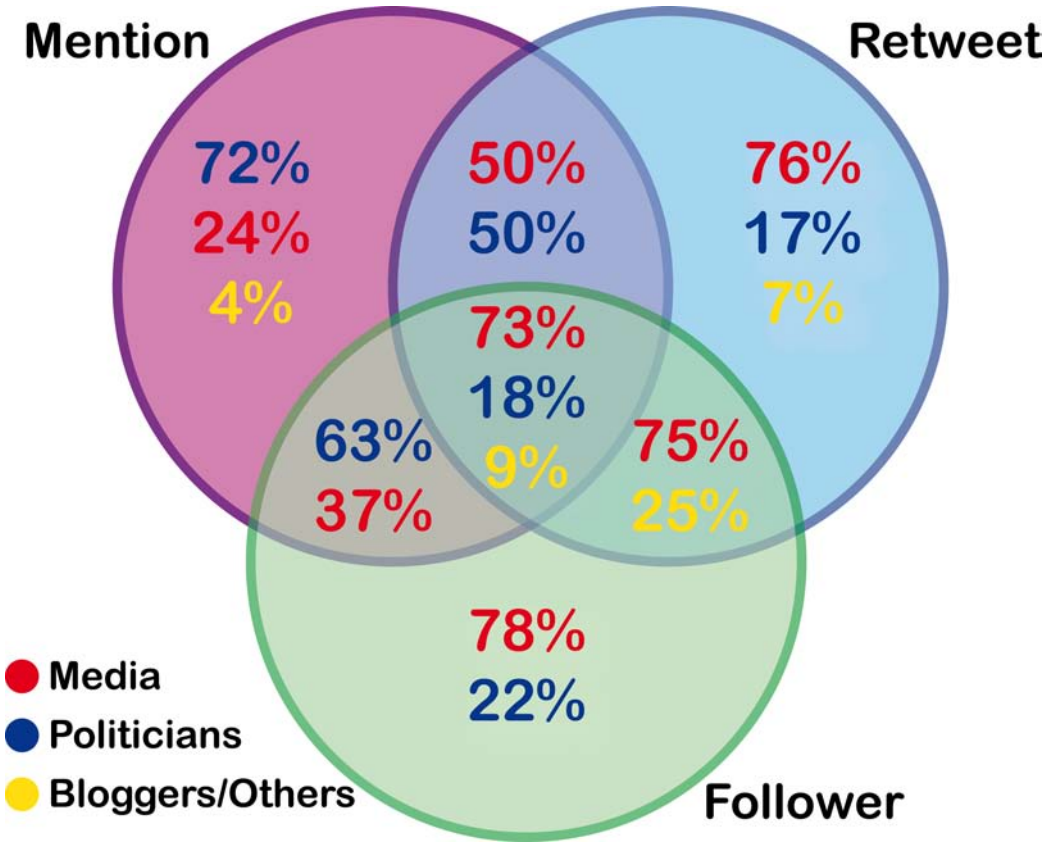

Figure 6.14: Representation of the Venn Diagram for the follower, mention and retweet layers. The percentage of accounts in each region belonging to media (red), politicians (blue), and bloggers (yellow) have been indicated for the $20N$ dataset.

varies across them. Following, we analyze the community structure of the layers, identifying the leaders of the top communities at each layer. Finally, we explore the existing relations among the communities of the different layers. To illustrate the results we will focus on the $20N$ dataset, although similar conclusions could be derived from the Venezuelan conversation following the same procedure.

We can begin to understand how the collective attention has been distributed in the three layers by identifying the top 50 in uential users (those with higher in-degree) in each one. More particularly, we have studied the role played by politicians and traditional media on each layer and on the overlapping among layers. Thus, we have classified each of the top 50 in uential accounts at each layer as either politician, media or blogger/others. We have visualized the results on Figure 6.14 by representing the Ven diagram. On each region of the diagram we have indicated the percentage of accounts belonging to each collective: politicians, media, and bloggers. As the figure shows, the relevance of politicians and media varies according to the considered layer or layers. Overall, traditional media tend to be the most in uential. However, when just considering the mention, or the overlap between the mention and follower layers, politicians are the most important. On the other hand, if just considering retweets, media emerge again as the top in uential collective. While when considering the overlap between mention and retweet both elite seem to be equally popular. This highlights how conclusions can significantly vary from layer to layer, and therefore when just considering one layer these conclusions should limit to the considered layer, rather than general for the entire Twitter.

Next, to further understand the impact that media and politicians have on the structure of the different interaction layers, we analyze their community structure. For this, we have performed a community structure analysis (using the map equation algorithm [RB08]) of the follower layer and compare the results to those reported in [BMBL14] for the mention and retweet layers. Figure 6.15 B shows the top 10 communities of the follower layer, together with the main relationships among them. In this layer, communities are large and contain several in uential accounts, related to a same collective-like Mass Media, Political Parties, Social Activism, or geographical region. For example, the largest community holds several important Spanish media (panel D). As it can be seen, various hubs stand out above the crowd. These hubs correspond to accounts of the main Spanish media, such as Europa Press, El Pais, or ABC. Another important community was formed around popular

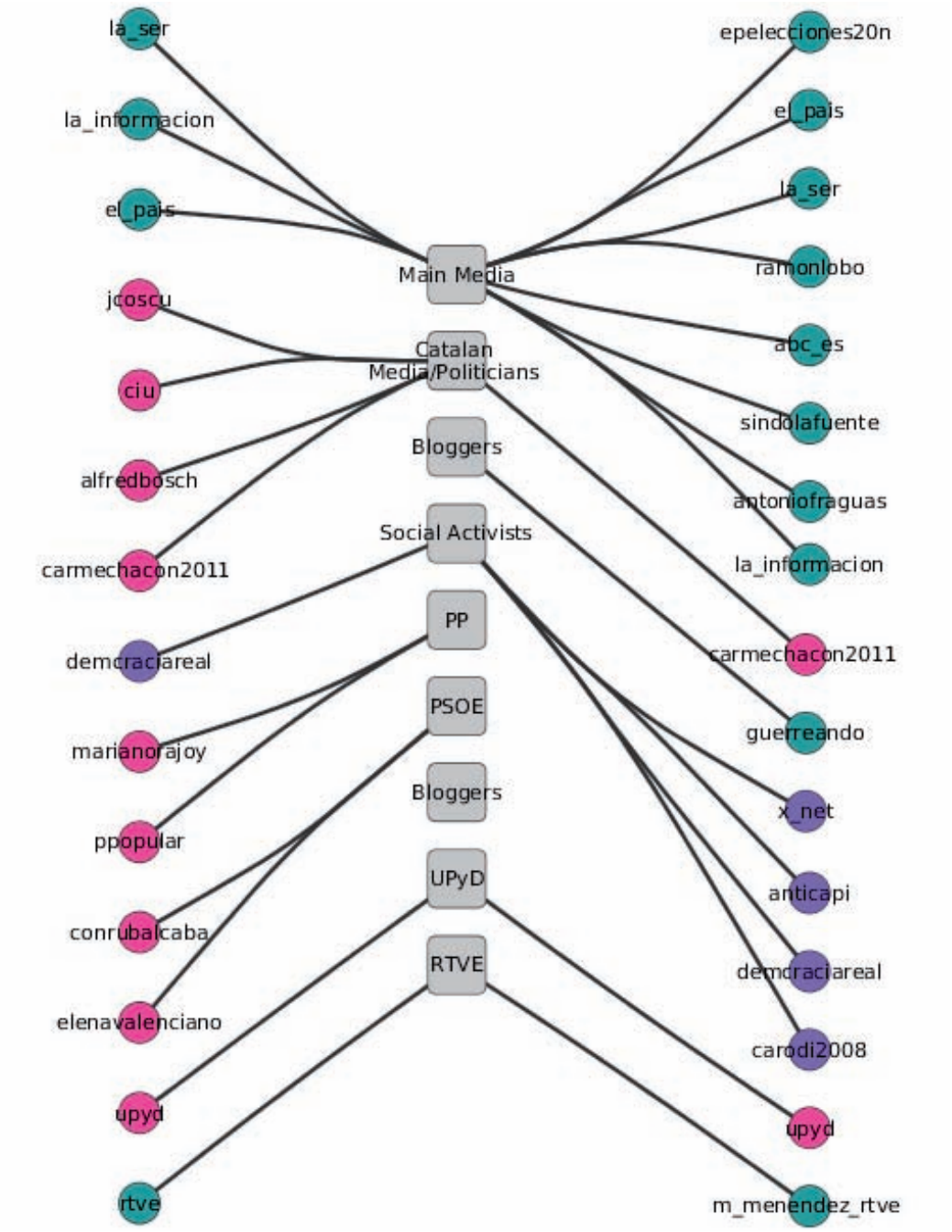

Figure 6.15: For the 20N conversation. Sketch, showing how the large communities of the follower layer split into several smaller communities in the mention and retweet layers.

politicians and media from Catalonia (panel C). The main characteristic of this community is the use of the Catalan language. This community holds a majority of users ($\sim 66\%$) that preferentially tweeted in Catalan. In fact, the online Spanish political debate is segregated by language [BMBL14]. Other large communities clustered together users holding the similar political ideology. These communities were formed around a single political party accounts, and therefore exhibit a highly segregated partisan structure. Despite a small fraction of links across opposed ideology communities, users tend to interact with those holding a similar ideology. This phenomenon is illustrated in Panel E that visualizes the communities of the two dominant parties: Partido Popular (PP) and Partido Socialista Obrero Español (PSOE). In this panel PP has been colored in blue, while PSOE has been colored in red. As can be appreciated there is a high political polarization. This behavior is similar to that observed for the United States, on Twitter [CRF+11a] and blogs[AG05].

Finally, we explore the relation among the community structure of the different layers. For this, we identified the top accounts of each community at the mention and retweet levels, and determined the community that they belonged in the follower level. We found that hubs embedded on large followers communities emerge as leaders of smaller and sparser ones in the mention and retweet layers. This showing that the large followers communities brake down into several smaller ones, formed around hubs belonging to a same follower community, at the active layers. Hence, users are more selective when extra effort is required to interact, holding less links and clustering in smaller and more selective groups. For example, while accounts belonging to different media may be classified on a same follower community, this phenomenon is not repeated in the retweet level. At this level media accounts belonging to a same community, also belong to a same media. This re ects that while users may passively follow different medias, they always relay on the same one to propagate information. This issue has been illustrated in Figure 6.15 A, where we have sketched the main relationships between the followers communities and the mention and retweet communities. The links highlight how the top communities in the reweet (right side) and mention (left side) layers grow around high visible accounts belonging the top communities in the follower layer. In the visualization communities of the mention and retweet layers have been colored according to the collective leading the community (media in green, politicians in pink, and bloggers in purple). While most of the top communities in the retweet layer grow around media accounts, the top mention communities were formed around politicians. Hence, this showing again that politicians are the main characters in the mention layer, while traditional media accounts were the preferred source of information

from which to propagate news at the retweet level.

All these showing that the communication patterns associated to each interaction mechanism are considerably different, what re ects the need to study Twitter as a multilayer network.

6.8. Discussion

In this chapter we have analyzed the communication patterns of the Spanish political debate taking place on Twitter. The three differentiated Twitter interaction mechanisms, follower, mention, and retweet, define three layers through which individuals receive and diffuse information. Hence, the Twitter information diffusion process does not take place through a single channel, but three. In order to fully understand the process we have to simultaneously analyze all three channels. For example, the propagation of messages via retweets is strongly conditioned by the topology of the follower layer, as it establishes the substratum through which individuals receive information. Additionally, users establish conversations or refer to each other using the third available channel, the mention. We found that the Spanish political conversation is centralized around a small fraction of in uential accounts. Politicians are the main characters, since their accounts were the most mentioned, and captured most of the collective attention. However, despite their accounts are still in uential in the retweet network, users tend to propagate information from the same sources as in the o ine world. Thus, traditional media accounts were the most retweeted. Therefore we can affirm that, on the light of our results, despite social media should allow more voices to be heard, the political communication is still driven by a minority of political parties and elite media.

Analyzing the mention network by means of community structure, we have been able to map the ow of political information going through Twitter during the campaigns. We show it to be considerably polarized by political ideology, as users crowded around a single political party accounts, and preferentially communicated with those of their same political stance. However, there were a small fraction of users, more exposed to political disagreement, who sustained the exchange of information among these polarized communities. Similar conclusions can be made from the retweet analysis, where we found that despite the countless sources of information available, users do not take full advantage of it, tending to just rely on their preferred one. In this regard users mostly retweet from just one traditional media official accounts with whose editorial line they feel identified. Hence, we can affirm that the social network in which individuals are enmeshed is ideologically

homogeneous. Such networks might be too insular, limiting the opportunities to learn about politics and contrast information. However, they represent effective information shortcuts to access information [Mar87][Mut06][Lup94].

The global directed multiplex network does not present a rich-club ordering, as politicians presided large communities of regular users in the mention layer; while media accounts were the sources from which people retweeted information. However, when considering reciprocal interactions the rich-club ordering emerges, as elite accounts preferentially interacted among themselves and largely ignored the crowd. The rich-club was mainly composed by politicians, media, and well-known bloggers. Hence, we identified the top 50 in uential users at each layer, and classified them as media, politicians, or bloggers. Despite an slight overlapping among the top in uentials at each layer, the relevance of the three different collectives significantly varied from one layer to another. The relevance of media and politicians at the follower level seems to be balanced. However, politicians clearly stand out in the mention layer, while media stand out in the retweet layer. A high degree in the mention layer is usually associated as a high value name, *i e* a famous and popular account, while the gain of retweets is associated to producing high value content tweets. Our results show that media were the sources of information, while politicians were the main characters of both conversations. Moreover, it suggests that politicians in general were not capable of producing high quality content tweets that got highly retweeted. All these resulted on users clustering around politicians in the mention layer, and around media accounts on the retweet layer. Hence, the leaders emerging at each layer vary significantly, and one can not claim neither politicians ruled the media or vice versa. It all depends on what kind of interactions we are considering and what effect we are trying to understand.

Finally, we have also analyzed the impact that the structure of the follower network has on the efficiency with which users are able to transmit messages and ideas through Twitter. We found that the activity is fed by a small group of very active users, while the large majority hardly participates. As part of this activity there are interactions that determine the collective attention, which we found to be dominated by a very small group of highly in uential users. However, if any, the rest of users gain in uence in proportion to the activity they employ. However, for the large majority of users the efforts are usually higher than the results. We have proposed a methodology to measure this bonding between actions and reactions, as the ratio between the retransmissions gained by a user and the activity she employed for it. We understand this ratio as the efficiency spread messages in the network in which individuals are embedded and hence, it can be considered as a measure of in uence. We found the distribution of this measure to be universal

across several Twitter conversations, following a lognormal distribution with a larger density of users at the higher orders,. We have also proposed a model to explain the emergence of the efficiency distribution, based on biased independent cascades taking place through the followers networks. The simulation of the model unveiled the effects that topology and individual behavior have on the emergent dynamical patterns. More particularly, it revealed that the emergence of a small fraction of highly efficient users results from the heterogeneity of the network in which users are embedded, rather than the differences in individual strategies. In fact, we found that in an homogeneously organized society we would need a much larger population to find the same level of in uence to diffuse information than complex and heterogeneous organized society. We conclude, that although individuals may employ different activity strategies, it is the position they occupy in the network what will primarily determine their level of in uence

Chapter 7

The impact of diverse languages of the political landscape

In this chapter we discuss the impact that the existence of several co-official languages has on the Spanish political debate. To this end, we first identified the language in which each tweet was posted and explored to which extent the diversity of languages limits the communication among the different sides. Furthermore, we propose a method to determine the proximity between the different languages and political alignments.

7.1. Language detection

To automatically determine the language in which each user preferentially tweeted we have taken advantage of guess-language [Gue]. This is a Python library that determines the natural language of a given text. We determined the language in which each tweet was posted and assigned each user the language in which most of her tweets were written.

7.2. Language Polarization

Here, we explore the impact that the existence of several co-official languages has on the Spanish politics and its polarized landscape. For this we complement the $20N$ dataset with a more local political conversation: the 25th of November Catalan elections ($25N$). We classified each user according

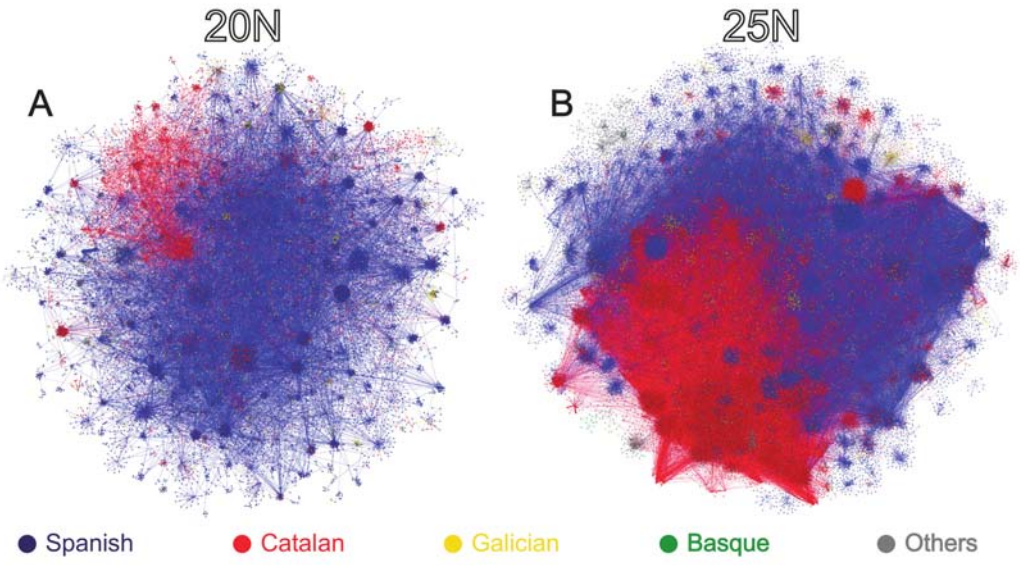

Figure 7.1: This figure visualizes the $20N$ (A) and $25N$ (B) retweet networks. Each node represents a user and has been colored according to her most used language.

to their most written language (see section 7.1 for details) and analyzed the communication patterns among the different languages.

Twitter re ects the Spanish linguistic diversity, as Tweets were posted in all the different official and co-official languages. This is illustrated in Figure 7.1 where we have visualized the $20N$ (A) and $25N$ (B) retweet networks coloring each node according to their language. In Figure 7.2 we present the percentage of users tweeting in each language for both datasets, and compare these results with those observed when considering all the raw Twitter data coming from Spain and Catalonia [MAN+13]. Regarding the general elections, Spanish was the most used language (78%), followed by Catalan (18%). However, when comparing this data to the statistics presented in [MAN+13] (see Figure 7.2), we found that the share of users tweeting in Catalan is above its expected value (3.11%). We also observed this same behavior during the Catalan elections, where Catalan presents a share of 54%, quantify far above its expected value (28%). This suggests that the use of Catalan on Twitter increases in politicized contexts such as electoral campaigns.

Next, we explore the communication patterns among users tweeting in different languages. For this we measure the language polarization by calculating the assortativity [New03] by language. This measure quantifies the tendency for users on Twitter to communicate to other users that tweet in

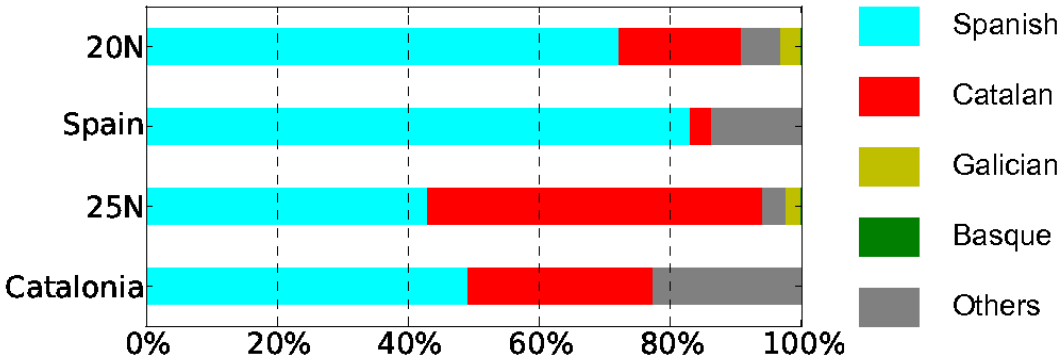

Figure 7.2: Language share for the two analyzed datasets, $20N$ and $25N$, together to the overall share found for Spain and Catalonia in general Twitter conversations [MAN+13].

their same (or not the same) language and is given by the following expression:

$$r_L = \frac{\mathrm{Tr}\,\mathbf{e} - \quad \mathbf{e^2}}{1 - \quad \mathbf{e^2}} \tag{7.1}$$

where $\mathbf{e}$ is the language mixing matrix, whose elements e_{ij} quantify the number of retweets going from users posting on language i to users posting on language j; Tr stands for trace; and $\quad \mathbf{e} \quad$ represents the sum of all elements in the matrix $\mathbf{e}$. This measure ranges between -1 and 1. A value of 1 means that users only retweet those tweeting on their same language. On the contrary, -1 means that users only retweet users who tweet in a different language. In between, $r_L = 0$, indicates users retweet others regardless their language.

Our results show the existence of a clear trend to retweet only those who tweet in the same language (see Table 4.1) and hence, in Spain, language represents a constrain for political communication. This tendency becomes more extreme for the Catalan conversation, meaning that for more local conversations the segregation by language becomes higher.

7.3. Language Preferences

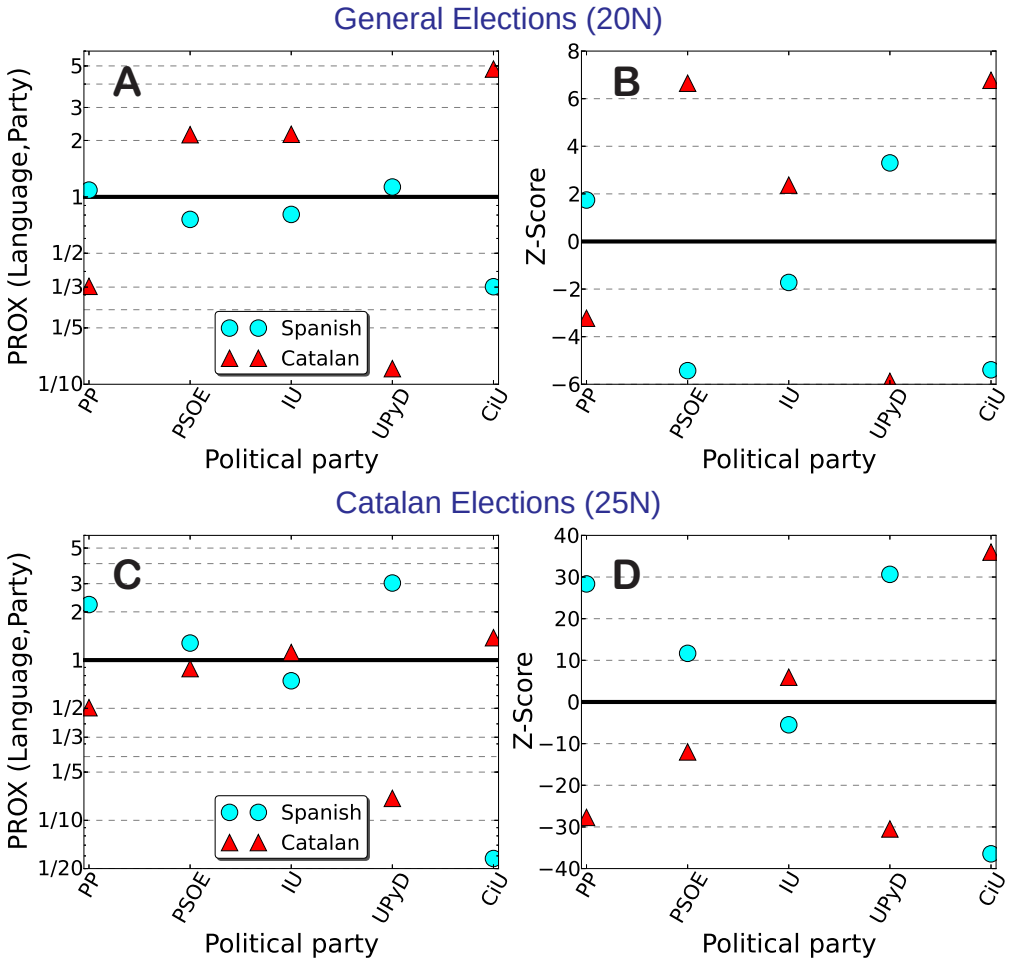

Figure 7.3: Measuring proximity between languages and political parties. Circles (cyan) correspond to Spanish, while triangles (red) correspond to Catalan. Panel (A) shows the Proximity measures for the $20N$ dataset. Their corresponding z-scores are plotted in panel (B). Panel (C) shows the Proximity measures for the $25N$ dataset. Their corresponding z-scores are plotted in panel (D).

Nationalist currents are known to have a big impact in the Spanish politics. In fact, in some regions, such as Catalonia, the nationalist component of several parties can be much more in uential than their ideology (liberal/conservative). Here, we intend to answer the following question: what is the relation between the language in which users tweet and the political party they support?

As a proxy to determine the party that each user supports we use the retweet network, since retweet is the Twitter interaction mechanism that exhibits the highest segregated partisan structure [BMLB12] [CGFM12]. We define the Proximity between a language (L_i) and a political alignment(A_j), $PROX(L_i\ A_j)$, as the relative frequency of retweets going on from users posting in language, L_i, to official accounts of party, A_j, observed in the empirical network compared to its expected value in the randomized version. Hence, it can be expressed as:

$$PROX(L_i\ A_j) = \frac{P(L_i\ A_j)}{\langle P_r(L_i\ A_j)\rangle} \tag{7.2}$$

where $P(L_i\ A_j)$ is the probability that a user tweeting in language, L_i, retweets a message originally posted by a politician affiliated to party, A_j. Analogously $\langle P_r(L_i\ A_j)\rangle$ represents the expected value (averaged from 100 realizations) of this same probability but for the randomized networks. When randomizing the networks we have preserved all the individual characteristics of each user, meaning that each node preserves her preferred language, and number of outgoing and ingoing retweets. In order to randomize the networks we used the Markov-chain algorithm that repeatedly exchanges randomly chosen pairs of connected nodes. By comparing to randomized networks we compensate for the effects of differences in party popularity and in number of users tweeting on each language. Finally, to determine the significance of each Proximity measure we can calculate its corresponding z-score, that is given by the following expression:

$$Z(L_i\ A_j) = \frac{P(L_i\ A_j) - \langle P_r(L_i\ A_j)\rangle}{{}_r(L_i\ A_j)} \tag{7.3}$$

where ${}_r(L_i\ A_j)$ is the standard deviation of $P_r(L_i\ A_j)$, computed out of 100 randomized networks.

In Figure 7.3 we present the Proximity between the most voted political parties (PP, PSOE, IU, UPyD, CiU) and the two most used languages (Spanish and Catalan), together with its corresponding z-score, for both datasets. The results corresponding to the general elections ($20N$) are displayed on panels A and B, while the Catalan elections ($25N$) are displayed on panels

C and D. Deviations from the standard preference line (PROX=1) imply that users preferentially ($PROX > 1$) or not preferentially ($PROX < 1$) retweeted specific parties. The deviations found for both datasets measure the closeness in ideology of each party to the Catalan nationalist current. For example, CiU is a Catalan party with a remarkable nationalist current, thus, its sympathizers preferentially tweeted in Catalan and hardly did it in Spanish. Conversely users tweeting in Catalan hardly retweeted PP or UPyD, as these parties strongly defend the unity of Spain.

7.4. Discussion

The results of this paper also speak about the importance that nationalist currents have in the Spanish political landscape, where at some regions the nationalist component of the parties becomes much more in uential than the ideology. It is widely known that the strong feeling of membership to the autonomous community, that exists in several regions of Spain, is frequently used as a political argument. This is re ected on Twitter in several ways: i) Catalan tends to be overused in political conversations. ii) Despite the vast majority of speakers of a co-official language are bilingual, the conversation is highly segregated by language. iii) There is an obvious relationship between the political alignment of users and the language in which they tweet.

However, the utility of the proposed method to determine the proximity between languages and political parties is not particular of Spain, but generalizable to other countries. For example, let's think about the United Sates. Although English is the main language, there is a diversity of cultures co-existing inside the country that speak different languages. Hence, the proposed methods could be used to estimate the proximity between the Chinese or Hispanic communities and the liberal or conservative parties. Thus, being useful to estimate electoral outcomes.

Of course, one can not directly extrapolate from the Twitter information diffusion process to the overall opinion formation of the population, as users of the social network may not be representative of the whole society. Even if they were, there are other channels through which individuals receive information, such as mainstream media (e.g.TV, radio) or personal relations. However, as the percentage of the population engaged to these platforms is rapidly growing, they have become the latest new medium being exploited for decisive competitive advantage. This, making the understanding of the communication patterns behind these platforms an increasingly important topic to further uncover opinion formation process of individuals.

Chapter 8

Characterizing politicians activity

8.1. Introduction

In this chapter we interpret politicians behavior during the 2011 electoral campaign by analyzing the interactions going on between the different political parties. Our target is to find patterns that help us understand how politicians interacted during the campaign.

In order to more deeply understand the online Spanish political landscape, we have analyzed the political filtered follower, mention and retweet networks. For this matter we have filtered the three Twitter networks by only remaining official politician's accounts from parties with over 10 participants in the conversation. Our results show that politicians do not relate to each other randomly, rather they use the different interaction mechanisms with specific purposes.

Finally we introduce a model based on the heterogeneous preferential attachment formalism [SB07] capable of growing political conversations and illustrate it by reproducing the mentions and retweets taking place in Twitter among politicians.

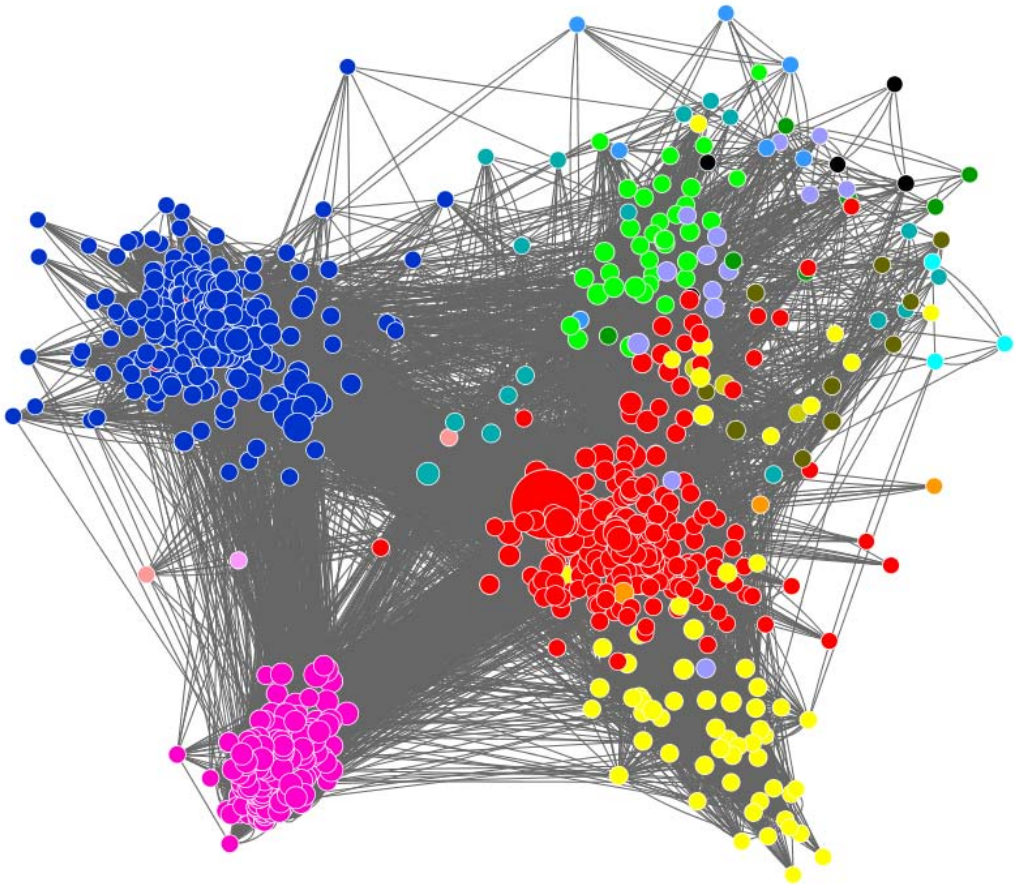

Figure 8.1: Visualization of the politicians follower network. Only accounts that belong to an oficial political party are shown. colors indicate the political party of each node.

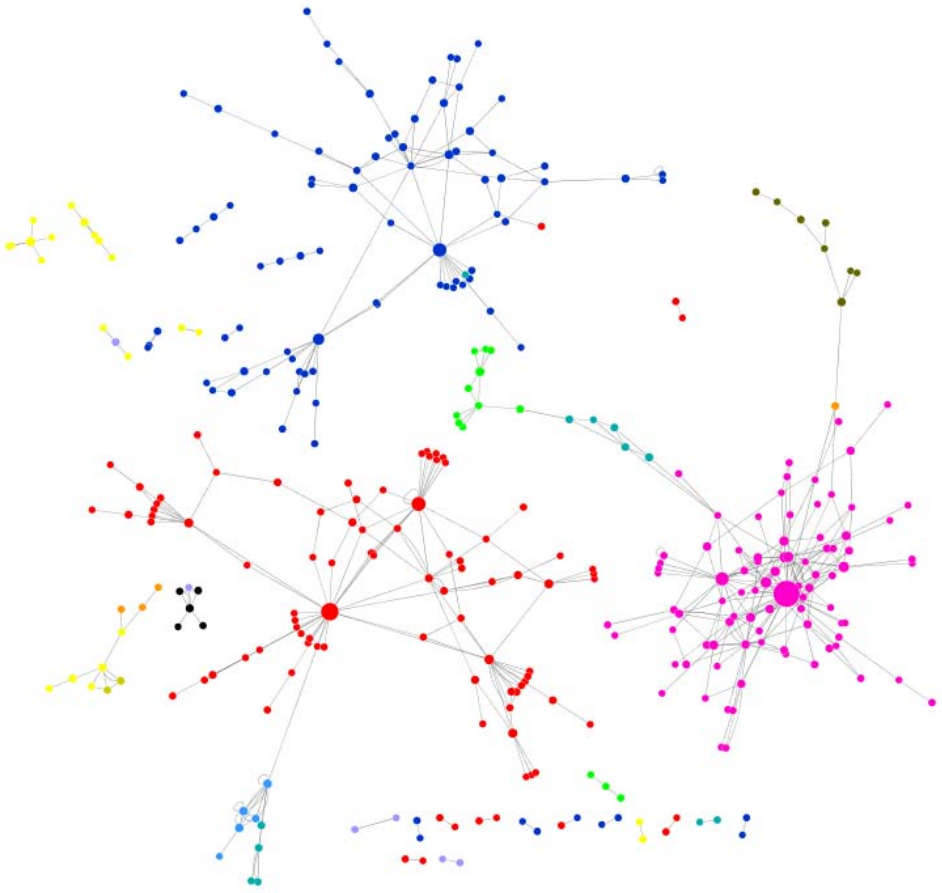

Figure 8.2: Visualization of the politicians retweet network. Only accounts that belong to an oficial political party are shown. colors indicate the political party of each node.

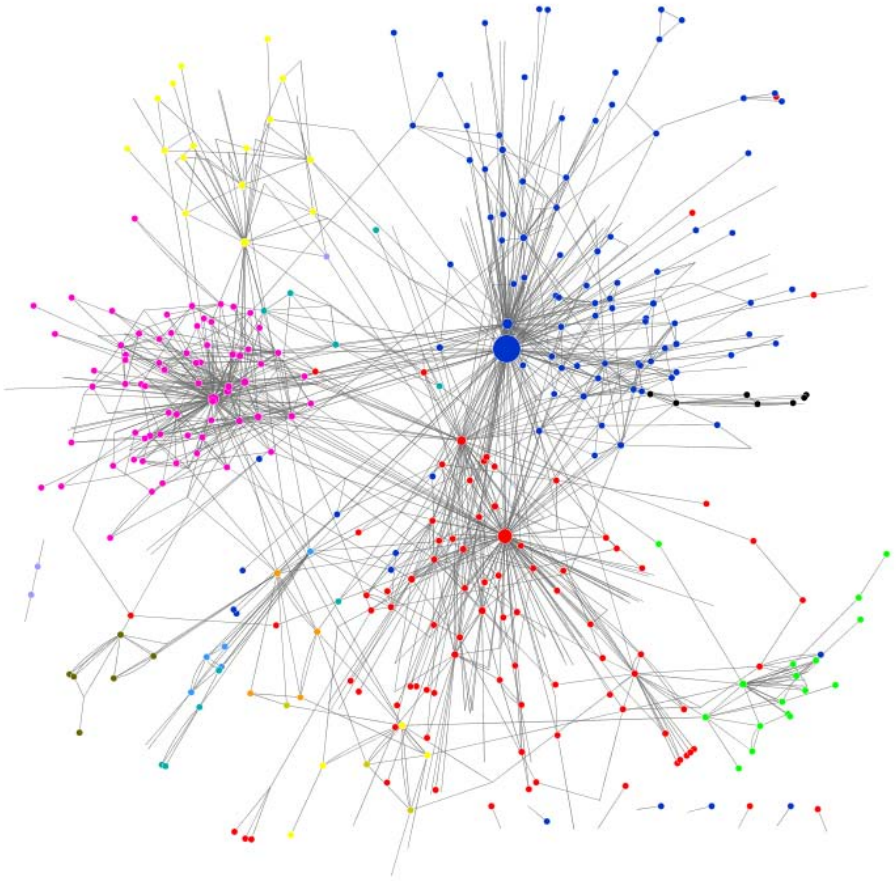

Figure 8.3: Visualization of the politicians mention network. Only accounts that belong to an oficial political party are shown. colors indicate the political party of each node.

8.2. Who follows who?

We begin our analysis by representing and analyzing the politicians filtered follower network. This network represents the follower relations among politicians Due to the reduced size of the network (500) a visual inspection of the network is meaningful. We have visualized the network on Figure 8.1. As can be seen politicians tend to follow those of their same political party. Thus, the network presents a high modularity where each cluster represents a different political party. In order to quantify the segregation by political party, we have first classified all nodes according to the political party they belong to. Next, we have calculated the assortativity coefficient r introduced by Newman in 2003 [New03]. This coefficient measures the fraction of edges going from one party to another and is given by the following equation:

$$r_P = \frac{\mathrm{Tr}\,\mathbf{e} - \mathbf{e^2}}{1 - \mathbf{e^2}} \tag{8.1}$$

where $\mathbf{e}$ is the political party mixing matrix, whose elements e_{ij} quantify the number of links going from party i to politicians that belong to party j. The formula ranges between -1 and 1, where a value of 1 implies that politicians only follow those of their same party.

The politicians follower network is assortative as the coefficient took a value of $r = 0\,83$. This value, implies that politicians tend to follow those of their same party, and barely follow other political parties. In fact, as Figure 8.1 shows the three bigger clusters correspond to PP (blue) PSOE (red) and UPyD (purple). PP and UPyD form two independent cluster that exhibit minimum mixing with other parties. However, the PSOE cluster has a significant amount of links with IU (yellow), re ecting that both of them are left-wing political parties. Finally, we can distinguish a less clustered set of politicians (top right corner of the figure) that form the catalan community. These accounts mainly belong to political parties or politicians from Catalonia, such as ICV, CiU or ERC.

8.3. Communication among political parties

Next, we represent and analyze the politicians filtered mention and retweet networks. We begin the analysis by computing the strength distribution for the mention and retweet networks. The strength distributions of both networks are shown in Figure 8.6 the collective attention was very heterogeneously distributed with the presence of hubs that captured most of the mentions and retweets. These hubs correspond to the top accounts of the

Table 8.1: Comparison of the assortative mixing by political party for the mention and retweet networks filtered by politicians official accounts and the proposed model results.

Network	**Experimental r**	**Modeled r**
Mention	0 905	0 86 0 03
Retweet	0 991	0 989 0 005

larger and more active political parties, such as *conrubalcaba*, *marianorajoy*, *ppopular* or *upyd*.

To further understand the communication patterns among the Spanish political parties, we have analyzed the politicians filtered mention and retweet networks. To do so, we have calculated the assortative mixing matrix and computed its asortativity coefficient expressed in eq. 8.1. The results are presented in Table 8.1. We found both networks to be extremely assortative as r took a value very close to 1 on both networks ($r_M = 0\ 905$ and $r_R = 0\ 990$). A visualization of each network is presented on Figures 8.3 (mention) and 8.2 (retweet). The retweet network is so extremely segregated that it is divided into several independent components. Thus, PP PSOE and UPyD form three isolated islands. However, mentions across parties were slightly more frequent than retweets, and therefore, the mention network does present a giant component holding all the parties together. As can be seen on the visualization of this network, politicians clustered around their party and candidate official accounts, but still there were some few links across parties. On this electoral campaign retweets were the most segregated Twitter interaction mechanism. This result goes in agreement with the analysis of the 2010 U.S. Congress elections, where retweets were also found to be more ideologically polarizing than mentions among regular users [CRF+11b].

To summarize, the interpretation of these results is that politicians communicated mostly with their own partisans. Hence, at the light of our analysis there was a considerable lack of online debate between the Spanish political parties. The results also re ect that the strategy of all the parties was just to publicize their own partisans and ignore the remaining politicians.

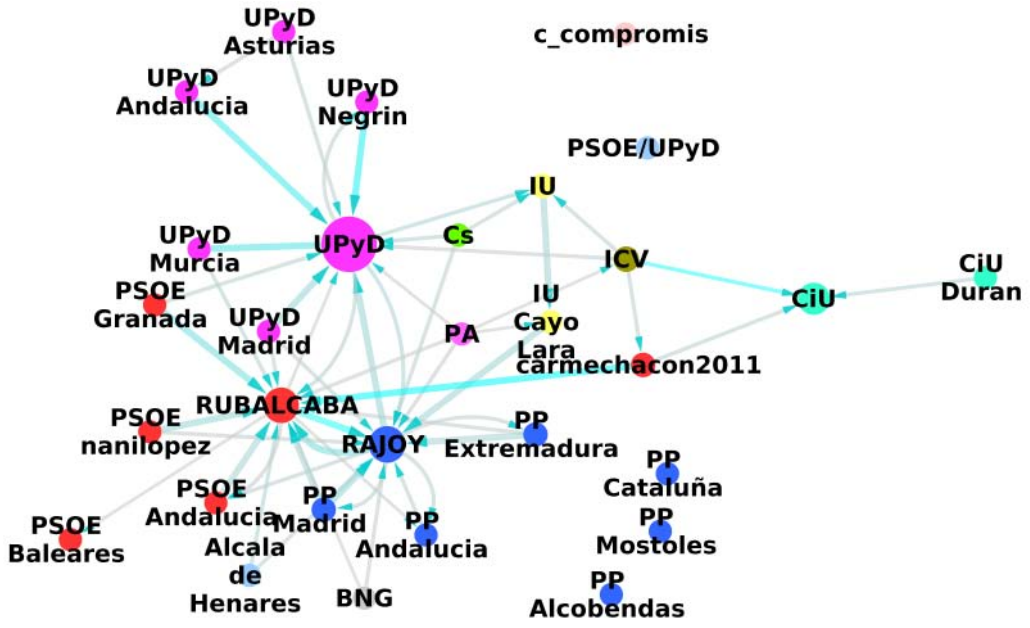

Figure 8.4: Visualization of the community structure of the politicians mention network. Only accounts that belong to an oficial political party are shown. colors indicate the political party of each node.

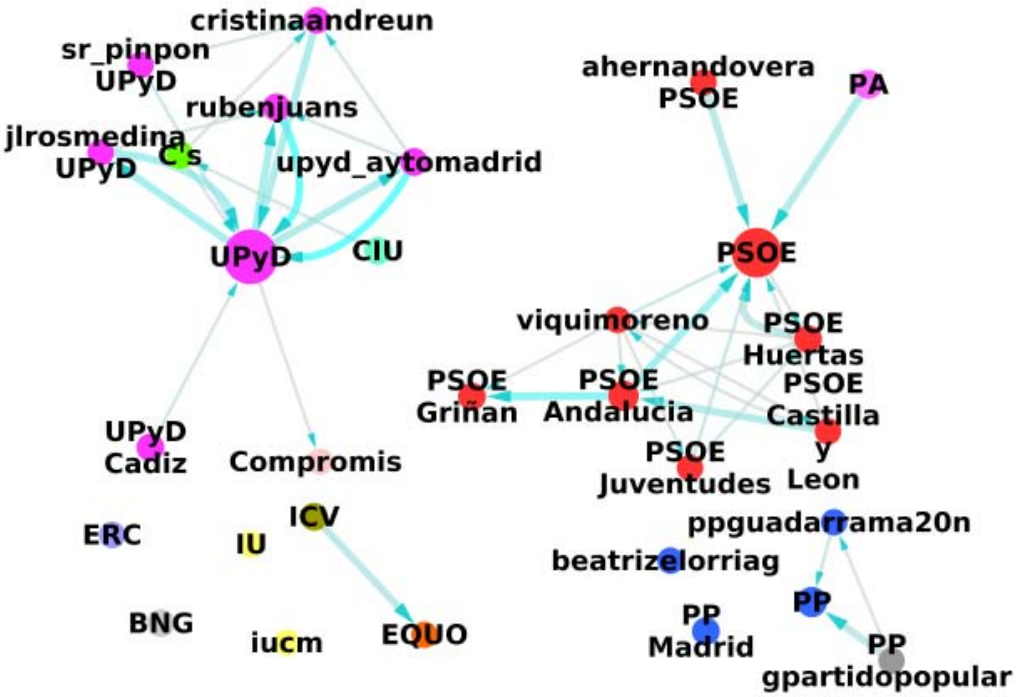

Figure 8.5: Visualization of the community structure of the politicians retweet network. Only accounts that belong to an oficial political party are shown. colors indicate the political party of each node.

8.4. Community structure

Next, we analyzed these networks by means of community structure by applying the mapequation algorithm [RB08]. The results are presented in figures 8.5 (mentions) and 8.5 (retweets) where each node corresponds to a community and the label to the political party or region to which the majority of accounts with the community belong. When analyzing the community structure of the political filtered mention and retweet networks, we observe both differences and similarities between them. On the one hand the way in which politicians group into communities is the same for both networks. Communities are formed almost entirely by politicians belonging to the same political party, as the percentage of politicians of a same party in each community is over 90%. Going into further detail, when a political party breaks up into several communities it follows the next pattern: Small communities corresponding to geographical regions grow around the central main community. On the other hand the ux between communities follows different laws for the mention and the retweet networks. Whereas on the mention network the three main political parties are linked, in the retweet network we can distinguish three main islands corresponding to UPyD, PP and PSOE. A relevant feature to stand out is that geographical communities hardly interact with each other, however they work as bridges linking political parties in the mention network. Two major conclusions can be pulled out from this analysis: (i) the community structure of the networks re ects the lack of debate between political parties and the importance of the autonomies in the Spanish politics; (ii) Mentions, although in a small scale (fraction) and contrary to retweets, are going from one party to another.

8.5. A model that reproduces political Interactions

To further explain the structural features found in the interactions between politicians, we propose a model based on the heterogeneous preferential attachment formalism [SB07]. The idea behind this formalism, is that the probability of a node i interacting with a node j not only depends on their respective degree, but also on the affinity between them. In our model nodes (politicians) are classified according to discrete characteristics (political parties). Thus, the probability of appearance of a new interaction from any politician, i , belonging to party A, to a politician j, who belongs to a party

B, is given by the following expression:

$$P_{ij} = \frac{S_j}{\sum_{j\ B} S_j} f_{AB} \tag{8.2}$$

where S_j is j's strength, and f_{AB}, is the affinity value from A to B. The first factor of equation 8.2 corresponds to the local connection at micro-scale, and the second one to the mesoscale. We implement this model in the mesoscale by grouping all the politicians of the same party in a supernode labeled with the name of the party. The properties of these supernodes are determined by those of the nodes inside them, and the number of interactions, N, between them can be obtained from the e_{ij} matrix [New03]. In this way the affinity value between two politicians will be determined only by their political parties. We represent the directional affinity between political parties by an f matrix whose elements quantify the relative ux of interactions from A to B. Using this notation we can model the structure of the f matrix as:

$$f_{AB} = \frac{N_{AB}}{N_A} \tag{8.3}$$

where N_{AB} is the total ux going from A to B and N_A is the total ux going out from A. After understanding the structure of the system's mesoscale, it's time to model the micro-scale. In this scale the probability rule for interactions between politicians is based on the preferential attachment model [BAJ99]. In the sense that the likelihood of a node, belonging to party B, to receive a new interaction, increases with the node's strength.

To implement the model we have calculated the experimental f matrix of each network, considering only those parties with over 10 politicians participating in the conversation. Also we have assigned a random number of outgoing edges to the newly added nodes, following power law distributions of exponents: $\gamma_M = 1\ 3$ for the mention network and $\gamma_R = 1\ 6$ for the retweets one. In this way we modeled the resulting distributions by simulating the heterogeneity found in the users behavior and using the same micro-scale connection rule for both interaction mechanisms. In Figure 8.6 we present the resulting cumulative strength function distribution for both networks, after having averaged over 1 000 realizations. It can be noticed that the model reproduces perfectly the strength function distribution for both networks, and maintains the assortative mixing levels as presented in Table 8.1.

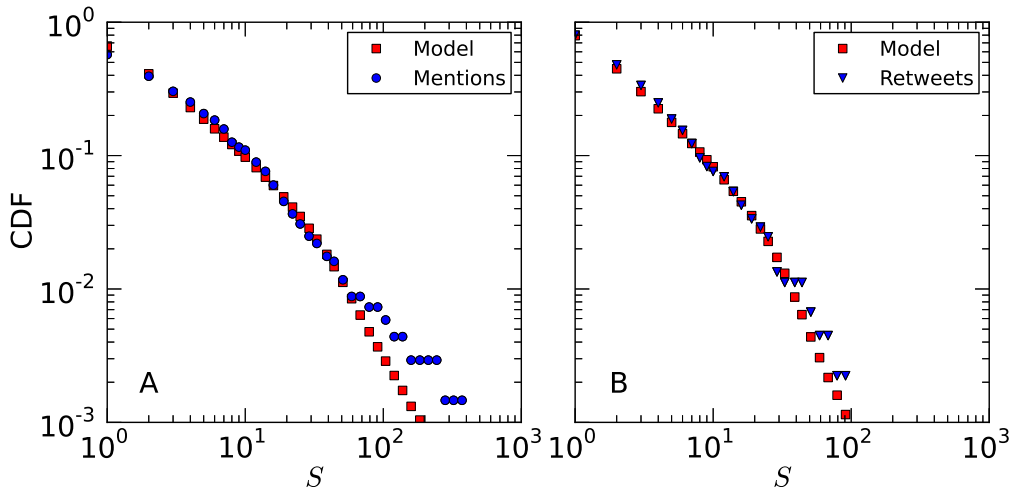

Figure 8.6: Comparison of the strength function cumulative distribution for the filtered by politicians and political parties official accounts mention (A) and retweet (B) networks (Blue) and the results obtained for the proposed model (Red).

8.6. Discussion

The perfect political campaign strategy has been eternally chased by politicians. To that effect we have tracked and analyzed politicians behavior. We found a scale-free organization and a profound segregation and lack of debate among political parties. Politicians communicated mostly with their own partisans and the most segregated interactions were retweets, followed by mentions and follows. The extremely high segregation observed in the retweet network goes in agreement with the results reported for other countries [CRF+11b]. Retweets are a sign of endorsement and imply propagating a tweet posted by someone else. Thus, in principle it makes no sense to publicize your competition. However, the high segregation measured on the mention network implies a lack of debate, and that although mentions are a suitable mechanism to debate and exchange ideas, politicians did not take advantage of it to debate with their competitors.

Moreover, the degree of each account in the network reveals the popularity of the politician behind the account. Consequently, the most connected hubs correspond to Rajoy, Rubalcaba, PP and PSOE. However, politicians not only preferentially communicated with those of higher authority, but also preferentially interacted with those of their same party. Hence, the mention and retweet networks follow the preferential attachment rule, but with a clear

bias of nodes to connect with those of their same ideology. On this chapter we have also proposed a network growth model based on heterogeneous preferential attachment to explain the emergence of such segregated modules in the politician's networks. The proposed model reproduces the empirical mention and retweet networks, as it maintains its degree distribution and modular structure.

Further research should be done on exploring the impact that the segregation by political party among the Spanish politicians has on the overall Spanish political landscape. On chapter 6 we showed that the global political landscape seems to be highly polarized by ideology. Hence, relating the results of that chapter to the structure of the politicians network would be of great interest. An approach that would be appropriate for such purpose would be to apply the model and methodologies that we propose on [MBLB15] to the $20N$ dataset. On that paper, that we fully describe on the next chapter (9), we propose a methodology consisting of a model and an index to quantify political polarization. On this methodology, we first propose a model that estimates the opinion of a population on Twitter from a minority of in uential users with a defined ideology, such as the politicians we have analyzed on this chapter. By using retweets as a proxy of in uence or homophily, we propagate the ideology of the elite nodes through the retweet network to the remaining users. Hence, the result of the model is distribution of ideologies from which to measure polarization. Thus, by calculating the polarization index presented in eq 9.8 that can be found in the next chapter (9) we measure the final polarization. Applying the model to this dataset to obtain the polarization between the sympathizers of the two dominant parties, PP and PSOE, would be straightforward and the results would be meaningful.

Chapter 9

Measuring political polarization

9.1. Introduction

From a sociological point of view, polarization is a social phenomenon that appears when individuals align their beliefs in extreme and con icting positions, with few individuals holding neutral or moderate opinions [Ise86, Sun02]. Thus, as a process it is the increase of such divergence over time when people evaluate issues of diverse nature [BB07, BG08, PD13], like politics or religion. In words of John Turner: 'Like polarized molecules, group members become even more aligned in the direction they were already tending'[Tur87].

in this chapter, we propose a methodology to study the emergence of political polarization and quantify its effects. To this end, we introduce a model to estimate opinions, and a polarization index that quantifies to which extent the resulting distribution of opinions is polarized. We say that a population is perfectly polarized when divided in two groups of the same size and with opposite opinions. Hence, our measure of polarization is inspired by the electric dipole moment - a measure of the charge system's ove e distance between the charges. Analogously, the polarization of two equally populated groups depends on how distant their views are.

As Downs argued in 1957 [Dow57a], political discussion among individuals minimizes the cost of becoming politically informed. In other words, sensible individuals tend to rely in the opinions of experts instead of analyzing information by their own. In fact, several observational studies support this theory and suggest that the expertise distribution within a social network affects the political communication patterns [Huc01]. Hence, by controlling the opinion of a minority of in uential individuals and mapping the communication uxes among the population we can estimate their distribution of opinions. To this end, we propose a model based on DeGroot model [Deg74].

The original model proposed by DeGroot describes how a group of individuals might reach a shared opinion, by iteratively updating their opinion as the average of their current opinion with the opinions of their neighbors. Such global coordination, without centralized control, can also be efficiently achieved when individuals adopt the majority state of their neighbors, even in the presence of noise or complex topologies [Ama04]. Recently, the DeGroot model has been used to study the conditions under which consensus is achieved [Ace11, Gol10, Jac10]. However, as consensus is rarely reached in real world [Kra09] [Ben08], variants of this model can held to a diversity of opinions [Bin11, Ace13, Kra00, Fri90].

In contrast to opinion generation models, such as the voter model [CLI86, Hol75, Suc05], we do not aim to study the evolution of opinions, but to infer a distribution of opinions formed on a social network from which to measure polarization. In our model, a minority of in uential individuals propagate their opinions through a directed network in uencing the remaining individuals. Thus, each individual iteratively updates her opinion according to her incoming neighbors-those in uencing her. Hence, by taking advantage of complex network analysis [New03], we are able to estimate the opinion of the whole majority that a priori was unknown. The behavior of the in uential minority is similar to zealots in the voter model [Mob03, Mob07], but their impact in the model's dynamics is different. In our model, zealots, rather than preventing consensus, allow us to infer the opinions of all the nodes in the network. Contrary to the voter model where opinions are binary (0 or 1), the opinions in our model represent a continuous distribution. In absence of polarization, the expected resulting distribution of opinions would be a narrow distribution centered at a neutral opinion. However, as polarization emerges, the resulting distribution shifts to a bimodal distribution with two peaks emerging around the two dominant and confronted opinions [DW07].

How can political polarization be detected and therefore be fixed? Nowadays, digital traces of human collective behavior [LPA+09] represent an opportunity to detect and measure in real time different phenomena, such as polarization. In fact political segregation has already been observed on political blogs [AG05] or Twitter [CRF+11a, BMLB12]. Recent research has shown that the most prominent and politically active users mainly interact with their own partisans [AG05, CRF+11a, BMLB12], leaving little space for real debate and cross ideological interactions. However, segregation does not necessarily imply polarization, as two separated groups of people that share the same opinion can not be considered as polarized. Hence, in order for a population to be polarized, the opinions of the two groups should also be con icting or opposed[Gue13]. In the latter part of this paper we show how to apply our methodology to online data gathered from Twitter

in order to estimate individuals opinions and to measure the emergent political polarization. Twitter provides an interesting context in which to study polarization as it represents a wide variety of different types of communications, going from personal to those coming from traditional mass media. In this platform, a minority of *elite* users concentrate much of the collective attention, but still a big fraction of the content they produce reaches the mass through intermediaries or 'opinion leaders' [Bur99]. In other words, the 'two-step- ow' of communication is still valid on Twitter [WHMW11].

We begin this chapter by proposing a model to estimate opinions in which a minority of in uential individuals propagate their opinion through a social network in uencing the opinions of the remaining individuals. Thus, the result of the model is a probability density function $p(X)$, that determines the fraction of individuals holding an opinion X. Next, we introduce the polarization index to measure the political polarization from the resulting opinion distribution. To illustrate the power of the methodology, we apply it to a Twitter conversation regarding the death announcement of the Venezuelan President (Hugo Chávez). Finally, we contrast the results with o ine data.

9.2. Estimating Opinions

We present a model to estimate the opinions of individuals who interact on a social network, in order to obtain their opinions distribution. In it we distinguish two types of individuals, *elite* and *listeners*. The first ones have a fixed opinion and act like seeds of in uence, while the opinion of the second ones depends on their social interactions. The model is fully specified by the following assumptions:

1. **Initial Conditions:** The world is abstracted by a directed network, G, in which each individual is represented by a node and links account for in uence rather than friendship or other kind of relationship. We define two different subset of nodes, S accounting for *elite*; and L, accounting for *listeners*. Additionally we endow each *elite* with a parameter, X_s, that determines her opinion value and that will remain constant for the duration of the model. X_s lies in the range, $-1 \quad X_s \quad 1$, where 1 and -1 represent the two extreme and confronted poles. Finally we set an initially neutral opinion, $X_l(0) = 0$ to all *listeners*.

2. **Opinion Generation:** At each iteration, *elite* nodes, S, propagate their opinions through the established network, G, in uencing *listeners*, L. Hence, each listener iteratively updates her opinion value as the mean opinion value of her incoming neighbors. Thus the opinion at time step, t, of a given

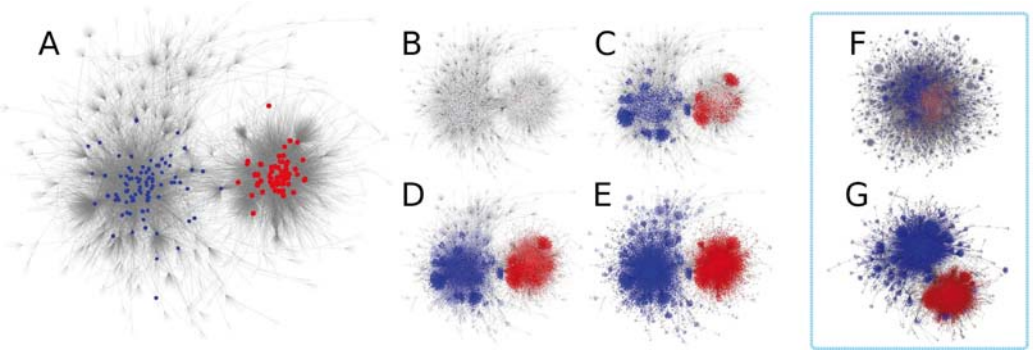

Figure 9.1: Schema of the in uence spreading process in the opinion estimation model. (A) Displays the seed nodes in the network, colored according to their respective ideology. (B) Displays the network at $t = 0$, before seeds start to propagate their in uence. (C) Shows the state of the network at $t = 1$. (D) shows the state of the network at $t = n$ 2. (E) Displays the final state of the network at $t = n$. (F) and (G) Visualizations of two examples of the result of the opinion formation model to the Venezuelan dataset for non polarized (F) and polarized (G) days.

listener, i, is given by the following expression:

$$X_i(t) = \frac{{}_j A_{ij} X_j(t-1)}{k_i^{in}} \tag{9.1}$$

where A_{ij} represents the elements of the network adjacency matrix, which is 1 if and only if there is a link from j to i, and k_i^{in} corresponds to her in-degree. The process is repeated until all nodes converge to their respective X_i value, lying in the range -1 X_i 1. Thus, the results of the model are given in a density distribution of nodes' opinion values $p(X)$. Note that the opinions of individuals do not depend on their opinion in the previous step. This is because we are estimating their opinion that a priori was unknown, rather than studying the evolution of opinions.

The dynamics of the model is illustrated in Fig. 9.1, where we present an schema of the in uence spreading process. Panel A visualizes the instantiation of the model where each *elite* node has been colored according to her opinion (red, $X_s = -1$; and blue, $X_s = +1$). Panels B-E show the dynamics of the in uence process from the initialization (B) to the final converged state (E). Panels (F) and (G) visualize two empirical networks corresponding to a non polarized (F) and a polarized (G) case.

9.3. Introducing a new measure of polarization in opinion distributions: the polarization index

We say that a population is perfectly polarized when divided in two groups of the same size and with opposite opinions. Hence, we propose a measure of polarization that quantifies both effects for the resulting X distribution obtained from our model. This definition is inspired by the electric dipole moment- a measure of the charge system's overall polarity. In the simplest case of two point charges of opposite signs ($-q$ and $+q$) the electric dipole moment is proportional to the distance among the charges. This is analogous to a simple scenario consisting of two persons with different ideologies, thus the polarization depends on how con icting their points of view are (*i.e.* the distance among the two ideologies).

We begin by calculating the population associated with each opinion (positive and negative). To this end, we define A^- as the relative population of the negative opinions ($X < 0$). By the same token, we define A^+ as the relative population of the positive opinions ($X > 0$). Hence, both variables can be expressed as:

$$A^- = \int_{-1}^{0} p(X)dX = P(X < 0) \tag{9.2}$$

$$A^+ = \int_{0}^{1} p(X)dX = P(X > 0) \tag{9.3}$$

So we can express the normalized difference in population sizes, A , as:

$$A = A^+ - A^- = P(X > 0) - P(X < 0) \tag{9.4}$$

Next, we quantify the distance between the positive and negative opinions. In other words we measure how differing the opinions of the two sides are. To this end we determine the gravity center of the positive and negative opinions that can be written as:

$$gc^- = \frac{\int_{-1}^{0} p(X)XdX}{\int_{-1}^{0} p(X)dX} \tag{9.5}$$

$$gc^+ = \frac{\int_{0}^{1} p(X)XdX}{\int_{0}^{1} p(X)dX} \tag{9.6}$$

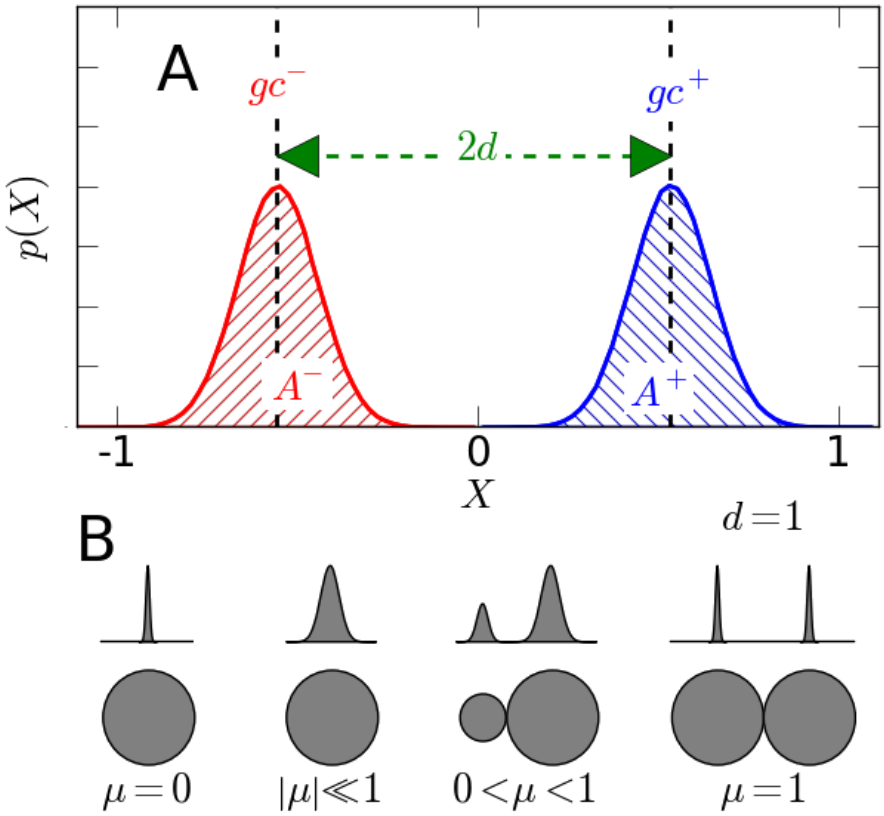

Figure 9.2: Schema explaining polarization and the proposed index . (A) Density distribution of opinions. *gc* stands for the gravity center of each pole, A stands for the area associated to each ideology, and d stands for the pole distance. (B) Visualization of the polarization index, , given in eq. 9.8, for four situations.

and define the pole distance, d, as the normalized distance between the two gravity centers. Hence, it can be expressed as:

$$d = \frac{gc^+ - gc^-}{X_{max} - X_{min}} = \frac{gc^+ - gc^-}{2} \tag{9.7}$$

This formula gives $d = 0$ when there is no separation between the gravity centers, *i.e.* there are no longer two differentiated groups and everyone shares a similar opinion; and $d = 1$ when the two opinions are extreme and perfectly opposed.

Finally, we can use eqs. 9.4 and 9.7 to write down a general formula to measure polarization as a function of the difference in size between both populations A and the poles distance d. Thus, we define the *polarization index*, , as:

$$= (1 - A)d \tag{9.8}$$

This formula gives = 1 when the distribution is perfectly polarized. In this case the opinion distribution function is two Dirac delta centered at -1 and $+1$ respectively. Conversely, = 0 means that the opinions are not polarized at all, and the resulting distribution of opinions would either take the form of a single Dirac delta centered at a neutral opinion, or be entirely centered in one of the poles, implying that the population (A) of the other pole would be reduced to zero and $A = 1$. Notice that for non-uniform distributions centered in a neutral opinion, 1, but still presents a minimum polarization due to a small separation between gravity centers, that depends on the standard deviation . In the case of a Gaussian distribution centered at zero, = $\overline{2}$.

In between, polarization can lie within the range, $0 <$ < 1, for three reasons: i) The population sizes associated to each opinion are equal, but the pole distance d is lower than 1. ii) Despite d being equal to 1, the population sizes associated to each opinion are different and therefore there is a majority sharing a similar opinion. iii) A combination of i and ii. Fig. 9.2A illustrates the basic concepts of the proposed index of polarization, as it visualizes the area associated to each opinion, their corresponding gravity centers and the pole distance for a standard case of a perfect bimodal distribution. In panel B of this figure, we have visualized non polarized distributions (= 0 and 1), a perfectly polarized one (= 1) and a case in between.

9.4. Twitter data: The Venezuelan case

In this section, we apply our model and polarization index to Twitter data regarding the late Venezuelan President Hugo Chávez. We downloaded

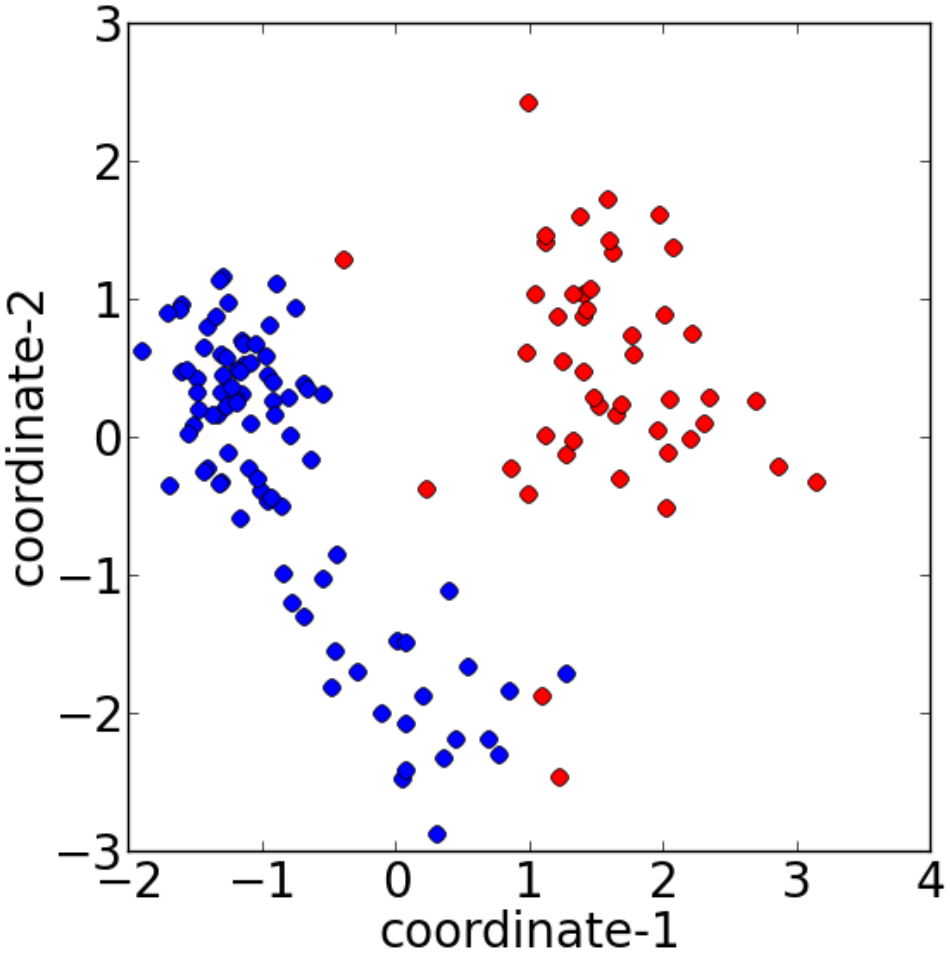

Figure 9.3: Projection in a two-dimensional space of the distribution of *elite* users according to the similarity of their content. Dots represent users and colors indicate the community they belong to in the *elite* network: red for the officialism and blue for the opposition. The distance between users is inversely proportional to the similarity of their content.

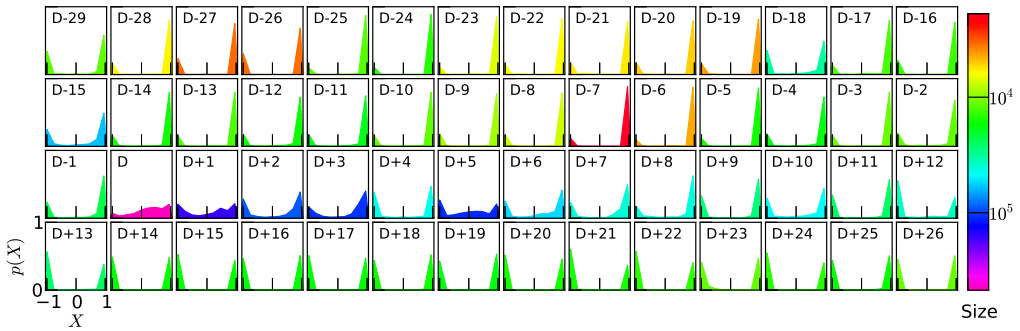

Figure 9.4: Time evolution of ideological value (X_i) probability density functions ($p(X)$) for the Venezuelan conversation. Labels indicate the day of observation, D standing for the day of the Presidents death. Colors indicate the number of participants.

over 16,383,490 messages written by 3,173,090 users from 02/04/2013 to 05/04/2013. This period covers one month preceding his death, the announcement of the death, and the schedule for new elections. We use retweets as a proxy for in uence [Boy10, Dan10, Jae12, Ste13, Lie10, Sha11, Web14], and build a weighted and directed network accounting for the adoption of ideas among Twitter users for each day. Whenever a user i retweets a message originally posted by user j, we assume that i is being in uenced by j's ideas. Hence, a new directed link (j i) is created. Note that in this case the direction of the links is opposite to the retweet networks built in the previous chapters. We constructed an individual retweet network for each day of the observation period, which is a total of 56 networks. More details about the dataset and the retweet networks can be found in sections 4.3 and 9.7 respectively.

In order to apply the model to these daily networks, we begin by defining a set of *elite* users. We denote as *elite* those users who gained a noticeable amount of retweets and actively participated in the conversation along the observation period. The distribution of users according to the total amount of retweets obtained (S_{out}) and participation rate ( ) is shown in Fig. 9.9 of section 9.7. In this case, we considered a very small set (0.02%) of in uential users who participated most of the observation period ($> 89\%$) and obtained a very high number of retransmissions ($S_{out} > 1000$).

The *elite* users mainly correspond to politicians, journalists and mass media accounts, whose political position and editorial tendency are publicly known and who belong to both sides of the Venezuelan political spectrum.

In order to assign them an ideology value, X_s, we first studied their network of interactions. In the *elite* network, nodes represent the *elite* users, and links are created and accumulated whenever an *elite* user i retweets an *elite* user j. This network is polarized in a well defined two-community structure, with modularity $Q = 0\ 38$. In each community, users share political ideology and hardly interact with users from the other pole. In fact, the assortative mixing [New03] by political ideology is very high ($r = 0\ 88$).

In order to further understand the *elite* polarization, we analyzed the content of their messages. For this purpose, we abstracted each *elite* user as a high-dimensional vector, where each element represents the number of times that the user posted each of the 500 mostly used words from all the *elite*'s messages. Then, we reduced the high-dimensional space into a two-dimensional one, by applying a multi-dimensional scaling algorithm [MDS05]. In this algorithm, users are mapped into a new space by preserving the distance between them in the original one. This means that the distance between users is inversely proportional to the similarity of their posted contents. In Fig. 9.3, we present the projection of the users in the new two-dimensional space. Dots represent users and colors are assigned according to the community they belong to in the *elite* network. It can be noticed that these users are not homogeneously distributed in the new space. Instead, they are separated from each other in agreement with our previous classification. This means that the use of language is polarized among the *elite* users.

After identifying the *elite* users, we assigned them ideology values of $X_s = -1$ to the officialism side and of $X_s = 1$ to the opposition. The remaining users (99 98%) were assigned the role of $listeners$ and $X_l = 0$. After running the model we obtained an ideology probability density function $p(X)$ for each day. The resulting $p(X)$ for each network are presented in Fig. 9.4. The label indicates the day of observation, D representing the day of the death. The color indicates the network size in terms of the number of participants. As can be seen the days with largest participation (purple and blue) correspond to the most important announcements: the presidents death (day D), and call for election (day $D + 6$). Next, we calculated the polarization index ( ), pole distance (d) and populations sizes for the resulting distributions of each day and plotted the results in Fig. 9.5.

We identify day D as a turning point which ended up polarizing even more the conversation. During the days preceding the announcement (from $D - 29$ to $D - 1$), X presents a bimodal distribution in which the officialism population (negative side of the X distribution) is considerably smaller than the opposition (positive side of the X distribution). This means that during this period the conversation was still polarized, but practically monopolized by the opposition. Hence, despite the fact that the pole distance reached val-

ues over 0 9, the polarization index just averaged under 0 4. Then a shift in the conversation emergent patterns took place on the day of the President's death announcement (day D). During this day X lost its bimodal distribution, and the resulting $p(X)$ was centered around neutral values, minimizing the pole distance. All these meaning that the conversation was not so polarized and that the network does not have a two-island structure anymore. Therefore, the polarization index decreased, 0 25. This behavior is due to the bursty growth of the conversation at day D (see Fig. 9.7 in the section 9.7). As a consequence, the previously segregated modules combined into a single-island structure, many times larger than the usual network size. Besides a large amount of users from all around the globe joined to the conversation, making the topic international, rather than local from Venezuela. In fact, during this day the percentage of users tweeting from Venezuela (20%) was very low in comparison to the rest of the days (average around $> 80\%$). Hence, our set of Venezuelan *elite* were not capable of polarizing this majority of worldwide users. However, from there on the conversation recovered its bimodal distribution of opinions. Moreover, the polarization reached its maximum from day $D + 12$ (marked with the dashed line) onwards, day that the officialism new leader entered the conversation. From this day onwards X presents a bimodal distribution, where the populations of both sides are similar. Therefore, the polarization index averaged values around 0 9.

9.5. Twitter shows the two sides of Venezuela

Next we evaluate our model and the validity of Twitter data by comparing the geographic distribution of the polarized users with o ine data regarding the Venezuelan socioeconomic and political landscape. More specifically, we analyze the geographical density of geolocated tweets in Caracas, the capital city of Venezuela, taking the results obtained from the most polarized days in section 9.4 as a proxy of their ideology. For this purpose, we have built the density functions that a tweet associated with the officialism or the opposition had been posted by a geolocated user at a given position (longitude and latitude). We considered a grid of 100 cells between longitudes [-67.12^o, -66.71^o] and latitudes [10.31^o, 10.57^o] and counted the number of tweets in each cell, identified with each ideology. Then, we normalized both counts by their respective total number of tweets. The resulting functions are two surfaces on top of the map, which we show in Fig. 9.6 as contour plots (red for the officialism and blue for the opposition) that indicate lines of equal value in the 2-D probability density function. These contour lines are superimposed

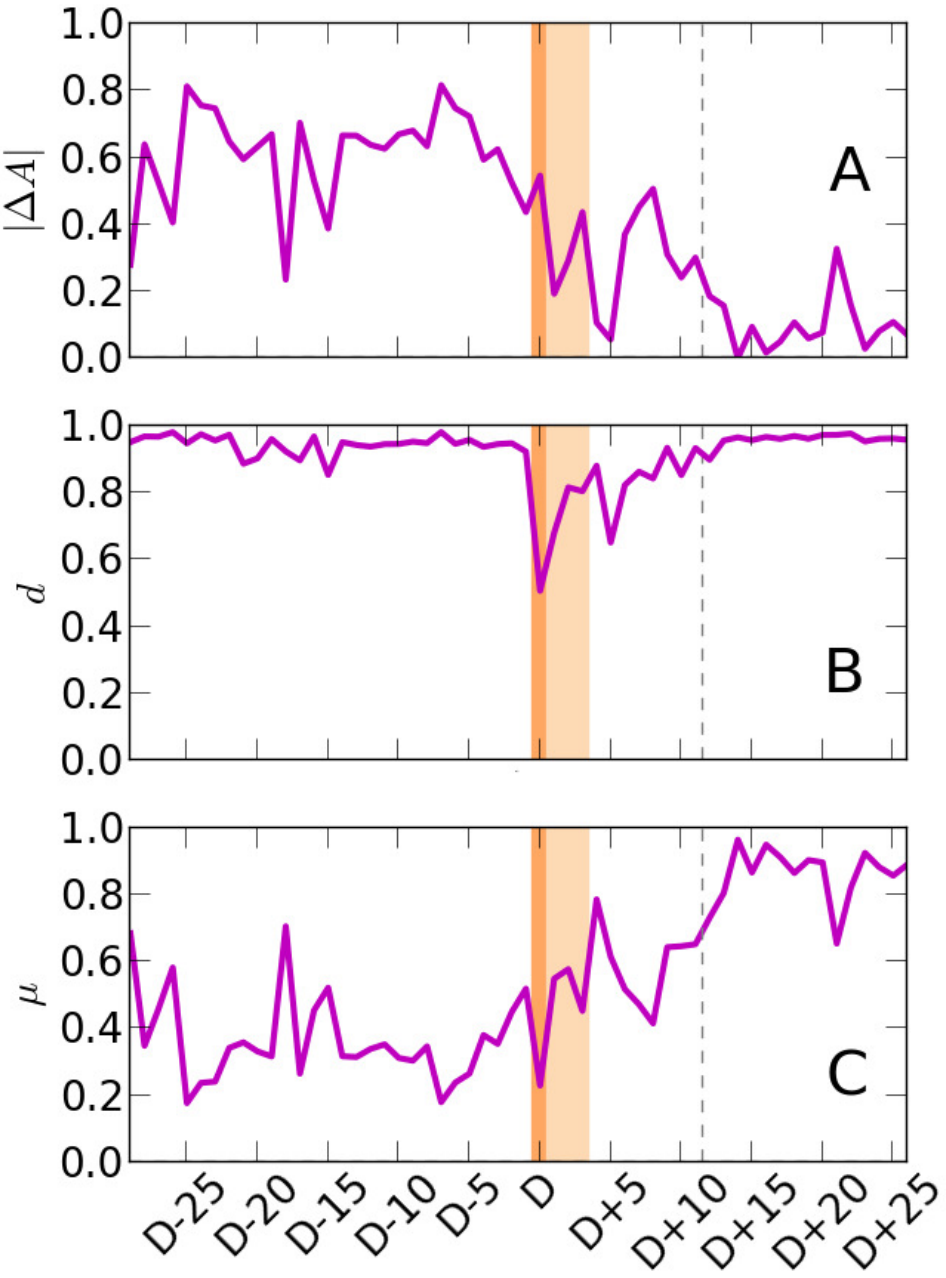

Figure 9.5: Time evolution of the polarization index (C), and the variables associated with it: difference in population sizes (A) and pole distance d (B) for the Venezuelan conversation.

on a map of the municipalities composing the city of Caracas. There are five of them, bordered in green. The labels correspond to the municipality name, and the color indicates the ruling party-like the officialism in Libertador and the opposition in Chacao, Sucre, Baruta and El Hatillo. Additionally, urbanized areas are colored in yellow and poorer regions (slums) in pink. Notice that the West region is characterized for having lower income and governed by the officialism, while the East part is wealthier and governed by the opposition.

It can be noticed that the regions where each pole concentrates most of their tweets are well separated from each other, showing that the city presents a clear geographical polarization. In fact, there is a good correspondence between the results of our model and o ine evidence, such as electoral results or socioeconomic factors. Those municipalities governed by the opposition contain the highest concentration of users identified with this pole, and the same effect occurs for the officialism side of the political spectrum. We also have to remark that the areas with higher concentration of users aligned with the officialism, correspond to the parts of the city with the largest concentration of poorer neighborhoods (pink areas). Conversely, the opposition users concentrate in urban developed regions. All these suggesting that the basis of the Venezuelan popular polarization resides in socioeconomic factors and that the political con ict in Venezuela presents a strong territorial facet.

9.6. Conclusions

Modern democracies have to represent the con icts existing in our society, while at the same time maintain the social stability [Dia90]. However, as polarization emerges, the few most powerful parties tend to capitalize the whole of the public attention and support, silencing the moderate opinions and under representing minorities. Consequently, todays' society is concerned about polarization, as a politically polarized society implies several risks. These risks include the appearance of radicalism or civil wars. In fact, one of the actual challenges and a cutting edge topic is how to detect the emergence of political polarization and how to fix it.

We state that the possibility to gather user generated data from social media platforms [LPA+09], together with network science [Lin02], represents an opportunity to detect political polarization. In this work, we have proposed a methodology to study and measure the emergence of polarization from social interactions. We have used it, to analyze the political polarization in one of the most polarized countries: Venezuela [Ell04, MLB12]. We have done this, by applying our methods to a Twitter conversation about the

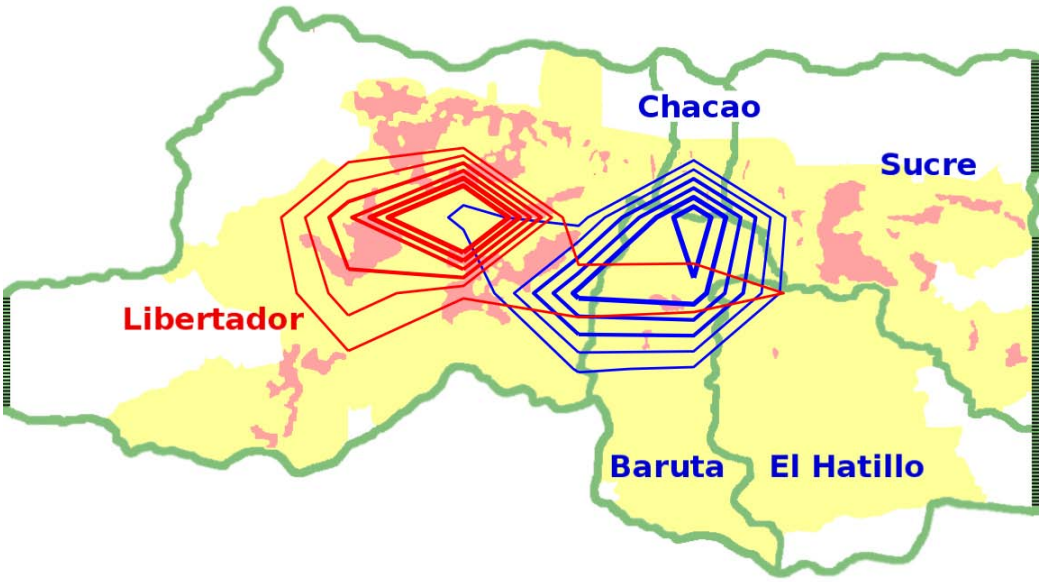

Figure 9.6: Geographical polarization in the city of Caracas. Contour lines represent the density functions of the probability that a tweet associated with the officialism (red) or the opposition (blue) had been posted by a geolocated user at a given position (latitude and longitude). These contours have been superimposed to the map of Caracas, Venezuela. From inside out, contours indicate the following values: [0.175, 0.15, 0.0125 0.10, 0.075, 0.05]. The green lines border the five municipalities composing the city. Labels indicate the name of the municipality and the color indicate the ruling party according to the 2013 Venezuelan local elections (red for the officialism party and blue for the opposition parties). White represents unpopulated areas, yellow urbanized areas and pink the poorer neighborhoods.

late Venezuelan president Hugo Chávez. We have shown that our methodology is able to detect different degrees of polarization in the conversation, depending on the participants' behavior, given by the structure of the network. Finally, we have contrasted our results against o ine data, such as municipality governments or socioeconomic factors, finding a good correlation between the online and o ine polarization. Hence, we conclude that online data seem to be a good proxy to detect politically polarized societies, as the online polarization that we found is a re ection of the Venezuelan political, territorial and social polarization.

Another relevant question is: Can social media platforms help reduce political polarization as more voices could be heard? Although we do not answer this question, our results show that a minority of *elite* users were able to in uence the whole online social network, resulting in a highly politically polarized conversation. However, these Venezuelan local in uential accounts were not capable of polarizing the network when the conversation stopped being local of Venezuela and turned to be international. This opens two questions that can be studied from a social media analysis perspective: i) How does online political polarization change at different scales-like city, country, continent or whole world? ii) How could we target interventions in control strategies on social media that might be implemented to reduce polarization?

9.7. Additional Methods: Networks

We have built one retweet network for each day of the observation period (56 networks). A retweet network emerges from user-to-user interactions during the message retransmission process provided by Twitter. Nodes represent users and links are created between users i and j, when i forwards the content previously posted by j. Edges are weighted in proportion to the frequency that i retweeted j's messages, and directed in the sense of the ow of information from the message source j to the retweeter i.

A single network contains several retransmission cascades, seeded and propagated by the conversation participants. When these cascades are aggregated, several disconnected network components emerge. Among these components, there is a single one called Giant Component (GC) whose size is in the same order of the whole network. As part of the GC, there is a set of nodes that are reachable from the set of in uential *elite*, that represent about 50% of the GC's size (Fig. 9.7A). For most of days, the amount of reachable nodes uctuated around 10,000 users and explosively grew to almost 500,000 users during day D (Fig. 9.7B). This behavior is typical of breaking news

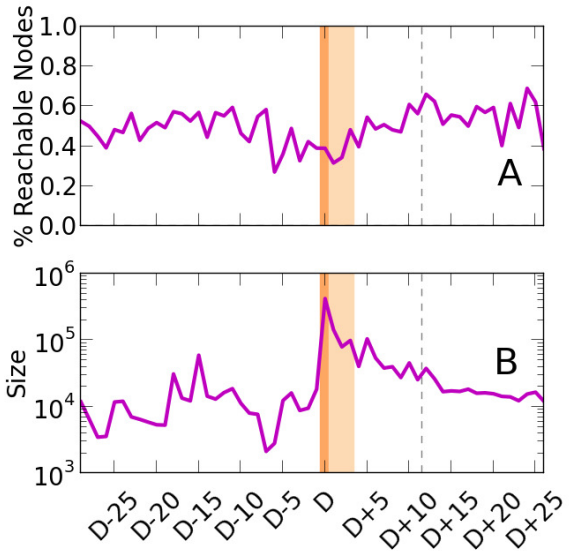

Figure 9.7: Time evolution of the relative number of reachable nodes in comparison to the GC (A) and size of the reachable nodes' networks (B).

and critical events, with a bursty increase during the main occurrence and a slow decay that may last for several days.

The retweet networks characterize the way that the collective attention is organized during an event on Twitter. The out strength (s_{out}) indicates the amount of retweets gained by a participant, while the in strength (s_{in}) indicates the number of retweets made by the participant. In Fig. 9.8 we have superimposed the out strength (top) and in strength (bottom) complementary cumulative density functions (CCDF) for each of the constructed networks, in log-log (left) and linear-log (right) scales. In both cases, the distributions display heterogeneous behavior, being the out strength distributions broader than the in strength distributions. In order to compare, whether these distributions behave like an exponential rather than a power law, we calculated the likelihood ratio statistical test. We found that the probability of these distributions to follow an exponential curve, instead of a power law, has a p-value $< 0\,01$ for more than 98% of the outgoing distributions and 75% of the incoming distributions, where over 87% of the distributions have a p-value $< 0\,05$.

From a dynamical point of view, the power law distributions imply a preferential attachment mechanism [Lin02], where the chances of being retweeted increases with the number of retweets previously gained. These dynamics re-

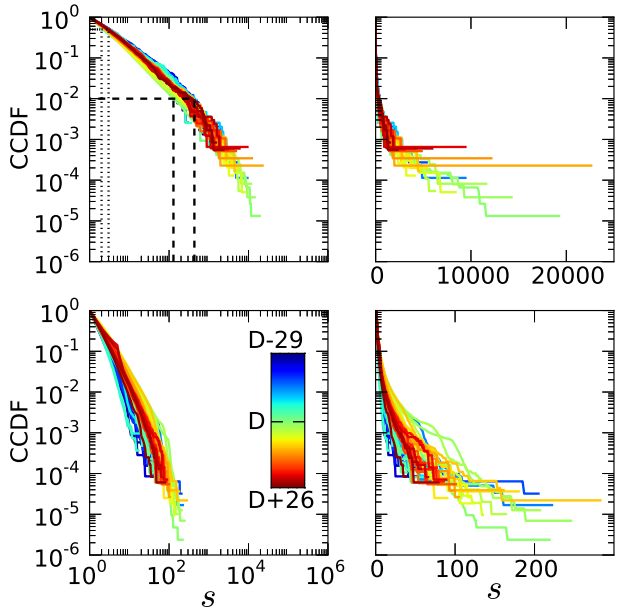

Figure 9.8: Complementary cumulative density function (CCDF) of the retweet networks out strength s_{out} (top) and in strength s_{in} (bottom), from the Twitter conversation about the Venezuelan President Hugo Chávez, in log-log (left) and linear-log (right) scale. The colors indicate the corresponding day of the observation period. The dotted line indicates the range of s_{out} for 50% of the population, while the dashed lines indicate the range of s_{out} for 1% of the population.

sult in heterogeneous distributions where the great majority of users receive a very small amount of the collective attention, while some scarce users receive a disproportionally larger amount of it. For example, at all days 50% of the population gained between 2 or 3 retweets at most (dotted lines in the top left panel of Fig. 9.8), while the 1% of most retweeted participants gained from 130 to 430 retweets as minimum (dashed lines in the top left panel of Fig. 9.8).

To further understand the relationship between the individual activity and the attention received, we will aggregate the observation period by characterizing the individuals according to their rate of participation and total amount of retweets gained. The participation rate is defined as:

$$= \ _i \ T \tag{9.9}$$

where $_i$ is the number of days that the user i actively participated in the retweet process and T is the total length of the observation period. The total number of retweets gained by user is measured as:

$$S_{out} = \sum_{t=0}^{T} s_{out}(t) \tag{9.10}$$

where $s_{out}(t)$ is the out strength of the node i at day t. If the user did not actively participated at day t, then $s_{out}(t) = 0$.

The joint probability density function of the accumulated out strength S_{out} and the participation rate , $P(S_{out} \ )$, is shown in Fig. 9.9. This distribution indicates the total amount of attention received by users according to their participation rate. It can be noticed that the largest density of users (red and orange dots in Fig. 9.9) participated less than 20% ($< 0\ 2$) of the days and present a small out strength value ($S_{out} < 10$), which means that most of them received a little amount of the collective attention. However, there is a very small set of users at the upper right corner in Fig. 9.9, who participated almost every day and present an extremely high S_{out}. This minority of highly in uential users captured most of the collective attention throughout the observation period, and define the *elite* users considered in section 9.4.

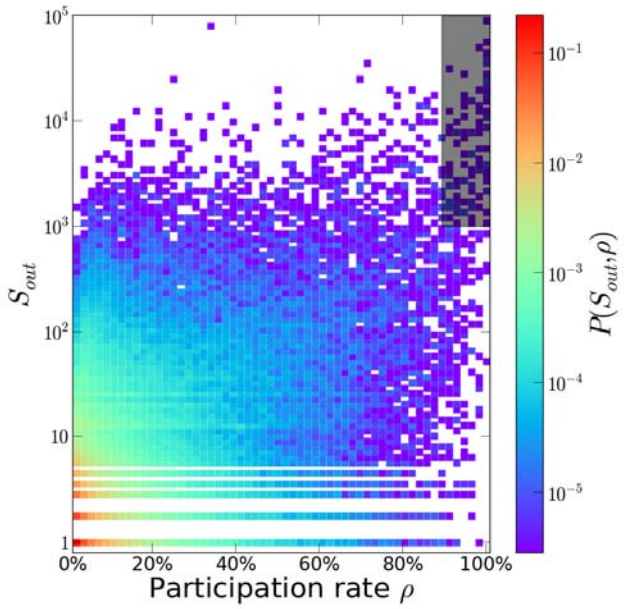

Figure 9.9: Joint probability density function of the accumulated out strength (S_{out}) and the participation rate ( ), from the Twitter conversation about the Venezuelan President Hugo Chávez. The colors correspond to the density of users. The black square at the top right corner indicates the *elite* users defined in section 9.4.

Chapter 10

Conclusions

In this thesis we have presented our contributions to a few problems in the field of network science that enhance our knowledge about society. More particularly, we have abstracted society as a network, what has enabled us to analyze the dynamical processes that occur within it and further understand how the social networks in which we are embedded shape our behavior.

Next, we summarize the main conclusions that derive from this thesis:

1. In the first part of this document we have presented an agent based model to explore how the structure of the social networks in which we are embedded can limit the meritocracy of societies. A society is said to be meritocratic if the compensation and power available to individuals is determined by their abilities and merits. As opposed to meritocracy, in this thesis we have introduced the term *topocracy*. We define a system as *topocratic* if the compensation and power available to an individual is determined primarily by her position in a network. In the model, individuals produce and sell content, but also distribute the content produced by others when they belong to the shortest path connecting a buyer and a seller. The production and distribution of content defines two channels of compensation: a meritocratic channel, where individuals are compensated for the content they produce, and a topocratic channel, where individual compensation is based on the number of shortest paths that go through them in the network.

   We have shown that the meritocracy of the model decays as the network becomes sparse, giving rise to a topocratic regime, in which the compensation received by individuals is explained primarily by the position they occupy in the network. Thus, we have demonstrated that the structure and connectivity of the networks where markets are embedded represent the main factor determining whether the system is in

a topocratic or meritocratic regime. This conclusion implies that theories that assume away the existence of networks are implicitly assuming away the possibility that markets can be non-meritocratic.

a) We have shown that social networks can not only be instrumental drivers of inequality in centrally planned economies, but also in free markets, where connections to business elites can take the role that connections to party leaders have in autocratic regimes. Networks, thus, affect the functioning of decentralized economies and limit the often desired equality of opportunity since they help determine the information and resources available to each individual. In this thesis we have gain insight in the role that social networks play. We have done so by exploring how the structure of the social network can limit the meritocracy of the economic activity taking place on it. In the context of equality of opportunities, John Rawls argues that equality of fair opportunity will only be satisfied in a society where the same native talent and the same ambition have the same prospects of success [Raw99, Raw01]. Policy-makers who adhere to Rawls ideas have emphasized the field-leveling role of inheritance taxes, education and anti-discriminatory policies in the labor market. Yet, opportunities are not constrained only by talents, education and property, but also by the connections available to each individual, which cannot be taxed. Hence, we have shown that a thorough understanding of the meritocracy of market mechanisms cannot be achieved without understanding the effects of an individual's position in a network and its relative effect with respect to other forms of advantage where field leveling policies do exist.

b) In a 21$^{\text{st}}$ century context the results of this model also speak about the social changes that are implied by recent changes in technology. In recent years the emergence of the internet has given rise to a world in which it is much easier for individuals to market directly to each other, or at least, through one large intermediator (such as iTunes, Amazon or eBay). Our model predicts that these changes should increase the meritocracy of society since they help reduce the long chain of intermediations that consume valuable payoffs in a poorly connected society.

c) However, our results do not strictly mean that denser networks are unambiguously preferable to sparser networks. Making such a judgement would require weighing the effects that network den-

sity has on meritocracy with its effect on other social and economic outcomes. Social networks do not only affect the distribution of payoffs among content producers and middlemen, but also are known to affect the outcome of coordinated collective action. For instance, evolutionary game theory suggests that cooperative strategies are more likely to emerge in networks that are not highly connected. In a public good game sparse network prevents free-riders from prospering because free-riders cannot sustain enough links to exploit multiple neighbors [SSP08, PS08, GGnCFM07]. Thus, in a sparse network, the same agents that benefit from their position as middleman might be the same agents that play a crucial role enhancing cooperation. Therefore, making a judgement on whether a denser or sparser network is more beneficial for society in general, is a matter that cannot be answered easily, since it requires weighing the effects of the network structure on meritocracy and cooperation, but also, on other relevant outcomes, from the preservation of cultural diversity to the spread of disease.

2. In the past, the absence of large detests that capture the structure of social networks has limit our ability to understand how social relations shape our behavior. However, our recently acquired ability to gather, store and analyze large datasets of human generated data has enabled the explosion of network science during the beginning of this century. Hence, by analyzing big data by means of network science we are able to further understand human behavior and how the structure of the social networks in which we are embedded shapes our lives. On the latter part of this thesis we have used Twitter data to proxy social relations and studied the communication patterns of several political conversations taking place on Twitter. Our main conclusions are summarized in the following paragraphs.

   a) We found that the collective attention is highly heterogeneously distributed, as there is a minority of extremely in uential accounts. In fact, the ability of individuals to propagate messages or ideas through the platform is constrained by the structure of the follower network underlying the social media and the position they occupy on it. Hence, although people have argued that social media can allow more voices to be heard, our results suggest that Twitter is highly topocratic, as only the minority of well positioned users are widely heard.

   b) We have proposed a methodology to capture the relationship be-

tween individual activity and the impact this activity has on the remaining individuals of the network. Accordingly, we have introduced a mesure to quantify the efficiency of users to propagate messages. We have defined efficiency as the ratio between retweets gained and activity employed for it. This ratio can be understood as a measure of in uence of an individual in the social network in which she is embedded.

c) We found the distribution of efficiency to be universal across several Twitter conversations, following a lognormal distribution but with a larger density of users at the highest orders. We have also proposed a model to explain the emergence of the efficiency distribution, based on biased independent cascades taking place through the followers network. The simulation of the model unveiled the effects that topology and individual behavior have on the emergent dynamical patterns. More particularly, it revealed that the emergence of a small fraction of highly efficient users results from the heterogeneity of the network in which users are embedded, and independently of individual activity strategies.

d) The three differentiated Twitter interaction mechanisms, follower, mention, and retweet, define three layers through which individuals receive and diffuse information. Hence, the Twitter information diffusion process does not take place through a single channel, but three. In order to fully understand the process we have to simultaneously analyze all three channels.

e) The propagation of messages via retweets is strongly conditioned by the topology of the follower layer, as it establishes the substratum through which individuals receive information. Additionally, users establish conversations or refer to each other using the third available channel, the mention.

f) We have also analyzed the communication patterns of the Spanish political debate taking place on Twitter. We found that the Spanish political conversation is centralized around a small fraction of in uential accounts. Politicians are the main characters, since their accounts were the most mentioned, and captured most of the collective attention. However, despite their accounts are still in uential in the retweet network, users tend to propagate information from the same sources as in the o ine world. Thus, traditional media accounts were the most retweeted. Therefore we can affirm that, on the light of our results, despite social media

should allow more voices to be heard, the political communication is still driven by a minority of political parties and elite media.

g) Analyzing the mention network by means of community structure, we have been able to map the ow of political information going through Twitter during the campaigns. We show it to be considerably polarized by political ideology, as users crowded around a single political party accounts, and preferentially communicated with those of their same political stance. However, there were a small fraction of users, more exposed to political disagreement, who sustained the exchange of information among these polarized communities. Similar conclusions can be made from the retweet analysis, where we found that despite the countless sources of information available, users do not take full advantage of it, tending to just rely on their preferred one. In this regard users mostly retweet from just one traditional media official accounts with whose editorial line they feel identified. Hence, we can affirm that the social network in which individuals are enmeshed is ideologically homogeneous. Such networks might be too insular, limiting the opportunities to learn about politics and contrast information. However, they represent effective information shortcuts to access information [Mar87][Mut06][Lup94].

h) The collapsed directed multiplex network does not present a rich-club ordering, as politicians presided large communities of regular users in the mention layer; while media accounts were the sources from which people retweeted information. However, when considering reciprocal interactions the rich-club ordering emerges, as elite accounts preferentially interacted among themselves and largely ignored the crowd. The rich-club was mainly composed by politicians, media, and well-known bloggers. Hence, we identified the top 50 in uential users at each layer, and classified them as media, politicians, or bloggers. Despite an slight overlapping among the top in uentials at each layer, the relevance of the three different collectives significantly varied from one layer to another. The relevance of media and politicians at the follower level seems to be balanced. However, politicians clearly stand out in the mention layer, while media stand out in the retweet layer. A high degree in the mention layer is usually associated as a high value name, *i e* a famous and popular account, while the gain of retweets is associated to producing high value content tweets. Our results show that media were the sources of information, while politicians were the

main characters of both conversations. Moreover, it suggests that politicians in general were not capable of producing high quality content tweets that got highly retweeted. All these resulted on users clustering around politicians in the mention layer, and around media accounts on the retweet layer. Hence, the leaders emerging at each layer vary significantly, and one can not claim neither politicians ruled the media or vice versa. It all depends on what kind of interactions we are considering and what effect we are trying to understand.

i) The results of our political analysis also speak about the importance that nationalist currents have in the Spanish political landscape, where at some regions the nationalist component of the parties becomes much more in uential than their ideology. It is widely known that the strong feeling of membership to the autonomous community, that exists in several regions of Spain, is frequently used as a political argument. This is re ected on Twitter in several ways:

1) Catalan tends to be overused in political conversations.
2) Despite the vast majority of speakers of a co-official language are bilingual, the conversation is highly segregated by language.
3) There is an obvious relationship between the political alignment of users and the language in which they tweet.

The utility of the proposed method to determine the proximity between languages and political parties is not particular of Spain, but generalizable to other countries. For example, let's think about the United Sates. Although English is the main language, there is a diversity of cultures co-existing inside the country that speak different languages. Hence, the proposed methods could be used to estimate the proximity between the Chinese or Hispanic communities and the liberal or conservative parties. Thus, being useful to estimate electoral outcomes.

j) We have introduced a methodology to study and measure the emergence of political polarization from social interactions. To this end, we have first proposed a model to estimate opinions in which a minority of in uential individuals propagate their opinions through a social network. The result of the model is an opinion probability density function. We have also proposed an index to quantify the extent to which the resulting distribution is polarized.

Finally, we have illustrated our methodology by applying it to Twitter data.

In this document we combine the sociological concept of embeddedness developed by Granovetter with network science to gain understanding of our society. By doing so we have been able to first, propose and solve a model that relates the structure of the social network in which a society is embedded and the meritocracy of its economy. In a world where personal data is increasingly available, the results of the analytical model introduced in this work can be used to enhance meritocracy and promote policies that help to build more meritocratic societies. Secondly, we have conducted several empirical studies that characterize the main properties of the new online social networks. These results, are key to understand the new data-driven society that is emerging. In particular, we have presented relevant information that can be used to benchmark future models for online communication systems or can be used as empirical rules characterizing our online behavior.

Appendix A

The Spanish political system

Spain is a parliamentary monarchy since the approval of the last constitution in 1978, and belongs to the European Community since 1986. The country is organized into 17 autonomous communities (AC) and two autonomous cities. In Spain there is a strong feeling of membership to the AC. In fact despite Castilian being the official language for the whole territory, several ACs (Galicia, Basque Country, Catalonia, Valencia and Balearic Islands) have their own co-official language. Moreover in some of them there is strong nationalist current, with the existence of nationalist parties represented in parliament, above all at Catalonia, Basque Country, and Galicia.

In a general elections context we can distinguish two kinds of political parties, those of national scope, and others of regional character also standing for election. The most important parties can be found in Table A.1, where they are brie y characterized.

Regarding the national scope, we can distinguish two main ideologies: the conservative represented by Partido Popular (PP), and the liberal represented mainly by Partido Socialista Obrero Español (PSOE), and the smaller and far more left-wing party Izquierda Unida (IU). Recently a new party, Union Progreso y Democracia (UPyD), whose political posturing is difficult to classify in the classic conservative/liberal framework has broken in the parliamentary spectrum. The Financial Times and The Economist classified it as of center ideology, as it combines social liberalism with a defense of 'the unity of Spain'.

Within the autonomic plane we can also distinguish liberal and conservative parties, though their nationalist component is much more in uential.

Table A.1: Information about the most important Spanish political parties that took part on the 20th of November of 2011 general elections. Name of the party; acronym; level at which the party stood to the elections, nationally (N) or just in one Autonous Community (R); important official Twitter accounts.

	ï¿½		Official Accounts:	
Political Party	Acronym	Level	Candidate	Others
Partido Popular	PP	N	marianorajoy	ppopular, esperanza-guirre, sorayapp
Partido Socialista Obrero Español	PSOE	N	conrubalcaba	psoe, carmechacon2011, elenavalenciano
Izquierda Unida	IU	N	cayo_lara	iunida, gllamazares
Union Progreso y Democracia	UPyD	N		upyd, tonicanto1
EQUO	EQUO	N	isabanes	proyectoequo
Convergencia i Unio	CiU	R (Catalonia)	ciuduran2011	ciu
AMAIUR	AMAIUR	R (Basque Country)		bildueh, aralarnafarroa
Partido Nacionalista Vasco	PNV	R (Basque Country)	jerkoreka	eaj_pnv
Ezquerra Republicana de Catalunya	ERC	R (Catalonia)	bosch	esquerra_erc
Bloque Nacionalista Gallego	BNG	R (Galicia)		bng, obloque
Iniciativa per Catalunya	ICV	R (Catalonia)	jcoscu	
Partido Andalucista	PA	R (Andalusia)		pandalucista, p_andalucista

Bibliography

[AA70] G. A. Akerlof. *The market for lemons: Quality uncertainty and the market mechanism.* Quarterly Journal of Economics, 84 (3): 488, 1970.

[Ace11] D. Acemoglu and A. Ozdaglar. *Opinion dynamics and learning in social networks.* Dynamic Games and Applications, 1 (1): 3-49. 2011.

[Ace13] D. Acemoglu, G. Como, F. Fagnani, and A. Ozdaglar. *Opinion fluctuations and disagreement in social networks.* Mathematics of Operations Research, 38 (1): 1- 27, 2013.

[AG05] L. A. Adamic and N. Glancee. *The political blogosphere and the 2004 U.S. election: divided they blog.* In Proceedings of the 3rd International Workshop on Link Discovery, LinkKDD '05, pages 36 43, New York, NY, USA, 2005.

[Ama04] A. A. Moreira, A. Mathur, D. Diermeier and L. A. N. Amaral. *Efficient system-wide coordination in noisy environments.* Proceedings of the National Academy of Sciences of the United States of America, 101 (33): 12085-12090, 2004.

[AMM11] E. Aramaki, S. Maskawa and M. Morita. *Twitter Catches the Flu: Detecting Influenza Epidemics Using Twitter.* In Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '11, pages 1568-1576, Association for Computational Linguistics Stroudsburg, PA, USA, 2011.

[AP09] S. Allesina and M. Pascual. *Googling Food Webs: Can an Eigenvector Measure Species' Importance for Coextinctions?.* PLOS Computational Biology, 5 (9): e1000494+, 2009.

[AZBA08] L. A. Adamic, J. Zhang, E. Bakshy and M. S. Ackerman. *Knowledge sharing and yahoo answers: everyone knows something.* In Proceedings of the 17th international conference on World Wide Web, pages 665 674. ACM, 2008.

[BA99] A. L. Barabási and R. Albert. *Emergence of scaling in random networks.* Science, 286 (5439): 509 512, 1999.

[BAJ99] A. L. Barabási, R. Albert and H. Jeong. *Mean-field theory for scale-free random networks.* Physica A: Statistical Mechanics and its Applications, 272 (1-2): 173 187, 1999.

[Bar04] M. Barthelemy. *Betweenness centrality in large complex networks.* European Physical Journal B, 38: 163, 2004.

[Bar12] A. L. Barabási. *Network science.* University of Chicago Press, 2012.

[BB07] D. Baldassarri and P. Bearman. *Dynamics of Political Polarization.* American Sociological Review, 72 (5): 784 811, 2007.

[BBC+12] I. Bordino, S. Battiston, G. Caldarelli, M. Cristelli, A. Ukkonen and I. Weber. *Web search queries can predict stock market volumes.* PloS ONE, 7 (7): e40014, 2012.

[BBC+14] S. Boccaletti, G. Bianconi, R. Criado, C.I. del Genio, J. Gómez-Gardeñes, M. Romance, I. Sendiña Nadal, Z. Wang and M. Zanin. *The structure and dynamics of multilayer networks.* Physics Reports, 544 (1): 1 122, 2014.

[BBPSV04] A. Barrat, M. Barthélemy, R. Pastor-Satorras and A. 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.

[BBRSH14] J. Borondo, F. Borondo, C. Rodriguez-Sickert and C. A. Hidalgo. *To Each According to its Degree: The Meritocracy and Topocracy of Embedded Markets.* Scientific Reports, 4: 3784, 2014.

[Ben08] I. J. Benczik, S. Z. Benczik, B. Schmittmann and R. K. P. Zia. *Lack of consensus in social systems*, Europhysics Letters, 82: 48006, 2008.

[BFJ+12] R. M. Bond, C. J. Fariss, J. J. Jones, A. D. I. Kramer, C. Marlow, J. E. Settle and J. H. Fowler. *A 61-million-person experiment in social influence and political mobilization.* Nature, 489 (7415): 295 298, 2012.

[BG08] D. Baldassarri and A. Gelman. *Partisans without Constraint: Political Polarization and Trends in American Public Opinion.* American Journal of Sociology, 114 (2): 408 446, 2008.

[BGL10] D. Boyd, S. Golder and G. Lotan. *Tweet, tweet, retweet: Conversational aspects of retweeting on twitter.* In HICSS, pages 1 10. IEEE Computer Society, 2010.

[BGLL08] V. D. Blondel, J. L. Guillaume, R. Lambiotte and E. Lefebvre. *Fast unfolding of communities in large networks.* Journal of Statistical Mechanics: Theory and Experiment, 2008 (10): P10008, 2008.

[Big93] G. Biglaiser. *Middlemen as experts.* The RAND Journal of Economics, 24 (2): 212-223, 1993.

[Bin11] D. Bindel, J. Kleinberg and S. Oren. *How bad is forming your own opinion?.* Foundations of Computer Science (FOCS), IEEE 52nd Annual Symposium, pp. 57-66, 2011.

[BLM54] B. R. Berelson, P. F. Lazarsfeld and W. N. McPhee. *Voting: A Study of Opinion Formation in a Presidential Campaign.* University of Chicago Press, 1954.

[BMBL14] J. Borondo, A. J. Morales, R. M. Benito and J. C. Losada. *Mapping the online communication patterns of political conversations.* Physica A: Statistical Mechanics and its Application, 414 (0): 403 413, 2014.

[BMLB12] J. Borondo, A. J. Morales, J. C. Losada and R. M. Benito. *Characterizing and modeling an electoral campaign in the context of Twitter: 2011 Spanish Presidential election as a case study.* Chaos, 22 (2): 023138 ST, 2012.

[Boy10] D. Boyd, S. Golder and G. Lotan. *Tweet, tweet, retweet: Conversational aspects of retweeting on twitter.* H. HICSS IEEE Computer Society, 1-10. 2010.

[BP98] S. Brin and L. Page. *The Anatomy of a Large-Scale Hypertextual Web Search Engine*. In Seventh International World-Wide Web Conference (WWW 1998), 1998.

[Bur87] R. S. Burt. *Social contagion and innovation: Cohesion versus structural equivalence*. American Journal of Sociology, 92 (6): 1287 1335, 1987.

[Bur99] R. S. Burt. *The social capital of opinion leaders*. The Annals of the American Academy of Political and Social Science, 566 (1): 37-54, 1999.

[Bur04] R. S. Burt. *Structural holes and good ideas.* American Journal of Sociology, 110 (2): 349, 2004.

[Bur09] R. S. Burt. *Structural holes: The social structure of competition.* Harvard University Press, 2009.

[CCP+14] G. Caldarelli, A. Chessa, F. Pammolli, G. Pompa, M. Puliga, M. Riccaboni and G. Riotta. *A Multi-Level Geographical Study of Italian Political Elections from Twitter Data.* PLoS ONE, 9 (5): e95809, 2014.

[CF09] N. A. Christakis and J. H. Fowler. *Connected: The Surprising Power of Our Social Networks and How They Shape Our Lives*. Little, Brown and Company, 2009.

[CFME11] M. L. Congosto, M. Fernández and E. Moro. *Twitter y política: Información, opinión y predicción?* Cuadernos de Comunicación Evoca, 4: 11 15, 2011.

[CFSV06] V. Colizza, A. Flammini, M. A. Serrano and A. Vespignani. *Detecting rich-club ordering in complex networks.* Nature Physics, 2 (2): 110 115, 2006.

[CGFM12] M. D. Conover, B. Gonçalves, A. Flammini and F. Menczer. *Partisan Asymmetries in Online Political Activity*. EPJ Data Science, 1 (6): 1 17, 2012.

[CGGZ+13] A. Cardillo, J. Gómez-Gardeñes, M. Zanin, M. Romance, D. Papo, F. del Pozo and S. Boccaletti. *Emergence of network features from multiplexity.* Scientific Reports, 3, 2013.

[CH03] R. Cohen and S. Havlin. *Scale-free networks are ultrasmall.* Physical Review Letters, 90: 058701, 2003.

[CHBG10] M. Cha, H. Haddadi, F. Benevenuto and P. K. Gummadi. *Measuring User Influence in Twitter: The Million Follower Fallacy.* ICWSM, 10: 10 17, 2010.

[CLI86] P. Lifford and A. Sudbury. *A model for spatial conflict.* Biometrika, 60 (3): 581-588, 1986.

[Col88] J. S. Coleman. *Social capital in the creation of human capital.* The American Journal of Sociology, 94: S95-S120. 1988.

[Cou12] Digital Policy Council. *World leader rankings on twitter.* Research Note, 2012.

[CRF+11a] M. D. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, A. Flammini and F. Menczer. *Political Polarization on Twitter.* Networks, 133 (26): 89 96, 2011.

[CRF+11b] M. D. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, A. Flammini and F. Menczer. *Political polarization on twitter.* In Proc. 5th International AAAI Conference on Weblogs and Social Media (ICWSM), 2011.

[Cul10] A. Culotta. *Towards Detecting Influenza Epidemics by Analyzing Twitter Messages.* In Proceedings of the First Workshop on Social Media Analytics, SOMA '10, pages 115 122, New York, NY, USA, 2010.

[CXMR07] P. Chen, H. Xie, S. Maslov and S. Redner. *Finding Scientific Gems with Google's PageRank algorithm.* Journal of Informetrics, 1 (1): 8-15. 2007.

[Dan10] S. Dann. *Twitter content classification.* First Monday, 15 (12), 2010.

[DDSRC+13] M. De Domenico, A. Solé-Ribalta, E. Cozzo, M. Kivelä, Y. Moreno, M. Porter, S. Gómez and A. Arenas. *Mathematical formulation of multilayer networks.* Physical Review X, 3: 041022, 2013.

[Deg74] M. H. DeGroot. *Reaching a consensus.* Journal of the American Statistical Association, 69 (345): 118-121, 1974.

[Dia90] L. J. Diamond. *Three Paradoxes of Democracy.* Journal of Democracy, 1 (3): 48 60, 1990.

[Dow57a] A. Downs. *An economic theory of political action in a democracy.* The Journal of Political Economy, 65 (2): 135-150, 1957.

[Dow57b] A. Downs. *An economic theory of democracy.* Harper, 1957.

[Dun98] R. I. Dunbar. *The social brain hypothesis.* Brain, 9: 10, 1998.

[DW07] A. K. Dixit and J. W. Weibull. *Political polarization.* Proceedings of the National Academy of Sciences of the United States of America, 104 (18): 7351-7356, 2007.

[EK10] D. Easley and J. Kleinberg. *Networks, Crowds and Markets: Reasoning about a Highly Connected World.* Cambridge University Press, 2010.

[Ell04] S. Ellner, D. Hellinger. *Venezuelan Politics in the Chávez Era: Class, Polarization and Conflict.* Lynne Rienner Publishers, 2004.

[Erd60] P. Erdös, A. Rényi. *On the evolution of random graphs.* Magyar Tud. Akad. Mat. Kutat Int. Kol., 5: 17-61, 1960.

[Fel91] S. L. Feld. *Why your friends have more friends than you do.* American Journal of Sociology, 96 (6): 1464-1477, 1991.

[FFGP10] J. G. Foster, D. V. Foster, P. Grassberger and M. Paczuski. *Edge direction and the structure of networks.* Proceedings of the National Academy of Sciences of the United States of America, 107 (24): 10815-10820, 2010.

[Fre96] L. Freeman. *Some antecedents of social network analysis.* Connections, 19 (1): 39 42, 1996.

[Fri80] M. Friedman and R. D. Friedman. *Free to choose.* Harcourt Press, 1980.

[Fri90] N. E. Friedkin and E. C. Johnsen, *Social influence and opinions*, Journal of Mathematical Sociology, 15 (3-4): 193-206, 1990.

[GA12] D. Gayo-Avello. *"I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper" – A Balanced Survey on Election Prediction using Twitter Data.* arXiv:1204.6441, April 2012.

[GGJ+10] A. Galeotti, S. Goya, M. O. Jackso, F. Vega-Redondo and L. Yariv. *Network games.* The Review of Economic Studies, 77:218, 2010.

[GGnCFM07] J. Gómez-Gardeñes, M. Campillo, L. M. Floría and Y. Moreno. *Dynamical organization of cooperation in complex topologies.* Physical Review Letters, 98: 108103, 2007.

[GHKV07] M. C. González, H. J. Herrmann, J. Kertész and T. Vicsek. *Community structure and ethnic preferences in school friendship networks.* Physica A: Statistical mechanics and its applications, 379 (1): 307 316, 2007.

[GJE+12] P. A. Grabowicz, J. J. Ramasco, E. Moro, J. M. Pujol and V. M. Eguiluz. *Social Features of Online Networks: The Strength of Intermediary Ties in Online Social Media.* PLoS ONE, 7 (1): e29358, 2012.

[GKK01] K. I. Goh, B. Kahng and D. Kim. *Universal behavior of load distribution in scale-free networks.* Physical Review Letters, 87 (27): 278701, 2001.

[Gle11] J. Gleick. *The information: A history, a theory, a flood.* Harper-Collins UK, 2011.

[GLM01] J. Goldenberg, B. Libai and E. Muller. *Talk of the network: A complex systems look at the underlying process of word-of-mouth.* Marketing Letters, 12 (3): 211-223, 2001.

[GMP+09] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski and L. Brilliant. *Detecting influenza epidemics using search engine query data.* Nature, 457 (7232): 1012 1014, 2009.

[GN02] M. Girvan and M. E. J. Newman. *Community structure in social and biological networks.* Proceedings of the National Academy of Sciences of the United States of America, 99 (12): 7821 7826, 2002.

[Gol10] B. Golub and M. O. Jackson. *Naive learning in social networks and the wisdom of crowds.* American Economic Journal: Microeconomics, 2 (1) 112-149. 2010.

[Gra85] M. Granovetter. *Economic action and social structure: the problem of embeddedness.* American Journal of Sociology, 91: 481, 1985.

[Gsm13] GSMA Intelligence, Mobile Economy Latin America 2013, http://www.gsmamobileeconomylatinamerica.com/ENG_LatAmME_v10_WEB_FINAL.pdf. 2013

[Gue] Guess Language, https://pypi.python.org/pypi/guess-language.

[Gue13] P. H. C. Guerra, W. Meira Jr, C. Cardie and R. Kleinberg. *A measure of polarization on social media networks based on community boundaries.* Seventh International AAAI Conference on Weblogs and Social Media. 2013

[HH09] C. Honeycutt and S. C. Herring. *Beyond microblogging: Conversation and collaboration via twitter.* In HICSS, pages 1 10. IEEE Computer Society, 2009.

[Hol75] R. A. Holley and T. M. Liggett. *Ergodic theorems for weakly interacting infinite systems and the voter model.*. The annals of probability, 3 (4): 643-663. 1975.

[HRW08] B. Huberman, D. Romero and F. Wu. *Social networks that matter: Twitter under the microscope.* First Monday, 14 (1), 2008.

[HS91] R. Huckfeldt and J. Sprague. *Discussant Effects on Vote Choice: Intimacy, Structure and Interdependence.* The Journal of Politics, 53 (01): 122 158, 1991.

[HS95] R. Huckfeldt and J. Sprague. *Citizens, politics and social communication: Information and influence in an election campaign.* Cambridge studies in political psychology and public opinion, Vol 60. Cambridge University Press, 1995.

[HS08] D. A. Hojman and A. Szeid. *Core periphery in networks.* Journal of Economic Theory, 139 (1): 295-309, 2008.

[Huc01] R. Huckfeldt. *The social communication of political expertise.* American Journal of Political Science, 45 (2): 425 438, 2001.

[Huc09] R. Huckfeldt. *Interdependence, Density Dependence and Networks in Politics.* American Politics Research, 37 (5): 921 950, 2009.

[HW09] H.B. Hu and X.F. Wang. *Disassortative mixing in online social networks.* Europhysics Letters, 86 (1): 18003, 2009.

[IM11a] J. L. Iribarren and E. Moro. *Branching dynamics of viral information spreading.* Physical Review E, 84:046116, 2011.

[IM11b] J. L. Iribarren and E. Moro. *Affinity paths and information diffusion in social networks.* Social Networks, 33 (2): 134 142, 2011.

[Ise86] D. J. Isenberg. *Group polarization: A critical review and meta-analysis.* Journal of Personality and Social Psychology, 50 (6): 1141, 1986.

[Jac10] M. O. Jackson. *Social and economic networks.* Princeton University Press, 2010.

[Jae12] K. Jihie and Y. Jaebong, *Role of Sentiment in Message Propagation: Reply vs. Retweet Behavior in Political Communication.* International Conference on Social Informatics (SocialInformatics2012), pp.131-136, 2012. (doi: 10.1109/SocialInformatics.2012.33)

[JN68] M. Kent Jennings and R. G. Niemi. *The transmission of political values from parent to child.* American Political Science Review, 62 (1): 169 184, 1968.

[JN12] F. Toro, J. Nigel. *Facebook gives a platform to the challenger of Chávez.* Foreign Policy. July 26, 2012.

[JSB09] M. K. Jennings, L. Stoker and J. Bowers. *Politics across Generations: Family Transmission Reexamined.* The Journal of Politics, 71 (03): 782, 2009.

[JSFT07] A. Java, X. Song, T. Finin and B. Tseng. *Why we twitter: understanding microblogging usage and communities.* In Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis, pages 56 65. ACM, 2007.

[JW96] M. O. Jackson and A. Wolinsky. *A strategic model of social economic networks.* Journal of Economic Theory, 71 (1):44-74, 1996.

[JZSC09] B. J. Jansen, M. Zhang, K. Sobel and A. Chowdury. *Twitter power: Tweets as electronic word of mouth.* Journal of the American Society for Information Science and Technology, 60 (11): 2169 2188, 2009.

[Kat57] E. Katz. *The Two-Step Flow of Communication: An Up-To-Date Report on an Hypothesis.* Public Opinion Quarterly, 21 (1): 61 78, 1957.

[Ken92] C. B. Kenny. *Political participation and effects from the social environment.* American Journal of Political Science, 36 (1): 259 267, 1992.

[KISB11] W. Ket, G. Iyenga, R. Seth and S. Bowle. *Inequality and network structure.* Games and Economic Behavior, 73 (1): 215-226, 2011.

[KJ09] R. Koop and H. J. Jansen. *Political blogs and blogrolls in Canada: Forums for democratic deliberation?* Social Science Computer Review, 27 (2): 155, 2009.

[KLPM10] H. Kwak, C. Lee, H. Park and S. Moon. *What is Twitter, a Social Network or a News Media?.* In Proceedings of WWW 2010, 112: 591 600, ACM Press, 2010.

[Kno90] D. Knoke. *Political Networks: The structural perspective.* Structural analysis in the social sciences, 4. Cambridge University Press, 1990.

[KR11] K. Zickuhr and L. Rainie. *Wikipedia, past and present.* Pew Research Center. January 13, 2011.

[Kra00] U. Krause. *A discrete nonlinear and non-autonomous model of consensus formation.* Communications in difference equations. Proceeding of the fourth International Conference on Difference Equations, pp. 227 236, 2000.

[Kra09] D. Krackhardt. *A plunge into networks.* Science, vol. 326, pp. 47-48.

[LAH07] J. Leskovec, L. A. Adamic and B. A. Huberman. *The dynamics of viral marketing.* ACM Transactions on the Web (TWEB), 1 (1): 5, 2007.

[LBG44] P. F. Lazarsfeld, B. Berelson and H. Gaudet. *The People's Choice: How the Voter Makes Up His Mind in a Presidential Campaign.* Columbia University Press, 1944.

[LGRC11] J. Lehmann, B. Gonçalves, J. J. Ramasco and C. Cattuto. *Dynamical classes of collective attention in twitter.* WWW '12 Proceedings of the 21st international conference on World Wide Web Pages 251-260, 2011.

[Lie10] D. van Liere. *How far does a tweet travel?: Information brokers in the twitterverse.* In Proceedings of the International Workshop on Modeling Social Media (MSM '10). ACM, New York, NY, USA,2010.

[Lin02] A. L. Barabási and J. Frangos. *Linked: The New Science Of Networks.* Basic Books. 2002.

[LKG+08] K. Lewis, J. Kaufman, M. Gonzalez, A. Wimmer and N. Christakis. *Tastes, ties and time: A new social network dataset using Facebook.com.* Social Networks, 30 (4): 330 342, 2008.

[LM99] Y. Li. *Money and Middlemen in an economy with private information.* Economic Inquiry, 37 (1): 1-12, 1999.

[LM06] A. N. Langville and C. D. Meyer. *Google's PageRank and Beyond: The Science of Search Engine Rankings.* University Press, Princeton, 2006.

[LPA+09] D. Lazer, A. Pentland, L. Adamic, S. Aral, A. L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy and M. Van Alstyne. *Life in the network: the coming age of computational social science.* Science, 323 (5915): 721-723, 2009.

[LSAA11] A. Livne, M. P. Simmons, E. Adar and L. A. Adamic. *The party is over here: Structure and content in the 2010 election.* Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, 2011.

[Lup94] A. Lupia. *Shortcuts Versus Encyclopedias: Information and Voting Behavior in California Insurance Reform Elections.* American Political Science Review, 88 (1): 63 76, 1994.

[MAN+13] D. Mocanu, A. Baronchelli, N. Perra, B. Gonçalves, Q. Zhang and A. Vespignani. *The Twitter of Babel: Mapping World Languages through Microblogging Platforms.* PLoS ONE, 8 (4): e61981, 2013.

[Mar87] P. V. Marsden. *Core Discussion Networks of Americans.* American Sociological Review, 52 (1): 122 131, 1987.

[Mar08] K. Marx. *Critique of the Gotha program.* Wildside Press LLC, 2008.

[MBLB14] A. J. Morales, J. Borondo, J. C. Losada and R. M. Benito. *Efficiency of human activity on information spreading on Twitter.* Social Networks, 39 (0): 1 11, 2014.

[MBLB15] A. J. Morales, J. Borondo, J. C. Losada and R. M. Benito. *Measuring political polarization: Twitter shows the two sides of Venezuela.* Chaos: An Interdisciplinary Journal of Nonlinear Science, 25 (3): 033114, 2015.

[MDS05] I. Borg and P. Groenen. *Modern Multidimensional Scaling: theory and applications* (2nd ed.). New York: Springer-Verlag. pp. 207-212.2 005.

[Mer68] R. K. Merton. *The matthew effect in science.* Science, 159: 56, 1968.

[Mil67] S. Milgram. The small world problem. *Psychology Today*, 2 (1): 60 67, 1967.

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

[MLA+11] A. Mislove, S. Lehmann, Y. Y. Ahn, J. P. Onnela and J. N. Rosenquist. *Understanding the Demographics of Twitter Users.* In Proceedings of the 5th International AAAI Conference on Weblogs and Social Media. ICWSM, 2011.

[MLB12] A. J. Morales, J. C. Losada and R. M. Benito. *Users structure and behavior on an online social network during a political protest.* Physica A: Statistical Mechanics and its Applications, 391 (21): 5244 5253, 2012.

[Mob03] M. Mobilia. *Does a single zealot affect an infinite group of voters?.* Physical Review Letters, 91 (2): 028701, 2003.

[Mob07] M. Mobilia, A. Petersen and S. Redner. *On the role of zealotry in the voter model.* Journal of Statistical Mechanics: Theory and Experiment, 8: 08029. 2007.

[Mut06] D. C. Mutz. *Hearing the Other Side: Deliberative Versus Participatory Democracy.* Cambridge University Press, 2006.

[MV12] E. Minaya and K. Vyas. *When Chávez tweets, Venezuelans listen.* The Wall Street Journal, April 25, 2012.

[New03] M. E. J. Newman. *The structure and function of complex networks.* SIAM Review, 45 (2): 167-256, 2003.

[New05] M. E. J. Newman. *Power laws, pareto distributions and zipf's law.* Contemporary Physics, 46 (5): 323 351, 2005.

[New06] M. E. J. Newman. *Modularity and community structure in networks.* Proceedings of the National Academy of Sciences of the United States of America, 103 (23): 8577 8582, 2006.

[New11] M. E. J. Newman. *Communities, modules and large-scale structure in networks.* Nature Physics, 8 (1): 25 31, 2011.

[NF00] P. Nieuwbeerta and H. Flap. *Crosscutting social circles and political choice. Effects of personal network composition on voting behavior in The Netherlands.* Social Networks, 22 (4): 313 335, 2000.

[NG03] M. E. J. Newman and M. Girvan. *Finding and evaluating community structure in networks.* Physical Review E, 69 (2): 16, 2003.

[OBRS10] B. O'Connor, R. Balasubramanyan, B. R. Routledge and N. A. Smith. *From tweets to polls: Linking text sentiment to public opinion time series.* Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, ICWSM, pp: 122 129, 2010.

[PD13] P. Dandekara, A. Goelb, D. T. Lee. *Biased assimilation, homophily and the dynamics of polarization.* Proceedings of the National Academy of Sciences of the United States of America, 110 (15): 5791â5796, 2013.

[Pea09] Pearanalytics. *Twitter study.* http://pearanalytics.com, August 2009.

[Pew11] Pew Research Center. *Global Digital Communication: Texting, Social Networking Popular Worldwide.* December 20, 2011.

[PMGV13] F. Pedroche, F. Moreno, A. González and A. Valencia. *Leadership groups on Social Network Sites based on Personalized PageRank.* Mathematical and Computer Modelling, 57 (7 8): 1891 1896, 2013.

[Pop91] S. L. Popkin. *The Reasoning Voter: Communication and Persuasion in Presidential Campaigns*, 2nd ed. University of Chicago Press, 1991.

[Pös11] J. Pöschko. *Exploring Twitter Hashtags.* arXiv:1111.6553, 2011.

[Pri76] D. S. Price. *A general theory of bibliometric and other cumulative advantage processes.* Journal of the Association for Information Science and Technology, 27:292, 1976.

[PS08] M. Perc and A. Szolnok. *Social diversity promotion of cooperation in the spatial prisoner's dilemma game.* Physical Review E, 77:011904, 2008.

[PSVV01] R. Pastor-Satorras, A. Vázquez and A. Vespignani. *Dynamical and correlation properties of the Internet.* Physical Review Letters, 87 (25): 258701, 2001.

[Raw99] J. Rawls. *A Theory of Justice.* Cambridge: Belknap of Harvard UP, 1999.

[Raw01] J. Rawls. *Justice as Fairness: A Restatement.* Cambridge: Belknap of Harvard UP, 2001.

[RB08] M. Rosvall and C. T. Bergstrom. *Maps of random walks on complex networks reveal community structure.* Proceedings

of the National Academy of Sciences of the United States of America, 105 (4): 1118 1123, 2008.

[RCM+11a] J. Ratkiewicz, M. Conover, M. Meiss, B. GonÃ alves, A. Flammini and F. Menczer. *Detecting and Tracking Political Abuse in Social Media.* In Proceedings of the 5th International AAAI Conference on Weblogs and Social Media. ICWSM, 2011.

[RCM+11b] J. Ratkiewicz, M. Conover, M. Meiss, B. GonÃ alves, S. Patil, A. Flammini and F. Menczer. *Truthy: Mapping the Spread of Astroturf in Microblog Streams.* In Proceedings of the 20th International Conference Companion on World Wide Web, WWW '11, pages 249 252, 2011.

[RGAH11] D. M. Romero, W. Galuba, S. Asur and B. A. Huberman. *Influence and passivity in social media.* In Proceedings of the 20th International Conference Companion on World Wide Web, WWW '11, pages 113-114, 2011.

[Rob76] J. P. Robinson. *Interpersonal influence in election campaigns: Two step-flow hypotheses.* Public Opinion Quarterly, 40 (3): 304 319, 1976.

[RTU11] D. M. Romero, C. Tan and J. Ugander. *Social-topical affiliations: The interplay between structure and popularity.* arXiv:1112.1115, 2011.

[RW87] A. Rubinstei and A. Wolinsky. *Middlemen.* The Quarterly Journal of Economics, 102: 581, 1987.

[SB07] A. Santiago and R. M. Benito. *An extended formalism for preferential attachment in heterogeneous complex networks.* Europhysics Letters, 82 (5): 58004, 2007.

[SBn03] M. A. Serrano and M. Boguña. *Topology of the world trade web.* Physical Review E, 68: 015101, 2003.

[SBV09] M. A. Serrano, M. Boguña and A. Vespignani. *Extracting the multiscale backbone of complex weighted networks.* Proceedings of the National Academy of Sciences of the United States of America, 106 (16): 6483 6488, 2009.

[Sem12] Semiocast. *Twitter reaches half a billion accounts more than 140 millions in the U.S. http://semiocast.com/en/publications/*

[Sha11] C.T. Li, S. D. Lin and M. K. Shan. *Exploiting endorsement information and social influence for item recommendation.* In Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval (SIGIR '11), 1131-1132. 2011.

[Sim71] H. A. Simon. *Designing Organizations for an Information-Rich World.* In Martin Greenberger, Computers, Communication and the Public Interest, Baltimore, MD: The Johns Hopkins Press, 1971.

[Sim55] H. A. Simon. *On a class of skew distribution functions.* Biometrika 42 (3-4): 425-440. 1955.

[SOM10] T. Sakaki, M. Okazaki and Y. Matsuo. *Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors.* In Proceedings of the 19th International Conference on World Wide Web, WWW '10, pages 851 860, New York, NY, USA, 2010.

[SSP08] F. C. Santo, M. D. Santo and J. M. Pacheco. *Social diversity promotes the emergence of cooperation in public goods games.* Nature, 7201: 213, 2008.

[Ste13] D. Stewart. *When Retweets Attack: Are Twitter Users Liable for Republishing the Defamatory Tweets of Others?.* Journalism and Mass Communication Quarterly, 90 (2): 233-247, 2013.

[Sti12] J. E. Stiglitz. *The price of inequality: How today's divided society endangers our future.* W.W. Norton & Co., 2012.

[Suc05] K. Suchecki, V. M. Eguíluz and M. San Miguel. *Voter model dynamics in complex networks: Role of dimensionality, disorder and degree distribution..* Physical Review E, 72 (3): 036132, 2005.

[Sun02] C. R. Sunstein. *The law of group polarization.* Journal of political philosophy, 10 (2): 175 195.2002.

[SVR07] S. Goyal and F. Vega-Redondo. *Structural holes in social networks.* Journal of Economic Theory, 137 (1): 460-492, 2007.

[Ten12] Tendencias Digitales. *La penetración de Internet en Venezuela.* 2012. http://tendenciasdigitales.com/1433/la-penetracion-de-internet-en-venezuela-alcanza-40-de-la-poblacion/.

[TSSW10] A. Tumasjan, T. O. Sprenger, P. G. Sandner and I. M. Welpe. *Predicting elections with twitter: What 140 characters reveal about political sentiment.* Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, ICWSM, pp. 178â185, 2010.

[Tur87] J. C. Turner. *Rediscovering the Social Group: A Self-categorization Theory.* B. Blackwel, 1987.

[Uzz96] B. Uzzi. *The sources and consequences of embeddedness for the economic performance of organizations: The network effect.* American Sociological Review, 61: 674, 1996.

[Uzz97] B. Uzzi. *Social structure and competition in interfirm networks: The paradox of embeddedness.* Administrative Science Quarterly, 42(1): 35, 1997.

[Wat02] D. J. Watts. *A simple model of global cascades on random networks.* Proceedings of the National Academy of Sciences of the United States of America, 99 (9): 5766 5771, 2002.

[Wea82] M. S. Weatherford. *Interpersonal Networks and Political Behavior.* American Journal of Political Science, 26 (1): 117 143, 1982.

[Web14] Z. Liu and I. Weber. *Predicting ideological friends and foes in Twitter conflicts.* In Proceedings of the companion publication of the 23rd international conference on World wide web companion (WWW Companion '14), 575-576, 2014.

[WG10] W. Galuba, K. Aberer, D. Chakraborty, Z. Despotovic, W. Kellerer. *Outtweeting the twitterers - predicting information cascades in microblogs.* In Microblogs 3rd Workshop on Online Social Networks, WOSN, 2010.

[WHAT04] F. Wu, B. A. Huberman, L. A. Adamic and J. R. Tyler. *Information flow in social groups.* Physica A: Statistical Mechanics and its Applications, 337 (1): 327 335, 2004.

[WHMW11] S. Wu, J. M. Hofman, W. A. Mason and D. J. Watts. *Who says what to whom on twitter.* In Proceedings of the 20th international conference on World wide web, WWW '11, pages 705 714, 2011.

[Won11] Y. Y.Wong and R. Wright. *Buyers and Sellers.* Working Paper. Federal Reserve Bank of Minneapolis. Research Department. http://www.minneapolisfed.org/research/wp/wp691.pdf, 2011.

[WS98] D. Watts and S. Strogatz. *Collective Dynamics of Small-World Networks,.* Nature, 393: 440 442, 1998.

[WXYM07] D. Walker, H. Xie, K.K. Yan and S. Maslov. *Ranking Scientific Publications Using a Simple Model of Network Traffic.* Journal of Statistical Mechanics: Theory and Experiment, 2007: P06010, 2007.

[Yul25] G. U. Yul. *A mathematical theory of evolution and based on the conclusions of Dr. J. C. Willis F.R.S.* Philosophical Transactions of the Royal Society of London. Series B, 213: 21-87, 1925.

[ZJ01] E. W. Zuckerman and J. T. Jost. *What makes you think you're so popular? Self-evaluation maintenance and the subjective side of the "friendship paradox".* Social Psychology Quarterly, 64 (3): 207 223, 2001.

[ZM04] S. Zhou and R. J. Mondragón. *The rich-club phenomenon in the internet topology.* Communications Letters, IEEE, 8 (3): 180 182, 2004.